Szövegalapú nyelvi elemző kiértékelése gépi beszédfelismerő hibákkal terhelt kimenetén

Tündik Máté Ákos[1], Szaszák György[1]

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
e-mail:{tundik, szaszak}@tmit.bme.hu

Kivonat A cikkünkben felvázolt vizsgálat fókuszában az áll, hogy kiderüljön, milyen mértékű szintaktikai elemzést képes végrehajtani a „magyarlánc" nyelvi elemző a beszédfelismerő által kibocsájtott, hibákkal terhelt szövegeken, és ez az elemzés mennyiben „hasonlít" a hibátlan referenciaszövegen futtatotthoz, illetve azonosítható-e az elemzésnek olyan szintje, részeredménye, amely nagyban korrelál a hibátlan szövegével. A feladathoz egy híradós adatbázis 535 mondatból álló részhalmazát használtuk fel. Ezen a „magyarlánc" nyelvi elemzővel szintaktikai elemzést hajtottunk végre, mely meghatározta a mondatokra a szófaji és függőségi címkéket. Ezt követően a szintaktikai / szemantikai elemzések elemi részekre (szavakra) történő azonosítása és felbontása következett, majd az ezek halmaza felett megvalósított bag of words reprezentáció vizsgálata, melyet a korreláció, hasonlóság mérésére használtuk fel. További összehasonlítás történt a kinyert szófaji és dependencia tagek távolságszámításával is, a szóhibaarány számításával analóg módon. Az eredmények alapján elmondható, a beszéd-szöveg átalakítással nyert szövegeken végzett elemzés nagyban korrelál a hibáktól mentes referenciaátiraton végzettel. [1]

Kulcsszavak: gépi beszédfelismerés, nyelvi elemzés, információkinyerés

1. Bevezetés

A csupán írott szöveget felhasználó tartalomelemző, jelentéskivonatoló, kulcsszókereső alkalmazásokból számosat ismerünk, melynek társadalmi-gazdasági haszna megkérdőjelezhetetlen (pl. az adatbányászat területén). Ugyanezen funkciók beszéden történő megvalósítása nagyobbrészt még várat magára (leggyakrabban a kulcsszó alapú keresés az egyetlen elérhető funkció), holott jelentős társadalmi-gazdasági haszna feltételezhető, hiszen számos archívum vagy egyéb adathalmaz csak beszélt nyelvi adatokat tartalmaz, legépelése, beszéd-szöveg átalakítása nem történik meg.

A beszédfelismerő rendszerek „csak" szöveggé konvertálják a beszédet, azonban egyre inkább előtérbe kerül, hogy legyen emögött valamilyen beszédértést, a

[1] A szerzők köszönetüket fejezik ki a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatalnak, amely a PD-112598 projekt keretében az itt ismertetésre kerülő kutatást támogatta.

tartalom elemzését, értelmezését megvalósító funkció is. Napjainkban a rendelkezésre álló adatok egyre nagyobb hányada hangfelvétel, így a probléma megoldása egyre inkább hangsúlyossá válik. A hanganyagokon használható elemzők köre jóval szűkösebb, illetőleg két lehetőség kínálkozik: közvetlenül a beszédből nyerjük ki az információt, illetve a beszédet szöveggé alakítva szöveges elemzőket, keresőket alkalmazunk.

Az utóbbi lehetőség azt a potenciált is magában rejti, hogy az írott nyelvre már rendelkezésre álló elemzőeszközöket is felhasználhatjuk. Ennek során alapvető problémaként jelentkezik, hogy a beszédfelismerővel átalakított szöveg többkevesebb felismerési hibát – szócserét, -törlést vagy -beszúrást – tartalmaz. A kutatás első lépéseként azt vizsgáltuk, mennyire működik hatékonyan a nyelvi elemzés a beszédfelismerő által szolgáltatott kimeneten. Munkánkhoz a „magyarlánc" magyar nyelvű, függőségi nyelvtan alapú szintaktikai elemzőt [1] használtuk, egy médiából származó híranyagokon [2] alkalmazott beszéd-szöveg átalakítási lépés után. A beszédfelismerő közel 35% szóhibaaránnyal működött a választott anyagokon.[2]

Vizsgálatunk arra irányult, hogy kiderüljön, milyen mértékű szintaktikai elemzést képes végrehajtani az elemző a hibákkal terhelt szövegen, és ez az elemzés mennyiben „hasonlít" a hibátlan szövegen futtatotthoz, illetve azonosítható-e az elemzésnek olyan szintje, részeredménye, amely nagyban korrelál a hibátlan szövegével[3].

