Kísérlet automatizált szövegelemzési módszerek kialakítására a szóhangsúlyok meghatározásához

Tamm Anne, Olaszy Gábor

MTA Nyelvtudományi Intézet, 1068 Budapest, Benczúr u. 33.

Kivonat. Az automatikus szövegelemzés bonyolult kérdésköréből egy részfeladat vizsgálatát tűztük ki célul, nevezetesen a hangsúly-kategóriák szavankénti kijelölését meghatározott mondatokban. Az eredményeket a gépi beszédszintézis prozódiai támogatáshoz tervezzük felhasználni. A hangsúly-kategóriákat úgynevezett címkékkel jelöljük a szó előtt a szövegben. A célkitűzést két irányból közelítjük: A klasszikus módszernél a címkéket nyelvészeti mondatelemzés eredményéből nyerjük. A másik eljárás lényege nem nyelvészeti központú, hanem egyfajta egyszerű felszíni szövegelemzés, melyben nem használunk nyelvészeti módszereket, csupán szólistákat, táblázatokat, egyszerű szabályokat. Mindkét elemzési formánál alapkövetelmény az algoritmizálhatóság. Az elemzésekhez ugyanazokat a hangsúly-kategóriákat használjuk, így mód nyílik arra, hogy közvetlenül összehasonlíthassuk a nyelvészeti elemzés eredményét a nem nyelvészeti központú eljárásból kapott hangsúly jelölésekkel. Rávilágítunk mindkét elemzésnél, hogy mely problémák miatt nem kaphatunk teljes értékű eredményt sok esetben.

1 Bevezetés

Gépi szövegfelolvasás során a mondatok prozódiája akkor lehet természetes, ha képesek vagyunk egyrészről a mondat dallamának, másrészről a szavak hangsúlyozásának, azaz a prozódiát alkotó két leglényegesebb összetevőnek a szövegbe való jelzés szintű beillesztésére. A mondatok prozódiája egyrészről a szintaktikai szerkezet függvénye, de szemantikai szempontok is meghatározóak lehetnek. Az elmúlt két évtized mondattani, fonológiai és fonetikai eredményei alapján [1], [2], [3], [4], [7] jó eséllyel meg tudjuk jósolni, hogy egy mondatban milyen lesz - vagy lehet - a hangsúlyok eloszlása, és a hangsúlymintázatra milyen intonációs dallam épül. Jelen kutatásban a hangsúlymintázattal, annak is az automatizált megállapításával foglalkozunk. Rámutatunk számos problémára is, ami gátolja a teljes értékű elemzés megvalósítását.

2 Anyag és módszer

A kutatás nyelvi korpuszát médiából gyűjtött híranyag és időjárás jelentés mondatai (500-500 db) képezik. A hangsúlykijelölés alapegysége a mondat. Az algoritmizálás-

ból fakadó kiindulási elv, hogy a mondat minden szavára teszünk hangsúlyjelzést. Ötféle hangsúly kategóriát használunk a szavak besorolására. Ezek jelzései és tartalmuk a következő: [:F]=fókusz, [:E]=kiemelt, [:W]=normál, [:N]=neutrális (hangsúlytalan), [:-] = erősen hangsúlytalan, (esetleg redukált). A fenti kategóriákból három hangsúlycsoport következik: azok a szavak, amelyeken van valamilyen hangsúly (F,E,W jelek), azok, amelyeken nincs hangsúly (N jel), és azok amelyek hangsúlytalan elemeknél is redukáltabban vesznek részt a hangsor felépítésében (negatív jelűek). A hangsúly jelzését a szó elé tesszük az elemzés során. Így a mondat minden szaván lesz jelzés. (A megadott szövegpéldák némelyikében más intonációs jeleket is szerepeltetünk (/26, //11 stb.). Az ilyen jeleket nem kell figyelembe venni, mivel ezek az automatikusan generált prozódiai elemzés nem hangsúlyhoz tartozó jelei.)

Kétféle elemzési elvet vizsgálunk: (A) a szövegek felszíni tanulmányozásából kialakított egyszerű elemző, amelyik nem használ nyelvészeti eszközöket, csak speciális szó- és szókapcsolat-listákat, valamint egyszerű szabályokat; (B) nyelvészeti eszközrendszert felhasználó elemző. A két elemzési módszert összehasonlítottuk olyan formában, hogy vizsgáltuk a szavak elé tett hangsúly jelek helyességét, illetve helytelenségét. Az összehasonlításához egységes alapot teremtettünk percepciós teszttel. A hibákat a zárójelbe tett, helyes jelzéssel érzékeltettük a szöveges anyagban (1), majd összegeztük. A hibákat két fokozatba soroltuk.

Súlyos hiba, ha az elemző a szóra [:E,F] jel tesz, ugyanakkor a szónak [:N], vagy [:-] jelűnek kellene lenni.

Közepes hiba, ha [:W] helyett a szót [:-] jelűként kell jelölni, továbbá, ha [:N] helyett [:W]-t kell szerepeltetni.

