Görgey Eszter – Jékel Pál – Papp Ferenc

Automatikus szövegelemzés a legalsó nyelvi szinteken

0. Bevezetőben körülírjuk a címben használt s az alábbiak számára fontos fogalmakat, majd dolgozatunk rövid vázát ismertetjük.

„Automatikus"-on azt értjük, hogy valamely folyamat lejátszódik anélkül, hogy e folyamat egyes pontjaiban be kellene kívülről avatkozni a lefutásba. „Folyamat" lehet például az elemzés. „Elemzés"-en azt értjük, hogy valamely természetben adott dolgokat (esetünkben: szövegeket, szövegdarabokat) azok nevével helyettesítünk, tehát tulajdonképpen egy természetesen adott szintről egy metanyelvi szintre térünk át. (A „szintézis" ezzel szemben az a folyamat, amikor kiindulásul bizonyos metanyelvi elnevezések vannak s ezeket helyettesítjük tárgynyelvi megfelelőikkel.)

A „legalsó nyelvi szintek" a betűk-fonémák, illetőleg a szövegszavak szintjét jelentik, tehát a morfológia és a szintaxis előtti szinteket. Kulcsfontosságú számunkra a „szöveg" fogalma. „Szöveg" lehet a jeleknek egy automata által létrehozott véges sorozata. E sorozatban az egyes jelek tetszőleges ismétlődése meg van engedve. Másrészt a jelek bizonyos csoportokat alkotnak, bizonyos hierarchia van közöttük. E csoportokat, csak a szöveg felől szemlélve a dolgot, részint a jelek eloszlása alapján állapítottuk meg, részint — ami nehezebb — szemantikájuk alapján. A jelek ilyen csoportokra való bontása tette egyáltalán lehetővé a szövegelemzést e legalsó szinteken — vagyis azt, hogy a szöveg egészét valamely, a további feldolgozás számára kiindulásul szolgálható, darabokra osszuk. A szöveg maga igen hosszú lehet (vagy nagyon rövid — ez nem lényeges), de ahhoz, hogy egyáltalán felsőbb szintű elemzésnek vethessük alá, elengedhetetlen ezen alsó szinteken a darabokra bontás.

Az írott szövegekben látszólag egyszerű a dolgunk: mondjuk, egy-egy írógépen leütött karakter egy-egy betű (jól tudjuk, hogy például a magyar íráselmélet számára nem ilyen egyszerűen áll e kérdés); egy-egy „space" között álló jelsorozat a (szöveg)szó; egy-egy pontértékű jel között álló szósorozat a mondat. A valóságban, mint alább látni fogjuk, távolról sem volt ilyen könnyű dolgunk, amikor természetes (magyar) nyelven írt szövegeket kívántunk automatikusan elemeztetni. Dolgozatunk célja éppen az, hogy megmutassa, milyen metaapparátus szükséges az analízis leírásához.

A számunkra kulcsfontosságú „szöveg" fogalma tekintetében itt még egy kitérést teszünk. Nevezetesen: az alábbiakban csak magyar természetes nyelven í r t szövegekről lesz szó. Nem szabad megfeledkeznünk arról, hogy — például — fennállhatna a magyar természetes nyelven e l h a n g z ó szövegek automatikus elemzésének a kérdése is — sőt, ezt a kérdést egészen biztosan fel is fogja előbb-utóbb vetni a korszerű termelési-technikai gyakorlat. És akkor, ha hangzó szövegekről van szó, bizonyos értelemben új feladatok fognak előttünk állni. Amíg ugyanis az írás mintegy kvantáltan adja elénk a szöveget, addig a hangzó beszédben vannak ugyan határjelek a szöveg bizonyos darabjai között, ám lényegében az a szöveg kontinuusan kerül elénk, ott még a legkisebb egységek (a fonémák vagy azok megkülönböztető jegyei) kitaglalása is minden bizonnyal az itteninél komolyabb problémát fog jelenteni. Másrészt, és főleg: „szöveg"-ről szólván távolról sem szabad csupán (írott vagy hangzó) magyar, német, orosz stb. természetes nyelvi termékekre gondolnunk. A szöveg fent adott körülírása lehetővé, a szemiotika mai állása szükségessé teszi, hogy mindig szemünk előtt tartsuk: szöveg lehet b á r m e l y, automata által létrehozott jelsorozat. Tehát, például (vö. Zaliznyak 1962): a vallás-grammatika által létrehozott vallásos szövegek állhatnak bizonyos állandó és változó fényjelekből (a templom ablakai, vitrázsa és enteriőrje; a szertartás szerint elvégzett mozgások; a hivők mozgásai mint vizuális élmény); bizonyos hangjelekből (zene — orgona, ének, a pap hangja, egyéb hangok); bizonyos illatjelekből (pl.: tömjén, az égő gyertya illata); bizonyos tapintható és ízlelhető jelekből (pl. az ún. „áldozásnál" a száraz ostyának, más szertartásokban az ostyának és a vizezett bornak a tapintása-ízlelése stb.). Mindezen külsőre bonyolult jelenségek tehát egyetlen szerves szöveget alkothatnak (a vallásgrammatika automatája hozta őket létre, jelek, nem összevissza, hanem meghatározott rendben követik, körítik stb. egymást — ezért nem szöveg viszont például egy virágzó rét, bár nagy hasonlóságot mutat fel külsőleg egy templom-enteriőrrel: virágok, szél, méhzümmögés stb.). Fennállhat természetesen annak a szükségessége, hogy e bonyolultabb fizikai megjelenésű szövegeket is automatikusan elemezni tudjuk a mondott értelemben. Itt tehát tudatos leszűkítés eredményeképpen foglalkozunk csupán a mondott egyszerű felépítésű szövegekkel. Ugyanakkor meg vagyunk róla győződve, hogy lényegében mindenféle, tehát az itt leírtnál bonyolultabb fizikai megjelenésű szövegek leírásához, elemzéséhez az itt ismertetendőhöz nagyon hasonló metarendszerre, fogalmakra stb. lesz szükség; igyekeztünk is rendszerünket úgy leírni, hogy az esetleg szélesebb keretben is alkalmazható legyen.

