Számítógépes összehasonlító szövegelemzés ügyfélszolgálati tájékoztatók legfontosabb prozódiai elemeinek a meghatározására

Abari Kálmán[1], Tamm Anne[2], Gábor Kata[3], Olaszy Gábor[4]
[1] Debreceni Egyetem, Pszichológia Intézet és Matematikai és Számítástudományi Doktori Iskola
abarik@delfin.unideb.hu
[2] Észt Nyelvtudomány Intézet, Tallinn és Firenzei Egyetem, Finnugor Szektor,
anne.tamm@eki.ee, anne.tamm@unifi.it
[3] MTA Nyelvtudományi Intézet
gkata@nytud.hu
[4] BME Távközlési és Médiainformatikai Tanszék
olaszy@tmit.bme.hu

Kivonat: A szövegelemzéssel történő hangsúlykijelölés bonyolult feladat. Jelenleg nincs olyan elemző algoritmus, amelyik gépi úton képes a magyar mondatokban a hangsúlyok kijelölésére. Alapvető célunk, hogy az eddig elért és hozzáférhető elméleti nyelvészeti eredményeket, valamint kész mondatelemző algoritmusokat egyetlen, jól körülhatárolható struktúrált számítógépes rendszerré fejlesszük tovább automatikus hangsúlykijelölési kísérletek végzése céljából. Ebben a tanulmányban munkánk végeredményét, egy futtatható programrendszert (elemző) mutatjuk be. Az elemző bemenetére a mondat szöveges formája kerül, majd a feldolgozás során a mondat minden szavát egy hangsúlycímkével látja el. A feldolgozást két szinten végezzük: (i) tagmondatokra bontás, (ii) a hangsúly kijelölése tagmondatonként. Öt hangsúlykategóriát definiáltunk az elemzéshez: (F)=erős hangsúly, (E)=kiemelt, (W)=normál, (N)=hangsúlytalan és (-)= erősen hangsúlytalan (redukált) címkék. Az elemző 12 modulból épül fel, melyben mindegyik modul azonos koncepcióra épül. Az elemzőt egy szűk témakört leíró szöveges állományra fejlesztettük. A hangsúlykijelölés hatásfoka összességében 85%.

1 Bevezetés

A szövegelemzéssel történő hangsúlykijelölés bonyolult feladat. Nincs is olyan elemző algoritmus a magyarra, amelyik gépi úton képes a feladat megoldására. Alapvető célunk, hogy az eddig elért és hozzáférhető elméleti nyelvészeti eredményeket egyetlen, jól körülhatárolható struktúrált számítógépes rendszerré fejlesszük további elemzési kísérletek végzése céljából. A munkánk köztes eredményéről korábban már beszámoltunk [7], jelen tanulmányban az algoritmus fejlesztésének eredményét, egy futtatható programrendszert (elemző) mutatjuk be. Az elemzőt egy szűk témakört leíró szöveges állományra fejlesztettük.

Az elmúlt két évtized mondattani, fonológiai és fonetikai eredményeire alapozva [2], [3], [5], [6], [8] állítottuk össze azt az algoritmust, amelyikkel jó eséllyel meg tudjuk jósolni, milyen lesz - vagy lehet - a hangsúlyok eloszlása a mondatban. Az elmélet szerint a magyar mondat topik részre és predikátum részre oszlik. A topik 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 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.

2 Az elemző felépítése

Az elemző bemenetére a mondat szöveges formája kerül, majd a feldolgozás során a mondat minden szavát egy hangsúlycímkével látja el a program. A feldolgozást két szinten végezzük: (i) tagmondatokra bontás, (ii) a hangsúly kijelölése tagmondatonként. Az elemzőt a következő hangsúlykategóriák meghatározására terveztük: erős hangsúly (F címke), közepesen erős hangsúly (E), normál hangsúly (W), hangsúlytalan a szó (N), erősen hangsúlytalan a szó (-). A megvalósított algoritmus alapvetően elemző modulokra, szabályokra és listákra támaszkodik. Elemző modulok a következők: szófajmeghatározó és NP elemző. Ez utóbbi egy korábban fejlesztett általános algoritmus, amelynek a magyarított és az adott feladatra adaptált változatát használtuk (Gábor 2007). Ez, mint futtatható önálló modul áll rendelkezésre. Az algoritmus összes többi eleme saját fejlesztés. Az elemző 12 modulból épül fel, melyben mindegyik modul azonos koncepcióra épül. Az elemzési algoritmus szabálymoduljait és a listákat alább ismertetjük az algoritmus folyamatának bemutatásával (1. ábra).

