Statisztikai és hibrid módszerek párhuzamos korpuszok feldolgozására

Laki László János[1], Prószéky Gábor[1,2]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar
1083 Budapest, Práter u. 50/a
laklaja@digitus.itk.ppke.hu
proszeky@itk.ppke.hu
[2] MorphoLogic
1116 Budapest, Kardhegy u. 5.
proszeky@morphologic.hu

Kivonat: Előadásunkban foglalkozunk a statisztikai gépi fordítás minőségének javításával, az egyre mélyebb hibridizáció alkalmazásával, majd az angol–magyar kísérletek mellett olyan, morfológiailag közelebb álló nyelvpárok bevonásával, mint a lovári cigány nyelv és a magyar. Az előadás második felében egy tisztán statisztikai alapon működő szövegannotáló rendszer létrehozásával és kiértékelésével foglalkozunk.

1 Bevezetés

Az informatika fejlődése szinte az összes tudományág számára új lehetőségek halmazát nyitotta meg, és ez nem volt másképp a nyelvészetben sem. Napjaink számítógépei segítségével képesek lettünk óriási méretű szöveges anyagok gyors és hatékony kezelésére, valamint feldolgozására. Könnyen belátható, hogy a szabályalapú módszerekkel nagyon nehéz a nyelvekben kifejezett kapcsolatokban meglevő törvényszerűségeket megfogalmazni, viszont kézenfekvő megoldásnak tűnik statisztikai módszereket használni ezen feladatok megoldására. Jelen munkánk célja a magyar és más nyelvek közti átalakítások vizsgálata, gyakorlati megvalósításuk, illetve a meglévő módszerek javítása és a témában lévő lehetőségek felmérése.

2 Statisztikai gépi fordítás

A statisztikai nyelvfeldolgozás elterjedt alkalmazása a gépi fordítás. A statisztikai gépi fordító (SMT) módszer nagy előnye a szabályalapú fordítással szemben, hogy az architektúra létrehozásához nem szükséges a nyelvek grammatikájának ismerete. A rendszer tanításához csupán egy kétnyelvű korpuszra van szükség, amelyből statisztikai megfigyelésekkel nyerjük ki a szabályokat. Az idegen nyelvpárokon elért komoly eredményekben bízva választottuk ezt a módszert, hogy létrehozzunk egy fordítórendszert – először az angol–magyar nyelvpárhoz.

A fordítás során az egyetlen, amit biztosan tudunk, az a mondat, amit le szeretnénk fordítani (forrásnyelvi mondat). Ezért a fordítást úgy végezzük, mintha a célnyelvi mondatok halmazát egy zajos csatornán átengednénk, és a csatorna kimenetén összehasonlítanánk a forrásnyelvi mondattal. Az a mondat lesz a rendszerünk kimenete, amelyik a legjobban hasonlít a fordítandó mondatra. Ez a hasonlóság lényegében egy valószínűségi érték, amely a nyelvi modellből és a fordítási modellből számolható. Rendszerünk felépítéséhez egy angol–magyar párhuzamos korpuszt használunk, amely a Hunglish [3] korpusz két alkorpuszából áll: a Literature és a Magazines nevű részekből (a továbbiakban LitMag). A LitMag korpusz 654 939 mondatot és 9 425 911 szót tartalmaz, így kisméretű korpusznak tekinthető.

Több módszert is megvizsgáltunk, melyek képesek párhuzamos korpuszból információt kinyerni. Végül az IBM modellek mellett döntöttünk, mivel, hatékony, viszonylag pontos, és a feladatnak nagyon jól megfelelő algoritmusnak bizonyultak. Ezen okból kezdtük használni a Moses keretrendszert [5], amelyik implementálja ezeket a modelleket. Ebben a rendszerben megtalálható a párhuzamos korpusz előfeldolgozása, a fordítási és nyelvi modellek létrehozása, a dekódolás, valamint a BLEU-pontra optimalizálás. A kiértékeléshez a BiLingual Evaluation Understudy (BLEU) módszert használtuk, amelynek lényege, hogy a fordításokat referenciafordításokhoz hasonlítja, majd pontozza őket. Eredményként egy 0 és 1 közötti valós számot kapunk. A fentebb jelzett tanítás után alaprendszerünk 0,1085 (10,85%) BLEU-pontot kapott.

