II. VARGA GYULA

IGEKÖTŐS SZÓALAKOK MORFOLÓGIAI ELEMZÉSE SZÁMÍTÓGÉPPEL[1]

ABSTRACT: (The morphological analysis of the verbs with prefix by computer)
The extract is a part of a more detailed study on the theme of creating a special software, simple and quick, which is able to recognize the Hungarian verbs with prefix, infinitives and participle forms, and to separate the prefixes. The software itself is based on the adaptation of phonotactical rules.

A magyar nyelv igekötős szóalakjait morfológiai szempontból két irányból kell elemeznünk: jobbról a toldalékokat, balról pedig az igekötőket. A jobbról történő elemzés tulajdonképpen megegyezik a többi magyar ige ilyen természetű vizsgálatával, s az erre készült modelleket, kísérleteket Prószéki Gábor részletesen ismerteti (Számítógépes nyelvészet, Bp., 1989, 503--9.).

A balról történő formális elemzéssel nyilvánvalóan több kutató és gyakorlati szakember találta már magát szemben, bár ennek publikálásával eddig még nem találkozhattunk. A téma feldolgozását több tényező is szükségessé teszi.

a) A kvantitatív kutatási módszerek nagymértékű térhódításával egyre tarthatatlanabb az a felemás (sőt megtévesztő, hibás) gyakorlat, miszerint szófajstatisztikai vizsgálatokban csak az elváló igekötőket regisztrálják. Eszerint, ami elválik, az igekötő, ami nem, az meg nincs is, illetve az adott igével, igenévvel, névszóval együtt egy szövegszó. Pedig attól, hogy elválik, az igekötő sem funkciójában, sem természetében nem változik meg. Az ige is ugyanúgy igekötős marad: *elmondja -- el is mondja, mondja el.* A kutatónak el kell döntenie, hogy önálló szónak, előtagnak vagy prefixumnak tartja-e az igekötőt, és az egész vizsgálandó szövegben következetesen aszerint kell eljárnia. Az ilyen statisztikai vizsgálatokban segítségünkre lehet egy olyan felismerő automata beiktatása, amely az igekötős szóalakokat tudja kezelni.

[1] Elhangzott a Magyar Nyelvtudományi Társaság egri csoportjának felolvasóülésén 1991. április 16-án.

Itt kell szólnunk arról az új helyesírási hibatípusról, amely a nyomtatott sajtóban ütötte fel a fejét. A fényszedésre való áttéréssel, illetve az elektronikus szövegszerkesztők használatával újságjainkban elburjánoztak a hibás sorvégi elválasztások. (Legutóbb Fábián Pál és F. Kovács Ferenc bírálatait olvashattuk az Édes Anyanyelvünk XIV. évfolyamának 1. számában.) Az előbb említett és a továbbiakban ismertetendő műveletek segítségével elkerülhető a **fe-ladat*, **me-géri*, **bek-rémez* stb. szóalakok hibás elválasztása.[2]

b) Igekötős szerkezeteink viselkedésének explicit leírását az **elméleti nyelvtudomány** is igényli. Egy mondatszerkezetbe (-vázba) bekerülve egy igekötős ige másképpen viselkedik, mint egy szimplex ige, bizonyos logikai műveletek és pragmatikai tényezők hatására az igekötős igéken morfoszintaktikai változások mennek végbe. Fel kell tárnunk azokat a szabályokat, amelyek a mondatgenerálás során egy folytonos szóalakból egy nem folytonos -- más szóhasználattal: analitikus -- szóalakot hoz létre, s le kell írnunk ezen szóalakok létrehozásának módját.

c) Az igekötős alakulatok kezelése megoldandó feladatcsoportot jelent a maga sajátos módszereivel és eszközeivel dolgozó **számítógépes nyelvészet** számára is. A mondatszintetizálás, az automatikus szótárkészítés, a gépi fordítás, a különféle interaktív rendszerek stb., stb. igénylik az igekötős szerkezetek működésének formális leírását.

Szükség van tehát egy olyan algoritmusra, melynek során programunk fölismeri a beérkező szóalakokban lappangó igekötőket, és -- a további feldolgozás számára -- automatikusan leválasztja őket igéjükről, igenevükről.

Az analízis műveletsora onnan indul, hogy egy szótárból vagy szövegből különféle betűsorokat (szavakat) kapunk. Az elemzés két részből áll: a) igekötő-keresés és leválasztás; b) a visszamaradt szórész ("maradék") azonosítása. Az 1. lépés viszonylag egyszerű: listát kell adni az igekötőkről, s a szókezdeteket megvizsgálva leválasztani őket. Rögtön kiderül, hogy nem az igekötő megtalálása és leválasztása nehéz, hanem a téves igekötő-azonosítások kiszűrése. Formai szempontból -- vagyis csupán a betűket (karaktereket) vizsgálva -- ugyanis "igekötőket" találunk nemcsak a *megint*, *elül*, *felette*, *betűz* stb. homonímákban, hanem az *átkoz*, *becsül*, *felel*, *kisebbít*, *les*, *rág* stb. igékben is.