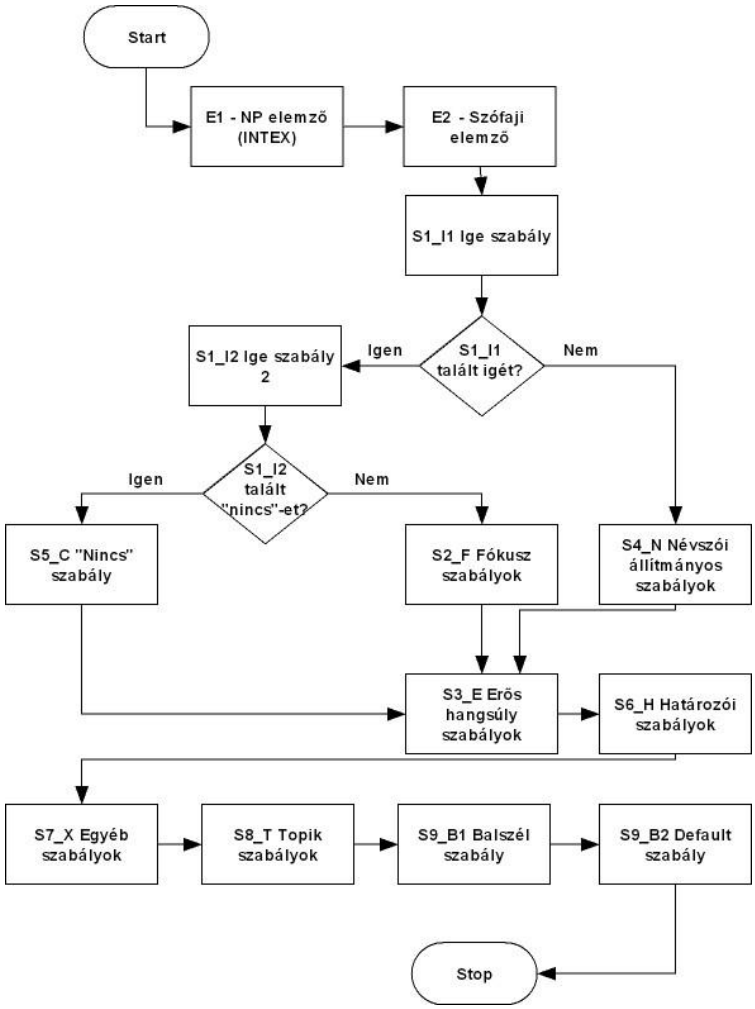

Fig 1. A szöveg alapján működő hangsúlykijelölő elemző moduljai és működési folyamatábrája

E1 – a frázisok meghatározása. Ezt a feladatot az MTA Nyelvtudományi Intézetben magyarított INTEX elemző parancssorból futtatható változata végzi [7]. Ez a modul tartalmaz tokenizálót (az Intexbe beépített funkcióként, ám magyar-specifikus és szöveg-specifikus szabályokkal kiegészítve), szótárat a szöveg lexikai elemzéséhez, valamint a feladathoz szükséges szintaktikai elemző nyelvtanokat. A parancssori alkalmazás egyrészt megjelöli a mondatban található NP-k (főnévi csoportok), melléknévi csoportok (AdjP) és névutós kifejezések kezdetét és végét (beágyazottakét is), másrészt megjelöli a mondat- és tagmondathatárokat. Az elemző kimeneteként egy köztes szövegállomány jelenik meg amelyben {S}, <NP> és <AdjP> címkék jelölik az előbbi szerkezethatárokat. A főnévi csoport határait azért releváns megkeresni,

mert az elmélet szerint [2] a frázisra adott erősebb hangsúly egy főnévi csoportban csak az első „tartalmas” szóra esik. A többi szó az NP-ben semleges hangsúlyt kap.