Szembetűnő eredmény, hogy a más nyelvpárok közti fordítógépek eredményei sokkal jobbak az angol–magyar nyelvű fordítógépekhez képest: angol–francia 32%, spanyol–katalán > 40%, angol–spanyol 30% [2]. Ha jobban megvizsgáljuk az eredményeket, az is látszik, hogy az előbb felsorolt nyelvek között nagy hasonlóság van mind nyelvtanilag, mind a szavak szintjén. Kis eltérés van ugyanis a szórendben és a nyelvtani szerkezetekben. Ennek köszönhetően a gépi fordító rendszer nagyobb biztonsággal képes létrehozni a transzformációkat a két nyelv között, valamint a kapott fordítások is sokkal jobban fognak hasonlítani a referenciafordításokhoz. Ezzel szemben a magyar és az angol nyelv között igen jelentős a formai eltérés. Ennek köszönhetően, ha ugyanazt a rendszert használjuk angol–francia vagy angol–magyar viszonylatban, jelentős minőségi különbség figyelhető meg. A pusztán statisztikai rendszer helyett tehát célszerű valamilyen szabályalapú elemet is tartalmazó, ún. hibrid rendszert alkalmazni a fordítás minőségének javítása érdekében.

3 Hibrid rendszerek

3.1 Szótár hozzáadása a korpuszhoz

A fordítórendszerek kiértékelésénél megfigyelhető, hogy a szóösszerendelő nehezen találja meg az összetartozó szövegrészeket, ha azok a nyelvtani szerkezet miatt messze vannak egymástól, vagy nagyon eltérnek. Az első ötlet a minőség javítására, hogy az eredeti korpuszhoz hozzáadunk egy kétnyelvű szótárat. Ettől azt az eredményt reméltük, hogy a kifejezések pontos fordítása nem csak segít az összerendelőnek megtalálni az összetartozó kifejezéseket a mondatban, de csökkenti a lefordítatlan szavak számát is.

Ehhez a feladathoz egy egyszerű angol–magyar szótárat használtuk [10], melyet először átalakítottunk, hogy egy kifejezésnek csak egyetlen megfelelője legyen. Így 344 924 darab kifejezéspárt kaptunk. Az elkészült szótárat többször is beleraktuk a korpuszba abból a célból, hogy a helyes előfordulások minél nagyobb súllyal forduljanak elő a fordítási modellben. Ezzel párhuzamosan viszont folyamatosan csökken az eredeti korpusz fontossága, csökken a többszavas kifejezések súlyozása a fordítási modellben, és romlik a nyelvi modell minősége. Ennek érdekében meg kell találni azt a számot, hogy hányszor éri meg a szótárat hozzáfűzni a korpuszhoz. E célból oly módon tanítottuk be az SMT rendszert, hogy az eredeti korpuszhoz egyszer, kétszer, háromszor, négyszer és ötször hozzáadtuk a kétnyelvű szótárat. Így a következő eredményeket kaptuk:

1. táblázat: Különböző rendszerek BLEU-eredményei.

Rendszer	**BLEU-érték**
Alaprendszer fordítása:	10,85%
Alap+1xszótár rendszer fordítása:	11,18%
Alap+2xszótár rendszer fordítása:	11,01%
Alap+3xszótár rendszer fordítása:	10,88%
Alap+4xszótár rendszer fordítása:	10,87%
Alap+5xszótár rendszer fordítása:	10,87%

Az 1. táblázatból látszik, hogy az alaprendszer (10,85% BLEU) értékéhez képest 0,33%-os javulás figyelhető meg, amely mértéke a behelyezett szótárak számától függően folyamatosan csökken. Ez a görbe azért tetőzik az első esetben, mert a szótár mérete összemérhető az eredeti korpusz méretével (fele az eredeti korpusznak), és emiatt annak ismétlése viszonylag hamar eltolja a súlyokat. A többszörös szótárhozzáadástól várt javuláshoz szükséges lenne egy nagyobb méretű párhuzamos mondatszintű korpusz is, de erőforrási problémák miatt nem tudtunk ilyet használni.

A teszthalmazból kiválasztott példamondat fordításait a 2. táblázat tartalmazza. Az első sorban az eredeti angol mondat olvasható, a másodikban ennek a referenciafordítása, majd az alaprendszer, végül a több szótárral kiegészített SMT fordítások eredményei.

