# A szóról és a szófajokról (a számítógépes nyelvfeldolgozás kapcsán)

Bibok Károly

Szegedi Tudományegyetem, Orosz Filológiai Tanszék
6722 Szeged, Egyetem u. 2.
kbibok@lit.u-szeged.hu

Jelen cikk a számítógépes nyelvfeldolgozásnak három, a morfoszintaktikai elemzéssel kapcsolatos kérdéskörét vizsgálja. Ezek a következők: 1. a szövegszavakra bontás (tokenizálás) nehéz esetei, 2. a szövegszavak és a morfoszintaktikai elemzés bemenetének viszonya, valamint 3. a hagyományos szófajok közé be nem illeszthető szavak osztályainak létrehozása.

## 1 Bevezetés

A Szegedi Tudományegyetem Informatikai Tanszékcsoportja és a MorphoLogic Kft. 2000–2002-ben elkészítette a Szeged Korpusz első változatát (http://www.inf.u-szeged.hu/lll), amely különböző nyelvhasználati területekről összesen 1 millió szövegszónyi, morfoszintaktikailag elemzett (MSD szerint kódolt) és egyértelműsített magyar nyelvű szövegeket tartalmaz [1]. A korpusz előállítási munkálatai közben egy sor olyan szövegfeldolgozási problémával találkoztunk, amelyeknek teljes körű megoldását akkor nem vállalhattuk fel. Ezek közül fogok jelen cikkemben hármat tüzetesen megvizsgálni.

## 2 Mi a szó?

A nyelvészetben a szó meghatározásával kapcsolatban lehetséges az az álláspont, hogy nem a szót általában, hanem részfogalmait definiáljuk. A nyelvi részrendszereknek megfelelően beszélhetünk fonológiai, morfológiai, szintaktikai és lexikai szóról [2: 76–79]. Egy másik szempontot követve, a szövegszót, a morfoszintaktikai szót és a lexémát különíthetjük el [6: 190–191]. Itt most részletesen nem hasonlíthatom össze a terminus technicusok fenti két halmazát, de a következőket szükséges megjegyezni. A *lexéma* helyett használható a *lexikai szó*. A *morfoszintaktikai szó*-ban nem válik külön a morfológiai és a szintaktikai jelleg. A *szövegszó*-nak nincs megfelelője az első felsorolásban, mert a szövegszó nem a nyelv (langue), hanem a beszéd (parole) egysége. Mivel azonban a számítógépes

nyelvfeldolgozás szövegekkel operál, számunkra mégis ez képezi a kiindulópontot, mégpedig úgy, ahogy a nyelv írott formájában, az írásbeli szövegekben megvalósul.[1]

Kezdjük Papp Ferenc következő meghatározásával: a szövegszó az írás szemszögéből a szövegnek két szóköz határolta darabja [7: 76]. Ha el is tekintünk attól a partikuláris problémától, hogy a szöveg kezdő és végső eleme nem szóközök között áll, marad egy sokkal általánosabb nehézség, amelyre maga Papp Ferenc is utal. Ha egy számítógép az előbbi definíció alapján végezné a szöveg szegmentálását, akkor bizonyos szövegszavak elején és végén írásjeleket, pl. idézőjelet, vesszőt, felkiáltójelet, találnánk, mert ezek szóköz nélkül kapcsolódnak az első, ill. az utolsó betűhöz [7: 77]. Ezért teljesebb Papp Ferencnek az a meghatározása, amely szerint a szövegszó a szövegnek az a részlete, amely két szóköz között helyezkedik el, leszámítva az írásjeleket [6: 190]. Sokszor azonban még ez a meghatározás sem vezet a kellő eredményre és módosításra szorul a szövegszavak két csoportja miatt.

Az első csoportba azok a szövegszavak tartoznak, amelyeknek **a szélein le nem választható/választandó írásjel(ek)** van(nak). Mielőtt rátérnék ezekre a szövegszavakra, megjegyzem, hogy az ilyenek feltételezése egyáltalán nem ellentétes a definíció magját képező „a szövegnek az a részlete" kitétellel, vagyis nincs kikötve, hogy a szövegszavaknak csak betűkből kell állniuk. Tehát az első csoportba sorolhatók a következő esetek. Először, bizonyos rövidítések végére és az arab vagy római számmal írt sorszámnevek[2] után pontot teszünk. Másodszor, az olyan esetben, mint pl. *írásjel(ek)*, a záró zárójel leválasztása hibás elemzést adna (vö. még: *„Hét évszázad magyar versei"-ben*). Harmadszor, a *majd'; bel- és külkereskedelem; gépgyártó, -szerelő és -javító üzem;* stb. kifejezésekben az írásjelek elhagyása megtévesztő lenne. Negyedszer, az egyéb, nem a központozást, tagolást szolgáló írásjeleket sem kell levágni (pl. *a), 10%, 3°, °C, +2, –2*).

