DOI:10.15774/PPKE.ITK.2015.004

# *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.

DOI:10.15774/PPKE.ITK.2015.004

DOI:10.15774/PPKE.ITK.2015.004

# 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 ötletelésekért 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.

DOI:10.15774/PPKE.ITK.2015.004

## Abstract

Phrase-based statistical machine translation systems rely on statistical observations derived from phrase alignments automatically extracted from parallel bilingual corpora. The main advantage of applying SMT is its language-independence. The phrase-based model works well for language pairs with similar syntactic structure and word order. However, phrase-based models fail to handle great grammatical differences adequately.

The first part of my work deals with improving statistical machine translation between grammatically distant languages. It is almost impossible to create a high quality machine translation to agglutinative languages with purely statistical methods. The main problems are the data sparseness problem, generating the surface form of the word in agglutinative languages, or the different word number between a sentence pair. In this work a hybrid translation system is described that is an extension of the baseline statistical methods by applying syntax- and morphology-based preprocessing steps on the training corpus and morphological postprocessing during translation. Effects of my improvements were demonstrated using English-to-Hungarian translation. The goal was to transform the source side English sentences to a syntactic structure that is more similar to that of the target side Hungarian sentences. I concentrated on syntactic structures that have systematically differing realizations in the two languages. In this work several experiments were performed on English–Hungarian machine translation. First of all different syntax-motivated reordering rules were applied as preprocessing steps; secondly a morphological generator was used to generate the correct surface form of a word; and thirdly three morpheme-based translation system were presented. The results showed that readability and accuracy of the translation are improved by the process of reordering the source sentences prior to translation, especially in the cases when the somewhat fragile POS tagger-parser chain does not lead to wrongly reordered sentences, which has a deteriorating effect on translation quality. Although automatic evaluation assigned the morpheme-based system a significantly and consistently lower score than the baseline system, the human evaluation confirmed that applying reordering and morphological segmentation does improve translation quality in the case of translating to an agglutinating language like Hungarian. I found that several linguistic phenomena can be translated with a much better accuracy than using a traditional SMT system.

DOI:10.15774/PPKE.ITK.2015.004

The second part of my work focuses on a really important task for computational linguistics, namely marking texts with syntactic and/or semantic information, or the morphological analysis of the language. Complete morphological disambiguation is the process to find the lemma and identify the morphosyntactic label of each word of a sentence in one step. Nowadays, only few of them carry out complete morphological disambiguation, which is essential in the case of morphologically rich languages. Furthermore, there are only a few POS taggers that achieve high accuracy amongst grammatically different languages. The aim of this work is to introduce a new approach for complete morphological disambiguation tool, that performs POS tagging and lemmatization simultaneously based on the Moses framework. This tool can be used for different sorts of languages, while producing accuracy scores competing with the ones of language dependent systems. The presented system employs a trie-based suffix guesser, which effectively handles the problem of out-of-vocabulary words, typical for morphologically rich languages like Hungarian. The performance of the system was compared to the state-of-the-art language dependent and language independent systems for annotating Hungarian and five other languages (English, Croatian, Serbian, Bulgarian and Portuguese). The presented method outperforms most of the language independent systems that were compared with mine. Furthermore, the accuracy of the system is comparable with language dependent ones.

DOI:10.15774/PPKE.ITK.2015.004

# Kivonat

A kifejezésalapú statisztikai gépi fordítórendszerek a párhuzamos kétnyelvű korpusz szóösszekötései alapján készített statisztikai megfigyelések alapján működnek. Alkalmazásuk legfőbb előnye nyelvfüggetlen mivoltukban rejlik. A kifejezésalapú modell jó eredménnyel működik hasonló szintaktikai struktúrájú és szórendű nyelvpárok esetén, de a számottevő grammatikai különbségeket nehezen kezeli.