(1) */23[:W]reagan /13[:N(**W**)]évek [:N]óta [:pause 1]/24[:E(**N**)]nem [:N]mutatkozott /26[:-]a [:W]nyilvánoss/15ág [:N]előtt[:pause 2800].*

A helyes jelzést a példában kiemelt betűvel jelöltük.

A vizsgálatok eredményeit emberi hangon megszólaló beszédszintetizátorral, hangzó formában is előállítottuk percepciós tesztek végzése céljából.

3. Automatikus hangsúlykijelölés mondatokban

Az automatikus hangsúlykijelölésnél a szöveget vizsgáljuk, és ennek alapján döntjük el az adott szóról, hogy milyen hangsúlyozási fokozatba tartozik. Mindig csak egy mondatot vizsgálunk, mondatok közötti összefüggéseket nem. Mivel nyelvi elemzést végzünk számolnunk kell azzal, hogy lesznek olyan esetek, amelyekben nem tudunk egyértelmű döntést hozni, kompromisszumot kell kötnünk. Erre a következőkben számos példát fogunk látni.

3.1 Hangsúlykijelölés nem nyelvészeti központú megközelítéssel (A)

Ennél a módszernél a szövegek felszíni tanulmányozása alapján alakítunk ki szabályokat, listákat. A szabályok kialakításához felhasználjuk a szövegek felolvasásának hanganyagán végzett fonetikai elemzések (1. ábra) eredményeit is (Olaszy et al. 2001).

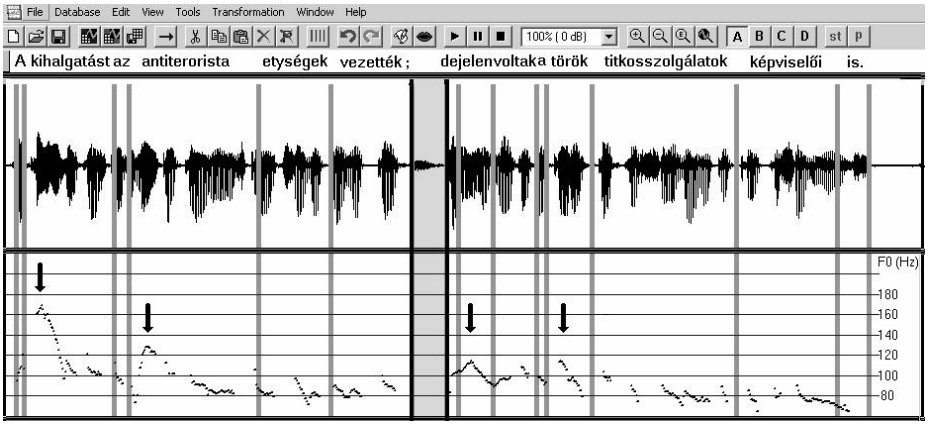

Fig. 1. A hangsúlyos szavak megjelölése a szövegben az alapfrekvencia-görbe (alul) szövegre való visszavetítésével. A nyilak jelzik a hangsúlyokat a szavak első szótagján, a függőleges vonalak a szóhatárokat

Az 1. ábra alapján az elemzett mondatban a szavakra a következő hangsúly-jelöléseket lehet tenni. [:-]A [:W]kihallgatást [:-]az [:W]antiterrorista [:N]egységek [:N]vezették, [:-]de [:W]jelen [:N]voltak [:-]a [:W]török [:N]titkosszolgálatok [:N]képviselői [:N]is.

Hipotézis: A szóhangsúlyozási fokozatok jó hatásfokkal meghatározhatók és jelölhetõk a szövegben szintaktikai elemzésnél egyszerűbb módszerrel, szövegelemek vizsgálatával, és egyszerű szabályok megfogalmazásával is.

A nem nyelvészeti központú hangsúly-meghatározásnál két dolgot kell kiemelni. Az egyik, hogy nem törekszünk teljességre. Elvünk az, hogy a hangsúlyos szavak többségét megtaláljuk a mondatban, továbbá az, hogy lehetőleg ne tegyünk hangsúlyt olyan szavakra, amelyek hangsúlytalanok a kiejtésben. A másik fontos szempont, hogy erősen támaszkodunk a gyakorisági adatokra. Ez azt jelenti, hogy a gyakoribb eseteket vesszük szabálynak, a szabály alóli kivételeket pedig esetlegesen listákban, vagy magában a szabályban adjuk meg. Az eljárásban tehát gyakran kell kompromisszumot kötni.

A szöveg felszíni vizsgálatában a szavakat két kategóriára osztjuk: hangsúly szerinti **tartalmas** szavak, illetve **nem értékes** szavak. Ez utóbbiak azok, amelyekre nem kerülhet hangsúly, vagyis a [:-] jelű szavak (ilyenek például a névelők, a kötőszók). Mindkét kategóriához tartoznak kivételek. A vizsgálatban fontos szerepet tulajdonítunk a mondatban elhelyezett szeparátoroknak (vessző, pontosvessző, kettőspont stb.). A hangsúly-meghatározási eljárás két lépcsős. Az első lépcsőben a megállapított jelöléseket helyezzük el a szavakra, a másodikban az egymás utáni szavakra tett jelzések együttes vizsgálatával (szabályok alapján) az első lépcsőben meghatározott jelöléseket hagyjuk jóvá, illetve változtatjuk meg. Így alakul ki a végleges hangsúly-térkép a mondatban.