A cikkünk az alábbi struktúra szerint épül fel: elsőként bemutatjuk a korpuszt, illetve a „magyarlánc" nyelvi elemzőből kinyert adatokon végzett, összehasonlíthatóságot célzó utófeldolgozást. Ezután sor kerül a beszédfelismerő kimenetének és a referenciaszöveg feldolgozásának ismertetésére, amely több lépést is magában foglal. Végül a vizsgálati eredményeket ismertetjük az elemzések részletes, többszintű összehasonlítása mellett. A puszta korreláció mérése mellett további összehasonlításokat, hasonlósági és távolsági metrikákat is megadunk, a kinyert szófaji (POS) és függőségi (DEP) adatokra.

2. Anyag és módszer

2.1. A felhasznált híradatbázis

A kísérleteinkhez magyar nyelvű televíziós hírműsorok felvételeit használtuk fel. A felvételek két közszolgálati és két kereskedelmi csatornáról származtak, hozzávetőleg egyenletes eloszlásban, mondatszinten leiratozva. Összesen 535 mondatnyi anyagot választottunk ki vizsgálatra véletlenszerűen, de a hírblokkok egységét megtartva.

[2] A hivatkozott beszédfelismerő ennél lényegesen kisebb szóhibaarányt szolgáltat, esetünkben szándékosan állítottunk be ezt a magasabb értéket.

[3] A választott anyagokra nem áll rendelkezésünkre „gold standard" elemzés, ugyanakkor nem is célunk a „magyarlánc" elemző abszolút pontosságának mérése, munkánkhoz elegendőnek tartjuk a helyes referenciaszöveggel való összevetést.

Egy-egy hangfájl jellemzően egy hírblokkot tartalmazott, amelyet valós idejű médiafeliratozásra fejlesztett beszédfelismerő rendszerrel [2] szöveggé alakítottunk. A felismerést ezúttal szándékosan viszonylag magas, átlagosan 35%-ot közelítő szóhibaarányt szolgáltató akusztikai és nyelvi modell kombinációval végeztük, a felismert anyagokon pedig a szóhibaarány viszonylag nagy szórást mutatott (lásd 6. ábra), ami szempontunkból a teljes körű analízishez és az egyes eredmények szóhibaarány függésének megadásához kedvező beállítás.

2.2. Utófeldolgozás és adatreprezentáció

A referenciaszövegen és a beszédfelismerő kimenetét tartalmazó szöveges átiraton a „magyarlánc" nyelvi elemzővel végrehajtottuk a szintaktikai elemzést, mely meghatározta a mondatokra a szófaji és függőségi címkéket [1].

A feladat a beszédfelismerő kimenetének és a referenciaszöveg normalizálásával kezdődött, kézi központozással. A beszéd-szöveg átalakítás másik nehézsége a szóhibákon túl, hogy az írásjelek, központozás sem minden esetben megoldott. Jelen munkában ettől eltekintünk, és a központozást kézzel pótoljuk, amire különösen azért van szükség, mert a nyelvi elemző erre nagymértékben támaszkodik.

A szintaktikai / szemantikai elemző kimenetén előállt szófaji és dependencia tageket információ-visszakereső rendszerekben használt vektortér modellbe (vector space model) transzformáltuk. Ez a modell magában foglalja a szózsák (bag of words)-megközelítést is. Az információkeresésben ismert szózsák modellben a szavak dokumentumon belüli előfordulási gyakorisága az, ami számít, nem a sorrendjük. Ebben a modellben az *a fa zöld* és az *a zöld fa* rövid dokumentumok azonosan fognak viselkedni. Világos, hogy eltérő jelentésűek, azonban az is igaz, hogy mindketten relevánsak a fákat és a zöld színt kulcsként tartalmazó lekérdezésekre.

