STATISZTIKAI GÉPI FORDÍTÁS MÓDSZERÉNEK ALKALMAZÁSA

EGY- ÉS TÖBBNYELVŰ NYELVTECHNOLÓGIAI PROBLÉMÁK HATÉKONY MEGOLDÁSÁRA

DOKTORI (PH.D.) DISSZERTÁCIÓ

Laki László János

Témavezető:

Dr. Prószéky Gábor,
az MTA doktora

Pázmány Péter Katolikus Egyetem
Információs Technológiai és Bionikai Kar
Multidiszciplináris Műszaki és Természettudományi
Doktori Iskola

Budapest, 2015.

1. Bevezetés és kutatói célok

A humán nyelvtechnológia egyik legfontosabb feladata a nyelvi diverzitás okozta akadályok áthidalása, vagyis a számítógépek alkalmassá tétele a különböző nyelvek közti fordítások megvalósítására. Az elmúlt néhány évben az információtechnológia robbanásszerű fejlődése lehetővé tette a számítógépes nyelvészet számára, hogy megoldást nyújtson erre a problémára. Napjainkban az erre a célra leginkább alkalmazott módszer a statisztikai gépi fordítás (SMT – statistical machine translation). Az SMT rendszer egy teljesen nyelvfüggetlen eszköz, amely felügyelt gépi tanulásos módszerek segítségével tanítható, valamint megfelelő mennyiségű tanítóanyag birtokában a fordítás minősége is elfogadható. A módszer hátránya azonban, hogy a nyelvtanilag nagyon különböző, illetve a gazdag morfológiájú nyelvek esetén a szimplán statisztikai módszer nem elégséges a feladat jó minőségű megoldására. Ezeknél a nyelveknél ugyanis fellépnek a mondatok szószámbeli különbségéből, a forrás- és célnyelvi szavak mondaton belüli eltérő pozíciójából, és az adathalmazban nem megfelelő mennyiségben előforduló szavak esetén az adathiány-problémából eredő nehézségek. **Munkám első felében a gazdag morfológiájú nyelvek fordításánál fellépő nehézségekre kerestem megoldást a fordítórendszer hibridizációjával. Célom volt egy olyan architektúra kidolgozása, mely képes csökkenteni az adathiány-probléma okozta negatív hatásokat, valamint képes a nyelvtanilag helyes szóalakok előállítására.**

A szövegfeldolgozáshoz elengedhetetlen az írott szövegek megértése, és azok elemzése. A szövegelemzési lánc egyik első lépése az úgynevezett teljes szófaji egyértelműsítés, melynek feladata a szavak szótövének meghatározása, és besorolása az egyes szófaji kategóriákba. A szófaji egyértelműsítés feladata nem minden esetben egyértelmű, hiszen számos olyan szóalak létezik, mely több csoportba is tartozhat, és csak az adott szó szövegkörnyezete, valamint a mondatban elfoglalt pozíciója alapján dönthető el, hogy éppen melyik osztályba kell sorolni. A feladat megoldására számos alkalmazás létezik, de ezek közül kevés végez egyidejűleg lemmatizálást és morfoszintaktikai elemzést, illetve még kevesebb az olyan, amely ezt nyelvfüggetlen módszerekkel oldja meg. Mivel a statisztikai gépi fordítás feladata két nyelv közti transzformáció megvalósítása, emiatt alkalmas lehet a teljes morfológiai egyértelműsítés feladatának elvégzésére. Ebben az esetben az eredeti elemzendő szövegről a szófajilag egyértelműsített, szótövesített szóalakok közti fordítást kell megvalósítanunk. **Munkám második felében célom egy statisztikai gépi fordítás módszerén alapuló nyelvfüggetlen teljes morfológiai egyértelműsítő rendszer kidolgozása és bemutatása, mely eléri vagy meghaladja a már létező nyelvfüggő és nyelvfüggetlen rendszerek eredményességét.**

2. Módszerek és eszközök

Munkám során a *statisztikai gépi fordítórendszer* különböző feladatokra történő alkalmazását vizsgáltam. Foglalkoztam továbbá a már meglévő rendszerek fejlesztésével, tökéletesítésével, illetve az eredmények javításával is. A statisztikai gépi fordítás alapötlete, hogy a rendszer párhuzamos kétnyelvű tanítóanyag segítségével felügyelt módon tanulja meg a fordításhoz szükséges modelleket. A *párhuzamos kétnyelvű korpusz* egy olyan, mondatpárokból álló, szöveges adathalmaz, amelyben a forrásnyelvi mondatokhoz hozzá van rendelve azok célnyelvi fordítása. Az algoritmus könnyű és gyors implementálhatósága, valamint nyelvfüggetlen alkalmazhatósága nagymértékben hozzájárult ahhoz, hogy a módszer napjainkra a leginkább elterjedt gépi fordító architektúra legyen. Munkám során az SMT módszer legtöbbet hivatkozott eszközét a MOSES nevű [1] keretrendszert használtam, ami összegyűjti a meglévő eszközöket, valamint implementálja a gépi fordításhoz szükséges algoritmusokat. Az SMT architektúra jó eredménnyel működik hasonló szintaktikai struktúrájú és szórendű nyelvpárok esetén. Ezzel ellentétben a kifejezésalapú modellek számára nehezen kezelhetők a számottevő grammatikai különbségek. Dolgozatomban bemutatok egy hibrid fordítórendszert, mely az alapvető statisztikai metódusok mellett szintaxis- és morfológia-vezérelt elő- és utófeldolgozási lépéseket alkalmaz a tanítóhalmazon, valamint morfológiai utófeldolgozást végez a fordítás során. A fejlesztések hatásait az angol-magyar nyelvpár közti fordítás segítségével mutatom be. A létrehozott architektúrában a statisztikai alapú dekódert *morfológiai generátorral* helyettesítettem, ami esetemben a HUMOR [2] volt.

Fontos kérdés a létrehozott rendszerek kiértékelése. Az SMT rendszer automatikus értékelésének alapvető módszere a lefordított mondatnak egy referenciamondathoz való hasonlítása különböző jellemzők mentén. Napjaink legnépszerűbb *kiértékelő módszere* a BiLingual Evaluation Understudy (BLEU) [3], mely megoldást kínál a szavak sorrendjéből adódó probléma kezelésére is. Lényege, hogy a vizsgált rendszer által lefordított mondat kifejezéseit keresi a referenciamondatban. Minél nagyobb a hasonlóság a két mondat között, annál több pontot kap érte. A BLEU metrika előnyei ellenére több publikáció is figyelmeztet arra, hogy számos esetben az algoritmus nem korrelál az *emberi kiértékelés*sel. Emiatt a létrehozott fontosabb rendszereket több emberi kiértékelő segítségével is megvizsgáltam.