3.1.1 Az [:F] jelű kiemelt hangsúlyozású szavak meghatározása

Az [:F] jelű hangsúly a legerősebb a hangsúlyozott szavak között. Ezen szavakat lista alapján jelöljük (például: *nem, ne, nagyon, nincs, soha, semmi, senki, jó, szép, minden, meg kell, mikor, milyen, stb.).* A hangsorban ezek a szavak képviselik a leg-

erősebb hangsúlyt. Fontos szempont, hogy az [:F] jelű szavak után az „irtó" szabályhoz hasonló műveletet hajtunk végre. Ez azt jelenti, hogy ha esetleg a hangsúlykiosztás első fázisában egy [:F] jelű szó utáni szó [:W] jelet kapott, akkor azt törölni kell és [:N]-re kell változtatni. Ugyanilyen szabály vonatkozik az [:F]-jelű szó előtti szóra is. További szabály, hogy [:F] jelzés két egymást követő szón nem lehet. Ha ilyen előfordul, akkor mindig az első tartja meg az [:F] jelzést, a következő [:N]-re íródik át.

3.1.2 Az [:E], illetve a [:W] jelű szavak meghatározása

Az [:E] és [:W] jelzésű szavaknál a hangsúlyozást megvalósító Fo kiemelkedés mértékében van csupán különbség. Az előbbi erősebb hangsúlyt képvisel, mint az utóbbi. A normál szóhangsúlynak a [:W] jelzést tekintjük. Az [:E] jelzést szintén lista alapján osztjuk ki. A [:W] jelzés meghatározásához listát is és szabályokat is alkalmazunk. Ezen szavak kijelölésének a legbonyolultabb a szabályrendszere, ezek empirikus szabályok. Számos olyan tény van, amelyik meghatározza, hogy egy szó az esetek többségében normál hangsúllyal ejtendő. A [:W] szóhangsúly jellel jelöljük meg a listában megadott szavakat, a számok elemeit a *száz, ezer, millió* kivételével, a névelők (*a, az*) utáni szót, a vessző utáni első tartalmas szót, a mondat első tartalmas szavát, a [:-] jelzésű szavak utáni szót, bizonyos szóösszetételek meghatározott szavait (listából), a tulajdon neveket, a személy neveket és a mozaik szavakat (például MTA).

3.1.3 Az [:N], illetve a [:-] jelű szavak meghatározása

Az [:N] jelű szavak jelzésének fizikai jelentése az, hogy a hangsúlyozandó szón nem hajtunk végre sem Fo-, sem intenzitás-emelést. A szó a kiejtés szempontjából tehát neutrális (hangsúlytalan), csak a mondatdallamnak engedelmeskedik. Az [:N] jelű szavak kijelölését az ide tartozó lista szerint, valamint a maradék kitöltése elv alapján végezzük. Miután minden jelzést elhelyeztünk a mondatban és a jelzések véglegesítése is megtörtént, akkor az addig nem jelölt szavakra [:N] jelzést teszünk. A [:-] jelű szavak jelzésének fizikai jelentése az, hogy a hangsúlyozandó szón csökkentjük az Fo értékét, az intenzitását, valamit a hangok időtartamát. Így egy általános redukciót hajtunk végre a szó hangsorban elfoglalt szerepe szempontjából. A [:-] jelzésű szavakat listából jelöljük ki. Talán ezek a szavak rendelkeznek a legstabilabban a [:-] jelzéssel, mint a hangsúlyozási hierarchia legalsó elemei.

3.2 Vizsgálati eredmények nem nyelvészeti elemzővel

A fenti szabályokkal megvalósított hangsúly-meghatározó algoritmus a Profivox magyar beszédszintetizátorban [6] került megvalósításra a BME Távközlési és Telematikai Tanszékén. A hangsúly-meghatározó szabályok száma mintegy 390 a rendszerben. Az algoritmus döntéseit jelen vizsgálatban egy szűkített korpuszon (50 mondat, összesen 756 szó) vizsgáltuk meg (ugyanezt a korpuszt használtuk a (B) elemző értékelésére is). Minden mondatban elemeztük az algoritmus által a szavakra tett jelzéseket és azok hibás, illetve helyes voltát. Az ítéleteket a mondatok szintetizált formáinak a meghallgatásával végeztük. A mondatokat 3 személy (2 férfi és egy nő, 25, 62, 47 évesek) hallgatta meg és értékelte. A feladatuk az volt, hogy meg kellett hallgatni az adott mondatot, ezzel párhuzamosan tanulmányozhatták a szavakra tett jelzéseket is. Ezután döntötték el, hogy mely jelzések hibásak a mondatban. A hibás

jelzést kijavítva a szövegben a mondatot újra szintetizálták és újbóli meghallgatással ellenőrizték, hogy a rossznak vélt címkékben a javítások a hangzásban is javulást okoztak-e. Példaként bemutatunk egy ilyen tesztelési eredményt. A javított három címkét félkövérrel jelezzük.