A vektortér modell eredeti ötlete szerint minden egyes dokumentumot (a mi esetünkben a dokumentumpárok egy-egy mondatpárnak felelnek meg) unigram szógyakoriságok vektoraként ábrázolnak. Ezt a modellt felhasználva, a szófaji tagek gyakoriságát és a dependencia tagek gyakoriságát vizsgáltuk az egyes mondatpárokra, valamint az indikátorvektorokat is megadtuk, ami prezencia / abszencia jellegű viselkedést ír le. Megfontolásaink szerint ugyanis egy információkinyerést célzó felhasználásban is legfőképpen a szófaji és a függőségi elemzésre való támaszkodás dominál [3].

Ezen kívül a mondatokban előforduló szófaji és függőségi címkék helyett azok szófaji és függőségi címkelistában elfoglalt sorszáma szerint is megcímkéztük a tokeneket (szavakat), melynek adatreprezentációja az 1. ábrán látható, egy adott példamondatpárra vonatkoztatva.

Prezencia vagy abszencia vizsgálata esetén az előbbi vektorok 0 értékei 0, 0-nál nagyobb értékei 1 értékűek lennének. A 2. ábrán a gyakorisági szózsák (pontosabban szófajzsák) reprezentáció látható ugyanerre a mondatra a szófaji címkék uniója alapján.

```
Referenciaszöveg: Szerbiában a pravoszláv karácsony állami ünnep.
Felismert szöveg: Szerbiában a pravoszláv karácsony áll    aminek.

Szófaji gyakorisági maszk
[N, V, A, P, T, R, S, C, M, I, X, Y, Z, 0]
Referencia szófaji gyakoriság
[3, 0, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Felismerésre vonatkozó szófaji gyakoriság
[2, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

1. ábra. Szófaji alapú gyakorisági reprezentáció egy példamondatpárra