Nem foglalkozunk itt azzal a kérdéssel sem, hogy hogyan lehetne automatikusan megkülönböztetni a szöveget a nem-szövegektől, a jeleket a rájuk megszólalásig hasonlító nem-jelektől. (Erről valamivel bővebben vö. Papp 1968). Egyszerűen ismertnek tekintjük azt, hogy az adott fizikai jelenségek nem egyszerűen önmaguk, hanem valamely jelek s valamely automata (esetünkben a magyar természetes nyelv grammatika automatája) által lettek szöveggé szőve.

A dolgozatunk elé kitűzött cél megvalósítása érdekében előbb röviden ismertetjük az ALGOL 60 (a továbbiakban: ALGOL) nyelven írt szövegek

struktúráját (**1.**), majd a magyar természetes nyelven írt szövegek elemzését a fonémaszinten (**2.1.**) és a szövegszók (a továbbiakban csak: szók, szavak) szintjén (**2.2.**).

1. ALGOL-szövegek struktúrája

Némi tréfálkozással azt mondhatnánk, hogy az ALGOL általában „Write only" nyelv: nem szoktunk ALGOLul olvasni, ALGOLból fordítani (csak olyan, kivételesnek tekinthető esetekben, amikor az ALGOLban írt program szövegében keresünk hibát). ALGOLról a gép fordít magának egy speciális fordító program segítségével. ALGOLul *írni* szoktunk, tehát explicit formában az ALGOLt ismertető, leíró stb. dokumentumokban nem az analízis, hanem a szintézis szabályait közlik. Ugyanakkor világos, hogy az ALGOL szabályai szerint megalkotott szövegeknek egyértelműen elemezhetőknek kell lenniük, különben nem működhetnének az említett fordító programok. Ezért az ALGOL szövegek szerkezetének utólagos áttekintése igen hasznos lehet a számunkra; az ilyen nyelven készült szövegek mintegy mintául szolgálhatnak egyéb (például természetes) nyelveken kapott szövegek jó leírásához.

Az ALGOL szövegek szerkesztésére (írására) vonatkozó *metanyelv* (az ALGOL 60-ban) a következő. Az ALGOL szövegek szintézisének szabályait definíciósorral írjuk le. A definíciókat definiáló egyenlőségek alakjában rögzítjük. A definiáló egyenlőség tartalmaz: (i) metanyelvi alapjeleket: definiáló (meta) zárójelet: ⟨ ⟩; definiáló egyenlőségjelet: : : =, amit úgy kell olvasni, hogy „lehet"; függőleges vonalat: |, melyet úgy kell olvasni, hogy „vagy (akár)"; (ii) természetes nyelvi vagy egyéb jeleket.