A statisztikai gépi fordítás nemcsak a természetes nyelvek közötti fordításra alkalmas, hanem tetszőleges szövegek közti transzformációra is. Dolgozatomban bemutatom, hogy a teljes szófaji egyértelműsítés feladata megfogalmazható, mint fordítási feladat. Az általam létrehozott SMT-alapú egyértelműsítő rendszerhez *végződésfa-alapú ajánlórendszert* ([4]–[6]), valamint *morfológiai elemzőt* (HUMOR) integráltam.

3. Új tudományos eredmények

A dolgozatomban bemutatott eredmények két téziscsoportba sorolhatók. Az első téziscsoportban a nyelvtanilag távoli nyelvek közötti gépi fordítás minőségét javítottam a tisztán statisztikai fordítórendszer hibridizációjával. A második téziscsoportban bemutattam a statisztikai gépi fordítórendszer teljes szófaji egyértelműsítés céljából történő alkalmazását.

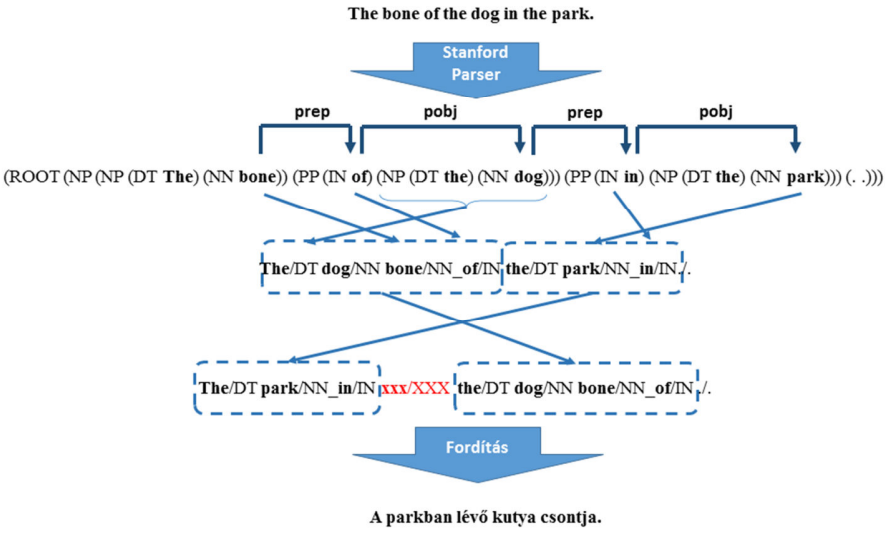
I. TÉZISCSOPORT

Ebben a téziscsoportban az agglutináló nyelvek gépi fordítása során jelentkező nehézségek megoldására kerestem módszereket. A problémák közül a legjelentősebbek az agglutináló nyelvek esetében az adathiány-probléma és a szóalakok statisztikai módszerrel történő előállítása. Nehézséget okoz továbbá az egymástól nyelvtanilag távol álló nyelvek közti fordítás során a gyakran jelentős szórendbeli és szószámbeli különbség. Munkám során a tisztán statisztikai szóalapú gépi fordítórendszert a forrásnyelv és célnyelv közti nyelvtani különbségek kezelésére irányuló algoritmusokkal egészítettem ki, melyek integrálásával javítottam a fordítás minőségét.

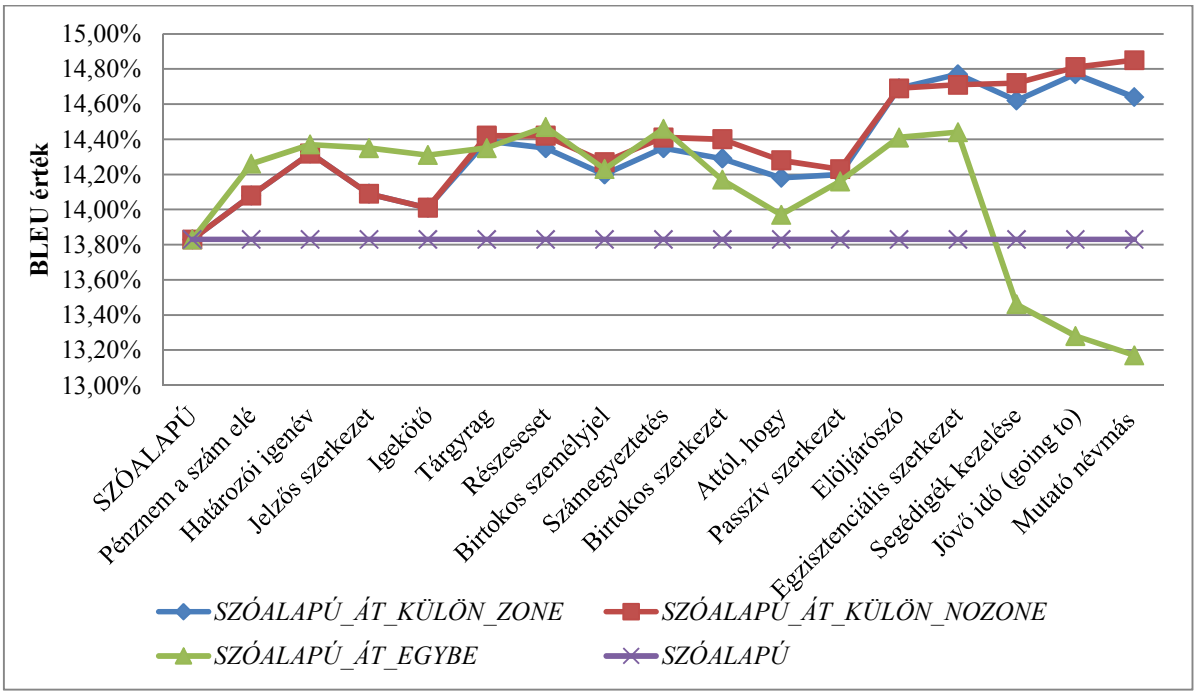
1. tézis: ***A tisztán statisztikai alapú gépi fordítórendszert hibridizáltam az eltérő szórendet okozó nyelvtani sajátosságok alapján definiált nyelvpár-specifikus (esetünkben: angol-magyar) átrendező szabályok alkalmazásával, melynek során az alaprendszer teljesítményéhez képest javulást értem el a fordítás minőségében.***

A tézishez kapcsolódó publikációk: [Laki_1], [Laki_4], [Laki_8]

A dolgozatban beláttam, hogy a szimplán statisztikai gépi fordítórendszerek nem elégségesek az egymástól grammatikailag távoli nyelvpárok fordításának megoldására. Ez elsősorban a nyelvek közt felmerülő jelentős szórendbeli különbségből fakad, mivel az általánosságban használt dekóder-implementáció csak lokális szórendi átrendezésekre képes. Emiatt létrehoztam egy olyan hibrid fordítórendszert, amely általam megfogalmazott szintaxismotivált szabályokat alkalmaz előfeldolgozásként a forrásnyelvi angol szövegen. Az átrendezendő szerkezetek megtalálása a vizsgált mondatok közvetlen összetevős elemzése valamint a függőségi relációk alapján történt. Az angol-magyar nyelvpárra alkotott szabályok segítségével az alap fordítórendszer eredményeihez képest javulást értem el.