Az NP elemzés végeredménye erősen befolyásolja a további modulok helyes döntéseit. Az NP keresés számos esetben bizonytalan lehet. Például függhet a szöveg tartalmától és szerkezetétől. Az általános írott szövegek feldolgozására készült NP elemzőt több ponton adaptálni kellett az ügyfélszolgálati szövegek kezelésére, egyfelől a távközlési témájú szövegek speciális szókincse, másfelől a szöveg beszélt nyelvihez közelítő szerkezete miatt. A szókincs megfelelő kezeléséhez szükség volt a szótár kibővítésére. Ezen felül a szövegre jellemző, hogy mondatai nagy arányban tartalmaznak számos, nyílt tokenosztályba tartozó entitást (pl. telefonszámok, egy- vagy többszavas márkanevek, kódok), melyeket nem lehet a szótárban kezelni, így helyes címkézésükről reguláris nyelvtanokkal kell gondoskodni. A reguláris nyelvtanok egy részét az NP-elemzés előtt, a tokenizálást végző modul részeként alkalmazzuk a szövegre (pl. telefonszámok felismerése). A többszavas márkaneveket a szótárban kezeljük, így a feldolgozás későbbi lépéseiben ezek is külön tokenként kezelhetők. A nyílt tokenosztályba tartozó entitások kezelésekor a felismerésen túl az NP-nyelvtant adaptálni kellett a speciális viselkedést mutató, de a névszói kategóriába sorolt entitásokhoz (pl. idegen szó mint főnévi fej esetén gondoskodni kell a fejhez kötőjellel csatolt esetrag felismeréséről is).

A szókincs és különösen a nyílt tokenosztályba tartozó elemek mellett a másik nehézséget a beszélt nyelvihez hasonlító jellegzetességek jelentik, melyek leginkább gondolatjeles közbevetések, valamint hiányos mondatok és tagmondatok formájában mutatkoznak meg. Az ilyen szerkezetek egyfelől megnehezítik a tagmondatra bontást, másfelől újabb hibafaktort jelentenek az NP-k felismerésében is a szerkezeti homonímia miatt.

E2 - szófaji elemző.

Az NP keresés számos esetben bizonytalan lehet. Például függhet a szöveg tartalmától és szerkezetétől. Esetünkben az ügyfélszolgálati szövegek némely pontjának helyes feldolgozására be kellett tanítani az elemzőt. Felsorolunk néhány ilyent. A feldolgozott témakör mondatai nagy arányban tartalmaznak gondolatjeles közbevetéseket, valamint számos, nyílt tokenosztályba tartozó entitást (pl. telefonszámok, egy- vagy többszavas márkanevek, kódok), melyeket nem lehet a szótárban kezelni, így helyes címkézésükről reguláris nyelvtanokkal kell gondoskodni. A nyílt tokenosztályba tartozó entitások kezelésekor a felismerésen túl az NP-nyelvtant adaptálni kell azokhoz (pl. idegen szó mint főnévi fej esetén gondoskodni kell a fejhez kötőjellel csatolt esetrag felismeréséről is). Az NP elemzés végeredménye erősen befolyásolja a további modulok helyes döntéseit.

E2 - szófaji elemző. Ez a modul [4] végzi az ige, a létige, az igekötő, a határozószó, a kötőszó és néhány esetben a szótő felismerését. A szófaji elemző címkékkel látja el a mondat szavait.

S1 – igeazonosítás. A mondatban az ige képezi az első döntési pontot. Két igeszabály végzi a lehetséges esetek teljes feldolgozását. Az első (S1-I1) eldönti, hogy van-e ige a mondatban és ennek megfelelően ágazik el. A második (S1-I2) a talált igét

osztja két kategóriába (tagadás a *nincs* szóval, illetve más ige), e szerint válik ketté a további feldolgozás menete.