```
Szózsák maszk
[N, V, A, P, T ]
Referencia szózsák
[3, 0, 2, 0, 1 ]
Felismerésre vonatkozó szózsák
[2, 1, 1, 1, 1 ]
```

2. ábra. Szófaji alapú szózsák reprezentáció az előző példamondatpárra

2.3. Az adatsorok összehasonlításához használt mértékek

Az adatsorok összehasonlításánál több szempontot is figyelembe kellett venni, tekintettel arra, hogy kategorikus adatokról van szó. Az egyik fő megközelítés a **prezencia / abszencia** szempontú vizsgálat, amely azon alapul, hogy mely POS és DEP tagek fordulnak elő az egyes adatsorokban. Továbbhaladva, számításba vettük az egyes kategóriák előfordulásának **gyakoriság**át is, erre kétféle hasonlóságot vetettünk be; egyrészt az összes címke halmazát használtuk fel, másrészt az aktuális referenciamondat és beszédfelismerő kimenet szófaji, illetve függőségi címkéinek unióját vettük, és azon halmaz felett hajtottuk végre az összehasonlítást. Hasonlóságot kerestünk az adatsorok között úgy is, hogy a kategorikus címkéket az előre definiált címkelistában elfoglalt sorszámukkal helyettesítettük a vektorban, és így vetettük össze a vektorokat. Ez utóbbi eljárásra **sorrendi** összehasonlításként utalunk a továbbiakban.

Elsőként az adatsorok Pearson-korrelációját határoztuk meg. A korreláció jelzi azt, hogy két tetszőleges érték nem független egymástól. Az ilyen széles körű használat során számos együttható, érték jellemzi a korrelációt, alkalmazkodva az adatok fajtájához:

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2} \sqrt{\sum_i (Y_i - \bar{Y})^2}}, \tag{1}$$

ahol a felülvonásos betűk a várható értéket, X és Y pedig az adatsorokat jelölik.

A hasonlóság másik lehetséges mértéke a felismert- és a referenciavektor skaláris szorzata. Geometriailag a két vektor skaláris szorzata az általuk bezárt

szög koszinusza, azaz ha két ilyen vektor koszinuszát maximalizáljuk (amennyiben azonos kvadránsban találhatók), akkor az általuk bezárt szög nullához közeli lesz. Ezen alapul az úgynevezett koszinusz-hasonlóság számítása:

$$sim(d_j, d_k) = \frac{d_j d_k}{|d_j||d_k|} = \frac{\sum_i w_{i,j} w_{i,k}}{\sqrt{\sum_i (w_{i,j})^2}\sqrt{\sum_i (w_{i,k})^2}} \quad (2)$$

A koszinusz-hasonlóságot és a Pearson-korrelációt a referenciában és a beszédfelismerő kimenetéből kinyert POS bigramok prezencia / abszencia és gyakoriság alapú vektorreprezentációira is kiszámítottuk.

Az adatsorokat megvizsgálva gyakran előfordult, hogy a beszédfelismerő kimenetén megjelenő szóbeszúrás vagy szókihagyás miatt az adatsorok hasonlósága rosszabb értéket mutatott annál, mint amivel intuitívan „ránézésre" rendelkezett, hiszen a hasonlósági mértékeink esetén főként az egyes címkék páronkénti összehasonlítására koncentráltunk. A POS-tagek alapján történő összehasonlításnál ennek kapcsán felhasználtuk a bioinformatikában használt Needleman-Wunsch globális szekvencia-illesztő algoritmust [4], melyet 4 pontozási értékkel súlyoztunk. Ha megegyeztek a i. indexen talált karakterek, ez 1 pontot ért, ha pedig nem, akkor az 0-t. Ha egy új hézagot kellett nyitni az igazításhoz, azt -0.5 ponttal büntette az algoritmus, ha pedig meghosszabbítani kellett, azt -0.1 ponttal. Az algoritmus hasonlít a sztringek összehasonlításához használt Levenshtein-távolsághoz [5], de annyiban meghaladja azt, hogy konkrét illesztési eredményeket szolgáltat a szekvenciákra, melyből a legnagyobb pontszámot választjuk ki, mivel ott a legnagyobb az egyezés.

Alább egy példát közlünk, ahol jól látszódik az igazítás haszna. Vegyük az alábbi referenciamintára és a beszédfelismerő kimenetére futtatott szófaji elemzést:

```
NRVTNS
CNRVTNS
```

Így Needleman-Wunsch igazítás nélkül a páronkénti összehasonlításból adódó korreláció értéke: -0,934 lesz. Ugyanakkor, ha felhasználjuk az igazító algoritmust, az alábbi rendezést kapjuk:

```
-NRVTNS
CNRVTNS
```

Így a korreláció értéke máris a valósághoz közelebb esően alakul: 0,895.

A függőségi címkékre ezt az illesztési módszert nem alkalmaztuk, helyette a nemzetközileg is használt kiértékelési paramétereket választottuk, azzal a kényszerrel élve, hogy csak az egyező hosszúságú mondatokra határoztuk meg. A LAS (Labeled Attachment Score) esetében azok a függőségi ívek érnek pontot, ahol a beszédfelismerő kimenetén lévő adott ív mind a szülőobjektumot tekintve (ez egy sorszám), mind az ívre írt függőségi élcímke megegyezik a referencia átirat függőségi ívéhez viszonyítva, míg az ULA (Unlabeled Attachment Score) esetében elégséges a szülő csomópont egyezése (itt nem számít hibának a rossz élcímke).

```
Referenciaszöveg: [...] amit látok, az tényleg megtörténik [...]
Felismert szöveg: [...] amit látok, azt tényleg megtörténik [...]