---

[2]Bizonyára ezeket a hibákat is ki fogja küszöbölni a magyar helyesírási programcsomag, melyen -- értesüléseim szerint -- két munkacsoport is dolgozik. Eredményeikről azonban jelenleg nincs tudomásom.

A művelet nehezét tehát a második metszet, a **maradék** ellenőrzése adja. Ez megoldható lenne olyan módszerrel, amilyennel a (készülő) helyesírás-ellenőrző programcsomagok dolgoznak. Vagyis mellékelünk egy szótárt, mely minden egyes magyar lexéma összes ragos-jeles szóalakját tartalmazza, így egyszerűen elvégezhető lenne az azonosítás. Csakhogy nekünk egy könnyen kezelhető, kis háttérinformációt igénylő programra van szükségünk, a szótár óriási anyaga meg elnehezítené, fölösleges és hosszú kereséseivel lelassítaná a műveletet.

A maradék ellenőrzésére külön algoritmust kell kidolgoznunk. Ez a lépéssor fonotaktikai szabályoknak betűkre adaptálásával indul. Négy "betűtilalmi" szabályt állítunk fel:

1. a magyar nyelvben ige egyetlen betűből nem állhat,
2. két azonos betű nem lehet szókezdő helyzetben,
3. szó elején csak meghatározott betűkapcsolat állhat,
4. adott igékben bizonyos betűkombinációk szókezdeten nem fordulhatnak elő.

Az első szabály kiszűri az olyan téves azonosításokat, mint az *aláz, lep, lesz, ráz* stb., a második szerint nem igekötős szó a *berreg, lebben, lennék* stb. Programunk ugyanis -- a karakterek alapján -- igekötőnek fogja találni az *alá, le, rá, be* szórészeket, így azokat leválasztja, a megmaradt szegmentumokról meg föltételezi, hogy igék vagy igenevek. Az első négy szóban a maradék egyetlen mássalhangzó, a többiben pedig a metszés után azonos mássalhangzók kerülnek szókezdő helyzetbe. Vagyis mindkét esetben kiderült a tévedés, így vissza kellett állítani az eredeti alakot.

A harmadik szabály egy kicsit bonyolultabb. Lényege, hogy **a magyar ábécé betűi nem állhatnak bármilyen kombinációban a szó elején.** (Pl.: nincs *dg-, akb-, chi-* stb. kezdetű szavunk.) Különösen érvényes ez a mássalhangzókra. Vegyük szóalakjaink első két betűjét. A magyar mássalhangzókat jelölő 26 betű így -- tehát kettesével -- 650 variációt ad. Kassai Ilona kimutatta, hogy nyelvünk -- a fonémákat tekintve -- ebből mindössze 44 (!) kapcsolatot használ föl (Kassai 1981, 73--74). A betűket vizsgálva meglepetéssel tapasztaljuk, hogy a magyar igékben szókezdő helyzetben összesen 27 mássalhangzó-kombináció használatos (sőt ebből néhány kapcsolattípus csak egy-két idegen szó elején fordul elő, pl.: *pszichologizál, szcenírozz, glorifikál, flörtöl* stb.).

A harmadik szabály tehát azt jelenti, hogy ha a második metszet (tkp. igekötős ige alapigéje) két mássalhangzóval[3] kezdődik, az csak a következő lehet:

pr	br	tr	dr
kr	gr	fr	sr
pl	bl		kl
gl	fl	sl	
kv			sm
sp			st
sk			

Így kerülhetők el az olyan téves bontások, mint **le-bzsel, *le-ndít, *rá-zkódik* stb.

A betűkapcsolatok ellenőrzésével azonban még mindig nem tudjuk az összes hibalehetőséget kiiktatni. Vannak olyan igéink, amelyeknek szókezdetei véletlenül egybeesnek valamely igekötő írásképével. A *beszól* pl. nyilvánvalóan igekötős ige, de a *beszél* nem az, pedig a betűszerkezetük megegyezik. Ezeket is -- szerencsére nem sokan vannak -- össze kell gyűjtenünk, listájuknak lehetőleg teljesnek, jól rendezettnek és minél rövidebbnek kell lennie.