1. ábra: A példa bemutatja az angol elöljárószós módosító szerkezet átrendezését. Abban az esetben, ha a prepozíció része egy birtokos szerkezetnek ez a magyarban csak egy „lévő”-s szerkezetre fordítható. Ez alapján az elöljárószavas szerkezetet a birtokos szerkezet elé helyezem, míg a birtokos szerkezetet az előző szabály (birtokos szerkezet átrendezési szabály) alapján rendezem át. Továbbá a magyar mondatban szereplő extra „lévő” melléknévi igenév jelölésére a két szerkezet közé beillesztek egy „xxx” sztringet.



2. ábra: Az egyes szabályok hozzáadási sorrendjét, illetve az adott szabállyal kiegészített rendszer szemlélteti. Az ábra egy pontja bemutatja az aktuális és a már előtte szereplő átrendezési szabályok összhatását az alaprendszerhez (*SZÓALAPÚ*) képest. A Moses rendszer dekódolójában lehetőség van arra, hogy a fordítandó mondatban definiáljunk olyan szócsoportot, amit a dekódoló egy egységként (*SZÓALAPÚ_ÁT_KÜLÖN_ZONE*) kezel. Ennek ellentéte a *SZÓALAPÚ_ÁT_KÜLÖN_NOZONE* rendszer, ahol az átrendezett kifejezéseket önálló fordítási egységbe csoportosítottam. A *SZÓALAPÚ_ÁT_EGYBE* rendszerben a fordítás folyamán a lemmát és a hozzá tartozó funkciószókat összekapcsoltam.

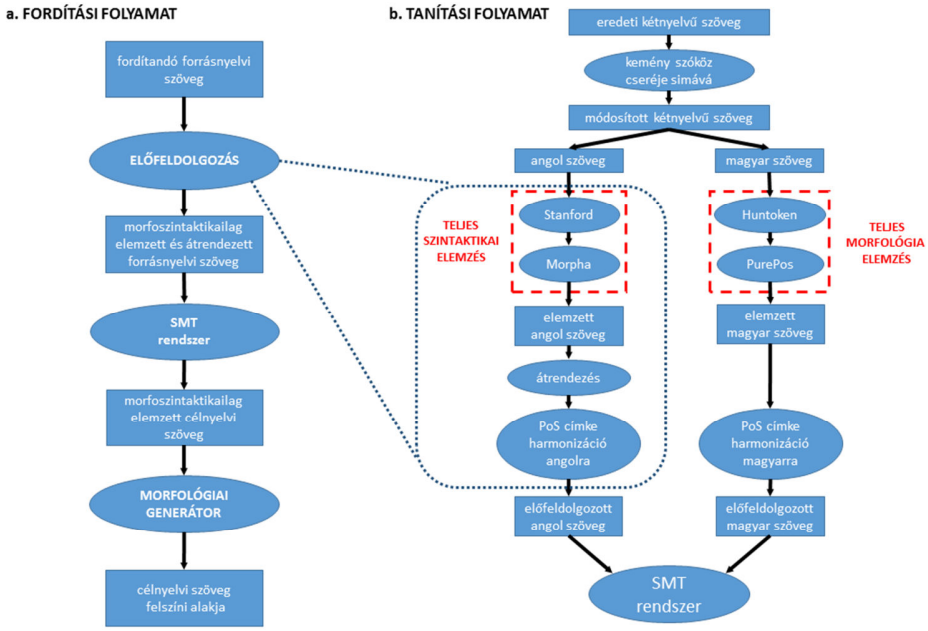
Ezekkel a transzformációkkal a legjobb rendszer 14,85% BLEU pontot ért el, ami 1,02%-os BLEU és 7,38%-os relatív javulás a *SZÓALAPÚ* alaprendszer 13,83%-os pontosságához képest. Továbbá számos olyan jelenség helyesen fordítható ezzel a módszerrel, melyet a hagyományos statisztikai gépi fordítórendszer nem tud kezelni. Természetesen, a megfogalmazott szabályok nem fedik le az összes szórendbeli különbséget okozó jelenséget, ezek finomítása további kutatás témája lehet.

2. tézis: ***Létrehoztam egy morfológiai generátorral kiegészített morfémaalapú SMT fordítási láncot, melynek alkalmazása során a magyar nyelvben gyakori homonímia kezelése érdekében a szóalakok helyett azok szótő+toldalékcímke alakú reprezentációját vezettem be.***

A tézishez kapcsolódó publikációk: [Laki_1], [Laki_4], [Laki_8]

A statisztikai dekóder számára egy agglutináló nyelvre történő fordítás az adathiány-probléma miatt rendkívül nehéz feladat. Ez is közrejátszik abban, hogy az agglutináló nyelvre történő SMT rendszer általi fordítás messze alulmarad a más nyelvek közti fordításhoz képest.

A morfológiailag bonyolult nyelvek alakjának előállítása nagy nehézséget jelent a fordítórendszer dekódere számára, ugyanis a dekóder nem képes a tanítóanyagban nem szereplő szavak előállítására. Létrehoztam egy egyedülálló hibrid gépi fordítórendszer architektúrát, melyben a fordítást egy SMT-alapú rendszer végzi morfológiailag elemzett – forrás- és célnyelvi – szövegeken, a fordított szóalak pedig morfológiai generátor segítségével kerül előállításra (2. ábra). A morfológiai generátor a statisztika alapú dekóderrel szemben nagy pontossággal képes előállítani olyan toldalékolt szóalakokat is, amelyek nem szerepeltek a tanítóhalmazban. Az adathiány és a homonímia csökkentése érdekében a szavak toldalékmorfémái helyett az azoknak megfelelő morfoszintaktikai címkéket alkalmaztam, mivel így nagymértékben csökkent a lehetséges szóalakok száma. Az általam felépített morfológiai generátort alkalmazó architektúrák az emberi kiértékelés számára könnyebben érthető, gördülékenyebb fordítás előállítására voltak képesek a szóalapú statisztikai dekódert használó fordítórendszerekkel szemben.