Az eredeti mondat:

(2) *A hajnali pára- és köd feloszlását követően ma is sok lesz a napsütés és sokfelé meghaladja a hőmérséklet a 20 fokot.*

A felcímkézett mondat:

(3) */23[:-]a [:W]hajnali [:N**(W)**]pára- [:pause 1]/13[:-]és [:W]köd [:N]feloszlását [:N]követően [:pause 48][:N**(W)**]ma [:pause 1][:-]is [:pause 1]/24[:E**(N)**]sok [:-]lesz /23[:-]a [:W]napsütés [:pause 1]/24[:-]és [:W]sokfelé /23[:W]meghaladja [:pause 1]/24[:-]a [:W]hőmérséklet /26[:-]a [:W]h/15úsz [:N]fokot.*

A szavak száma a mondatban: 21
A hibás jelzések száma:3
A hibák és fajtáik
[:N(W)]pára
[:N(W)]ma
[:E(N)]sok [:-]lesz súlyos hiba

A későbbi (B) eljárás eredményeinek értékelésénél az (A) elemzés percepciós tesztjéből származó jó jelzéseket vettük alapul. Az összesített eredmények szerint a gépi elemzés eredményeit vizsgálva a 756 szóból 97 szón találtak a tesztelők hibás jelzést, ami a teljes szóállomány 12,8 %-a. Az esetek 87,2 %-ában a nem nyelvészeti központú elemző tehát jó hangsúly-kategóriát állapított meg a mondatok szavaira. Ezzel igazolódott a hipotézis. Megvizsgáltuk a hibák összetételét is. Három kategória szerint osztályoztuk: a) amikor a szóra nem tett hangsúlyt a rendszer, noha kellett volna (tipikus esetben az [:N] jelzést [:W]-re kell cserélni); b) amikor a szóra tévedésből hangsúlyt tett, de ez igen zavaró (tipikus eset, amikor a [:W] jelzést [:-]-ra kell változtatni; c) a b) eset enyhébb változata (tipikusan, amikor a [:W] jelzést [:N]-re kell változtatni). Az osztályozás eredménye a következő: a)-ból 77, b)-ből 10 és c)-ből 10 hibát vétett az elemző. A legtöbb esetben tehát a hibás döntés eredménye az volt, hogy a szóra nem tett hangsúlyt az elemző, azt neutrális szintűnek ítélte. Ez összhangban van azzal a korábbi kitétellel, hogy feltételezésünk szerint az a legzavaróbb, ha olyan helyre teszünk hangsúlyt a mondatban, ahová nem kéne. Ilyen hiba mindössze az esetek 2,6%-ában fordult elő.

3.3 A hangsúlyok kijelölése nyelvészeti megközelítéssel (B)

A magyar mondatoknak egy lényegében invariáns hierarchikus szerkezetet [2] tulajdonítunk, s ebből levezethetők a prozódiai szerkezet leglényegesebb komponensei. A mondat topik részre és predikátum részre oszlik. A topik tetszés szerinti (nulla, egy vagy több) igebővítményt és szabad határozót tartalmaz. Bizonyos típusú összetevők (pl. a határozók *szerencsére, valószínűleg, látszólag* típusú mondathatározók) csak a topik részben állhatnak. A topik rész összetevői mind gyenge hangsúlyt viselnek. A

mondat legerősebb hangsúlya és intonációs csúcspontja a predikátumrész első fő összetevőjére esik. A predikátumrész tetszés szerinti és számú (nulla, egy vagy több) disztributív kvantorral (azaz *mindenki, senki, minden előfizető, a posta is* típusú összetevővel) kezdődik. Ezek mindegyike főhangsúlyos. Őket követi a szintén főhangsúlyos, közvetlenül az ige előtti összetevő, mely akár fókusz (***A POSTÁS*** *csengetett be*), akár igekötő (***be-****csengetett*), akár névelőtlen főnév (***levelet*** *hozott*) lehet. Az ezt követő ige hangsúlytalan. Bizonyos mondattípusokban az ige előtti pozíciók üresen maradnak és maga az ige a predikátumrész kezdete: ilyen esetben az ige főhangsúlyos. A főhangsúlyos elemek hangsúlyának erőssége balról jobbra csökken. Az ige utáni fő összetevők attól függően hangsúlyosak, hogy ismert vagy új információt közölnek-e és hogy van-e fókusz a mondatban. Az ige utáni disztributív kvantorok akár hangsúlyosak, akár hangsúlytalanok lehetnek.

A fenti fő elvek alapján egy humán elemző minden lehetséges magyar mondathoz hangsúlyszerkezetet tud rendelni. Ugyanakkor az automatikus elemzést rendkívüli módon megnehezíti, hogy a magyar mondatban lényegében minden mondatpozíció maradhat üresen is, továbbá az igét és az ige előtti pozíciót kivéve a szerkezeti pozíciók több fő összetevővel is kitölthetők. A szerkezeti pozíciók azonosításában tehát nem segít a számolás; irreleváns, hogy egy összetevő hányadik helyen áll. Nem mindig könnyű feladat a fő összetevők határainak automatikus felismerése sem.