Ezen jeleket felhasználva a definiáló egyenlőség áll (a) egy bal oldalból, ez tulajdonképpen nem más, mint a definiálandó fogalom metazárójelek között (egy metanyelvi fogalom megnevezése a metazárójelekben); (b) a négypont-egyenlő definiáló egyenlőségjelből; (c) a jobb oldalból, melyben a bal oldal lehetséges realizációit írjuk le. A jobb oldal struktúrája többféle lehet, éspedig: I. egy jel vagy metazárójelbe tett fogalom; ezzel a bal oldali fogalomnak megfeleltetjük a jobb oldali jelet vagy fogalmat; II. több jel vagy metazárójelbe tett fogalom egymás mellé helyezett lánca; ekkor a bal oldali fogalmat az olyan jelcsoportokkal definiáljuk, amelyek a láncban megadott sorrendben összerakott elemekből állnak; III. több jel vagy metazárójelbe tett fogalom, vagy lánccal képzett összetétel függőleges vonalakkal elválasztva; ebben az esetben a függőleges vonalakkal elválasztott jelek vagy jelcsoportok mindegyike lehet azonos a bal oldali fogalommal. Egy példa: A tizedes tört formájú írásmód definiálása így történnék ezen a metanyelven:

⟨számjegy⟩ : : = 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 (1)
⟨előjel⟩ : : = + | − (2)
⟨természetes szám⟩ : : = ⟨számjegy⟩ | ⟨természetes szám⟩ ⟨számjegy⟩ (3)
⟨egész⟩ : : = ⟨természetes szám⟩ | ⟨előjel⟩ ⟨természetes szám⟩ (4)
⟨tizedes szám⟩ : : = ⟨egész⟩ · ⟨természetes szám⟩ (5)

Figyeljük meg a (3) egyenlőséget, itt az írásmód nagyfokú alkalmazhatósága, tömörsége és rugalmassága jól kidomborodik: természetes szám lehet akár egy számjegy, akár ha egy természetes szám után újabb számjegyet írunk, akkor ismét természetes számot kapunk. Az (1)—(5) definíciók értelmében a következő számok mindegyike tizedes szám: —5.8 0023.45 +0.5 56.96 —000005.2 — de nem tekinthető a fenti definíció alapján tizedes számnak: +.78 —55 + —45.67 23. stb.

Íme, néhány fontosabb ALGOL-definíció az előző bekezdésben érintett metanyelv segítségével leírva:

⟨azonosító⟩ : : = ⟨betű⟩ | ⟨azonosító⟩ ⟨betű⟩ | ⟨azonosító⟩ ⟨számjegy⟩ (1)
⟨változó⟩ : : = ⟨egyszerű változó⟩ | ⟨indexes változó⟩ (2)
⟨egyszerű változó⟩ : : = ⟨azonosító⟩ (3)
⟨indexes változó⟩ : : = ⟨tömbnév⟩ [⟨indexlista⟩] (4)
⟨tömbnév⟩ : : = ⟨azonosító⟩ (5)
⟨indexlista⟩ : : = ⟨aritmetikai kifejezés⟩ | ⟨indexlista⟩, ⟨aritmetikai kifejezés⟩ (6)

A fenti definíciósor az ALGOL egy fontos fogalmát, a változót határozza meg. A definíció során csak a „betű", a „számjegy" és az „aritmetikai kifejezés" tisztázatlan, de ezek pontos definíciója nélkül is mindenki követheti a többi definíciót, csak e kettő helyére a közhasználatban is így nevezett dolgokat képzelje bele (például a betűkön a latin ábécét, az aritmetika kifejezésen pedig az aritmetikai műveleteket).

Az (1) azt határozza meg, hogy azonosítón olyan jelsorozatot értünk, amely csupán betűket és számjegyeket tartalmaz és betűvel kezdődik.

A (2) azt fejezi ki, hogy a változó tulajdonképpen egy gyűjtőfogalom, amivel két dolgot együttesen jelölünk.

A (3) és (4) megadja, hogy a változók két fajtája pontosan mi. Itt arra hívjuk fel azonnal a figyelmet, hogy a (4)-ben a metazárójeles fogalmak mellett a jobb oldal láncában két olyan jelet is láthatunk (a szögletes zárójelek), ami a definiált nyelvnek valóságos jele; ezeket a jeleket nem tesszük zárójelbe.

A (3) és az (5) összevetésénél láthatjuk azt, hogy a fogalmak megfeleltetése nem kétirányú azonosítás.

A (6)-ban az (1)-hez hasonló rekurzív definíciót láthatjuk szintén.