Munkám első része a nyelvtanilag távol eső nyelvpárok közti statisztikai gépi fordítás fejlesztésével, tökéletesítésével foglalkozik. Szimplán statisztikai módszerek alkalmazásával szinte lehetetlen magas minőségű fordítórendszert alkotni agglutináló nyelvek esetében, főleg ha az a célnyelv. Ebben legfőbb akadályt az adathiány-probléma, a szóalakok generálásának nehézsége és a mondatpárok eltérő szószáma jelenti. 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. Célom a forrásnyelvi angol mondat szintaktikai struktúrájának átalakítása volt, hogy az minél inkább megfeleljen a célnyelvi magyar mondat felépítésének. Főleg azokat a szintaktikai struktúrákat változtattam meg, melyeknek szisztematikusan különböző realizációi vannak a két nyelvben. Több kísérletet végeztem az angol-magyar gépi fordítás minőségének javítására. Egyrészt előfeldolgozó lépésként különböző kézzel írt szintaxismotivált átrendezési szabályokat alkalmaztam. Ezenkívül a helyes célnyelvi szóalak előállítása érdekében morfológiai generátort alkalmaztam a statisztikai gépi fordító dekódere helyett. Végül három morfémaalapú fordítórendszert építettem fel és mutatok be. Az eredmények megmutatták, hogy a fordítás az emberi kiértékelők szerint mind olvashatóság, mind pontosság szempontjából javult, valamint az automatikus kiértékelő módszer esetén is sikerült javulást elérni. Ez főleg azokban az esetekben volt megfigyelhető, amikor a szintaktikai elemzés során nem merült fel elemzési hiba, ami rossz átrendezéshez és helytelen fordításhoz vezetett. Habár az automatikus kiértékelés az általam készített morfémaalapú rendszereket jelentősen alulpontozta az eredeti kifejezésalapú SMT-hez képest, az emberi kiértékelés megerősítette, hogy az átrendezési szabályok alkalmazásával és morfológiai szegmentációval javítható az agglutináló nyelvekre történő fordítás minősége. Az elvégzett vizsgálatok megmutatták, hogy a hagyományos statisztikai gépi fordítórendszerhez képest rendszeremmel több nyelvi jelenség is nagyobb pontossággal fordítható.

DOI:10.15774/PPKE.ITK.2015.004

Munkám második felében a számítógépes nyelvészet egyik fontos kérdésével, a szöveg szintaktikai és/vagy szemantikai információval történő ellátásával, vagyis a nyelv morfológiai elemzésével foglalkozik. A teljes morfoszintaktikai egyértelműsítés feladata egy lépésben megtalálni a mondat szavainak lemmáit és morfoszintaktikai címkesorozatait. Napjainkban nagyon kevés olyan alkalmazás létezik, ami teljes morfoszintaktikai egyértelműsítést végez, ami alapvető probléma gazdag morfológiájú nyelvek feldolgozása esetén. Ezenkívül kevés olyan szófaji egyértelműsítő rendszer létezik, ami nyelvtanilag különböző nyelvek esetén is nagy pontossággal működik; ugyanis egy nyelvfüggő alkalmazás nagyon magas pontosságot képes elérni adott korpuszon. Munkám célja egy új megközelítéssel működő morfológiai egyértelműsítő eszköz bemutatása, mely egyidejűleg végez morfológiai elemzést és lemmatizálást. Az általam készített rendszer különböző típusú nyelvek elemzésére alkalmazható amellett, hogy pontossága eléri – de néhány esetben meg is haladja – a nyelvfüggő rendszerekét. A bemutatott rendszer egy végződésfa-alapú ajánlórendszert alkalmaz, amely egy tanítóhalamaz segítségével javaslatokat ad a tanítóanyagban nem szereplő szavak lehetséges szófajára. Ez megoldást nyújt a gazdag morfológiájú nyelvek esetén, mivel hatékonyan kezeli az ismeretlen szavak elemzésének problémáját. Az általam felépített rendszer teljesítményét több nyelv nyelvfüggő és nyelvfüggetlen egyértelműsítő rendszereinek eredményeivel hasonlítottam össze. Rendszerem eredménye meghaladja a legtöbb vele összehasonlított nyelvfüggetlen alkalmazás teljesítményét, valamint összemérhető a nyelvfüggő alkalmazások teljesítményével.