A másik csoportot, amely miatt pontosítani kell a fenti definíciót, a **belsejükben írásjelet tartalmazó szövegszavak** alkotják. Ahogy – a fentiekkel összhangban – arab vagy római számmal írt számok szerepelhetnek szövegszóban, ill. szövegszóként, és lehetnek a szövegszavak elején és végén írásjelek, úgy a definíció nem zárja ki azt sem, hogy a szövegszó belsejében írásjel legyen.[3] A kérdés az: Vonatkoztassuk-e ezekre az írásjelekre is a definíció azon részletét, hogy „leszámítva az írásjeleket"? A számítógépes korszak előtt született szövegek szavainak belsejében leggyakrabban a kötőjel fordul elő. A kötőjelnek, mint ismeretes, több funkciója lehet: használhatjuk összetett szavakban, az *-e* kérdőszós szavakban, toldalékok kapcsolására és a szavak sorvégi elválasztására is. Az utóbbi esetben – a központozási szerepet játszó írásjelekhez hasonlóan – nem tekinthető a szövegszó részének. Nyilvánvalóan az nem oldja meg a problémát, ha minden sorvégi kötőjelet eltüntetünk a szövegünkből, hacsak- nem előzőleg különböző gépi kötőjeleket alkalmazva rögzítettük a szöveget (a kettőzött kétjegyű betűkkel még ekkor is vigyázni kell). De mi a helyzet a számítógépes korszak szövegeiben előforduló szóbelseji írásjelekkel (l. pl. URL- és e-mail-címek)? Úgy gondolom, ugyanúgy kell

[1] A szakirodalomban a *szövegszó*-n kívül találkozhatunk a *szóelőfordulás* terminussal [5: 26], illetve a nyelv írott formájában előforduló szövegszót hívják ortográfiai szónak is [8: 1058].

[2] A szöveg részeként szereplő szám – a definíció szerint – szintén szövegszónak tekintendő.

[3] Azt a megszorítást azonban megtehetjük, hogy a szövegnek két szóköz közötti azon részlete, amely csak írásjelet (nevezetesen egy gondolatjelet) foglal magában, nem számít szövegszónak.

eljárni, mint a nem elválasztást szolgáló kötőjel esetében. Az ilyen írásjeleket nem kell kiiktatni a szövegszavakból. Ez azt is jelenti, hogy azt sem tartom elfogadható megoldásnak, ha a (szóköz nélküli) írásjelek mentén szétdarabolnánk őket. (Még szóköz nélküli pont esetén sem kell ezt tennünk.) Ugyanis nem kapnánk ezáltal a morfoszintaktikai elemzés számára értelmezhetőbb kifejezéseket. Ha figyelembe vesszük, hogy az újabb keletű, szóbelseji írásjeleket tartalmazó szövegszavak is elláthatók névszói ragokkal, akkor a sporthírekben olvasható olyan időeredményekre, mint pl. *1:20:36.7, 4:01,95* szintén igaz az a fenti kijelentés, hogy a szétdarabolás nem ad jobb elemzést.

## 3 A morfoszintaktikai elemzés bemenete

Prószéky Gábor szerint „az ortográfiai szavak definíciója egyértelmű definíciót ad a számítógépes morfológia bemenetére" [8: 1058]. Én azonban úgy vélem, hogy a szöveg szegmentálása útján kapott szövegszavak nem minden esetben felelnek meg a morfológiai elemzés, ill. a morfoszintaktikai kategóriák szerinti elemzés számára. Vannak olyan **szövegszavak, amelyeket szét kell bontani,** és vannak olyanok, **amelyeket egybe kell vonni a morfoszintaktikai elemzés előtt.** Másként megfogalmazva, egy szövegszó magában foglalhat két morfoszintaktikai szót, vagy több szövegszó tesz ki egy morfoszintaktikai szót.