Az ALGOL olyan mesterséges nyelv, amely arra szolgál, hogy egy feladat megoldását írjuk le vele. Egy feladat megoldását tartalmazó ALGOL szöveget programnak neveznek. A metanyelv felhasználásával szeretnénk a program szerkezetéről némi felvilágosítást adni:

⟨program⟩ : : = ⟨összetett utasítás⟩ | ⟨blokk⟩ (1)
⟨összetett utasítás⟩ : : = *begin* ⟨összetétel vége⟩ (2)
⟨összetétel vége⟩ : : = ⟨utasítás⟩ *end* | ⟨utasítás⟩; ⟨összetétel vége⟩ (3)
⟨blokk⟩ : : = ⟨blokkfej⟩; ⟨összetétel vége⟩ (4)
⟨blokkfej⟩ : : = *begin* ⟨deklaráció⟩ | ⟨blokkfej⟩; ⟨deklaráció⟩ (5)
⟨utasítás⟩ : : = ⟨feltétel nélküli utasítás⟩ | ⟨ciklus utasítás⟩ |
| ⟨feltételes utasítás⟩ (6)

⟨feltétel nélküli utasítás⟩ : : = ⟨értékadó utasítás⟩ | ⟨átirányító utasítás⟩
| ⟨eljárás utasítás⟩ | ⟨program⟩ (7)

Noha itt is van néhány olyan fogalom, ami még nincs definiálva, mégis, ha ezeket a kezdőbetűjük és esetleg egy szám által tekintjük meghatározottaknak, akkor például a következőt egy programnak fogadjuk el:
begin d2; d3; é2; e2; *begin* é3; é4; é5 *end*; á2; e3 *end*

Definíciónkban feltűnők az itt dőlten szedett szavak. Ezek az ALGOL 60-ban egyetlen jelnek felelnek meg. (Az ALGOL 68 ezt azzal is hangsúlyozza, hogy a fenti *begin* és *end* jelölésére megengedi a kerek kezdő és vég zárójelet. Az ALGOL 60 is ismer ilyen, a teljes szinonímiához hasonlítható jelenséget, például a logikai műveleti jeleknek kétféle jelölését engedi meg: $\wedge$ ill. *and*, $\vee$ ill. *or*.)
Ezek után egy valódi program álljon itt példának:

```
begin
  integer i,min,max,n;
  real m1,m2;
  read(n);
  begin
    array toemb[1:n, 1:2];
    read(toemb);
    m1:=m2:=0.0;
    min:=max:=1;
    for i:=1 step 1 until n do
      begin
        m1:=m1+toemb[i,1]×toemb[i,2];
        m2:=m2+toemb[i,2];
        if toemb[min,1]>toemb[i,1]
          then min:=i
          else
            if toemb[max,1]<toemb[i,1]
              then max:=i
        end;
      m1:=m1/m2;
      print(m1,toemb[min,1],toemb[max,1])
    end
end
```

A közölt programban jól kivehetők az egyes „mondatok" (az utasítások és a deklarációk), minthogy a definíciónk szerint a pontosvessző mindig utasítások és deklarációk után áll, vagyis azok egy-egy pontosvesszővel fejeződnek be. Ugyanakkor a definíciónkból az is kitűnik, hogy bizonyos helyen az utasítást nem a pontosvesszővel kell lezárni, hanem az *end* alapjellel. Például a 14—18. sorok egyetlen utasítást (feltételes utasítás) jelentenek, amit *end* zár le. Ugyanekkor ennek az *end*-nek a végén is áll pontosvessző, ami látszólag semmit sem zár le. Ha azonban a definíciónkat figyelembe vesszük, akkor látjuk, hogy minden *end* előtt kell egy *begin*-nek is szerepelnie, és az így meghatározott egység összetett utasítás, vagy blokk

lehet. A példa programunkban a 11. sorban levő *begin*-től a 19. sorban levő *end*-ig egyetlen összetett utasítást láthatunk, amiben a már említett feltételes utasításon kívül két értékadó utasítás is van. Maga a program egy blokk, ami az első sorban levő *begin*-nel kezdődik, és az utolsó sorbeli *end* zárja. A programon belül még egy blokk található: kezdődik az 5. sorban, és az utolsó előtti sorban fejeződik be. (A program tartalmilag kívülről megadható számú számpárt olvas be; ezeket mint egy kísérletsorozat mért eredményeit, illetőleg az eredmények súlyszámait kezeli és meghatározza a mérési adatok súlyozott számtani átlagát, valamint a legkisebb és a legnagyobb mérési adatot.)

2. Magyar természetes nyelven írt szövegek elemzése

2.1. Fonéma-szint

Erről a kérdésről másutt részletesebben írunk (Jékel—Papp 1974, Jékel 1973). Ezért itt elég, ha csak a következőket jegyezzük meg.