3. ábra: A morfológiai generátorral kiegészített fordítási, és az előfeldolgozó lepésekkel kiegészített tanítási folyamat bemutatása

3. tézis: ***Kidolgoztam a morfémákra bontott forrás- és célnyelvi szövegeken működő szóharmonizációs módszert, melynek során a két nyelv eltérő morfológiai viselkedését a morfémák számának egymáshoz közelítésével és a fordítás során történő megfeleltetésével kezeltem, ezáltal a fordított szöveg morfológiai komplexitása a forrásnyelvnek megfeleltethető maradt. Megmutattam, hogy a szóharmonizáció alkalmazásával a morfológiailag összetett nyelvek esetén javulás érhető el a fordítás minőségében.***

A tézishez kapcsolódó publikációk: [Laki_1], [Laki_4], [Laki_8]

Munkám során létrehoztam három olyan rendszerarchitektúrát, melyek az agglutináló és flektáló nyelvek mondatpárjaiban megfigyelhető szószám-különbségre képesek megoldást nyújtani. Bemutattam egy morfológiailag elemzett szövegen dolgozó szóalapú rendszert, ami az angol nyelvet agglutináló szerkezetűvé alakítja, valamint egy morfémaalapú fordítórendszert, ami a morfémákra bontott szövegek között végez fordítást. A harmadik rendszer egy faktoros fordítórendszer, amely az előző két rendszer előnyeit egyesíti. A módszer lényege, hogy párhuzamosan fordít lemmáról lemmára és toldalékmorfémáról toldalékmorfémára. A rendszer egyedisége, hogy a faktoros fordítás végén nem egy szóalakot kapunk kimenetként, hanem a lemmából és a hozzá kapcsolódó szófaji címkékből álló rekordot, melyből a 2. tézisben bemutatott morfológiai generátor állítja elő a feszíni szóalakot.

Több, morfémaalapú fordítás (szó-, morféma- és faktoralapú fordítási modellek) segítségével megoldottam az angol és a magyar mondatok között jelentkező szószám-különbségből adódó problémákat. Munkám során annak több fázisában végeztem automatikus kiértékelést a BLEU metrika szerint, de néhány esetet emberi kiértékeléssel is megvizsgáltam, ami igazolta azt, hogy az automatikusan mért alacsonyabb értékek nem feltétlenül jelentenek rossz minőségű fordítást. Ezzel bebizonyosodott, hogy a szóharmonizáció hatására az emberi kiértékelés számára jobb minőségű rendszereket hoztam létre a tisztán statisztikai alapon működőkkel szemben (1. táblázat).

Rendszer neve	**Emberi kiértékelés**	**w-BLEU**	**mm-BLEU**
referenciafordítás	88,33%		
MetaMorpho	76,30%	6,86%	50,97%
Google Translate	72,80%	15,68%	55,86%
Bing Translator	61,66%	12,18%	53,05%
MORFÉMAALAPÚ_ÁT_T6	**55,60%**	12,22%	**64,94%**
FAKTORALAPÚ_ÁT_T6_FIX	55,42%	10,88%	60,83%
MORFÉMAALAPÚ_T6	54,28%	12,19%	63,87%
FAKTORALAPÚ_T6_FIX	52,03%	9,91%	57,09%
SZÓALAPÚ_T6	51,33%	13,83%	59,32%
SZÓALAPÚ_ÁT_T6	50,89%	**14,83%**	58,06%
SZÓALAPÚELEMZETT_ÁT_T6	37,57%	13,05%	57,21%

1. táblázat: A morfológiai módosításokat tartalmazó fordítórendszerek emberi és gépi (BLEU) kiértékelése, valamint ezek összehasonlítása az általánosan használt fordítórendszerekhez képest. A táblázatból jól látható, hogy az emberi és a gépi kiértékelés nincs összhangban. Ezen kívül megfigyelhető, hogy a kevesebb BLEU pont ellenére a munkám során felépített rendszerarchitektúrák közül az emberi olvasó számára több is (*MORFÉMAALAPÚ_ÁT_T6, FAKTORALAPÚ_ÁT_T6_FIX, MORFÉMAALAPÚ_T6, FAKTORALAPÚ_T6_FXI)* jobb eredményt ért el az alaprendszerhez (*SZÓALAPÚ_T6*) képest.

4. tézis: ***Megmutattam, hogy a fordítás minősége javul, ha a tanítóhalmazt kiegészítem rövid kifejezések (szótári egységek, példaszerkezetek) pontos fordítását tartalmazó kétnyelvű kifejezéstárral, aminek megfelelő súlyozású figyelembe vétele javítja a hosszabb szegmenseket tartalmazó tanítóhalmazból számított statisztikát, robosztusabbá téve a fordítási modellt.***

Az SMT rendszer fordítása során gyakran előfordul, hogy a szóösszekötő nehezen találja meg az összetartozó szövegrészeket, ha azok a nyelvtani szerkezet miatt messze vannak egymástól, vagy ha nagyon különbözők. A túl hosszú mondatok is gyakran okoznak nehézséget, mivel gyakran előfordul, hogy a második tagmondat minden szavát egy szóhoz köti, vagy a többször szereplő, gyakori szavak párját nem jól találja meg. Ennek kiküszöbölése érdekében a tanítóhalmazt rövid, pontos fordítású kifejezéspárokkal egészítettem ki. A létrehozott szótárat többször egymás után hozzáadtam a tanítóhalmazhoz annak érdekében, hogy a pontos kifejezések előfordulása minél nagyobb súlyú legyen a fordítási modellben. Ezzel

párhuzamosan viszont folyamatosan csökkent az eredeti korpusz relevanciája, csökkent a többszavas kifejezések súlyozása a fordítási modellben, és ezáltal romlott a nyelvi modell minősége.

A tesztelések során kiderült, hogy akkor érek el legjobb eredményt, ha a szótárat egyszer adom a tanítóhalmazhoz. Ekkor az *ALAPRENDSZER* 10,85%-os BLEU értékéhez képest 0,33%-os BLEU és 11,18%-os relatív javulás figyelhető meg, ahogy azt a 2. táblázat is szemlélteti.