LAS[%]=94,12; UAS[%]=100; LA[%]=94,12.
```

3. ábra. Függőségi alapú összehasonlítás

A LA (Label Accuracy) esetén pedig a függőségi élcímkék egyezése számít [6]. A 3. ábrán egy példát is láthatunk.

Megállapítható tehát, hogy *az-azt* páros eltérése a függőségi kapcsolatokat az élcímkék szintjén befolyásolta, viszont maguk a függőségi ívek nem változtak. Láthatjuk, hogy ebben az esetben a globális illesztő függvény alkalmazása további alapos megfontolásokat igényelne (pl. milyen karakterrel jelöljük az igazítási hézagokat, és milyen címkét kapjanak?), így ezt nem alkalmaztuk.

A következő összehasonlítás a szóhibaarány (WER: Word Error Rate) mintájára történt. Ennek a képlete:

$$WER = \frac{S + D + I}{N}, \tag{3}$$

ahol S jelenti a szócserék számát, D a törölt szavak számát, I a szóbeillesztések számát, N pedig az eredeti szóhalmaz méretét. Ennek mintájára megalkottuk a csak szótövekre értelmezett SER, mint Stem Error Rate; a szófaji címkékre értelmezett PER, mint POS Error Rate; valamint a függőségi címkékre értelmezett DER, mint Dependency Error Rate mérőszámokat. Az utóbbi összefüggéseket a címkékből képzett bigramokra is meghatároztuk.

A szófaji és függőségi tagek vizsgálata mellett mondatszintű jellemzőket is meghatároztunk, úgy, mint pl. az átlagos Levenshtein-távolság és Jaro-Winkler távolság [7].

3. A vizsgálati eredmények

3.1. Szófaji címkék hasonlósága

Az 1. táblázatban láthatóak a referenciaszöveg és a beszédfelismerő hibákkal terhelt kimenetének Pearson-korreláció és koszinusz-hasonlóság alapú kapcsolatára vonatkozó összehasonlítás eredményei, melyeket az unigram szófaji címkék, valamint a belőlük képzett bigramok között határoztunk meg.

A továbbiakban értékeljük a különböző megközelítéseket, elsősorban az automatikus rendszerekben történő felhasználhatóság szempontjából. A prezencia / abszencia (1 / 0) jellegű értékek a kapcsolat erősségének szempontjából a leggyengébbek, hiszen nem veszik figyelembe az egyes szófaji címke gyakoriságokat. A gyakoriságot figyelembe vevő értékek közül az összes szófaji címkére vonatkozó a gyengébb a szózsák megközelítéshez képest, hiszen az összehasonlításnál az összes címke halmazán több közös abszencia adódik, ami így pozitívan súlyozza az eredményt, míg a szózsák modellt a referenciaszöveg és a beszédfelismerő

1. táblázat. Pearson-korreláció és koszinusz-hasonlóság szófaji címkékre és belőlük képzett bigramokra

Vizsgálattípus	Pearson-korreláció		Koszinusz-hasonlóság	
	unigram	bigram	unigram	bigram
Prezencia / abszencia jellegű (összes címke)	0,86	0,74	0,91	0,75
Gyakoriság jellegű (összes címke)	0,89	0,76	0,92	0,77
Páronkénti, sorrendi alapú	0,54	–	0,84	–
Páronkénti, sorrendi, Needleman-Wunsch	0,65	–	0,88	–