Amikor egy minimálisan előszerkesztett (ennek fogalmáról l. alább) magyar szöveget kapott a gép, akkor a programnak a következő fő feladatokat kellett megvalósítania ahhoz, hogy az írott szöveg jó közelítéssel a fonémák szintjét tudja reprezentálni: (a) meg kellett találni az első nem nulla jelet (a szalag elejét); (b) a szöveget meg kellett tisztítani a nem betű jelektől (ezekről az alábbiakban, l. **2.2.**, még bővebben lesz szó: ilyenek például az írásjelek, a pont, a vessző stb., amelyek nem képviselnek semmiféle fonémát); (c) az egy- és többjegyű betűket azonosítani kellett; (d) az így azonosított fonémákat gyűjteni kellett, bizonyos statisztikázást kellett velük kapcsolatosan elvégezni; (e) meg kellett találni a szöveg végét; (f) megfelelő (táblázatos) formában ki kellett írni az eredményeket. E műveletek közül a (c) a tulajdonképpeni analízis („azonosítani": épp ez jelentette azt, hogy a „természetben adott" leütéseket helyettesítette a program az egyes fonémákkal, e fonémák valamilyen módon leírt nevével), az (a)—(b) ezt lehetővé tevő előkészítés, a (d)—(f) már az analízis egy konkrét gyakorlati felhasználása, semmiképpen nem tartozik magához az analízishez. Hasonló előkészítés — tulajdonképpeni analízis-felhasználás (pl.: költői szóalaktár összeállítása) megy végbe a szövegszó szinten is, l. **2.2.**

A minimális előszerkesztésen azt értettük, hogy a lyukasztásra való előkészítés során a filológusnak meg kellett jelölnie mindazokat a helyeket, ahol a hagyományos helyesírási alakról a fonématikus megközelítésre való áttérés nem volt könnyen (memória igénybevétele nélkül) algoritmizálható. Két fő eset volt ezen belül: (a) az idegen nevek és hagyományos magyar nevek átírandók voltak „magyaros" alakra, így: *Volter, Sekszpír, eszpresszó, Battyányi, Verbőci;* (b) a szótár nélkül automatikusan nem szegmentálható betűsorok közé egy „/" (törtvonás) volt beiktatandó ott, ahol az adott szövegszóban a szegmentálást megkívántuk: *egész/ség, arc/szín, víz/sugár* stb. Az így előkészített szöveg csak megközelítőleg adott fonématikus eredményt, mert az egyedi esetek jelöletlenek maradtak. Például az *utca* lexéma minden szövegszóban *t-c* fonémákat mutat (a helyes hosszú *cc* helyett), a *tetszik t-sz* fonémákat (úgyszintén a hosszú *cc* helyett) stb. Kisebb mintákon való manuális utánaszámolás azt mutatta, hogy a közelítés elég finom

azon műveletek számára, amelyeket végezni kívántunk: nagyságrendekkel kisebb volt az eltérés a valódi fonémaalaktól annál, mint ahol a következtetéseket levontuk.

Lényeges megjegyeznünk, hogy e munkálat során már megkülönböztettük a szövegek különféle darabjait (szegmenseit). Éspedig (tekintettel arra, hogy eredetileg Ady-köteteket dolgoztunk fel): a szöveg egésze volt az életmű (ezt nem is jelöltük külön sehogy), ezen belül jelölve: kötet, ciklus, vers. A kötet állhat egy vagy több ciklusból; a ciklus egy vagy több versből. A különféle statisztikai összesítéseket e különféle szinteken kérhettük.

3. Magyar természetes nyelven írt szövegek elemzése

2.2. Szövegszó-szint

Bemenetként a **2.1.** alatt említett, minimálisan előszerkesztett magyar természetes nyelvi szövegeket tekintettük. Ahhoz, hogy e szövegek egyáltalán elemezhetők legyenek, az alább leírandó módon kellett egységeiket definiálnunk. Megjegyzendő, eredendően nehezebb volt a helyzetünk, mint az ALGOL esetében azért, mert az ALGOL eleve úgy készült, hogy egyértelműen lehessen tagolni, az ALGOL konvencióit az azt létrehozó matematikusok állították elő; ugyanakkor valamely természetes nyelv helyesírási, központozási stb. konvenciói hosszú történeti fejlődés eredményeképpen jöttek létre, a nyelvészek ebbe a fejlődésbe csak viszonylag későn szóltak bele s befolyásuk még akkor sem lehetett olyan abszolút, mint az ALGOL szuverén létrehozóié. (A magyar helyesírást nem Simonyi; az orosz helyesírást nem Fortunatov és Sahmatov hozta létre.)