DOI:10.15774/PPKE.ITK.2015.004

# Tartalomjegyzék

DOI:10.15774/PPKE.ITK.2015.004

DOI:10.15774/PPKE.ITK.2015.004

# Táblázatjegyzék

DOI:10.15774/PPKE.ITK.2015.004

# Ábrajegyzék

DOI:10.15774/PPKE.ITK.2015.004

# I. Alapozó fejezetek

## 1.1 Bevezetés és kutatói célok

A nyelvtechnológia egyik legfontosabb feladata a nyelvi diverzitás okozta akadályok áthidalása, vagyis a számítógépek alkalmassá tétele 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 erre a célra leginkább alkalmazott módszer a statisztikai gépi fordítás (SMT). Az SMT rendszer egy teljesen nyelvfüggetlen eszköz, ami felügyelt gépi tanulási módszerek segítségével tanítható, valamint megfelelő mennyiségű tanítóadat 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 betöltött 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 szimplán statisztikai fordítórendszer szintaxisalapú szabályokkal történő 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 morfoszintaktikai egyértelműsítés, melynek feladata a szavak szótövének meghatározása, és besorolása az egyes morfoszintaktikai 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 végzi. Mivel a statisztikai gépi fordítás feladata két nyelv közti transzformáció megvalósítása, emiatt alkalmas lehet a teljes morfoszintaktikai egyértelműsítés feladatának elvégzésére. Ebben az esetben az eredeti elemzendő szövegről a szófajilag egyértelműsített, szótövezett szóalakok közti fordítást kell megvalósítanunk. **Munkám második felében célom egy statisztikai gépi fordításon alapuló nyelvfüggetlen teljes morfoszintaktikai egyértelműsítő rendszer kidolgozásának bemutatása, mely eléri vagy több esetben meghaladja a már létező nyelvfüggő és nyelvfüggetlen rendszerek eredményeit.**

DOI:10.15774/PPKE.ITK.2015.004

## 1.2 Rövidítésjegyzék

BLEU: BiLingual Evaluation Understudy

BP: brevity penalty

CFG: környezetfüggetlen nyelvtani (context free grammar)

EM: expectation-maximization

HMM: rejtett Markov modell

IRSTLM: IRST Language Modeling Toolkit

mm-BLEU: morfémaalapú BLEU

MSD: Morpho-Syntactic Description

OOV: tanítóanyagban nem szereplő szavak (Out-of-Vocabulary)

PBSMT: kifejezésalapú gépi fordító

PER: Position independent word Error Rate

POS: Part-of-Speech

RANDLM: Randomised Language Modeling

RBMT**:** szabályalapú gépi fordítórendszer

SMT: statisztikai gépi fordítás

SRILM: SRI Language Modeling Toolkit

TBL: transzformáció-alapú gépi tanulás

TbSMT: faalapú fordítórendszer (Tree-Based Statistical Machine Translation)