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%

2. táblázat: A szótár hozzáadásával készült rendszerek eredményei

A tézishez kapcsolódó publikációk: [Laki_11], [Laki_12]

II. TÉZISCSOPORT

Dolgozatom második felében a teljes morfológiai egyértelműsítés egy teljesen új megközelítését mutattam be azáltal, hogy a feladat megoldására statisztikai gépi fordítórendszert alkalmaztam. Rendszerem a HuLaPos2 nevet kapta. Amellett, hogy a rendszer egyidejűleg végez lemmatizálást és szófaji egyértelműsítést, további előnye, hogy a nyelvfüggetlen moduloknak köszönhetően bármilyen nyelvre és morfoszintaktikai címkekészletre alkalmazható. A kiértékelés során bebizonyosodott, hogy teljesítménye legalább olyan jó, mint a többi létező nyelvfüggetlen rendszeré, sőt néhány esetben megközelíti az egyes nyelvfüggő rendszerek által elért eredményeket is.

Az ismeretlen szavak hatékony kezelésének céljából az elemzési folyamatba egy végződésfa-alapú ajánlórendszert (guesser) integráltam. Végül magyar nyelvre alkalmazva beláttam, hogy a guesser és a morfológiai elemző kombinálásával tovább javítható a rendszer eredményessége.

Megvizsgáltam több nyelvre (angol, portugál, bolgár, magyar, horvát és szerb) és morfoszintaktikai kódkészletre (MSD és HUMOR) a módszer hatékonyságát. Az eredmények vizsgálatából megállapítható, hogy az általam létrehozott rendszer legalább olyan jól teljesít, ráadásul sok esetben felülmúlja a már létező nyelvfüggetlen rendszerek minőségét. Néhány nyelv esetén még a nyelvfüggő rendszerek teljesítményét is megközelíti.

5. tézis: ***Létrehoztam egy a statisztikai gépi fordítás módszerén alapuló teljes, azaz lemmatizálást is végző morfológiai egyértelműsítő rendszert, és megmutattam, hogy a célnyelvi szótár méretének csökkentése nagy mértékben javítja a rendszer minőségét.***

A tézishez kapcsolódó publikációk: [Laki_2], [Laki_3], [Laki_5], [Laki_6], [Laki_7], [Laki_9], [Laki_10], [Laki_12]

Mivel a statisztikai alapú fordítórendszer tulajdonképpen két nyelv közti transzformációt valósít meg, emiatt alkalmazható a sima és annotált szöveg közti „fordítás" megvalósítására is. Munkám során egyedülálló módon ezt a tulajdonságot kihasználva létrehoztam egy SMT-alapú teljes morfológiai egyértelműsítő (POS – part-of-speech) rendszert, mely szimultán végez lemmatizálást és szófaji egyértelműsítést. A jó minőségű teljes morfológiai egyértelműsítés kulcsfontosságú az agglutináló nyelvek feldolgozása során.

A statisztikai rendszer számára a túlságosan specifikus címkék hatására gyengül a kontextuális információk relevanciája, és emiatt a rendszer nehezebben tudja feloldani a szavak szófaji többértelműségét. Ezen problémára megoldást jelent, ha a címkekészlet általánosításával csökkentjük a célnyelvi szótár méretét. Munkám során megvalósítottam, valamint összehasonlítottam több módszert a célnyelvi címkekészlet csökkentésére. Az általam használt első technika a címkékben tárolt információ mennyiségének csökkentésével egyszerűsített a feladat komplexitásán. Ez gyakorlatban elsőként a lemmatizálás, majd a kevésbé fontos POS alosztályok elhagyását jelentette a célnyelvi címkékből. Ezekről a rendszerekről (*ALAP_SZIMB_SZÁM_CSAKPOS, ALAP_SZIMB_SZÁM_FOPOS*) elmondható, hogy a nagymértékű információveszteség ellenére viszonylag kis mértékű volt az elemző rendszer minőségének javulása. A második megoldás a tárolt információ megőrzése mellett képes csökkenti az egyértelműsítő rendszer komplexitását. Ezt a célnyelvi szótövek kompaktabb formában történő eltárolásával oldottam meg. Orosz és Novák [6] megoldásához hasonlóan a szavak lemmáját egy olyan rekorddal reprezentáltam, melyek megadják azt a szükséges transzformációt, amit el kell végezni egy adott szón, hogy megkapjuk annak szótövét. Egy ilyen rekord <*töröl,csatol*> ahol a *töröl* a sztring végéről eltávolítandó karakterek számát adja meg, a *csatol* pedig az a karaktersorozat, amit illeszteni kell a „csonka szó" végére, hogy megkapjuk a szótövet. Az így felépített rendszer (*TÖRÖLCSATOL_SZIMB_SZÁM*) lemmatizálás és POS címkézés szempontjából is jobban teljesít az alaprendszerhez képest (3. táblázat). Az elért eredmények alapján bebizonyítottam, hogy a célnyelvi címkekészlet komplexitásának csökkentésével javítható az egyértelműsítő rendszer teljesítménye.

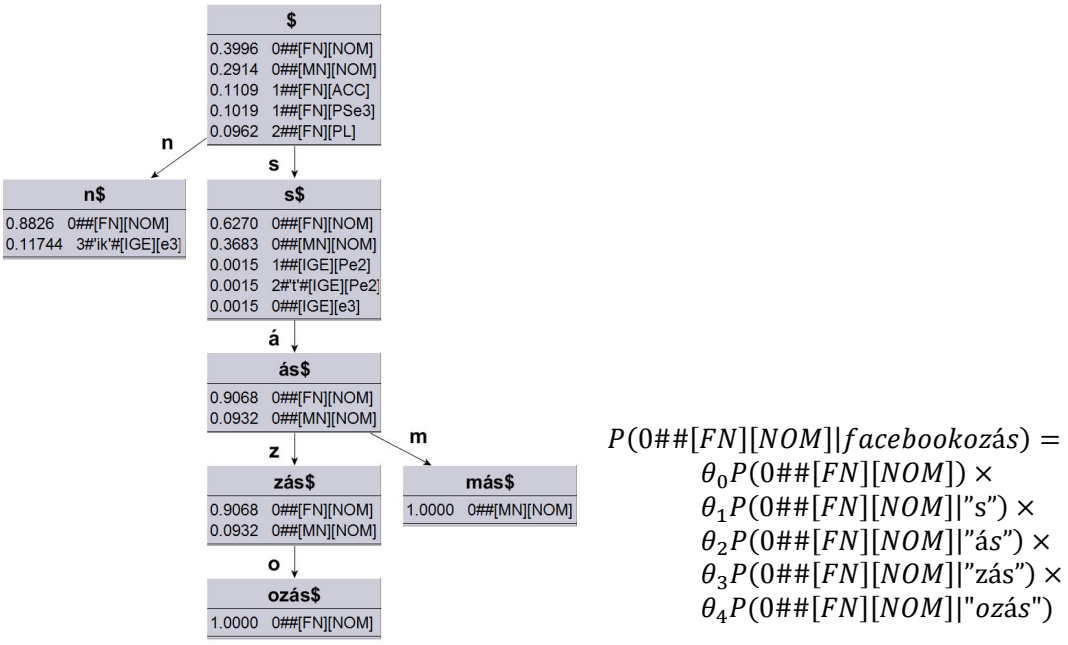
	Szószintű			Mondatszintű	
	POS	Lemma	Összes	POS	Összes
ALAP	91,281%	94,303%	91,257%	35,371%	35,294%
ALAP_SZIMB_SZÁM_CSAKPOS	91,534%	-	91,534%	37,071%	37,071%
ALAP_SZIMB_SZÁM_FOPOS	95,471%	-	95,471%	53,898%	53,898%
TÖRÖLCSATOL_SZIMB_SZÁM	91,496%	94,330%	91,447%	36,977%	36,684%