A l a p j e l e k:

⟨alapjel⟩ : : = ⟨betű⟩ | ⟨kvázi betű⟩ | ⟨szóköz⟩ | ⟨mondatvég⟩ | ⟨törlő⟩ | ⟨gépjel⟩ | ⟨természetes szám⟩

⟨betű⟩ : : = a | á | ä | b | c | d | e | é | f | g | h | i | í | j | k | l | m | n | o | ó | ö | ő | p | q | r | s | t | u | ú | ü | ű | v | w | x | y | z | A | Á | B | C | D | E | É | F | G | H | I | J | K | L | M | N | O | Ó | Ö | Ő | P | Q | R | S | T | U | Ü | V | W | X | Y | Z

⟨kvázi betű⟩ : : = - | ” | / | : | ,

⟨szóköz⟩ : : = sp | nl | cr | lf | (|)*

⟨mondatvég⟩ : : = ; | · | ! | ?**

⟨törlő⟩ : : = =

⟨gépjel⟩ : : = + | Ft***

* E rövidítések értelme: sp = 'space' (szóköz), nl = 'new line' (új sor), cr = 'carriage return' (kocsi vissza), lf = 'line forward' (soremelés). Ezeknek a pusztán metanyelvi kifejezéseknek a protokollon természetes nyomuk van: köz az egyes szavak között, új sor stb.; a szalagon egy-egy sajátos és csak nekik megfelelő lyukkombináció áll ezen a helyen.

** A „;” csupán jelen feldolgozásunkban lett mondatvéget jelölő elem, mert a szövegszó környezetét elég volt, ha a legelső pontosvesszőig (a legelső pontosvesszőtől) vettük figyelembe.

*** A „Ft” jelet az adatrögzítő egyetlen billentyű lenyomásával (mintegy ligatúraként) viszi a protokollra; a szalagon egyetlen meghatározott lyukkombináció felel

⟨természetes szám⟩ : : = ⟨az **1.** alatti példában definiált!⟩

Szó és írásjel definíciója:

⟨szóelem⟩ : : = ⟨betű⟩ | ⟨szóelem⟩ ⟨betű⟩ | ⟨szóelem⟩ ⟨kvázi betű⟩
⟨törölt szó⟩ : : = ⟨szóelem⟩ =
⟨szó⟩ : : = ⟨alapszó⟩ | ⟨végszó⟩
⟨alapszó⟩ : : = ⟨szóelem⟩ ⟨szóköz⟩ | ⟨alapszó⟩ ⟨írásjel⟩
⟨végszó⟩ : : = ⟨mondatvégszó⟩ | ⟨fejezetvégszó⟩ | ⟨szalagvégszó⟩
⟨mondatvégszó⟩ : : = ⟨szóelem⟩ ⟨mondatvég⟩ | ⟨alapszó⟩ ⟨mondatvég⟩ | ⟨mondatvégszó⟩ ⟨mondatvég⟩ | ⟨mondatvégszó⟩ ⟨írásjel⟩ ⟨mondatvég⟩
⟨fejezetvégszó⟩ : : = ⟨szóelem⟩ + | ⟨mondatvégszó⟩ +
⟨szalagvégszó⟩ : : = ⟨szóelem⟩ **Ft** | ⟨mondatvégszó⟩ **Ft**
⟨írásjel⟩ : : = ⟨szóköz⟩ | ⟨írásjel⟩ ⟨kvázi betű⟩ | ⟨írásjel⟩ ⟨írásjel⟩

Ennek alapján a szó úgy van definiálva, hogy szót jelent bármely, olyan, betűvel kezdődő, betűket és kvázi betűket tartalmazó jelsorozat, melyet szóköz jel, mondatvég jel vagy gépi jel zár le. A szó végén lehet tetszőleges számú írásjel is, de ez nem teljesen azonos a szokásos írásjellel, minthogy az írásjel itt szóközzel kezdődő kvázi betűkből és szóközökből álló tetszőleges sorozat. A törölt szó az adatrögzítés során hibásan leírt szavak megjelölésére szolgáló jelölésforma, mellyel a gépet arra lehet utasítani, hogy az így jelzett szavakat a feldolgozáskor hagyja figyelmen kívül.

Mondat definiálása:

⟨mondattörzs⟩ : : = ⟨alapszó⟩ | ⟨mondattörzs⟩ ⟨alapszó⟩
⟨mondat⟩ : : = ⟨alapmondat⟩ | ⟨végmondat⟩
⟨alapmondat⟩ : : = ⟨mondatvégszó⟩ | ⟨mondattörzs⟩ ⟨alapmondat⟩
⟨végmondat⟩ : : = ⟨fejezetvégmondat⟩ | ⟨szalagvégmondat⟩
⟨fejezetvégmondat⟩ : : = ⟨fejezetvégszó⟩ | ⟨mondattörzs⟩