A célkitűzésünk megvalósítására a nyelvészeti eszköztárakból a következőket használjuk: morfológiai elemző; NP elemző; fókusz szabályok (azon belül azonosító szabályok és hangsúlytörlő szabályok); egyéb erős hangsúlyt adó szabályok és környezetük; topik szabályok; határozói szabályok ; a szintaktikai egységeken, frázisokon belül működő balszél szabályok; listák (szavak, kifejezések) és egyéb szabályok (például: szöveg- vagy mondat típusától függő szabályok).

3.3.1 Morfológiai elemző (szószablya) - minden szóra megállapítja annak a kategóriáját (pl. hazarendelték = haza[PREF]+rendel[vrb]) és az alaktani alakját (pl. +[PAST INDIC DEF PL 3]). A morfológiai elemző különösen fontos a fókuszos mondatok elemzésében, ahol a ragozott ige és az igerészek helyétől függően a mondatban több frázison is törlődik a hangsúly.

3.3.2 NP elemző (INTEX-alapú, és a Nyelvtudományi Intézetben fejlesztett) – megadja a mondatok határait (a cimkéje: {S} a mondatok elején és a végén). Mondatok (esetleg a mondatrészek) határjelei között keresi a szövegben rejlő NP-ket és ellátja az azonosított elemeket az "NP" címkékkel (4).

(4) {S}*Váratlanul hazarendelték* [np *konzultációra* np] [np *Irakból* np] [np *az ország amerikai polgári kormányzóját* np]. {S}

Ha az NP elemző talál egy főnevet (névszót), akkor megállapítja ennek a közvetlen környezetéhez való viszonyát, és ezután dönti el, hogy az NP-hez tartozik ez a környezet. A főnévi csoport (az NP) határait azért releváns megkeresni, mert a frázisra adott erősebb hangsúly egy főnévi csoportban csak az első „tartalmas" szóra esik. A többi szó az NP-ben semleges hangsúlyt kap vagy azt a hangsúlyjelölést, amelyet a listák alapján előírják neki. Névelőtlen főnevek helye a ragozott igéhez képest viszont fontos az ige utáni frázisok hangsúlyadásban.

3.3.3. Ige kereső - a hangsúlyadás szempontjából fontos az ige helye bizonyos más szavakhoz képest, de azt a tényt is kell megállapítani, hogy van-e ige a mondatban és

ha van, akkor milyen az ige alakja. Tehát az egyik legfontosabb szabályunk az igeszabály. Az igeszabály egy ragozott igét keres, kétfajta kimenetet ad (talált ilyent, illetve nem). Ha a program talál egy ragozott igealakra utaló jelölést (5), akkor ez a mondat egy további igeazonosítás-szabály bemenete lesz (6).

(5) {S}*Váratlanul hazarendelték* **[vrb]** [np *konzultációra* np] [np *Irakból* np] [np *az ország amerikai polgári kormányzóját* np]. {S}

Ha a program nem talál ilyent, akkor a mondat elemzése a névszói állítmányos szabály alkalmazásával folytatódik (6).

(6)
morfológiai elemző → NP-elemző → igekereső → igeazonosítás-szabály
→ névszói állítmányos szabály

3.3.4 Fókusz kereső - az igeszabályokat követő lépésben a fókuszt és a fókusztól függő hangsúlyadást lehet megjelölni. A fókusz lehet egy vagy több szóból álló csoport (egy frázis). A fókusz (F, amit [:F] jellel jelölünk a szó elött) a legerősebb hangsúly a mondatban. A fókusz hat a környezetére is: hangsúlyt irt. Az őt követő ige mindig hangsúlytalan [2]. A fókuszt több ágon lehet meghatározni. A fókuszt kereső szabályok összetettek, egymást követően alkalmazhatók, amíg megtaláljuk a fókuszt. A fókuszt kereső szabályokból jelenleg 5 van. Az első szabály a névelőtlen főnév, igekötő vagy azzal azonos státusú igerész előfordulásnál alkalmazható. Gyakran van egy mondatban egy hátravetett igerész, igekötő vagy névelőtlen főnév. Ha névelőtlen főnév (NP, pl. ***könyvet***), igekötő (***be***) vagy azzal azonos státusú igerész (***haza-***) közvetlenül az ige után helyezkedik el, akkor az a frázis, amelyik az ige előtt helyezkedik el, fókusz (NAGY betűvel jelöltük a példákban) (7). A példában a szabály megtalálja a ragozott, igekötős ige (*berontott*) igekötőjét (a példában: ***be***) az ige után, azért az a frázis, ami az ige előtt van, fókusz (*valamivel 10 óra előtt*). Az [:F] jelzés helyes elhelyezését majd a későbbi balszél szabály fogja kijelölni.