3. táblázat: A célnyelvi címkekészlet csökkentésével felépített rendszerek eredményei

6. tézis: ***Az SMT-alapú egyértelműsítő rendszerhez integráltam a tanítóanyagban nem szereplő szavak kezelésére egy végződésalapú morfológiai ajánlót (guesser), aminek köszönhetően a többi létező nyelvfüggetlen rendszer eredményét felülmúltam.***

A tézishez kapcsolódó publikációk: [Laki_2], [Laki_3], [Laki_6]

Az egyértelműsítő rendszerek legnagyobb hiányossága az ismeretlen szavak (OOV –) elemzése. Ez különösen igaz az agglutináló nyelvek esetében, hiszen egy szótőnek akár több száz szóalakja is lehet. Ezek közül azonban nem mind szerepel a tanítóhalmazban, így az egyértelműsítő rendszernek semmilyen előzetes ismerete sincs ezekről a szavakról. A dolgozatban bemutatott mérések alapján megállapítható, hogy az ismeretlen szavak jól modellezhetők a tanítóhalmazban szereplő ritka szavak segítségével. Az első módszer az SMT-alapú egyértelműsítő rendszert hivatott támogatni azzal, hogy az ismeretlen szavakat nagyobb gyűjtőosztályokba sorolja. A módszer azzal a feltételezéssel él, hogy az ismeretlen szavak hasonlóan viselkednek a mondaton belül, mint a tanítóhalmazban szereplő hasonló pozícióban lévő társaik. A mondatban elfoglalt pozíciójuk, a környező szavak és azok szófaji elemzése, valamint az ismeretlen szavakon lévő toldalékok alapján lehet következtetni az OOV szó elemzésére. Ezt oly módon értem el, hogy az OOV szavakat *„unk_abcd”* formátumú karaktersorozattal (továbbiakban *unk-szuffix*) helyettesítettem, ahol az *„abcd”* a szó utolsó négy karakterét jelöli. A ritka szavakban rejlő információt a rendszer tanítása során használtam fel, mivel ezekkel a szavakkal modelleztem a rendszer számára az ismeretlen szavak viselkedését. Ez a gyakorlatban azt jelenti, hogy a tanítóhalmazban lévő ritka szavakat cseréltem le az *unk-szuffix* sztringre. Ennek hátránya azonban, hogy a rendszer az egyértelműsítés során a szavak fix hosszúságú szuffixumát veszi figyelembe.



4. ábra: A végződésfában való keresés folyamatának bemutatása egy példa segítségével

A fix hosszúságú szuffixumból fakadó nehézségek kiküszöbölésére az egyértelműsítési láncba előfeldolgozó lépésként szuffix-guessert integráltam. A Moses rendszerben lehetőség van a fordítandó szövegbe fordítási javaslatokat definiálni, amiket a dekóder a fordítás során figyelembe vesz. Ilyen módon az egyértelműsítendő szövegben az OOV szavakhoz a guesser címkézési javaslata, mint előfordítás megadható. Az általam használt szuffix-guesser a tanítóhalmaz szavaiból egy végződésfát épít, ahol a gráf csúcsaiban tárolja azt az információt, hogy az adott végződés esetén mekkora valószínűsége van az egyes annotációs címkéknek. Ezeket a valószínűségeket a tanítóhalmaz ritka szavai alapján tanítottam meg. A módszernek köszönhetően nagymértékben sikerült javítani az OOV szavak egyértelműsítésének pontosságát (4. táblázat).

	Szószintű			Mondatszintű	
	POS	Lemma	Összes	POS	Összes
TÖRÖLCSATOL_SZIMB_SZÁM	91,496%	94,330%	91,447%	36,977%	36,684%
TÖRÖLCSATOL_SZIMB_SZÁM_UNKSZUFFIX	96,025%	97,828%	95,383%	58,752%	54,284%
TÖRÖLCSATOL_SZIMB_SZÁM_GUESSER	96,511%	98,595%	96,177%	62,465%	59,692%

4. táblázat: Az OOV szavak kezelésére alkalmazott technikák segítségével felépített rendszerek eredményei

7. tézis: ***Megmutattam az SMT-alapú teljes morfoszintaktikai egyértelműsítő rendszer nyelvfüggetlen viselkedését. Ehhez a létrehozott elemzőt hét különböző nyelven, illetve morfoszintaktikai kódkészleten tanítottam, melynek eredménye összemérhetőnek bizonyult az adott nyelvekre létező más rendszerek teljesítményével.***

A tézishez kapcsolódó publikációk: [Laki_3], [Laki_6]

Összehasonlítottam az általam létrehozott nyelvfüggetlen teljes morfológiai egyértelműsítő rendszer eredményeit más nyelveken és kódkészleteken elérhető rendszerek teljesítményével. A vizsgálat során kiderült, hogy rendszerem eredménye összemérhető más – esetenként nyelvfüggő – rendszerek eredményeivel, sőt több esetben meg is haladja azokat (5. táblázat).