Gyakoriság jellegű (szózsákra, címkeunió)	0,73	0,30	0,92	0,77

kimenetének szófaji címkéinek uniójából képzett tér felett értelmeztük. A legnagyobb információértéket a páronkénti összehasonlítás képviselné, hiszen ezeknél a pozícióinformációt is figyelembe vettük, vagyis hogy a mondat megfelelő sorszámú tokenjei azonos szófajú címkével rendelkeznek-e. Jól látszik a Needleman-Wunsch algoritmus pozitív hatása, az adatsorok egymásra igazításával nőtt a Pearson-korreláció és a koszinusz-hasonlóság is. Ugyanakkor tudni kell, hogy a páronkénti összehasonlításon alapú értékek az átlagra és a szórásra érzékenyek, holott ez kategorikus adatoknál nem szabadna figyelembe venni, ezért az alábbi megközelítés matematikailag helytelen. Kijelenthető tehát, hogy jelenleg a mérőszámok közül a legoptimálisabb a gyakoriságon alapuló szózsák megközelítés a referenciában és a felismert szövegben előforduló POS tagek uniója felett.

3.2. Függőségi címkék hasonlósága

Táblázatba foglaltuk a függőségi címkékre vonatkozó hasonlósági értékeket is (2. táblázat). A szófaji címkéknél leírt, az összefüggés erősségét érintő megállapítások érvényesek a függőségi címkékre is. Ugyanakkor a nemzetközileg elterjedt LAS / UAS / LA mértékek a pozícióinformációt is figyelembe veszik, ezért ezek mutatják a legerősebb összefüggést (lásd 3. táblázat). A hátrányuk viszont az, hogy ezt a módszert megegyező mondatok „gold standard" szerinti elemzésének és függőségi elemző szerinti elemzésének összehasonlítására találták ki, amely magával vonja azt, hogy a függőségi címkék száma is egyenlő. Ezeknek a mértékeknek a különböző hosszúságú mondatokra történő adaptációja további megfontolásokat igényel.

2. táblázat. Pearson-korreláció és koszinusz-hasonlóság függőségi címkékre

Vizsgálattípus	Pearson-korreláció	Koszinusz-hasonlóság
Prezencia / abszencia jellegű (összes címke)	0,82	0,87
Páronkénti, sorrendi alapú	0,49	0,74
Gyakoriság jellegű (szózsákra)	0,63	0,89

3. táblázat. Függőségi címkék összehasonlítása

Vizsgálattípus	Hasonlóság
LAS	80,5 %
UAS	86,6 %
LA	84,3 %

3.3. Hibaarány jellegű jellemzők

Vegyük sorra először az elemzési egységre vonatkoztatott hibaarány alapú értékeket (4. táblázat). Látható, hogy az unigram megközelítésben a legszigorúbb a szóhibaarány alapú megközelítés, hiszen előfordulhat, hogy a szótövező ugyanazt rendeli a referenciában és a felismert szövegben előforduló tokenhez. Ha pedig a szótő sem egyezik, ettől még előfordulhat, hogy ugyanolyan szófajú tokenre téved a felismert szöveget feldolgozó nyelvi elemző, mint ami a referenciaszövegben van. Ugyanakkor látható, hogy a bigram megközelítésben is elviselhető hibaarány mutatható ki, közel esik a korábban említett, beszéd-szöveg átalakításból eredő 35% körüli szóhibaarányhoz. A **részletes** POS- és POS-bigram hibaarány meghatározására is lehetőségünk volt, mivel nemcsak a fő szófaji címkék álltak rendelkezésre, hanem „finomszemcsés" POS címkék is (pl. igék esetén az igeragozást is tartalmazza).