(7) *A gazdagréti bankfiókba VALAMIVEL 10 ÓRA ELŐTT rontott* ***be*** *a símaszkos rabló.*

Gyakran nincs a mondatban hátravetett igerész, igekötő vagy névelőtlen főnév, akkor nem tudjuk helyesen megállapítani az ige előtti és utáni hangsúlyeloszlást. Hosszabb, összetett mondatokban ez viszont lényeges. Olyan esetekben más "kapaszkodókat" használunk. A második szabály: ha a létige bővítménye közvetlen az ige után helyezkedik el, akkor az a frázis, ami az ige előtt helyezkedik el, fókusz. A harmadik szabály a negatívan minősítő határozószók esetén találja a fókuszt, a negyedik akkor, ha van egy frázis a kvantorok és az ige között, az ötödik akkor, ha van egy frázis kezdetén a "csak"-szó. Egy példának legyen itt a negatívan minősítő szavak fókuszszabálya. A negatívan minősítő szavak kevésre értékelt számosságú vagy kevésre értékelt mennyiségű dolgot, kis gyakoriságot, kis fokot, mértéket, vagy kevésre értékelt módot jelölnek (vö.. É. Kiss (1998: 48)). Az időjáráskorpuszban gyakran előforduló példák: *rossz, kevés, ritka, kevéssé, ritkán, rosszul*. A szabály szerint ha a negatívan minősítő határozószó, pl. *kevés, ritkán, rosszul*, áll közvetlenül az ige előtt, akkor ez a határozószó fókusz. Ugyanezzel a szabállyal lehet kijelölni a problémás névszói állítmányos

mondatokban vagy mondatrészekben is a fókuszt.(8).

(8) *.... a magas hőmérséklet miatt a lehullott csapadék hamar elolvad, így KEVÉS az esély a fehér karácsonyra.*

Segítség a fókusz megállapításához az is, ha van egy szó vagy szavak csoportja (egy frázis) a kvantorok és az ige között (9), illetve ha a frázis kezdetén a "csak"-szó van, akkor ez a frázis fókusz (10).

(9) *A hajnali pára- és köd feloszlását követően ma is SOK lesz a napsütés...*

(10) *...CSAK ELSZÓRT ZÁPOROK valószínűek.*

Ha semelyik fókuszkereső részprogram nem talált fókuszt, akkor fókuszhangsúlyt adó szabályokra nem kerül sor és tovább lehet lépni a nem-fókuszos főhangsúly, úgynevezett erős hangsúly (E, amit [:E] jellel jelölünk a szó előtt) keresésre. Ha megtaláltuk a fókuszt, akkor lehet tovább lépni a hangsúly-írtó szabályokra.

3.3.5. Erős hangsúly keresése - az erős hangsúlyú frázisok, az „E-elemek" a hangsúly szempontjából a fókuszra hasonlító, de nem teljes fókusz hatáskörrel felruházott frázisok. Az E-elem hangsúlya vagy a fókusz hangsúlya jelöli a predikátumrész kezdetét a mondatban, pl.: *János* ***mindig beteg.*** (a predikátumrészt félköver betűkkel jelöltük). Az E-elem hangsúlya abban hasonlít a fókuszhangsúlyhoz, hogy erős. Két szempontból viszont különbözik a fókusztól, az egyik, hogy hangsúlyirtó hatása nincs, a másik, hogy egy mondatban vagy mondatrészben több E-elem is előfordulhat. A fókuszszabályokkal ellentétben itt fontos az alkalmazási sorrend is. Ha a szabályok nem találtak egyetlen E-elemet se, csak akkor lehet az igét E-elemként címkézni (azaz az ige lesz az E-elem az utolsó azonosítószabály szerint). Az erős hangsúly keresésére is több szabály vonatkozik. Az egyik szabály például azt állapítja meg, hogy ha a névelőtlen főnév (*orvos*), igekötő (*meg-*) vagy azzal azonos státusú igerész (*haza-*) közvetlenül az ige előtt vagy az ige részeként helyezkedik el (11), akkor ez a frázis E-hangsúlyt kap, pl. ***Orvos*** *lett,* ***meg****lett,* ***haza****rendelték,* ***szükség*** *van*.

(11) *[:E]Boát [:-] loptak az állatkertből.*

Egy másik szabály azt állapítja meg, hogy ha egy disztributív kvantor található a mondatban, akkor ez a frázis E-hangsúlyt kap (*János* ***mindig*** *beteg*). Ha egy frázis egy „*is*" szót tartalmaz, akkor ez disztributív kvantor és erős hangsúlyt kap (***János is*** *beteg*). Ha a szó tagadószó, akkor erős hangsúlyt kap, de magába olvasztja a következő szót [2]. A létige ige előtti bővítménye (kiegészítője, pl. ***szerencsés*** *volt,* ***orvos*** *volt*) is erős hangsúlyt kap.

Az erős hangsúlyt adó szabályok után következnek "finomabb" szabályok, topikszabályok, határozószabályok és egyéb szabályok. A konkrét hangsúlyok adása mellett ezeknek a szabályoknak később, a "korrigáló" szabályoknál is fontos szerepük van.