---

meg neki (tehát nem a *F* és a *t* lyukkombináció). A „+"-ot és a „Ft" ligatúrát igen fontos metajelekként azért használhattuk, mert folyó szövegekben úgy vettük, hogy nem fognak előfordulni (matematikai szövegeket, ahol a „+" jel a maga természetes funkciójában bőven előfordulhat, egyelőre nem szándékoztunk feldolgozni). Nagyon lényeges megérteni e jelek metajel természetét a szövegben. Ha, ami a „+" jel értelme nálunk, valami ilyesfélét írnánk helyette: „itt fog kezdődni egy új vers", vagy „verskezdet", vagy ehhez hasonlót, akkor e kifejezések bármelyikét hiába tennénk, mint itt tettük, idézőjelbe: a gép reális szavaknak érzékelné őket, megállapítaná a fonéma összetételüket (más programjainkban a hangrendjüket) és így tovább, vagyis elrontaná velük a valóban következő tárgynyelvi szöveg adatait; másrészt nem érzékelné, hogy itt éppen őhozzá, a géphez szóltunk, őt akartuk valamire utasítani, vagyis, hogy a tárgynyelvi szövegbe metanyelvi elem épült e helyen. Hasonló módon a „Ft" ligatúra helyett nem lehetett volna ilyesféléket írni, e jel jelentésének megfelelően: „itt a szöveg vége", „állj meg és lass hozzá a statisztikai összesítésekhez, mert vége a szövegnek" stb. Vö. ezzel részben az ALGOLban dőlten szedett metanyelvi kifejezéseket, utasításokat; másrészt és különösen Jakobson rendkívül világos megkülönböztetését a kód és az üzenet különböző kombinációiról (1956): a „+" és a „Ft" rendszerünkben, Jakobson terminológiájával élve, M/C típusú üzenet.

⟨fejezetvégmondat⟩
⟨szalagvégmondat⟩ : : = ⟨szalagvégszó⟩ | ⟨mondattörzs⟩
⟨szalagvégmondat⟩

Fejezet definíciója:

⟨fejezet⟩ : : = ⟨alapfejezet⟩ | ⟨végfejezet⟩
⟨alapfejezet⟩ : : = ⟨fejezet eleje⟩ ⟨fejezetvégmondat⟩
⟨végfejezet⟩ : : = ⟨fejezet eleje⟩ ⟨szalagvégmondat⟩
⟨fejezet eleje⟩ : : = ⟨természetes szám⟩ | ⟨fejezet eleje⟩ ⟨alapmondat⟩

A fejezetek elején álló természetes számok a fejezetek (versek) megjelölésére szolgálnak.

Az egyszerre lelyukasztott szöveget tartalmazó szalag definíciója:

⟨szalag⟩ : : = + ⟨szalagvég⟩
⟨szalagvég⟩ : : = ⟨végfejezet⟩ | ⟨alapfejezet⟩ ⟨szalagvég⟩

E szintaxis alapján rögzített szövegeket a gép hasonló módon értékeli, mint a nyelvet ugyan nem ismerő, de az írás jelrendszerével tisztában levő ember tenné. Tekintettel az európai alapú írásra épülő rendszerek e szempontból rokon voltára, rendszerünk természetesen távolról sincs csupán a magyarhoz kötve, hanem eléggé univerzálisnak tekinthető. Az eddigi természetes nyelven írt szövegek feldolgozási kísérleteinek szükségszerűen vagy kellett (esetleg implicit formában, nem ilyen rendszeresen kifejtve) tartalmazniuk ezt a rendszert, vagy fiaskót kellett vallaniuk már ezen az egészen kezdeti stádiumon, illetőleg ismeretlen okokból — a valóságban e stádium kidolgozatlan voltánál fogva. Az egész rendszer még a legpozitivistább iskolán felnőtt filológus számára is túlságosan bonyolultnak és pontoskodónak tűnhet. A valóság azonban az, hogy a rendszer egy cseppet sem bonyolultabb annál, mint amit az eddig vizsgált magyar természetes nyelvi szövegek megkívántak. Néhány példa: (a) ha a „szó" fogalmát lazábban definiáljuk, akkor azt nyerjük, amit mi is kaptunk első kísérleteinkkor: a „—" (gondolatjel) külön szónak számíttatott és besoroltatott ennek megfelelő „betűrendes helyére"; (b) ha a „mondat" fogalmát lazábban definiáljuk, akkor például a „. . ." („három pont") esetében két-két pont között egy-egy mondatot fog a gép érzékelni és nem fog tudni azzal mit kezdeni; (c) miután a „mondat" fogalmát a (b) alatt mondott miatt finomabban definiáltuk, teljesen hibás eredményeket kaptunk például Szabó Lőrinc „A huszonhatodik év" szonettciklusa egyes szonettjeiben, nevezetesen olyankor, amikor egy pontértékű írásjel után nem egy pontértékű másik írásjel és nem egy igazi betű következett, hanem például egy idézőjel: a program emiatt teljes mondatokat szétdobott, belenyúlt jó mondatokba, kihagyott szavakat stb.