S2 – fókusz. A fókusz szabályok két részre oszlanak, a hangsúlyadó és hangsúlytörlő szabályokra. A hangsúlyadó szabály akkor lesz aktív, ha van fókusz a mondatban. A fókusz keresésére a következő négy szabályt alakítottuk ki. Ha névelőtlen főnév, igekötő vagy azzal azonos státusú igerész közvetlen az ige után helyezkedik el, akkor az a frázis, ami az ige előtt helyezkedik el, fókusz (S2-F1). Ha a létige bővítménye közvetlen az ige után helyezkedik el, akkor az a frázis, ami az ige előtt helyezkedik el, fókusz (S2-F2). Ha negatívan minősítő határozószó (Lista-10) áll közvetlenül az ige előtt, akkor ez a határozószó fókusz (S2-F3). Ha van egy frázis a kvantorok (Lista-17) és az ige között, akkor ez a frázis fókusz (S2-F4). A hangsúlytörlő szabály (S2-F5) azt mondja ki, hogy a fókusz után levő ige a fókusz szerkezetétől függően hangsúlytalan (N), illetve erősen hangsúlytalan (-) minősítést kap, a mondatban az utána levő frázisokon (végig) az (N) minősítést kell alkalmazni. Itt két részletszabállyal finomítjuk a végleges döntést. Ha többtagú a fókusz, akkor utána az ige (N) jelzést kap (S2_F5.1), ha egytagú a fókusz, akkor az ige (-) jelzést kap (S2_F5.2.).

S3 – nem fókuszos főhangsúly. Ez a hangsúlyfajta a főhangsúly enyhébb változata, abban különbözik a fókusztól, hogy az utána következő szövegrészben a törlő szabályok másképp működnek. Hat szabály határozza meg az (E) jelű hangsúlyokat. A frázis (E) hangsúlyt kap, ha a névelőtlen főnév (NP), igekötő vagy azzal azonos státusú igerész közvetlen az ige előtt vagy az ige részeként helyezkedik el (S3_E1), ha egy disztributív kvantor (Lista-17) található a mondatban, ha egy frázis disztributív kvantor, vagyis egy *is* szót tartalmaz (S3_E3). Ilyen hangsúlyt kap a tagadószó (S3_E4), illetve a létige ige előtti bővítménye (S3-E5). Ha nincs fókusz vagy más E-frázis, akkor az ige kapja az (E) hangsúlyt (S3_E6). A törlő szabályok (S3-E7) a nem fókuszos főhangsúly után a következők. Az utána levő névelőtlen főnév esetén az ige (N) jelzést kap, igekötő esetén pedig az ige (-) minősítésű, a mondat további részében az utána levő frázisokon végig (W) lesz a hangsúly.

S4 – predikatív mondatok szabály. A névszói állítmányos mondatokban nincs ige. Ezért problémás a hangsúlyok helyes kiosztása. Ennek a szabálynak a kidolgozása folyamatban van, a döntést ilyen esetekben jelenleg az S9 balszél szabály veszi át. Ha van (E) hangsúly a mondatban, akkor minden frázis, a balszél-szabály szerint kap hangsúlyt.

S5 – „nincs" igei szabály. A tagadószó „nincs" (F) főhangsúlyt kap.

S6 – egyéb kategóriák.

Ez a modul a hangsúlykiosztás maradék részeit próbálja megoldani a mondatban. Tíz szabályból áll, legtöbbjüknek lokális, lista-alapú jellege van. A szabályok a következők. Ha a szó kötőszó, akkor törlődik az esetleges hangsúly és (N) jelölésre változik (S6-X1). Az *egy* szó hangsúlyos lesz (W), ha utána vagy számnév a 2. lista szerinti szó, vagy mértékegység jön (S6-X2). Ha van kis fokozatot jelölő összetevő (lásd 14. lista), akkor törlődik a korábbi hangsúly és (N) lesz (S6-X3). A szemantikailag kiüresedett bővítmények (15. lista) esetében törlődik a korábbi hangsúly és (N) lesz (S6-X4). Ha címek és rangok vannak a tulajdonnevek előtt (16. lista), akkor törlődik a korábbi hangsúly és (N) lesz (S6_X5). Mértékegységek esetén (2. lista) törlődik a korábbi hangsúly (S6-X6). A hangsúlyszabályokat befolyásoló konstrukciók (18. lista), páros kötőszók) esetén alkalmazható a (W) hangsúly (S6-X7).