4. táblázat. Hibaarányok

Vizsgálattípus	Hibaarány értéke
Szóhibaarány	0,35
Szótőhibaarány	0,29
POS hibaarány	0,22
POS bigram hibaarány	0,34
Részletes POS hibaarány	0,29
Részletes POS bigram hibaarány	0,43
DEP hibaarány	0,25
DEP bigram hibaarány	0,39

Az 5. táblázatban néhány általános szövegszintű jellemzőt is megadunk a felhasznált korpuszra.

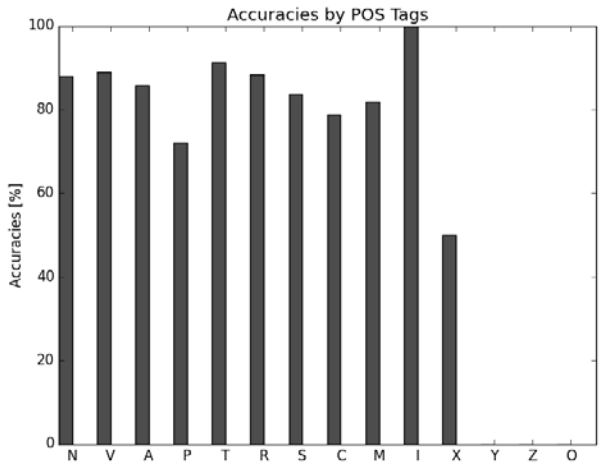
A 4. ábrán a szófaji címkék szerinti pontossági értéket ábrázoltuk, az egyező hosszúságú mondatokra, a „magyarlánc" elemző által használt szófaji kódokat megtartva [1]. Az indulatszó (I) kategória pontosságát figyelmen kívül hagyva - alacsony elemszáma miatt - a legjobb egyezéseket a főnév (N), ige (V) és a névelő (T) kategóriákban adta az elemző.

A függőségi viszonyokra vett kategóriánkénti megoszlás a 5. ábrán látható, az egyező hosszúságú mondatokra. Látható, hogy az elemző a legpontosabb a főnévhez tartozó névelő, a mondat gyökereként (ROOT) meghatározott igéhez tartozó igekötő (PREVERB), a helyhatározós (TLOCY) és a jelzős szerkezetek (ATT)

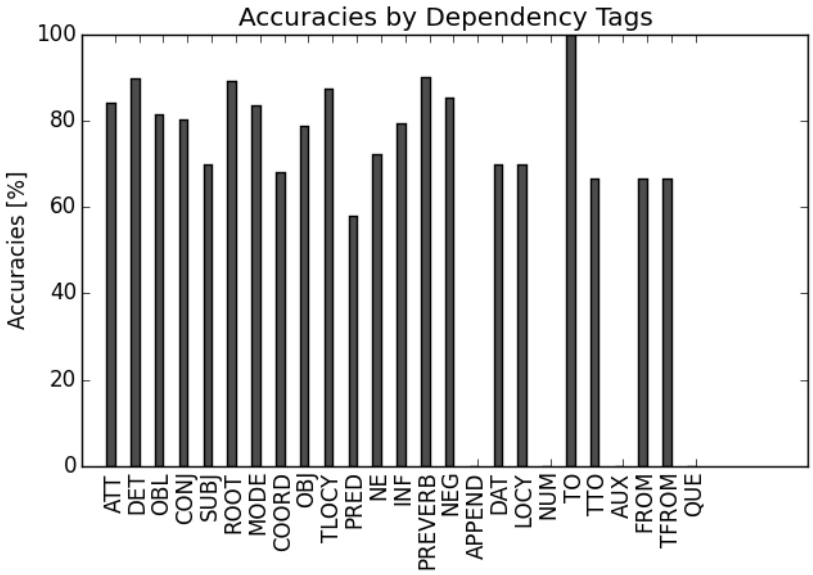
5. táblázat. Néhány szövegszintű jellemző

Vizsgálattípus	Érték
OOV-arány a referencia mondatokra	0,013
OOV-arány a felismert mondatokra	0,012
Referenciaszavak száma	2968
Felismert szövegben a szavak száma	3018
Mondatszintű Levenshtein-távolság	12
Mondatszintű Jaro-Winkler-távolság	0,92
Leghosszabb mondat (szóban mérve)	43
Átlagos mondathossz (szóban mérve)	12,17

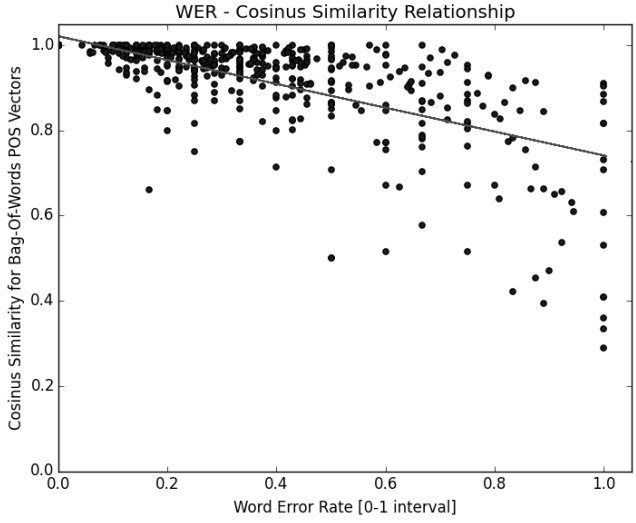
megállapításában volt. A szóhibaarány és a gyakoriság alapú, szófajokra tekintett koszinusz-hasonlóság közötti összefüggést a 6. ábra mutatja. Az összefüggés a várakozásoknak megfelelő, a szóhibaarány romlása a koszinusz-hasonlóság gyengülését vonja maga után, lineáris regressziót alkalmazva -0,279 meredekség és 1,02-es metszéspont adódott. Látható az adatokból, hogy az összefüggés közel lineáris, nem azonosítható tehát olyan szóhibaarány-küszöbérték, amelyet elérve a koszinusz-hasonlóság meredeken esni kezdene.



4. ábra. Az egyes szófaji kategóriák pontossága



5. ábra. Az egyes függőségi kategóriák pontossága



6. ábra. A szóhibaarány és a szófajokra vonatkozó, gyakoriság alapú koszinusz-hasonlóság összefüggése

4. Összegzés