**Az első esetre** két példát hozok:

1. Az *-e* kérdőszós szövegszavakban előforduló *-e* kérdőszót csak akkor tudjuk külön kategóriaként kódolni (vö. az ÉKsz.[2]-ben: határozószó (simuló kérdőszócska) [9]), ha az ilyen szövegszavakat egy újabb szegmentálás során kettévágjuk.
2. A sporthírekben gyakori kifejezés az olyan, mint pl. *Dortmund (német)-Barcelona (spanyol),*[4] amelyből a *(német)-Barcelona* a 2.-ben mondattak alapján egy szövegszó. Mivel ehhez nem tudunk „értelmes" kódolást rendelni, itt is először újabb szegmentálásra van szükség: három részre osztjuk, ugyanis a kötőjel is külön elem.

**A második esetre** jóval több példát tudok mutatni.

1. A több szövegszóból álló tulajdonnevek: pl. *Szegedi Tudományegyetem, New York-i.* Az utóbbi különösen jól szemlélteti a többtagú tulajdonnevek egyes tagjainak egybetartozását, hiszen a *York*-ból képzett *-i* képzős alak a *yorki.*[5]
2. Többtagú rövidítések: *i. sz., i. e., i. m.* stb.
3. A többtagú számnevek:

a) Hátulról számított hármas számcsoportok szerint szóközzel tagolt (arab) számok. Gondoljuk meg, milyen eredményre vezetne, ha a *3 000 000*-t mint három szövegszót elemeznénk morfoszintaktikailag. Ahhoz, hogy a *000*-ról értelmes dolgot mondhassunk, fel kell tételeznünk egy morfoszintaktikai osztályt a számok részeit képező számok számára. De ennek alapjául csak az az írásgyakorlat szolgálna, amely szerint a számmal írt számneveket hármas számcsoportokba tagoljuk.

---

[4] Szövegeinkben sajnos nincs különbség kiskötőjel és nagykötőjel között.

[5] Ezen az alapon a *New York-i,* ha nem is egy szövegszó, egy ortográfiai szónak tekinthető.

b) Az (arab) számokat és az *ezer, millió* stb. szavakat tartalmazó számnevek. Ha ezeket nem (külön írt) összetételeknek,[6] hanem jelzős szerkezeteknek tartanánk, akkor a csak betűvel, valamint a csak számmal írt számnevektől eltérően kezelnénk őket.

4. Kötőjeles kifejezés egyik eleme többtagú, pl. *Ferencváros-Vác FC.*
5. A vagylagosságot kifejező „/" jellel összekapcsolt kifejezés egyik eleme többtagú, pl. *főnevek/főnévi csoportok.*
6. Elmaradó közös tagú kifejezések, pl. *bel- és külkereskedelem; gépgyártó, -szerelő és -javító; Tömörkény- és Gárdonyi-szerű,* kivéve *Tömörkény- és később Gárdonyi-szerű.* Ezekben az esetekben tulajdonképpen a morfológiai és szintaktikai tulajdonságok ütközéséről van szó. A *belkereskedelem* mint összetétel egy morfológiai szó, de szintaktikailag nem egységként viselkedik: az elő- és az utótag elválhat egymástól a mellérendelés során. Ez a kettős jelleg lehetővé teszi, hogy kétféleképpen adjunk számot a *bel- és külkereskedelem*-ről. Vagy besoroljuk a morfoszintaktikai kategorizálás révén a főnevek közé, nem törődve a mellérendeléssel, amely – az alárendelő szerkezetekkel szemben – a szintaxisnak is másodlagos vizsgálati tárgya. Vagy szintaktikai szerkezetként kezeljük és feltételezünk egy, a szélein kötőjelet tartalmazó morfológiai osztályt. Erre az utóbbi megoldásra is van lehetőség a 4.-ben vázolandó klasszifikációs rendszer keretében. Ugyanis a valószínűleg nagyon ritkán előforduló *Tömörkény- és később Gárdonyi-szerű* és az ehhez hasonló kivételek miatt mindenképpen szükséges lesz bevezetni az elöl és/vagy a végükön kötőjelet tartalmazó szavak morfológiai osztályát (Hatvani Csaba, személyes közlés).