A betűtilalmi szabályok alkalmazásának, működésének bemutatására szolgáljon itt most a *lesz* ige. Ennek ugyanis minden paradigmatikus alakja és származéka -- a szó szoros értelmében vett formai szempontból -- "igekötőgyanús", hisz a szókezdet egybeesik a le igekötő írásképével. Ki kell tehát szűrni a sok *le* igekötős ige közül (*leszed, leszid, leszorít, leszűr* stb.) a *lesz* összes ragos-jeles alakját. Ez a következőképpen történik:

a/ A főnévi igenévi, a feltételes módú és a múlt idejű alakok metszetei fennakadnak az ellenőrzésen, mert a maradékok két azonos betűvel kezdődnek (*-nni, -nnék, -ttem*). Vagyis a vizsgált karaktersor nem tartalmaz igekötőt.

b/ Az előbbi eredményre jutunk a *lesz* esetében is: egy betűből álló igénk nincs. (Ha az *sz*-t két karakternek vesszük, akkor a c/ vonatkozik rá.)

c/ Nem azonosíthatók a *lesztek, lesznek* szóalakok sem, mert a táblázatunk nem tartalmaz *szt, szn* betűcsoportos szókezdeteket. (Vagyis *szt-, szn-* kezdetű (igekötős) igénk nincs; az egyetlen *szt-* kezdetű *sztrájkol* nem igekötős.)

d/ Láttuk, a betűtilalmi szabályok a metszetek kezdő karaktersorait ellenőrizték. Némely ige bizonyos szóalakjainak homográfiája azonban átcsap ezeken a szabályokon. A *lesz* igénél maradva, a *leszek, leszel, leszünk* és az összes felszólító

[3] A kétjegyűeket egy betűnek vesszük.

módú alak átmegy az ellenőrzésen -- vagyis igekötősnek minősül --, mert nem ütközik az általános tilalmi szabályokkal. Az ilyen igék esetében az egyes igékre vonatkozóan kell letiltani bizonyos karakterkombinációkat. A *-gyetek* és a *-szünk* metszetet le kell stoppolnunk, hisz ezekkel a betűsorokkal magyar ige nem kezdődhet. Más a helyzet a *leszek* és a *legyen* szóalakokkal. Van ugyanis *leszekerezik és legyengül/legyengít* igénk, így a *-szek** és a *-gyen** elfogadható alakok[4], viszont a *-szek* és a *-gyen* nem.

A szabályostól eltérő néhány ige kiegészítő ellenőrzése a következőképpen történik: a/ általános betűtilalmi szabályok alkalmazása és b/ egyedi betűkombinációk letiltása. A *lesz* ige esetében ez a következőképpen néz ki:

szóalak	1. fázis: keresés és bontás	2. fázis: azonosítás
lenni	*le-nni*	*nn-*
lennék, lennél stb.	*le-nnél stb.*	*nn-*
lettem, lettél stb.	*le-ttem stb.*	*tt-*
leszek	*le-szek*	*szek*
leszel	*le-szel*	*?*
lesz	*le-sz*	*sz*
leszünk	*le-szünk*	*szünk*
lesztek	*le-sztek*	*szt-*
lesznek	*le-sznek*	*szn-*
legyek	*le-gyek*	*gye* kiv.:*=ng-*
legyél	*le-gyél*	*é**
legyen	*le-gyen*	*ü**
legyünk	*le-gyünk*	
legyetek	*le-gyetek*	
legyenek	*le-gyenek*	

Látható, mennyire leegyszerűsödött e szóalakcsoport elemzése. Hasonlóképpen történik ez a többi igekötős ige és igenév esetében is. Még további 12 igekötőnk esetében fordul elő efféle véletlen egybeesés, s az ilyen részlegesen homográf alakoknak a száma megközelíti a 60-at.

A fenti táblázatban azonban egy szóalak még így is ellenőrizhetetlen maradt. A *leszel* ugyanis tisztán homográf (és egyben homoním is) a *szel* ige *le* igekötős párjával. Ezt tehát fel kell vennünk a homonímalistára, mely még további 13 szóalakot tartalmaz (*beles, betűz, elég, elöl* stb.). Róluk csak a felhasználó tudja eldönteni, hogy az adott környezetben mi az aktuális olvasatuk.

[4]A * itt azt jelenti, hogy még más betű is következhet a *k*, ill az *n* után.

Maga a program igen egyszerű: 4 adatállományt (igekötők, szókezdő msh-párok, kivételszótár, homonímaszótár) és 3 függvényt tartalmaz. Az 1. függvény valójában az igekötők keresését és leválasztását végzi el, a 2. pedig a metszet (maradék) betűkombinációit vizsgálja: ha az első két karakter -- kettősbetű esetében a 3. -- valamelyike nem mgh, akkor csak a megadott msh-párokat fogadja el. A 3. függvény a maradékot azonosítja a kivételszótár (a részlegesen homográf alakok) és a homonímaszótár adataival.

Ez a kis program egy olyan nagyobb műveletsor része, amely képes az analitikus igealakokból visszaállítani az eredeti (szintetikus) szóalakokat -- és fordítva: az elváló igekötős alakulatokból visszaállítani a kiinduló alakokat. A fenti részlet alkalmas az igekötős szóalakok elválasztására, az igekötők (statisztikai) vizsgálatára, a fő feladata pedig az, hogy előkészítse az igekötős igéket az analitikus szóalakok (elváló igekötős szerkezetek) automatikus létrehozásához.