w-BLEU : szóalapú BLEU

WMT: Workshops on Statistical Machine Translation

DOI:10.15774/PPKE.ITK.2015.004

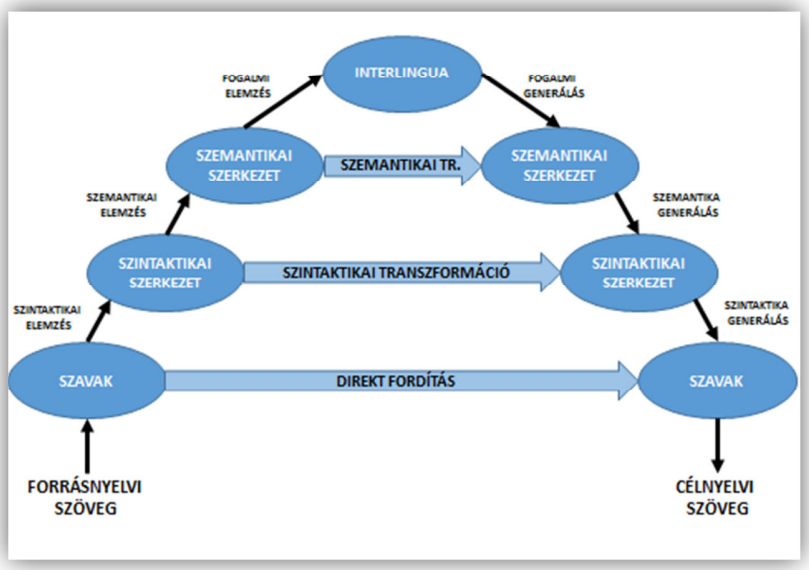
# 2 Elméleti háttér

## 2.1 A gépi fordítás típusai

A nyelvtechnológia egyik jelentős területe a soknyelvűség támogatása, amire a napjainkban igen hangsúlyos globalizációs törekvések miatt egyre növekvő igény van. Ebben nyújt támogatást a gépi fordítás, aminek módszerei nem csupán a nyelvek közötti transzformáció megvalósításáról szólnak, hanem tartalmazzák a szövegek elő- és utófeldolgozását, valamint a fordítások minőségének előzetes becslését, illetve azok kiértékelését is. A gépi fordítás tudománya egyidős az első számítógépek megjelenésével, és mind a mai napig a számítógépes nyelvészet egyik leginkább kutatott területe. Az elmúlt közel hatvan év során számos megközelítés született a természetes nyelvek közötti fordítás megoldására, amelyek közül jelen tanulmányban a legfontosabbakat mutatom be.

**A szabályalapú gépi fordítórendszer (RBMT – Rule-Based Machine Translation)** alapötlete, hogy a fordítandó szövegből a kinyerhető legtöbb információt felhasználja a fordítás során. A legegyszerűbb első implementációk az úgynevezett direkt fordítórendszerek. A módszer lényege, hogy a fordítandó szöveget egy szótár alapján szóról szóra fordítja le, majd a megfelelő sorrendbe rendezi. A módszer előnye, hogy viszonylag könnyen megvalósítható, viszont nem képes komplex nyelvtani szerkezetek kezelésére. Emiatt a fordítás minősége nem túl jó. A későbbi, bonyolultabb rendszerek a fordítandó szövegből elemzés segítségével állítanak elő egy köztes reprezentációt, amit előre definiált átviteli szabályok segítségével alakítanak át egy absztrakt célnyelvi reprezentációra. Végül ebből a reprezentációból generálják a célnyelvi szóalakokat. Ezeket a rendszereket az elemzés és generálás mélysége, valamint az átvitel helye alapján osztályozhatjuk, amit a Vauquois-háromszög szemléltet (1. ábra). Az ábrán látható, hogy minél mélyebb a nyelvi elemzés mértéke a fordítás során, annál közelebb áll egymáshoz a két nyelv reprezentációja, amik között a transzformációt végre kell hajtani. Egy szabályalapú gépi fordítórendszer, ha precízen megírt szabályokkal rendelkezik, nagy pontosságú fordítást képes előállítani, de az átviteli szabályok létrehozásához elengedhetetlen a jó minőségű szintaktikai és/vagy szemantikai elemző, ami csak nagyon kevés nyelv esetén áll rendelkezésre. Továbbá, mivel ezek a szabályok nyelvspecifikusak, minden nyelvpárra külön-külön kell létrehozni őket, ami megnehezíti a rendszer kiterjesztését újabb nyelvekre.