Nyelv	Rendszer	Szószintű pontosság		
		címkézés	szótövesítés	teljes
magyar (HUMOR)	PurePos	96,50%	96,27%	94,53%
	HuLaPos2	96,70%	98,23%	97,62%
	PurePos+MA	98,96%	99,53%	98,77%
horvát	**HuLaPos2**	**93,25%**	96,21%	90,77%
	HunPos+CST	87,11%	**97,78%**	-
szerb	**HuLaPos2**	**92,28%**	92,72%	86,51%
	HunPos+CST	85,00%	**95,95%**	-
bolgár	TnT [4]	92,53%	-	-
	gépi tanulás	95,72%	-	-
	gépi tanulás+morf.lexikon	97,83%	-	-
	HuLaPos2	97,86%	-	-
	gépi tanulás+morf.lexikon+szabályok	97,98%	-	-
portugál	**HuLaPos2**	**93,20%**	-	-
	HMM-alapú PoS tagger	92,00%	-	-
angol	TnT [4]	96,46%	-	-
	kifejezésalapú fordító [12]	96,97%	-	-
	HuLaPos2	97,08%	-	-
	Stanford tagger 2.0 [7]	97,32%	-	-
	SCCN [8]	97,50%	-	-

5. táblázat: Különböző nyelvű teljes morfológiai egyértelműsítő rendszerek eredményeinek összehasonlítása.

8. tézis: ***Megmutattam, hogy az általam létrehozott nyelvfüggetlen rendszer minősége tovább javítható morfológiai elemző integrálásával.***

A tézishez kapcsolódó publikációk: [Laki_3]

Bebizonyítottam, hogy a nyelvfüggetlen teljes morfológiai egyértelműsítő nyelvfüggő morfológiai elemzővel kiegészítve további minőségjavulást eredményez. Ezzel a módszerrel létrehoztam az egyik legnagyobb pontosságú rendszert magyar nyelvre, mely a lemmatizálást 99,12% pontossággal végzi, a tanítóanyagban nem szereplő szavak 84,82%-át helyesen elemzi, a teljes morfológiai egyértelműsítés tekintetében pedig 96,50% pontosságú (6. táblázat).

	Szószintű			Mondatszintű	
	címkézés	szótövesítés	teljes	címkézés	teljes
TÖRÖLCSATOL_SZIMB_SZÁM_UNKSZUFFIX	96,025%	97,828%	95,383%	58,752%	54,284%
TÖRÖLCSATOL_SZIMB_SZÁM_GUESSER	96,511%	98,595%	96,177%	62,465%	59,692%
TÖRÖLCSATOL_SZIMB_SZÁM_MORFLEXIKON	96,624%	**99,119%**	**96,498%**	63,236%	62,250%
PUREPOS2	96,350%	97,505%	95,101%	60,817%	51,294%
HUNPOS+CST_SZÓTÖVESÍTŐ	96,340%	96,512%	94,276%	61,279%	47,288%
MORFETTE	**96,751%**	96,048%	93,776%	64,591%	44,160%
NLTK_MAXENT+ CST_SZÓTÖVESÍTŐ	94,949%	95,439%	92,927%	51,402%	40,169%

6. táblázat: Az általam készített és a magyar nyelven elérhető rendszerek eredményeinek összehasonlítása

4. Az eredmények alkalmazási területei

A disszertációmban leírt munkák olyan feladatok megoldására irányultak, melyek elősegítik egyrészt a nyelvek közti fordítás minőségének, másrészt a teljes morfológiai egyértelműsítés pontosságának javulását. A hibrid gépi fordítással kapcsolatos eredményeim sikeresen integrálhatóak tetszőleges SMT rendszer architektúrába. Az elért eredmények alátámasztották, hogy a morfológiai információknak a fordítási láncban történő felhasználása pozitív hatással van a fordítás minőségére.

A második téziscsoportban bemutatott teljes morfológiai elemző rendszer képes nyelvfüggő valamint nyelvfüggetlen működésre. A leírt módszer alkalmas a szintaktikai elemzési láncba történő integrációra. Továbbá ahogy Orosz et al. [Orosz_1, Orosz_2] bemutatta, az SMT-alapú egyértelműsítő rendszer kifejezetten alkalmas arra, hogy különböző elveken működő egyértelműsítő rendszerek kombinációjával jelentősen javítsa azok pontosságát.

5. Köszönetnyilvánítás

Mindenekelőtt szeretnék köszönetet mondani témavezetőmnek, Dr. Prószéky Gábornak, akitől rengeteg segítséget és támogatást kaptam az elmúlt évek során. Hálás vagyok a szakmai irányításért, és hogy mindig felhívta figyelmem a kutatásaimmal kapcsolatos előadásokra, konferenciákra és publikálási lehetőségekre. Köszönöm Neki, hogy mindvégig baráti közvetlenséggel fordult felém, és minden munkámban sikerült meglátnia a jót. Nélküle ez a munka nem jöhetett volna létre.

Köszönöm a Pázmány Péter Katolikus Egyetem Multidiszciplináris Műszaki és Természettudományi Doktori Iskola korábbi és jelenlegi vezetőinek, Dr. Roska Tamás, Nyékyné Dr. Gaizler Judit és Dr. Szolgay Péter dékánoknak, hogy lehetőséget biztosítottak arra, hogy Ph.D. munkámat a Karon végezhessem.

Szeretnék köszönetet mondani Vincent Vandeghinstének, Frank Van Eyndének és Ineke Schuurmannak, a Leuveni Katolikus Egyetem professzorainak és doktorainak, hogy kaput nyitottak a statisztikai gépi fordítás világába, és felkeltették érdeklődésem a téma iránt.

Köszönöm legközelebbi munkatársaimnak, hogy a doktoranduszi évek alatt szakmailag és barátilag támogattak. Köszönettel tartozom elsősorban szerzőtársaimnak, Siklósi Borbálának, Orosz Györgynek és Novák Attilának, akik a kutatásaim és publikációim készítése alatt végig segítséget nyújtottak. Köszönet Dr. Wenszky Nórának a magyar és angol nyelvű lektorálásokért. További köszönet a PPKE ITK Nyelvtechnológiai Kutatócsoport tagjainak, többek közt Endrédy Istvánnak, Indig Balázsnak, Dr. Miháltz Mártonnak, Dr. Sass Bálintnak és Yang Zijian Győzőnek az ötleteléseikért és vidám légkörért.

Köszönöm többi volt és jelenlegi doktorandusztársamnak – elsősorban Laki Andrásnak, Bojárszky Andrásnak, Dr. Feldhoffer Gergelynek, Fülöp Tamásnak, Füredi Lászlónak, Gelencsér Andrásnak, Gergelyi Domonkosnak, Dr. Horváth Andrásnak, Dr. Kiss Andrásnak, Dr. Koller Miklósnak, Kovács Dánielnek, Dr. Nemes Csabának, Pilissy Tamásnak, Radványi Mihálynak, Dr. Rák Ádámnak, Stubendek Attilának, Dr. Tátrai Antalnak, Dr. Tibold Róbertnek, Tisza Dávidnak, Dr. Tornai Gábornak, Dr. Tornai Kálmánnak, Tóth Emíliának és Dr. Zsedrovits Tamásnak – a sok baráti beszélgetést és bíztatást.