Rögtön az első kifejezés elemzésénél feltűnik az *i wonder* fordításában való eltérés. Mind az alaprendszer, mind a legjobb eredményt nyújtó első rendszer *csak tudnám*-ra, míg a többi a *kíváncsi vagyok*-ra fordítja. Annak ellenére, hogy mind a két fordítás helyes, az automatikus kiértékelő mégis más eredményt ad a két fordításra, mivel a referenciafordításban a *kíváncsi vagyok* szerepel.

A következő érdekes kérdés a *teaching us* elemzése. A fordítás vizsgálatából kiderült, hogy az alaprendszer a *teaching*-et az *a tanítást*-ra fordította, ami a mondatbeli jelentéstől nem is áll messze. Ezzel szemben a szótárral kiegészített rendszerekben egységesen a *tanított nekünk* kifejezés érte el a legnagyobb valószínűséget, amely az *us* fordítását (*nekünk*) jobban tükrözi. Sőt, kissé elvont értelmezéssel az eredeti jelentéshez is közelebb áll, a szó szerinti fordítás viszont eltávolodott. A legnagyobb probléma itt is az, hogy nem egyezik meg a referenciával, ezért sem kap nagyobb BLEU-értéket.

2. táblázat: Különböző rendszerek eredményeinek összehasonlítása.

Angol referencia-fordítás:	" i wonder who 'll be teaching us ? " said hermione as they edged into the chattering crowd .
Magyar referenciafordítás:	- kíváncsi vagyok , ki tartja a tanfolyamot - morfondírozott hermione , miközben barátaival befurakodtak a tömegbe .
Alaprendszer fordítása:	- csak tudnám , ki lesz a tanítást ? - kérdezte hermione , mikor ő az .
Alap+1xszótár rendszer fordítása:	- csak tudnám , ki lesz tanított nekünk ? - szólt hermione , mikor elindult a jóvoltából .
Alap+2xszótár rendszer fordítása:	- kíváncsi vagyok , aki tanított nekünk ? - szólt hermione , mikor elindult a zsibongó tömeg .
Alap+3xszótár rendszer fordítása:	- kíváncsi vagyok , ki lesz tanított nekünk ? - szólt hermione , mikor elindult az összeverődött tömegen .
Alap+4xszótár rendszer fordítása:	- kíváncsi vagyok , ki lesz tanított nekünk ? - szólt hermione , mikor elindult az összeverődött tömegen .
Alap+5xszótár rendszer fordítása:	- kíváncsi vagyok , ki lesz tanított nekünk ? - szólt hermione , mikor elindult az összeverődött tömegen .

A *said* fordításánál hasonló jelenség figyelhető meg. Az alaprendszer *kérdezte*, míg a szótáras módszerek a *szólt* fordítást adták, amely annak tudható be, hogy a szótárban ez volt a megfeleltetése.

Nézzük a példa második részét. Látható, hogy az alaprendszer eredménye viszonylag gyenge (*mikor ő az* .). Ez amiatt van, hogy a szóösszerendelő a hosszabb mondatok második felét gyakran hozzákapcsolja valamelyik szóhoz, így torzul a fordítási modell. Ebből kifolyólag a dekódoló sem tud megbirkózni a hasonló szövegrészekkel. Így fordulhat elő, hogy a program „összecsapja" a fordítandó mondatok végét. Ezzel szemben a szótáras esetekben megfigyelhető változások bizonyítják a szóösszerendelő minőségének javulását. Az 1xszótár esetben már egy kerek mondatot kapunk, 2xszótár esetben megjelenik a *zsibongó tömeg*, 3xszótár után pedig a *mikor elindult az összeverődött tömegen* . kifejezés lett a rendszer szerinti legjobb fordítás.

A statisztikai gépi fordítórendszerek egyik nagy hiányosságát tükrözi, hogy az angolban az *into* prepozíció egy külön egységnek felel meg, de a fordító nem találja a helyes magyar fordítást. Mivel a magyar nyelv toldalékokat használ, ezért a főnévhez kapcsolódó különböző ragok más-más jelentéssel bíró szavakat hoznak létre, melyek közül a fordítómodul általában nem a helyes toldalékkal ellátottat választja ki. Ennek tudható be az a jelenség, hogy az *into* az első három esetben mintha nem is jelenne meg a fordításban (*tömeg*), a 3xszótáras rendszertől pedig megjelenik a *tömegen*, ami már ragozott alak, ám a program nem a helyes toldalékot találta meg.