## 4 Hány szófaj van?

Itt nem a szófajok definíciós kritériumairól és a szófajok elhatárolásáról folytatott vitáról kívánok szólni. Hadd kezdjem annak leszögezésével, ami már a 3.-ból is kitűnhetett. A morfoszintaktikai, ill. a morfológiai és a szintaktikai szó mögött a lexikai szó (lexéma) és a termékeny módon alkotott létező vagy potenciális szó húzódik meg (esetleg még számolhatunk a mellérendeléssel is).[7] Ezek pedig morfoszintaktikai osztályokba sorolhatók. Ugyanakkor az is nyilvánvaló a fentiekből, hogy ha egy korpusz különböző nyelvhasználati rétegekből tevődik össze, akkor a **hagyományosan szófajokon túl** és az eddig alkalmazott MSD-kódrendszert kibővítve **újabb osztályokat kell létrehozni,** hogy a klasszifikációt maradéktalanul elvégezhessük (vö. számítógépes szaknyelvbeli URL- és e-mail-címek vagy a sporthírekbeli számok és írásjelek kombinációjából álló meccs- és időeredmények). Mielőtt a nyolc javasolt osztályt és alosztályaikat bemutatnám, előrebocsátok három megjegyzést. Először, még ha – ugyanúgy, mint a hagyományos szófajok esetében – formális meghatározásra törekszünk is, nem kerülhető el a jelentésre való hivatkozás

[6] Abban, hogy ezeket összetételeknek minősítjük-e, nincs jelentősége a különírásnak. Vö.: ha a szám helyett is betűt használunk, akkor egybeírjuk az ilyen számnevet.

[7] A létező és a potenciális szavak közötti különbségtételről l. [3: 148].

sem.[8] Másodszor, az új „szófajok" – az önállóan használt toldalékok osztályának kivételével (pl. a *-ság* a *-ság képzős főnevek* kifejezésben) – nyitott osztályok, azaz elemei korlátlanul szaporíthatók, például a szóalkotáshoz hasonló módokon.[9] Harmadszor, mivel eddig az osztályok megnevezéseinél a mnemonika és a magyarul nem beszélő felhasználók szempontjai részesültek előnyben, csak körülíróan és példákkal szemléltetve tudom bemutatni a **javasolt osztályokat és alosztályaikat.**

1. Elektronikus:
a) e-mail-címek, pl. *A bubo@doktor.hu címről...,*
b) webhelyek, pl. *A www.huninet.hu-ról...,*
c) számítógépes útvonal, pl. *A C:\CONFIG.SYS nevű...,*
d) fájlkiterjesztés, pl. *A .DOC és a .RTF fájlok...,*
e) egyéb.
2. Számok:
a) (sport)eredmények (meccs-, időeredmények), pl.: *1:20:36.7, 4:01,95, 2:0, 2-0,*
b) (csak) előjelet tartalmazó (egész és nem egész) számok, nem tartozik ide a telefonszámban az ország hívószáma az előtte álló „+" jellel,
c) időpont-megjelölések, dátumok, amelyek szóköz nélküli pontot, kettőspontot, kötőjelet vagy „/" jelet tartalmaznak, pl. *10.35, 10:35* (= *10.35*), *2003-01-06, 2003.01.06., 01/06,* de nem tartozik ide: *1984/85-ben, 1984/1985-ben, 1984/85. (tanév), 1984/1985. (tanév),*
d) (csak) pontot tartalmazó szám, pl. *139.000* (= *százharminckilencezer*), *80.5 MB* (= *80,5 MB*),
e) százalékjelet („%") tartalmazó kifejezések, pl. *(+)10%, 40.2%,*
f) fokjelet („°") tartalmazó kifejezések, pl. *(+)3°,* de nem: *(+)3 °C,*
g) arány, pl. *1653kJ/1000g, 1653 kJ/100 g, 1:10,*
h) méret, pl. *1024x768,* ide tartozik még: *2x* (= *2-szer*),
i) képletek, aritmetikai kifejezések, pl. *2rπ, 2+2 =4, 2 + 2 = 4* (összevonás után),
j) egyéb.
3. Indexként bármilyen karaktert tartalmazó kifejezés:
a) alsó index,
b) felső index, pl. *dpi**-vel, quattro®,* kivéve: $m^2$, $cm^3$, $3°$.
4. Különféle azonosítók, pl. szabvány jelzete, igazolványszám, iktatószám, ügyiratszám, alvázszám, motorszám, rendszám *(ABC-123, ABC123),* ISBN, könyvtári jelzet, *1/a, 1/A, III/1* típusú jelzet (a *3/4, 10/100, 10/1974* törtszámnév is lehet), írásmű részének jelzése (*1.1.2., 1.1.2, a)*), utak, géptípusok jelzése *(E5, M0, MiG-27, T-34, TU-154,* tulajdonnévként kódoljuk, ha a betű- és/vagy számjelzés szóhoz tartozik, pl. *Apollo-11, Boeing XXX, Commodore 64),* telefonszám (*473-1470,* de ha évszám, akkor „sima" számnév), irányítószám („sima" számnév is lehet).
5. Szónál kisebb tokenek, pl. *A -ság képzős főnevek..., A -tól/-től ragos eset...,* de nem ide tartozik: ***bel- és külkereskedelem*** stb.