3.3.6. Topik szabályok. Meg kell jelölni a topikrész végét ahhoz, hogy a topikhangsúlyt adó szabályok és néhány egyéb mondatprozódiai szabály működni tudjon. Ezt a feladatot a topikszabályok látják el. A topik szűkebb értelemben egy olyan vonzat, amely a mondatban az ige előtt áll. Az itt alkalmazott "topikrész" alatt

viszont azt a részt értjük a mondatban, amely több frázist is tartalmazhat. A topikrész alatt a predikátumrész előtti részt, technikailag az első [:E] jelű szó előtt vagy a fókusz előtt levő szövegrészt értjük. Tehát, az első [:E] és [:F] jel előtti szövegrész a mondatrészkezdet-jelzésig topikrész, beleértve a határozókat is. A topikrészhez tartozik minden olyan NP és hatarozó, amely az ige előtt áll és nem egy [:E] jelű elem, és nem egy [:F] jelű elem. Ha „topikos" a mondat, akkor a frázis vagy a frázisok topikhangsúlyt kap(nak), vagyis a [:W] jelzés alkalmazható a mondat elején az első tartalmas szón, utána az [:N] a predikátumrész kezdetéig. Több frázist tartalmazó, hosszabb mondatokban, olyan mondatokban, amelyekben van fókusz, de az algoritmusnak nincs egyetlen formai „kapaszkodója" sem" (pl. a hátravetett igekötő vagy a "*csak*"-szó) a fókusz megtalálásához is releváns megjelölni a topikrészt. Például a mondatban nagy valószínűséggel van fókusz akkor, ha a mondat predikátumrész lényegesen „nehezebb" a topikrésznél, azaz több NP-ből és határozófrázisból áll (12).

(12) *Clinton otthon lábadozik.*

Topikrész: *Clinton otthon*, predikátumrész: *lábadozik*. →
Topikrész: *Clinton,* predikátumrész: [:F]*otthon lábadozik.*

3.3.7. Határozó-azonosítók – a főfeladatuk a szövegben rejlő határozók azonosítása, címkézése, hangsúlystruktúrájuk azonosítása és cimkézése. Ehhez 3 külön szemantikai-prozódiai leírással rendelkező listát használunk. Mondat- és módhatározókat különböztetünk meg: pl. *tényleg* mondathatározó, *gyorsan* viszont módhatározó. Ha a határozószó mondathatározó (*esetleg, állítólag*, *okvetlenül, feltétlenül, tényleg*), akkor a mondat topikrészében van. Ez azt jelenti, hogy a lista alapján már el lehetne dönteni, hogy a határozószó a mondat kezdeti topikrészhez – nem a predikátumrészhez – tartozó szavak csoportjai között van és semleges topikhangsúlyt kap vagy nem. Mondathatározó listákból kettő van: hangsúlyosak és hangsúly nélküliek. Ha egy hangsúly nélküli mondathatározó a mondat első szava, akkor – annak ellenére is, hogy ez a pozíció általában hangsúlyos és egy [:W] jelzést kap – a hangsúly nélküli mondathatározó nem kap a topik elején megjelenő hangsúlyt. Tehát egy mondat elején megjelenő mondathatározó nem lesz mindig [:W] , hanem [:N] (13).

(13) *[:N]Állítólag [:W] kínvallatás alkalmazását is engedélyezte az amerikai védelmi miniszter a guantánamói amerikai támaszponton fogva tartott feltételezett terroristákkal szemben.*

3.3.8. Az egyéb szabályok összetettek, legtöbbjüknek lokális, lista-alapú jellege van. Néhány példa: ha a szó kötőszó (*hogy, ami, de, hanem*), akkor törlődik a hangsúly ([:N] lesz), az *egy* szó hangsúlyos, ha utána mértékegység jön (egy *fok, másodperc, perc, óra, nap, hét, hónap, év, évtized, évszázad, évezred, kilométer, milliméter, centiméter, méter*-sorozat*, láb, mérföld, gramm-sorozat, deka,* stb.). A mértékegységeken viszont törlődik a hangsúly ([:N] lesz). Ha van kis fokozatot jelölő összetevő [7] szerint:, *néhány, némi, egy kicsi, néha, néhol, egyelőre, enyhén, kissé, némileg, valaki, valahol, valahogyan, valamennyi, némileg*), akkor törlődik a hangsúly ([:N] lesz). A szemantikailag kiüresedett bővítmények esetében [7] törlődik a hangsúly [:N] lesz, pl. *bizonyos, valóságos, szegény, kis*. Ha címek és rangok vannak a tulajdonnevek előtt, akkor törlődik a hangsúly, [:N] lesz: *úr, néni, bácsi, út, köz, utca,*

doktor stb. Páros kötőszók esetén alkalmazható a [:W] jelzés (*nem.. hanem, akár... akár, vagy .... vagy, mind ... mind*). Ha a frázisban található egy listázott hangsúlykerülő, akkor [:-] jelzésű hangsúlyt kell alkalmazni. A hangsúlykerülők pl. *akar, érint, fog, folyik, talál, kell, szabad, szeretnék,* stb.

A szövegtípusból adódó hangsúlyszabályok közé tartozik például, hogy nagyobb hangsúlyt, [:W]-t kap az ige, ha a predikátumrész rövidebb a topik résznél, azaz kevesebb NP-ből és határozófrázisból áll.