Miként a **2.1.** alatt röviden érintett „betű—fonéma"-azonosító program, úgy ez a program is bizonyos konkrét célokat is szolgált, sőt megfordítva: az itt vázolt elméletet azért voltunk kénytelenek kidolgozni, hogy a gép

meg tudja oldani a következő problémát. Szövegről kívántunk ún. KWIC [**K**eyword **i**n **C**ontext] indexeket készíteni szerzői, költői stb. szóalaktárak gépi előállítása céljából. Egy KWIC-index lényege az, hogy valamely dokumentumból kiválogatja az abban megjelölt szövegszókat, ábécébe rakja őket és meghatározott nagyságú bal és jobb környezetükkel, valamint előfordulásuk pontos helymeghatározásával együtt listán közli őket. (Mi nyelvészeti meggondolásokból minden egyes szövegszót megjelöltnek tekintettünk, tehát az olyan formaszókat is, mit az *a, az, és* stb., melyeket bármely bibliográfiai és információ-feltáró munka során természetesen ki kellene hagyni.) Itt e munkálatból csupán a szövegszószintű analízis szempontjából lényeges momentumokat ragadjunk ki: (1) a szó fenti körmönfont definíciójára éppen ezért volt szükség, hogy a gép az ezen elméletre épülő program alapján minden egyes szót megtaláljon és csak a szavakat találja meg (tehát ne találja például a gondolatjelet önálló szónak); (2) a megtalált szövegdarabokat (szavakat) egy bizonyos nagyság szerinti rendbe — betűrendbe — kellett állítania. E betűrendbe állítás egyrészt bonyolultabb volt, mint a szótárak hasonló rendezései, hiszen figyelembe kellett vennünk a kvázi betűket és mindenféle egyéb jeleket is; ezeknek is kellett sorrendi értéket tulajdonítanunk. A mi teljes ábécénk: sp nl cr lf () ; . ! — " / = : , a á ä b stb.; tehát előbb jön az a szövegszó, amely „space"-re végződött (pl.: *alma* sp), majd az, ahol ugyane szóalak sor végén állt (vagyis: *alma* nl, majd *alma* cr stb.); ezt követték ugyane szóalak különféle írásjelekkel megtoldott változatai; csak ez után jöttek azok az esetek, immár majdnem a hagyományos sorrendben, ahol az *alma* szóalakhoz valamely betűvel kezdődő elem járult space nélkül (pl. *almabor, almaként)*. Másrészt, és ezért volt sorrendünk még a betűk között is csak majdnem hagyományos: nem tudtuk elfogadni a HSzab. különféle raffinált utasításait a szoros ábécérendbe szedésre vonatkozóan éppen úgy, mint ahogy nem fogadhattuk el a betű hagyományos magyar felfogását sem (egy betű = ami egy fonémát jelöl). A sorrend nálunk: *eggyé egres egy egyet-mást egyetlen egzakt; meccs mecénás mecset messiás messze messze-messze mester meszel meszes mész* stb. Az eddig már sokszorosítva, nyomtatásban stb. elkészült (pl. az OMKDOK kiadásában megjelent) KWIC-indexek ugyanezt a gyakorlatot követik. A **2.1.** alatti előzetes munkánk eredményeképpen nekünk nem okozott volna különösebb nehézséget a hagyományos ábécérendbe állítás sem. De egyszerűen nem tudtunk vele egyetérteni és ezért nem kívántuk alkalmazni. Egészen biztosak vagyunk benne, hogy a „gépkorszak" gépi úton előállított nyelvi-nyelvészeti termékei, a mi KWIC-indexeinkhez — költői szóalaktárainkhoz hasonlóan pontosan így, az általunk is alkalmazott ábécérend szerint fognak készülni.

Kossuth Lajos Tudományegyetem,
Orosz és Szláv Nyelvészeti Tanszék
és
Egyetemi Számítóközpont,
Debrecen

Irodalom

R. Jakobson, Shifters, verbal categories and the Russian verb (1956): *Selected Writings* II, 130—147 (1971).

Jékel Pál: *Magyar nyelvű szövegek számítógépes feldolgozása — 1. (Fonéma-szint).* (Egyetemi doktori értekezés, Debrecen, 1973).

Jékel Pál—Papp Ferenc: *Ady Endre összes költői műveinek fonémastatisztikája.* (Bp., 1974).

Papp Ferenc, Filmszinkronizálás és szemiotika: *Rádió és Televízió Szemle* 1/4. 83—97 (1969).

A. A. Zaliznyak—Vjacs. Vsz. Ivanov—V. N. Toporov, О возможности структурно-типологического изучения некоторых семиотических систем: *Структурно-типологические исследования* (Moszkva, 1962) 134—143.