Igaz, hogy a BLEU módszerrel való kiértékelés hatékonysága vitatható, de SMT rendszerek összehasonlítására alkalmas. Emiatt megvizsgáltuk a különböző rendszerek 1-9-gramos kifejezésekre vonatkozó BLEU %-ait (3. táblázat). Ebből megfigyel-

hető, hogy az alaprendszerhez képest (1) a szótárral kiegészített rendszerek 1-4-gram esetén mind jobb eredményt értek el. Ez jól mutatja, hogy a szótárban túlnyomórészt egy-két, de maximum négy-öt szóból álló kifejezések voltak, és emiatt ezek fordítása is egyre jobb lett. Látható, hogy a legjobb eredményt elérő 1xszótár (2) rendszer szinte az összes mérési esetben jobb lett, mint az alaprendszer, tehát ebben az esetben közelítettük meg legjobban a korpusz és a szótár méretének optimális arányát. E szint felett kezdenek az egy-két szavas kifejezések túl dominánssá válni, ami lerontja a magasabb n-gram értékeket. Ezért van az 5xszótár (4) esetben, hogy az 1-gram értéke sokkal magasabb még az 1xszótáras rendszerénél is, de már 2-gram esetén alacsonyabb lesz nála, míg 5-gram esetén már az alaprendszernél is.

3. táblázat: Különböző rendszerek n-gramonkénti eredményeinek összehasonlítása (BLEU-%-ban).

	1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	8-gram	9-gram
1	47,05	16,29	7,07	3,54	1,94	1,14	0,74	0,57	0,46
2	47,60	16,62	7,35	3,78	2,02	1,19	0,75	0,57	0,43
3	47,55	16,46	7,25	3,75	2,09	1,25	0,81	0,60	0,46
4	47,32	16,33	7,09	3,64	1,94	1,09	0,68	0,47	0,33
5	47,74	16,43	7,19	3,63	1,93	1,08	0,68	0,51	0,39

Sorai: 1. Alaprendszer; 2.Alap+1xszótár rendszer fordítása; 3. Alap+2xszótár rendszer fordítása; 4. Alap+3xszótár rendszer fordítása; 5. Alap+5xszótár rendszer fordítása

Az eredmények fontossága abban rejlik, hogy rámutatnak: a szóösszerendelés javításával lehet javítani a fordítórendszer minőségét.

3.2 Joshua

Következő lépésként a további hibridizáció lehetőségeit vizsgáltuk. Látható, hogy a távoli nyelvek fordítása esetén – amilyen például a magyar és az angol nyelvpár – az egytagú kifejezések közti statisztikákon felül szükségünk van más fogódzók kihasználására is. Ilyen tulajdonságok lehetnek a nyelvtani szabályok.

A fenti célok elérésére nyújthat lehetőséget a Joshua keretrendszer [6], mely nem pusztán szó- vagy frázisszintű statisztikai valószínűségi modelleket használ, hanem bizonyos nyelvtani jellemzők előfordulását is figyelembe veszi. A Joshua rendszer további nagy előnye, hogy képes ezen generatív szabályok közti fordításra oly módon, hogy megadhatóak a szabályok mind a forrásnyelvre, mind a célnyelvre, valamint az is megadható, hogy mekkora valószínűséggel transzformálhatók át a szabályok egymásba. Ennek köszönhetően alkalmasabb egymástól morfológiailag és szintaktikailag távoli nyelvpárok közötti fordításra.

A feladat során a következő egyszerű szabályrendszert használtuk, hogy meg tudjuk becsülni, hogy a módszer mennyire alkalmas az elvárt feladat megoldására:

```
[S] ||| [X,1] ||| [X,1] ||| 0 0 0                          (1)
[S] ||| [S,1] [X,2] ||| [S,1] [X,2] ||| 0.434294482 0 0    (2)
```

A Joshua rendszert a 2. rész 2-es bekezdésben leírt korpusszal tanítottuk be, hogy össze tudjuk hasonlítani a Moses rendszer eredményeivel. Az eredményt a 4. táblázat mutatja.

4. táblázat: A Joshua rendszer eredményének összehasonlítása.