DOI:10.15774/PPKE.ITK.2015.004




**1. ábra: Vauquois-háromszög [1], [2]**

**A példaalapú fordítórendszer módszerének** alapötlete, hogy az aktuális fordításhoz felhasználja a már korábban lefordított mondatokat. A rendszer egy előre létrehozott fordítómemóriából kiválasztja a fordítandó mondat részeinek eltárolt fordításait, amik egyesítésével megkapjuk a kívánt fordítást [3]. Annak ellenére, hogy a rendszer nem tartalmaz komplex, nyelvspecifikus modulokat, fordítási minősége nem sokkal marad el szabályalapú társaitól. Alapvető hiányossága azonban, hogy a fordítómemóriában tárolt szegmenspárok elemi egységei (pl. morfémák, szavak, kifejezések, stb.) nincsenek összekötve. Emiatt annak ellenére, hogy a fordítórendszer tudja, hogy a memóriában tárolt forrásnyelvi szegmens melyik részében különbözik a fordítandó szegmenstől, nem tudja megmondani, hogy a célnyelvi oldalon ez melyik szavakra van hatással.

**A statisztikai gépi fordítórendszer (SMT – Statistical Machine Translation)** a példaalapú fordítórendszer általánosított változatának tekinthető, mivel képes javaslatokat tenni a fordítómemóriában nem szereplő szegmensek fordítására 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, amiben 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 legtöbbet hivatkozott gépi fordító architektúra legyen.

DOI:10.15774/PPKE.ITK.2015.004

**A hibrid gépi fordítórendszer** úgy alakult ki, hogy a kutatások során bebizonyosodott, a fent felsorolt rendszerek önmagukban nem képesek általánosan megoldani a természetes nyelvek közötti fordítás feladatát. A hibrid megoldások a különböző módszerek együttes alkalmazásával javítják a fordítórendszer minőségét. Napjaink legjobban teljesítő fordítórendszerei az SMT és az RBMT módszerek integrációjából létrejött hibrid architektúrák.

Mindezek alapján, kutatói célkitűzésem is a legígéretesebbnek bizonyuló hibrid megoldásokban rejlő lehetőségek vizsgálatára irányult. Munkám során az SMT módszert különböző szabályalapú modulok integrációjával hibridizáltam, melynek segítségével mind az emberi, mind az automatikus kiértékelés a fordítási minőség javulását igazolta. A következő fejezetben áttekintem az SMT módszer elméleti hátterét.

## 2.2 A statisztikai gépi fordítás elméleti háttere

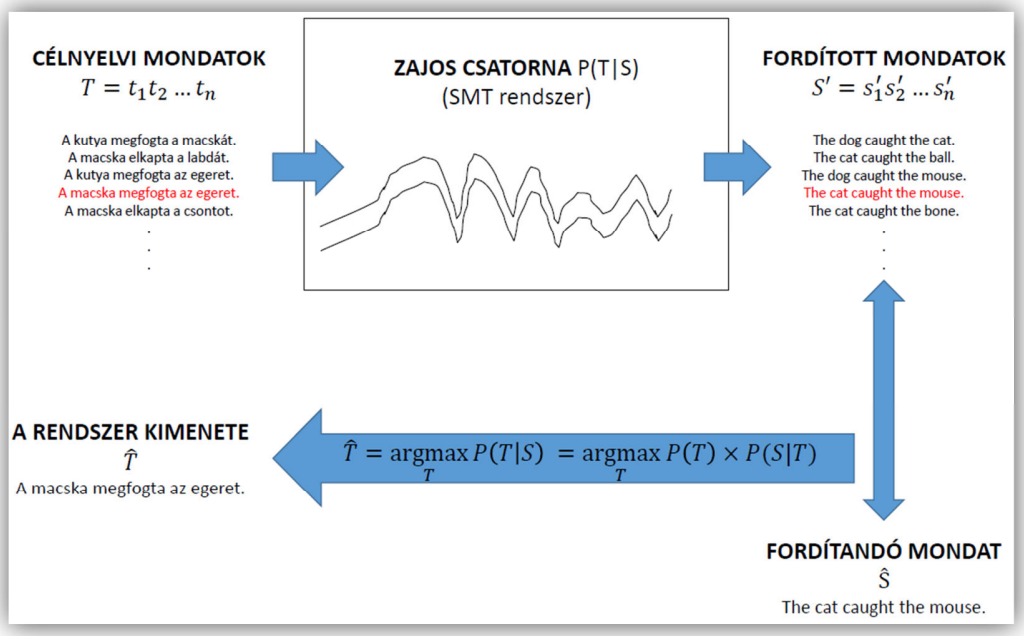
A statisztikai gépi fordítórendszer létrejötte a számítógépek növekvő teljesítményének, valamint a digitálisan hozzáférhető adatmennyiség robbanásszerű növekedésének köszönhető. Az 1990-es évek elejétől a számítógépek képessé váltak nagy mennyiségű adat gyors és hatékony kezelésére, ennek köszönhetően alkalmassá váltak arra, hogy gépi tanulási módszerekkel képesek legyenek a fordítási modell létrehozására. A módszer legnagyobb előnye, hogy a fordítórendszer felépítéséhez nem szükséges a nyelvek grammatikájának ismerete. A rendszer tanításához csupán egy párhuzamos kétnyelvű korpuszra van szükség, ebből tanulja meg a transzformációhoz szükséges komponenseket.