Egy vessző utáni „mondásige" (*mondta, döntött,* stb.) a következő hangsúlymintát kapja: az ige és az igemódosító [:-] jelet, a többi frázisba egy [:W] - [:N] típusú hangsúlyminta kerül (a hangsúlyt a frázis első tartalmas szava kapja a balszél-szabály szerint), a többi rész [:N jelet kap. Ez akkor is érvényes, ha egy hátravetett igemódosító és főnévi csoportok vagy határozók állnak mögötte. Ha nincs fókusz vagy E-elem a mondatban, akkor is a frázisokon balszél-szabály szerinti fenti hangsúlymintázat lesz a topik, a határozók és az erős hangsúlyú ige után.

Egyéb szabályok alkalmazásánál gyakran számít a sorrendjük. A balszélszabályt az elemzés legvégén alkalmazzuk. A frázisra adott erősebb hangsúly egy lineáris szavak csoportjában csak az első „tartalmas" szóra esik. A többi szó a frázisban semleges hangsúlyt kap vagy azt a hangsúlyjelölést, amelyet a listák alapján előírják neki. Ezek a szabályok sok esetben korrigálhatnak a lista alapú megközelítéssel kapott eredményt.

3.4 Vizsgálati eredmények a nyelvészeti elemzővel

A nyelvészeti központú hangsúlykijelölő elemzés jelöléseit humán erővel, az algoritmus figyelembevételével helyeztük el a vizsgált mondatokban. Itt is ugyanazt az 50 mondatot használtuk, amelyeket az (A) elemző értékelésénél (lásd 3.2 pont). Az értékelés során megvizsgáltuk, hogy a nyelvészeti elemzőből adódó jelzésekben hol van hiba (ez szintén humán elemzéssel történt). Az összesített eredmények szerint a nyelvészeti elemzés során a 756 szóból 90 szón találtunk hibás jelzést, ami a teljes szóállomány 11,9 %-a. A nyelvészeti központú elemző tehát közel annyiszor vétett, mint a gépi elemző. Itt is megvizsgáltuk a hibák összetételét is. Az a)-ból 66, b)-ből 5 és c)-ből 19 esetben vétett a nyelvészeti elemző. A legtöbb esetben tehát a hibás döntés eredménye itt is az volt, hogy a szóra nem tett hangsúlyt az elemző, azt neutrális szintűnek ítélte.

4 Összefoglalás

A kapott számszerűsített eredmények azt mutatják, hogy minkét elemző hasonlóan jó hatásfokkal végzi az elemzést. Az eredmények hasonlóságánál látnunk kell, hogy a számok mögötti hibák típusai különbözhetnek a két eljárásban. Ezért külön elemezzük az eredményeket. Az (A) eljárásnál tapasztalt hibák legtöbbje abból adódik, hogy nincs szisztematikus fókusz keresés, ezért az irtó szabály hatóköre sok esetben szűkebb, mint kellene, továbbá az ige meghatározás hiánya is sokszor hibát okoz. Az NP elemző és a balszél szabály nélkülözése túl sok felesleges [:W] hangsúlyhoz vezethet. A (B) típusú elemző hibái abból adódnak, hogy a jelenlegi szabályok túl általánosak, az NP elemzést finomítani kell. Ha az NP elemző rosszul azonosítja az NP határokat, akkor a többi, erre épülő elemzés hibás eredményt ad. Az elemzés hatékonyságának növeléséhez ezen kívül bővíteni kell a hangsúlykerűlő elemek szótárát. Be kell építeni továbbá mondatrészhatár azonosítót is a rendszerbe. A jelenlegi tapasztalatok alapján tehát úgy látjuk, hogy a meghallgatásos vizsgálatokra alapozva, valós elhangzó minták vizsgálatának támogatásával finomítani és bővíteni lehet az elméleti nyelvészeti kutatásokat is.

Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki az MTA Nyelvtudományi Intézet munkatársainak Gábor Katának és É. Kiss Katalinnak az elemzésekhez nyújtott segítségükért, továbbá a BME Távközlési és Médiainformatikai Tanszékről Kiss Gézának, aki az (A) elemző algoritmusából működő programot készített.

Ezt a kutatást az NKFP 2. programja (2/034/2004 sz.) támogatta.

Hivatkozások

1. É. Kiss, K.: The Syntax of Hungarian. Cambridge Syntax Guides. Cambridge: Cambridge University Press (2002)
2. É. Kiss, K., Kiefer, F., Siptár, P.: Új magyar nyelvtan. Budapest: Osiris (1998)
3. Hunyadi, L.: Hungarian sentence prosody and Universal Grammar. New York, Peter Lang (2002)
4. Olaszy, G.: The most important prosody patterns of Hungarian. Acta Linguistica Hungarica, Vol. 49 (3-4) (2002) 277-306
5. Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Zainkó, Cs., Gordos, G.: Profivox – a Hungarian TTS System for Telecommunications Applications. International Journal of Speech Technology. Vol 3-4. Kluwer Academic Publishers (2000) 201-215
6. Olaszy, G., Németh, G., Kiss, G.: Hungarian audiovisual prosody composer and TTS development tool. In: Prosody 2000. Editors: Puppel Stanislaw, Grazina Demenko. Poznan (2001) 167-178
7. Varga, L.: Intonation and Stress: Evidence from Hungarian. New York: Palgrave Macmillan (2002)