---

[8] A főnév és az ige formális alapú definiálására l. [4: 148–149, 209–210]. Ugyanakkor a melléknévnek a főnévtől és az igétől való szófaji elkülönítéséhez mindenekelőtt szemantikai kritériumokat kell keresni [4: 181].

[9] A nyitott és a zárt osztályokról a hagyományos szófajok kapcsán l. [2: 95].

6. Kötőjellel kezdődő vagy végződő tokenek az olyan kifejezésekben, mint pl. *Tömörkény- és később Gárdonyi-szerű.*
7. Nem magyar, vagyis idegen nyelvű szavak, ill. kifejezések.
8. Rossz helyesírással írt magyar szavak, amelyek nem homonimak más magyar szavakkal, pl.: *éccaka,* de nem tartozik ide: *aszt* stb.

Fontos kiemelni, hogy ugyanúgy, ahogy a hagyományos szófajoknál, itt is számtalan esetben találkozhatunk **ambiguitási problémával,** azaz azzal, hogy egy egység – a kontextus ismerete nélkül – különböző osztályokba, alosztályokba sorolható be. Ebből a szempontból l. még egyszer a különféle azonosítóknál említett példákat.

## 5 Összegzés helyett

A *Mondatszintaxis gépi tanulása (gépi tanulási módszerek a magyar nyelv szintaktikai szabályainak létrehozására)* c. IKTA-pályázat keretében (№ 37/2002, vezetője: Gyimóthy Tibor) az SZTE Informatikai Tanszékcsoportjában most folyik annak a számítógépes programnak a tesztelése, amelynek segítségével a fentebb vázolt szegmentálási és klasszifikációs problémák automatikusan kezelhetőkké válnak. Továbbá: Alexin Zoltán és Hatvani Csaba közreműködésével elkészítettem azt az útmutatót, amelynek alapján egyetemi hallgatók ellenőrizni tudják a Szegedi Korpusz e programmal történő elemzésének eredményeit és végrehajthatják a manuálisan elvégzendő feladatokat (szövegszavak összevonása és egyértelműsítés) is.

Végezetül megemlítem, hogy a jelen cikkben kifejtettek nemcsak a nyelvfeldolgozás területén, hanem helyesírás-ellenőrző programok tökéletesítésében és szótárak (pl. gyakorisági szótár) készítésében is hasznosíthatók.

## Irodalom

1. Alexin, Z., Csirik, J., Gyimóthy, T., Bibok, K., Hatvani, C., Prószéky, G., Tihanyi, L.: Manually Annotated Hungarian Corpus. In: Proceedings of EACL. Budapest (2003) 53–56
2. Kenesei I.: Szavak, szófajok, toldalékok. In: Kiefer F. (szerk.): Strukturális magyar nyelvtan I. Morfológia. Akadémiai Kiadó, Budapest (2000) 75–136
3. Kiefer F.: A szóképzés. In: Kiefer F. (szerk.): Strukturális magyar nyelvtan I. Morfológia. Akadémiai Kiadó, Budapest (2000) 137–164
4. Kiefer F.: Jelentéselmélet. Corvina, Budapest (2000)
5. Lengyel K.: A nyelvi egységek szinteződése. In: Keszler B. (szerk.): Magyar grammatika. Nemzeti Tankönyvkiadó, Budapest (2000) 24–33
6. Papp, F.: Morfologija. In: Papp, F. (szerk.): Kurs sovremennogo russkogo jazyka. Tankönyvkiadó, Budapest (1968) 189–423
7. Papp F.: Szövegszó, szóalak, lexéma. Magyar Nyelvőr 98 (1974) 76–82
8. Prószéky G.: A magyar morfológia számítógépes kezelése. In: Kiefer F. (szerk.): Strukturális magyar nyelvtan I. Morfológia. Akadémiai Kiadó, Budapest (2000) 1021–1063
9. Pusztai F. (főszerk.): Magyar értelmező kéziszótár. Második, átdolgozott kiadás. Akadémiai Kiadó, Budapest (2003)