Az SMT alapjait az IBM T. J. Watson Research Center munkatársai fektették le [4], akik a fordítás feladatát a beszédtechnológiában használatos Shannon-féle zajoscsatorna-modell [5], [6] segítségével közelítették meg. A későbbi kutatások eredményeként napjainkra a zajoscsatorna-modell kiegészített változatát alkalmazzák, az úgynevezett log-lineáris modellt [7], [8]. A következőkben bemutatom ezt a két modellt, valamint ezek kapcsolódását a gépi fordításhoz.

### 2.2.1 Zajoscsatorna-modell

A statisztikai gépi fordítás feladata megfogalmazható a Shannon-féle zajoscsatorna-modell [5], [6] segítségével, amit a 2. ábra mutat. Az elmélet alapja, hogy a fordítás során az egyetlen biztosan ismert információ a fordítandó szöveg. A fordítás úgy történik, mintha a célnyelvi szövegek halmazát egy zajos csatornán átengedve a csatorna kimenetén összehasonlítanánk a forrásnyelvi szöveggel. Az a célnyelvi mondat lesz a rendszer kimenete, amelyik a csatornán való áthaladás után a legjobban hasonlít a fordítandó mondatra.

DOI:10.15774/PPKE.ITK.2015.004




**2. ábra: Zajoscsatorna-modell**

Formálisan az SMT módszer a fordítás feladatát úgy tekinti, mint a fordítás pontosságát, valamint gördülékenységét reprezentáló modellek kombinációja által elérhető maximális valószínűségi értékhez tartozó szöveg meghatározása. A fordítás feladata tehát úgy fogalmazható meg, hogy keressük azt a célnyelvi mondatot ($\hat{T}$), amelyik a célnyelvi mondatok halmazából ($T$) a legvalószínűbb fordítása a forrásnyelvi mondatnak ($S$).

$$\hat{T} = \underset{T}{\operatorname{argmax}}\ P(T|S) \tag{1}$$

A $P(T|S)$ valószínűség azonban közvetlenül nem számolható, viszont önmagában modellezhető részekre bontható. A Bayes-tétel alapján az (1) egyenlet átalakítható a következőképpen:

$$\hat{T} = \underset{T}{\operatorname{argmax}} P(T|S) = \underset{T}{\operatorname{argmax}} \frac{P(S|T)P(T)}{P(S)} = \underset{T}{\operatorname{argmax}} P(S|T)P(T) \tag{2}$$

Mivel $T$ függvényében $P(S)$ konstans, ezért elhagyható. Az így kapott egyenlet két komponensből áll:

- $P(T)$ a nyelvmodell, ami a gördülékenységért felelős (2.2.3.1. fejezet)
- $P(S|T)$ a fordítási modell, ami a fordítás pontosságát biztosítja (2.2.3.2. fejezet)

A két modell kombinációjának maximális értékét a fordítórendszer dekóder komponense

DOI:10.15774/PPKE.ITK.2015.004

határozza meg. A (2) egyenlet legfontosabb jellemzője, hogy a fordítás feladatát kiszámítható egységekre bontja. Ráadásul e komponensek becslése egy- és kétnyelvű korpuszok segítségével automatikusan történik.

### 2.2.2 Log-lineáris modell

A zajoscsatorna-modell használata számos megszorítást vezet be, amik korlátozzák a fordítórendszer minőségét. Ilyen megszorítás például, hogy egy szó vagy kifejezés fordítását a környező szavaktól függetlenül tekinti, illetve hogy a nyelvmodell csak a néhány megelőző szót veszi figyelembe. Sajnos a megszorítások jelentősen csökkentik a fordítórendszer minőségét, ami miatt szükségessé vált, hogy a fordításhoz a nyelvmodellen és a fordítási modellen kívül egyéb tudás is felhasználható legyen. Erre ad megoldást a log-lineáris modell.

A log-lineáris modell a gépi tanulás tudományágának egyik gyakran használt módszere. Lényege, hogy egy feladatot egymástól független jellemzők $\left(h_i(x)\right)$ súlyozott szorzatával $(\lambda_i)$ ír le. Formálisan a modell a következő:

$$P(x) = \exp \sum_{i=1}^{n} \lambda_i h_i(x) \tag{3}$$

ahol $h_i(x)$ az $i.$ jellemző függvény, míg $\lambda_i$ a hozzá tartozó súly [7], [8]. Ennek köszönhetően a modell a zajoscsatorna-modell kiegészítésének tekinthető, mivel a nyelvmodell és a fordítási modell mellett tetszőleges számú komponenssel bővíthető.

### 2.2.3 A statisztikai gépi fordítórendszer által implementált eszközök és komponensek

Az SMT rendszer megvalósítására több implementáció létezik, ezek közül munkám során a Moses nevű keretrendszert [9] használtam, amely szabadon hozzáférhető, és a különböző komponensek több implementációja is megtalálható benne. Ebben a fejezetben az SMT módszernek a Moses rendszerbe integrált komponenseit mutatom be.

#### *2.2.3.1 A nyelvmodell*

Az SMT rendszer egyik alapkomponense a nyelvmodell, ami egy szósorozat adott nyelven való természetes előfordulásának valószínűségére ad becslést (azaz hogy egy szósorozat egy anyanyelvi beszélő számára mennyire hangzik természetesen). A modell feladata, hogy a fordítórendszer „gördülékenyen” olvasható szöveget adjon a kimenetén. Továbbá, a nyelvmodell szükséges a többértelműség és a szórendi problémák kezeléséhez is.

DOI:10.15774/PPKE.ITK.2015.004

A nyelvmodell a szöveg gördülékenységét úgy közelíti, hogy minden szóhoz megállapítja, hogy az mekkora valószínűséggel fordul elő az őt megelőző n db szó után. A modell létrehozása automatikusan, a célnyelvi korpuszból történik. Formálisan az n-gram nyelvmodell $\big(P(T)\big)$, az egyes szavakhoz tartozó $\big(T = [t_1, \dots, t_{|T|}]\big)$ valószínűségek szorzata, ahol minden szó valószínűsége az őt megelőző szavak sorozatából számolható a következőképpen:

$$P(T) = \prod_{i=0}^{|T|} P_{LM}(t_i|t_{i-1}, t_{i-2}, \dots, t_1) = \prod_{i=0}^{|T|} P_{LM}\big(t_i\big|t_1^{i-1}\big) \tag{4}$$