Rendszer	**BLEU-érték**
Alaprendszer	10,85%
LitMag+Joshua+OOV	9,85%
LitMag+Joshua	11,06%

A Joshua rendszer alapértelmezetten minden szót, amelyet nem tudott lefordítani, megjelöl az OOV (Out Of Vocabulary) jellel. Látható, hogy így a fordítás minősége rosszabb, mint az alaprendszeré. Ez annak tudható be, hogy például a tulajdonneveket is megcímzi a fordításban, és hiába lett volna helyes a fordítás, ezzel mégis elrontja azt. Ennek elkerülése érdekében utólag leszedtük ezeket a címkéket, és megkaptuk a 11,06%-os értéket, amely a rendszer nagymértékű javulását mutatja.

Az 5. táblázat jól szemlélteti, hogy akár már egy egyszerű szabály bevezetésével is hogyan változik a fordítórendszer eredménye. A *„For a little while only”* szerkezetet az alaprendszer egyszerűen *„egy darabig csak”*-nak fordította, míg a Joshua rendszer a rekurziós szabály alkalmazásával megtalálta a helyes *„csak egy kis ideig”* fordítást. Ebből a példából még az is látszik, hogy az emberi kiértékelő számára mind a két fordítás elfogadható, de a gépi kiértékelés számára az első minimális, míg a második fordítás maximális értéket kapott.

5. táblázat: Példafordítás a Joshua rendszerrel.

Angol referenciafordítás:	" for a little while only , " said the voice quietly .
Magyar referenciafordítás:	- csak egy kis ideig - mondta a hang csendesen .
Alaprendszer fordítása:	- egy darabig csak - mondta a hang .
Joshua rendszer fordítása:	- csak egy kis ideig nyugodtan - mondta a hang .

Annak ellenére, hogy rendkívül ígéretesnek tűnik ez az új rendszer, eredményeinkben mégis kevés helyen tüntettük fel. Ennek az az oka, hogy nagyobb korpusz esetén túlságosan nagy lett a rendszer erőforrásigénye, amit a közeljövőben szeretnénk megszüntetni.

3.3 Cigány-magyar SMT rendszer

Statisztikai fordítórendszerünket kipróbáltuk egy, a magyarhoz morfológiai gazdagságában közelebb álló nyelv esetében is. Korpuszként Vesho-Farkas-féle lovári cigány nyelvű Újszövetségét [7], illetve a Káldi-Neovulgáta magyar fordítást használtuk [8].

6. táblázat: A lovári-magyar rendszerek eredményeinek összehasonlítása.

Rendszer	BLEU-érték
Lovári-magyar (Moses)	30,53%
Lovári-magyar (Joshua)	29,20%
Magyar-lovári (Moses)	30,38%
Magyar-lovári (Joshua)	35,88%

A 6. táblázatból olvashatók a fordítórendszerek által elért eredmények. Megfigyelhető, hogy a magyar–angol nyelvpárhoz képest sokkal jobb eredményt sikerült elérni, aminek számos oka lehet. A legszámottevőbb ok, hogy a teszt- és a tanítókorpusz is egyaránt ugyanabból a szövegből (az Újszövetségből) került ki. Ebből következik, hogy a létrehozott fordítógép túlságosan téma- és stílusspecifikus lett. Ismeretes, hogy a négy evangélium esetében több alkalommal is előfordul tartalom- és szövegismétlődés, amikor az evangélisták ugyanazt a történetet írják le, sokszor nagyon hasonlóan. Emiatt fordulhat elő, hogy a tesztfordítások között 100%-os fordítás is van, mert egyszerűen más helyen ezek a mondatok benne voltak a korpuszban.

Ennek ellenére az eredmények vizsgálatából megfigyelhető (7. táblázat), hogy a fordítórendszer sokkal jobb fordításokat generált, és az emberi kiértékelés számára sokkal olvashatóbb eredményt kaptunk, mint az angol–magyar esetben.

7. táblázat: Példamondat-fordítás a különböző rendszerekkel.