Az S6-X8 szabályok a szövegtípusból adódó hangsúlykiosztást írják le a következők szerint. Hangsúlyos a mondat eleje, a topik (S6-X8.1). Normál hangsúlyt (W) kap az ige, ha a predikátumrész rövidebb a topikrésznél (S6-X8.2). Egy vessző utáni „mondásige", akkor is, ha hátravetett igemódosító áll mögötte + NP a következő hangsúlymintát kapja: ige (-), igemódosító (-), NP balszél-szabály szerinti hangsúly (W- N*) (S6-X8.3). A csillaggal többszöri megjelenést jelölünk, azt, hogy egy (N) egészen az adott frázis végéig alkalmazható. Ha a frázisban található egy listázott hangsúlykerülő (például *kell*), akkor (-) hangsúlyt kell alkalmazni (S6-X9). Ha a frázisban van egy tulajdonnév, akkor arra (W) hangsúlyt kell tenni (S6-X10). Ha nincs fókusz vagy E-elem a mondatban, akkor a balszél-szabály szerinti hangsúly (W-N*) alkalmazható a topik és a határozók után ha van topik vagy határozó (S6-X11).

S7 – határozófrázis szabályok. Fő feladatuk a szövegben rejlő határozók azonosítása, címkézése, hangsúlystruktúrájuk azonosítása és címkézése. Mondat- és módhatározókat különböztetünk meg. Két szabály írja le az ezekkel kapcsolatos elemzés menetét. A mondathatározókhoz két külön szemantikai-prozódiai leírással rendelkező listát használunk (Lista-12, 13). Ha a listákban szereplő mondathatározót találunk, akkor az topikrészben van és topikhangsúlyt kap (S7-H1). Ha módhatározó a 11. lista szerinti, azaz, ha pozitív értelmű a határozószó: fok, mód, gyakoriság (például. *nagyon, eléggé, sokszor, állandóan*) és a fókusz előtti pozícióban van, akkor (W) hangsúlyt kap. Ha mondathatározóként és módhatározóként van listázva a határozószó (S7_H3), akkor a viselkedése változó, elemzési szabály erre még nincs kidolgozva.

S8 – topik szabályok 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. Meg kell jelölni a topikrész végét ahhoz, hogy a topikhangsúlyt adó szabályok és néhány egyéb mondatprozódiai szabály müködni 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 határozó, 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 predikátumrésze lényegesen „nehezebb" a topikrésznél, azaz több NP-ből és határozófrázisból áll.

S9 – balszél szabályok. A frázishoz hozzárendelt hangsúlytípus (F, E vagy W) csak a frázis bal szélén marad meg. A következő szavakon a frázis végéig (N) hangsúly lesz (S9-B1).

Listák

L-1 klitikumok [8] szerint, azaz egy szótagú funkciószók: *a, az, egy, és, de, vagy, is, ha, én, ő, ez, már, még, csak*)

L-2 mértékegységek: *fok, másodperc, perc, óra, nap, hét, hónap, év, évtized, évszázad, évezred, milliméter, centiméter, méter-sorozat, kilométer, láb, mérföld, gramm-sorozat, deka, stb.*

L-3 számnevek: *egy, kettő* ..stb.

L-4 hangsúlykerülő, beférkőző igék: *akar, érint, fog, folyik, talál, kell, szabad, szeretnék…*

L-5 névmások: *én*...

L-6 névutók: *mellett, helyett, után..*

L-7 determinánsok: *a, az, e, ez, egy, eme, ama, ezen, azon, ez a, az a, ..*

L-8 névmásszerű főnevek, „üres szavak": *ország, ügy, kormány, (M/m)agyarország(i)*

L-9 vonzatszótár (opcionális)

L-10 negatív hatarozószók: fok, mód, gyakoriság: *csúnyán, rosszul, ritkán, kevésen, alig*

L-11 pozitív értelmű határozószók: fok, mód, gyakoriság: *nagyon, eléggé, sokszor, állandóan*

L-12 mondathatározók (N-jellel) *esetleg, állítólag*

L-13 mondathatározók (W-jellel) *okvetlenül, feltétlenül, tényleg*

L-14 kis fokozatot jelölő összetevők [8] szerint, azaz *néhány, némi, egy kicsi, néha, néhol, egyelőre, enyhén, kissé, némileg, valaki, valahol, valahogyan, valamennyi*