A „magyarlánc” nyelvi elemzővel elért eredmények alapján elmondható, hogy a beszéd-szöveg átalakítással nyert szövegeken végzett elemzés nagyban korrelál a beszédfelismerési hibáktól mentes (referenciaátirat) szövegen végzettel. A kapott eredmények tanúsága szerint a vizsgált híranyag korpuszon a szófaj, illetve a függőségi viszony tévesztését (megváltozását) is eredményező beszédfelismerési hibák száma az összes felismerési hiba mintegy 2/3-ára tehető. Bigram kapcsolatokat tekintve a szófajtévesztés valószínűsége nagyon közel esett a beszéd-szöveg átalakítás szóhibaarányához. Az automatikus információkinyerés és tartalmi kivonatolás szempontjából leginkább releváns főnévi és igei szófajkategóriák esetében a tévesztési arányok még kedvezőbbek, kevesebb, mint a beszédfelismerési hibák felére tehetőek. Jóllehet a józan megfontolás alapján is szoros összefüggést várnánk a szóhibaarány és a felismerési hibákkal terhelt, valamint a hibátlan szövegek szintaktikai elemzéseinek hasonlósága között, kísérletileg is megerősítettük, hogy ez az összefüggés közel lineáris, amiből az következik, hogy automatikus beszéd-szöveg átalakítással nyert szövegek nyelvi elemzésekor nem található olyan kritikus szóhibaarány érték, amelyen túl a nyelvi elemzés drasztikus, fokozódó leromlásával kellene számolni, és így a szóhibaarány alapján jól előrejelezhető az elemzés pontossága is. Az eredmények alapján érdemes lenne a beszéd-szöveg átalakítást, majd nyelvi elemzést megvalósító feldolgozási láncot közvetlenül „gold standard” referenciával összevetve kiértékelni. Ígéretes jövőbeli kutatási irány lehet továbbá a beszéd elemzéséből (pl. hangsúlydetekcióból vagy általánosabban prozódiaalapú elemzésből) származó, szavakhoz rendelt tagek továbbvitele akár a nyelvi elemzésbe, akár közvetlenül az információkinyerésbe.

Hivatkozások

1. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP (2013) 763-771
2. Tarján, B., Fegyó, T., Mihajlik, P.: A Bilingual Study on the Prediction of Morph-based Improvement. In: 4th International Workshop on Spoken Languages Technologies for Under-Resourced Languages, Saint Petersburg, Russia (2014) 131-138
3. Lioma, C. and Blanco, R.: Part of speech based term weighting for information retrieval. In: M. Boughanem, C. Berrut, J. Mothe and C. Soule-Dupuy (Eds.), ECIR, LNCS Vol. 5478 (2009) 412-423
4. Beddoe, Marshall A.: Network protocol analysis using bioinformatics algorithms, http://www.4tphi.net/ awalters/PI/pi.pdf (2004)
5. Singh, S. P., Kumar, A., Darbari, H., Chauhan, S., Srivastava, N., Singh, P.: Evaluation of Similarity metrics for translation retrieval in the Hindi-English Translation Memory. Int. Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 8 (2015)
6. Choi, Jinho D., Tetreault, J., Stent, A.: It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL. (2015)
7. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string metrics for matching names and records. Kdd workshop on data cleaning and object consolidation. Vol. 3. (2003)