Cigány referencia-fordítás:	le but manusha pale tele sharadine penge gada po drom , kavera pale kranzhi phagrenas tele pa kasht haj po drom rispisarnaslen .
Magyar referencia-fordítás:	a hatalmas tömeg pedig leterítette ruháit az útra , mások meg ágakat vagdostak a fákról és az útra szórták .
Moses fordítása:	a nép pedig le terítették ruháikat az úton , mások pedig ágakat phagrenas le a fa , és az úton rispisarnaslen .
Joshua fordítása:	a nép pedig le terítették ruháikat az úton , mások pedig ágakat phagrenas le a fa és az úton rispisarnaslen .

4 Statisztikai szövegelemző

Munkánk során egy másik témakörrel is foglalkoztunk, ugyanis az SMT rendszerrel végzett kísérleteink során szükségünk volt a korpuszunk morfológiai elemzésére, és ez adta az ötletet a statisztikai módszerek egy újabb felhasználási területére. A szövegelemzés feladata is megfogalmazható két nyelv közti transzformációként, ha rendelkezésünkre áll a sima szöveget és annak elemzését tartalmazó párhuzamos korpusz is.

Számos módszer létezik szövegelemzésre, melyek közül a két leggyakrabban használt a szabályalapú és a gépi tanuláson alapuló módszerek. Mind a két módszernek számos előnye van, de a szabályalapú rendszer számára – a gépi fordításhoz hasonlóan – rendkívül nehéz és körülményes megfogalmazni a megfelelő szabályokat. A gépi tanulásos megoldásnak is megvannak a nehézségei: igaz, hogy egyes

szabályokra nagyon pontosan betanítható, de ha az összes szabályra szeretnénk alkalmazni, túl komplex és lassú rendszert kapunk.

Ezzel szemben a statisztikai módszer a betanítás során az összes általa felismerhető szabályt figyelembe veszi. Ehhez csak egy megfelelő és elég nagy méretű korpuszra van szükség, cserébe egy online rendszert kapunk, ami viszonylag gyorsan képes ezután elemzést végezni. Ezen felbuzdulva megvizsgáltuk az SMT rendszer alkalmazhatóságát szövegelemzésre.

Ennek a rendszernek a felépítéséhez a Szeged Korpusz 2.0-t használtuk [1], melynek talán egyetlen hibája, hogy viszonylag kis méretű. Ennek ellenére alkalmasnak tűnt a felhasználásra. Mivel a szófaji címkék korlátozott számúak, elvben kisebb méretű korpuszban is elég nagy gyakorisággal szerepelhetnek. A rendszert kiértékeltük a BLEU módszerrel (8. táblázat), és kiszámítottuk a pontosságát is (9. táblázat).

8. táblázat: A szövegelemző rendszer automatikus kiértékelése I.

Rendszer	**BLEU-érték**
Szeged+Moses	90,97%
Szeged+Joshua	90,96%

9. táblázat: A szövegelemző rendszer pontossága I.

Rendszer	**Pontosság**
Szeged+Moses+helyes	90,29%
Szeged+Moses+helytelen	9,71%

A kiértékelésénél szembetűnt a rendszer néhány hibája. Az első és talán legfontosabb probléma a korpusz szerkezetéből fakad. Az elemzett korpuszban egymás után szerepelnek a szavak szótövei, amikhez hozzákapcsolódnak az elemzést tartalmazó címkék, de a több tagból álló kifejezések esetekben (pl. többtagú tulajdonnevek, igei szerkezetek) a címke csak a kifejezés utolsó szaván vagy utána helyezkedik el. Az egy szófaji egységbe tartozó kifejezések jelölésének hiánya a statisztikai módszerben félrevezető fordítási modellt eredményez. Ennek köszönhetően az első rendszer a tulajdonnévi címkéhez hozzácsatolt egy *„)_[PUNCT]”* szöveget, így gyengébb eredményt kaptunk. Az eredmény javítása érdekében minden önálló címkét hozzácsatoltunk az előtte álló szóhoz, így kaptuk a 10. táblázatban látható eredményt.

10. táblázat: A szövegelemző rendszer eredménye II.

Rendszer	**BLEU-érték**	**Helyes**	**Helytelen**
Moses	90,97%	90,80%	9,20%
Joshua	90,96%	90,72%	9,28%

A 10. táblázatból látszik, hogy változatlan BLEU-értékek mellett a rendszer pontossága 0,5–0,6 százalékkal javult. Ezt annak köszönhetjük, hogy nem kerültek a fordításba felesleges elemek, de a többtagú kifejezések fordítása továbbra sem megoldott. A probléma megoldásához elengedhetetlen ezeknek a kifejezéseknek az ösz-

szekapcsolása például a tulajdonnevek felismerésével. Nem volt célunk ilyen rendszer kifejlesztése, viszont az elmélet igazolása érdekében összekötöttük a korpuszban ezeket a kifejezéseket. A tanítás után a 11. táblázatban látható eredményt kaptuk.