L-15 szemantikailag kiüresedett bővítmények [8] szerint, azaz, balszélre kerülő melléknevek: *bizonyos, valóságos, szegény, kis*

L-16 címek és rangok a tulajdonnevek előtt vagy után: *néni, bácsi, út, köz, utca, doktor, stb*

L-17 disztributív (univerzális) kvantorok: *mind*-sorozat *(mind, minden, mindenki, mindegyik, valamennyi, az összes, minden alkalommal, mindig ...)*

L-18 hangsúlyszabályokat befolyásoló konstrukciók, például. a páros kötőszók: *nem… hanem, akár... akár, vagy ... vagy, mind ... mind*

3 Az elemző tesztelése

Az elkészült hangsúlyelemző tesztelését 580 mondaton végeztük manuális módszerrel. A mondatok összesen 6974 szót tartalmaztak. Tehát ennyi hangsúlyjelzést ellenőriztünk. Felkészültünk arra, hogy az algoritmus sok hibát fog elkövetni. Már a fejlesztés során kiderült, hogy az ilyen elemzők működése erősen függ a szöveg felépítésétől, tartalmától, tehát az ilyen elemzőket hozzá kell igazítani az elemzett szöveghez (általános nyelvi elemző készítése tehát egyelőre irreális célkitűzés). Példaként bemutatunk néhány elemzendő és elemzett mondatot. A hangsúly jeleket a számítógépes feldolgozás szögletes zárójelek közé teszi, kettősponttal kezdődően.

A következő WAP oldalon a tranzakció részleteinek megadásával, annak jóváhagyására kérjük, amit beállításai szerint vásárlási kódjával, vagy anélkül tehet meg.

[:-]A [:W]következő [:N]WAP [:W]oldalon [:-]a [:W]tranzakció [:N]részleteinek [:N]megadásával,
[:N]annak [:E]jóváhagyására [:N]kérjük,
[:N]amit [:W]beállításai [:N]szerint [:W]vásárlási [:N]kódjával,
[:N]vagy [:W]anélkül [:N]tehet [:N]meg.

A forgalmi díj – az 50 Mbyte feletti forgalom esetén – minden időszakban mindössze 1 Ft minden megkezdett 10 kbyte után.

[:-]A [:W]forgalmi [:N]díj [:-]az [:W]1000 [:N]Mbyte [:N]feletti [:N]forgalom [:N]esetén
[:E]minden [:N]időszakban [:W]mindössze [:W]1 [:N]Ft [:E]minden [:N]megkezdett [:N]10
[:N]kbyte [:N]után.

Átírás cégek közötti jogutódlás a Számlafizető elhalálozása esetén lehetséges, minden egyéb esetben Számlafizető módosítás történik.

[:W]Átírás [:W]cégek [:N]közötti [:N]jogutódlás
[:N]és [:-]a [:W]Számlafizetõ [:N]elhalálozása [:N]esetén [:W]lehetséges,
[:N]minden [:N]egyéb [:W]esetben [:W]Számlafizetõ [:E]módosítás [:N]történik.

A Budapest Bank mobilbank menüjében lekérdezheti számlatörténetét és bankszámlájának egyenlegét, módosíthatja bankkártyája limitösszegeit, letilthatja, vagy aktiválhatja azt, átutalásokat végezhet, árfolyamokhoz juthat, vagy egyenlege változásairól értesítést állíthat be.

[:-]A [:W]Budapest [:N]Bank [:F]mobilbank [:N]menüjében [:N]lekérdezheti
[:N]számlatörténetét [:N]és [:N]bankszámlájának [:N]egyenlegét,
[:N]módosíthatja [:W]bankkártyája [:N]limitösszegeit,
[:N]letilthatja,
[:N]vagy [:N]aktiválhatja [:N]azt,
[:E]átutalásokat [:N]végezhet,
[:E]árfolyamokhoz [:N]juthat,
[:W]vagy [:W]egyenlege [:N]változásairól [:F]értesítést [:-]állíthat [:N]be.