Köszönettel tartozom a Tanulmányi Osztály és a Gazdasági Osztály munkatársainak, valamint a könyvtárosoknak az évek során nyújtott segítségért.

Végül, de nem utolsósorban szeretném megköszönni egész családomnak az évek során nyújtott bíztatást, segítséget, és hogy minden lehetséges módon támogattak kutatásaim alatt.

6. A szerző publikációi

Folyóiratcikk:

[Laki_1] **Laki, László János**, Attila Novák, and Borbála Siklósi. 2013. "Syntax Based Reordering in Phrase Based English-Hungarian Statistical Machine Translation." *International Journal of Computational Linguistics and Applications* 4 (2): 63–78.

Könyvfejezet:

[Laki_2] **Laki, László János**, György Orosz, and Attila Novák. 2013. "HuLaPos 2.0 – Decoding Morphology." In: *Advances in Artificial Intelligence and Its Applications*, edited by Félix Castro, Alexander Gelbukh, and Miguel González. Lecture Notes in Computer Science Vol. 8265, 294–305. Springer: Berlin-Heidelberg.

Külföldi konferenciakötet:

[Laki_3] **Laki, László János**, and György Orosz. 2014. "An Efficient Language Independent Toolkit for Complete Morphological Disambiguation." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 26–31. Reykjavik, Iceland: European Language Resources Association (ELRA).

[Laki_4] **Laki, László János**, Attila Novak, and Borbála Siklósi. 2013. "English to Hungarian Morpheme-Based Statistical Machine Translation System with Reordering Rules." In: *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, 42–50. Sofia, Bulgaria: Association for Computational Linguistics.

[Laki_5] **Laki, László**. 2012. "Investigating the Possibilities of Using SMT for Text Annotation." In: *1st Symposium on Languages, Applications and Technologies*, 21:267–283. OpenAccess Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Hazai konferenciakötet:

[Laki_6] **Laki, László János**, and György Orosz. 2014. "HuLaPos2 - Fordítsunk morfológiát." In: *X. Magyar Számítógépes Nyelvészeti Konferencia*, 41–49. Szeged: Szegedi Egyetem.

[Laki_7] **Laki, László János**, and György Orosz. 2013. "Morfológiai egyértelműsítés nyelvfüggetlen annotáló módszerek kombinálásával." In: *IX. Magyar Számítógépes Nyelvészeti Konferencia*, 331–337. Szeged: Szegedi Egyetem.

[Laki_8] **Laki, László János**, Attila Novák, and Borbála Siklósi. 2013b. "Hunglish mondattan – átrendezésalapú angol-magyar statisztikai gépifordító-rendszer." In: *IX. Magyar Számítógépes Nyelvészeti Konferencia*, 71–82. Szeged: Szegedi Egyetem.

[Laki_9] **Laki, László János**. 2012. “SMT módszereken alapuló szófaji egyértelműsítő és szótövesítő rendszer.” In: *VI. Alkalmazott Nyelvészeti Doktorandusz Konferencia*, 121–133. Budapest: MTA Nyelvtudományi Intézet.

[Laki_10] **Laki, László János**. 2011a. “Statisztikai gépi fordítási módszereken alapuló egynyelvű szövegelemző rendszer és szótövesítő.” In: *VIII. Magyar Számítógépes Nyelvészeti Konferencia*, 12–23. Szeged: Szegedi Egyetem.

[Laki_11] **Laki, László János**. 2011b. “Angol-magyar statisztikai gépi fordítórendszer minőségének javítása.” In: *V. Alkalmazott Nyelvészeti Doktorandusz Konferencia*, 77–86. Budapest: MTA Nyelvtudományi Intézet.

[Laki_12] **Laki, László János**, and Gábor Prószéky. 2010. “Statisztikai és hibrid módszerek párhuzamos korpuszok feldolgozására.” In: *VII. Magyar Számítógépes Nyelvészeti Konferencia*, 69–79. Szeged: Szegedi Egyetem.

További publikációk:

[Laki_13] **Laki, László János**, and György Orosz. 2011. “VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2010. December 2–3.” *Magyar Terminológia* 4: 119–123.

[Orosz_1] Orosz, György, **László János Laki**, Attila Novák, and Borbála Siklósi. 2013. “Combining Language Independent Part-of-Speech Tagging Tools.” In: *2nd Symposium on Languages, Applications and Technologies*, edited by José Paulo Leal, Ricardo Rocha, and Alberto Simões, 29:249–257. OpenAccess Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Orosz_2] Orosz, György, **László János Laki**, Attila Novák, and Borbála Siklósi. 2013. “Improved Hungarian Morphological Disambiguation with Tagger Combination.” In: *Text, Speech, and Dialogue*, 8082:280–287. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg.

7. Irodalomjegyzék

[1] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180.

[2] A. Novák, "What is good Humor like?," in *I. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, 2003, pp. 138–144.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, USA, 2002, pp. 311–318.

[4] T. Brants, "TnT - A Statistical Part-of-Speech Tagger," in *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, USA, 2000, pp. 224–232.

[5] P. Halácsy, A. Kornai, and C. Oravecz, "HunPos: An open source trigram tagger.," in *Proceedings of the 45th Annual Meeting of the ACL*, Prague, Czech Republic, 2007, pp. 209–212.

[6] G. Orosz and A. Novák, "PurePos 2.0: a hybrid tool for morphological disambiguation," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Hussal, Bulgaria, 2013, pp. 539–545.

[7] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, Hong Kong, China, 2000, pp. 63–70.

[8] A. Søgaard, "Simple semi-supervised training of part-of-speech taggers," in *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, Sweden, 2010, pp. 205–208.

[9] R. Yeniterzi and K. Oflazer, "Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010, pp. 454–464.

[10] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, Jun. 1993.

[11] P. Koehn, *Statistical Machine Translation*, 1st ed. New York, NY, USA: Cambridge University Press, 2010.

[12] G. G. Mora and J. A. S. Peiró, "Part-of-Speech Tagging Based on Machine Translation Techniques," in *Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part I*, Girona, Spain, 2007, pp. 257–264.

[13] I. D. El-Kahlout and K. Oflazer, "Exploiting morphology and local word reordering in English-to-Turkish phrase-based statistical machine translation," *Audio Speech Lang. Process. IEEE Trans. On*, vol. 18, no. 6, pp. 1313–1322, 2010.