11. táblázat: A szövegelemző rendszer eredménye III.

Rendszer	BLEU-érték	Helyes	Helytelen
Moses	90,96%	91,05%	8,95%
Joshua	90,77%	91,07%	8,93%

A 11. táblázatból megfigyelhető, hogy az összetartozó szavak összekapcsolása tovább javította rendszerünk pontosságát, annak ellenére, hogy BLEU-értéke alacsonyabb lett, mint az elődeinek. A hibás fordítások vizsgálatából kiderült, hogy a hibák két fő csoportba sorolhatók. Az első, amikor a rendszer nem ad fordítást, hanem visszaadja az eredeti szót. Az esetek túlnyomó részében ezek a szavak nem szerepelnek a korpuszban, így a fordítási modellben sincs megfeleltetésük. A másik hibafajta a helytelen elemzések halmaza. Itt is két fő hibakategória különíthető el. Az egyik esetben a szófajt helyesen beazonosítja, csak annak a további elemzésében ront; a súlyosabb hiba pedig, mikor már a szófajt sem találja el.

Felmerül a kérdés, hogy egyértelműsítést tartalmaz-e a rendszer és alkalmas-e rá. A választ maga a statisztikai módszer tulajdonsága adja. Amelyik jelenségre van elegendő példa, arra az SMT rendszer nagyon jól fog működni. Ha minden többértelmű kifejezés elég sokszor szerepel a tanítóhalmazban, a rendszer helyesen fogja megítélni a többértelmű szavakat is. Sajnos ilyen korpusz egyelőre nem állt rendelkezésünkre, és valószínűleg a közeljövőben nem is fog. Mivel a felhasznált Szeged Korpusz viszonylag kis méretű, ezért ez a rendszer nagy valószínűséggel nem oldja meg jól az egyértelműsítést, tehát ha a kérdéses szövegkörnyezet nincs benne a korpuszban, általában a legvalószínűbb elemzést rendeli hozzá a szóhoz.

A 12. táblázatban egy példán keresztül bemutatott rendszer javításának további lehetőségeit is vizsgáltuk. Az előzőekben leírtakból következik, hogy megfelelő méretű korpusz segítségével bármilyen szabály jól betanítható. Mivel jelen esetben korpuszunk fix méretű, minőségi javulás akkor érhető el, ha csökkentjük a komplexitást. A mi esetünkben ez úgy érhető el, ha a sima szöveget csak a szófaji címkék „nyelvére” fordítjuk, tehát nem írjuk ki eléjük maguknak a szavaknak a szótöveit. Mivel ezekből a címkékből sokkal kevesebb van, mint a magyar szavakból, kisebb korpuszból is felépíthető egy relatíve pontos rendszer. Másfelől: ha elhagyjuk a szótöveket az elemzésből, és csak a címkékre fordítunk, sokkal nagyobb súllyal szerepel majd a szófajok mondatban elfoglalt sorrendje mind a nyelvi, mind a fordítási modell esetében.

A rendszer eredményei a 13. táblázatban láthatóak. Ismét megfigyelhető a BLEU-érték csökkenése az eredeti rendszerhez képest, ám a pontosság az eddigi legjobb lett (bár ezt szótövesítés nélkül oldottuk meg). A módszer azonban használható, mert ha egy másik rendszert a sima szöveg szótöves változatára tanítunk be, akkor hasonló jó hatásfokkal tudjuk fordítani, és a kapott eredményeket összekapcsolva az elemző minősége is javítható lehet.

12. táblázat: Példa a szövegelemző működésére.