4 A tesztelés eredményei

A hangsúlyjelzések ellenőrzésének eredményei szerint az 580 mondatból 98-ban nem volt hibás jelölés. A 6974 szóból összesen 1060 szón találtunk hibás jelölést. Hibák típus szerinti eloszlását az 1. táblázat mutatja. A négy hibatípus közül azokat számítjuk nagyobb hibának, amikor az adott szóra F, E, W kerül annak ellenére, hogy a szó hangsúlytalan, tehát az 1, 2, 3 kategóriákat. Ezekből összesen 302 esetet találtunk. Kevésbé zavaró hibának számítjuk a 4. kategóriát, mivel a hiányzó hangsúly kevésbé zavarja meg a hangzási képet, mint feleslegesen hangsúlyozott szó.

1. Táblázat: a hangsúly jelölési hibák eloszlása a vizsgált 580 mondatban

	A hiba típusa	A hibákszáma
1.	N kell F helyett	8
2.	N kell E helyett	64
3.	N kell W helyett	230
4.	W kell N helyett	228

Az eredmények tehát azt mutatják, hogy a súlyosabb hibából több van, mint a kevésbé zavaróból. A vizsgálatokból világossá vált, hogy a mondatelemzés legjelentősebb része a főnévi csoportok azonosítója. Egy ilyen mondatelemzési hiba több hangsúlyhibához is vezethet, mivel a hangsúlystruktúra a mondatszerkezetre épül. Végeredményben azt mondhatjuk, hogy az elemzési eredmények javítására egyrészről a főnévi csoportok azonosítóját kell pontosítani, más részről, az empirikus kutatást kell kiterjeszteni minél több mondatra. További vizsgálatokat kell végezni a szabályok sorrendjével kapcsolatban is, valamint azokat az eseteket is vizsgálni kell, ha két szabály esetleg ütközik egymással. Végül megjegyezzük, hogy a vizsgált szövegek nyelvi szerkezete sem mindig támogatja a sikeresebb elemzést.

7 Összefoglalás

Bemutattuk az első olyan mondatelemző algoritmust, amelyikkel magyar nyelvre végezhető hangsúly meghatározás. Legfőbb eredménynek azt tartjuk, hogy sikerült az eddigi, a témába vágó elméleti eredményeket algoritmikus formába önteni és működő programmá fejleszteni. Az automatikus elemzés hibázási aránya elég magas, a vizsgált szavak 15%-át hibás jelöléssel látja el. Ennek a hibaaránynak a csökkentése szerepel további céljaink között, valamint az, hogy kipróbáljuk az elemzőt különböző szövegtípusokon és meghatározzuk, hogy mennyi adaptációra van szükség, ha nem ügyfélszolgálati szövegeket kell elemezni.

Bibliográfia

1. Gábor, K.: Syntactic Parsing and Named Entity Recognition for Hungarian with Intex. In: Silberztein, Koeva, Maurel (eds.): Formaliser les langues avec l'ordinateur. Presses Universitaires de Franche-Comté, Besancon (2007) 353–366
2. É. Kiss, K.: The Syntax of Hungarian. Cambridge Syntax Guides. Cambridge: Cambridge University Press (2002)
3. É. Kiss, K., Kiefer, F., Siptár, P.: Új magyar nyelvtan. Budapest: Osiris (1998)
4. Kiss, G., Németh, G.: Tisztán statisztikai alapú szófaji címkéző használata a Szeged Korpuszon. IV. Magyar Számítógépes Nyelvészeti Konferencia MSzNy 2006., Szeged, (2006) 52–59

5. Koutny, I., Olaszy, G., Olaszi, P.: Prosody prediction fromt text in Hungarian and its realization in TTS conversion. International Journal of Speech Technology, Volume 3, Numbers 3-4. Kluwer Academic Publishers. (2000) 187–200
6. Olaszy, G.: The most important prosody patterns of Hungarian. Acta Linguistica Hungarica, Vol. 49 (3-4) (2002) 277–306
7. Tamm, A., Olaszy, G.: Kísérlet automatizált szövegelemzési módszerek kialakítására a szóhangsúlyok meghatározásához, in Z. Alexin, D. Csendes (szerk), III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, (2005) 383–393
8.Varga, L.: Intonation and Stress: Evidence from Hungarian. New York: Palgrave Macmillan (2002)

Ezt a kutatás-fejlesztést az NKFP 2. programja (szerződésszám: 2/034/2004) támogatta.