Sima szöveg:	mindenképp kötelességtudó szeretnék lenni , de azért nem olyan fanatikus szinten , mint egyes felnőttek , hogy még a családot is feláldozza a kötelesség miatt .
Referencia elemzés:	mindenképp_[Rg] kötelességtudó_[Afp-sn] szeret_[Vmcp1s---n] lesz_[Vmn] ,_[Punct] de_[Ccsp] azért_[Rd] nem_[Rm] olyan_[Pd3-sn] fanatikus_[Afp-sn] szint_[Nc-sp] ,_[Punct] mint_[Cssp] egyes_[Afp-sn] felnőtt_[Nc-pn] ,_[Punct] hogy_[Cssp] még_[Rx] a_[Tf] család_[Nc-sa] is_[Ccsp] feláldoz_[Vmip3s---y] a_[Tf] kötelesség_[Nc-sn] miatt_[St] ._[Punct]
SMT elemző:	mindenképp_[Rg] kötelességtudó_[Afp-sn] szeret_[Vmcp1s---n] lesz_[Vmn] ,_[Punct] de_[Ccsp] azért_[Rd] nem_[Rm] olyan_[Pd3-sn] fanatikus_[Afp-sn] szint_[Nc-sp] ,_[Punct] mint_[Cssp] egyes_[Afp-sn] felnőtt_[Nc-pn] ,_[Punct] hogy_[Cssp] még_[Rx] a_[Tf] család_[Nc-sa] is_[Ccsp] feláldoz_[Vmip3s---y] a_[Tf] kötelesség_[Nc-sn] miatt_[St] ._[Punct]

13. táblázat: A szövegelemző rendszer eredménye IV.

Rendszer	**BLEU-érték**	**Helyes**	**Helytelen**
Moses	88,65%	91,22%	8,77%
Joshua	88,57%	91,09%	8,91%

5 Összefoglalás

Kutatásunkban statisztikai módszerek és kétnyelvű párhuzamos korpusz segítségével igyekeztünk megoldani olyan feladatokat, ahol a cél megfogalmazható volt két nyelv közti transzformációként. Sikerült javítani az angol–magyar statisztikai gépi fordítórendszer minőségét részben kétnyelvű szótár hozzáadásával, részben a rendszer hibridizációjával. Sikeresen betanítottunk egy lovári–magyar statisztikai gépi fordító rendszert, amely eredményei segíthetnek az angol–magyar fordító rendszerek minőségét javítani. Eredményeink rávilágítanak, hogy bizonyos mértékű hibridizáció segítségével az SMT rendszerek minősége javítható. A kutatás folyamán létrehoztunk egy statisztikai szövegelemző rendszert is, és megvizsgáltuk annak minőségét. A kapott eredmények bíztatóak voltak, és rámutattak, hogy ebben a kutatási irányban is rejlenek további lehetőségek. Az eredmények figyelmeztettek arra is, hogy önmagukban a statisztikai módszerek nem elégségesek ezen feladat megoldásra: mindenképp szükséges valamilyen hibridizáció.

Hivatkozások

1. Csendes D., Hatvani Cs., Alexin Z., Csirik J., Gyimóthy T., Prószéky G., Váradi T. : Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. In: I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Egyetem (2003) 238–247
2. EuroMatrix-táblázat. http://www.statmt.org/matrix/ (2007)

3. Halácsy P., Kornai A., Németh L., Sass B., Varga D., Váradi T., Vonyó A.: A Hunglish korpusz és szótár. In: III. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Egyetem (2005)
4. Koehn P.: Moses – A Beam-Search Decoder for Factored Phrase-Based Statistical Machine Translation Models. User Manual and Code Guide (2009)
5. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions. Association for Computational Linguistics, Prague (2007) 177–180
6. Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W.N.G., Weese, J., Zaidan, O. F.: JosHUa: An Open Source Toolkit for Parsing Based Machine Translation. In: Proceedings of the Fourth Workshop on Statistical Machine Translation. Athens, Greece (2009) 135–139
7. Suntoiskirpe Nyevo Teshtamento (ford.: Vesho-Farkas Zoltán). Szent Jeromos Bibliatársulat, Budapest (2003)
8. Újszövetségi Szentírás a Neovulgáta alapján. Szent Jeromos Bibliatársulat, Budapest (1997)
9. Varga D., Halácsy P., Kornai A., Nagy V., Németh L., Trón V.: Parallel Corpora for Medium Density Languages. Recent Advances in Natural Language Processing Conference (2005) 590–596
10. Vonyó A.: A mindenki által keresett ingyenes angol–magyar magyar–angol köznapi, műszaki és szlengszótár. http://almos.vein.hu/~vonyoa/SZOTAR.HTM (1999)