A nyelvmodell építésére több eszköz is elérhető. Ezek közül a legismertebb a SRILM (SRI Language Modeling Toolkit) [10], ami a statisztikai nyelvi modellezés több implementációját is tartalmazza. Ez a rendszer jó minőségű nyelvmodellt hoz létre, aminek ára viszont a magas erőforrásigény. Az ingyenesen elérhető IRSTLM (IRST Language Modeling Toolkit) [11] és RANDLM (Randomised Language Modeling) [12] nevű alkalmazások hatékonyabban képesek nagyobb méretű tanítóanyagok feldolgozására, nagyjából hasonló pontosság mellett, viszont kisebb méretű korpuszok esetén a SRILM teljesít jobban [11]. Munkám során a SRILM-et használtam, mivel nem áll rendelkezésemre nagyméretű tanítóhalmaz.

#### *2.2.3.2 A fordítási modell*

A második komponens az úgynevezett fordítási modell, amely a fordítás tartalomhűségére ad becslést. A fordítási modell $\big(P(S|T)\big)$ valószínűségi értékeket tárol arról, hogy egy célnyelvi $(T = [t_1, \dots, t_m])$ szegmens mekkora valószínűséggel fordítása egy forrásnyelvi $(S = [s_1, \dots, s_n])$ szegmensnek. A fordítórendszer fajtája alapján ezek a szegmensek (fordítási egységek) lehetnek szavak (szóalapú SMT), kifejezések (kifejezésalapú SMT), tulajdonsághalmazok (faktoralapú SMT), vagy akár generatív szabályok is (szintaxisalapú SMT). A különböző módszereket a 2.2.4 fejezetben bővebben kifejtem. A fordítási modellt kétnyelvű párhuzamos korpusz segítségével tanítjuk.

#### *2.2.3.3 A dekóder*

A nyelvmodell és a fordítási modell mellett a dekóder a zajoscsatorna-modell alapú SMT rendszer harmadik alappillére. A dekóder hatékony, valós idejű keresési algoritmust valósít meg, ami a lehetséges fordítási javaslatok közül választja ki a legvalószínűbb fordítást. A feladat komplexitása a fordítandó mondat hosszától függ, hiszen a lehetséges fordítások száma a mondat hosszának exponenciális függvénye. Knight [13] bebizonyította, hogy a dekódolás feladata NP-teljes komplexitású, ezért a fordítórendszerekben használt algoritmusok nem érhetik el a tökéletes fordítást. Ehelyett heurisztikus kereséssel véges időn belül megközelítő megoldást adnak. A keresésre alkalma-

DOI:10.15774/PPKE.ITK.2015.004

zott rendkívül hatékony algoritmus a veremalapú nyalábolt keresési algoritmus (stack-based beam search), amit a Moses is használ [14], [15]. A nyalábolt keresés a heurisztikus keresési algoritmusok azon fajtája, amelyek gráfok segítségével találják meg az állapottérben a legnagyobb súlyú utat. A dekóder egy mondat fordítása során a fordítási modellben tárolt entitások alapján egy keresési gráfot épít, amiben a legkisebb súlyú út lesz a helyes fordítás. Belátható, hogy ez a keresési feladat a fordítási modell méretével, valamint a fordítandó mondat hosszával exponenciálisan arányos. Annak érdekében, hogy a dekóder valós időben működjön, szükség van az állapottér csökkentésére. A veremalapú dekóder [16], [17] lényege, hogy működése közben mindig csak a pillanatnyilag legjobb jelölteket veszi figyelembe, a többit figyelmen kívül hagyja.

#### *2.2.3.4 Szóösszekötő*

A statisztikai fordítórendszerben a fordítási modell (2.2.3.2. fejezet) tárolja a kifejezések fordításait és a hozzájuk tartozó valószínűségeket. Az összetartozó kifejezéseket a párhuzamos korpuszban a szóösszekötő rendszer segítségével határozza meg a rendszer (IBM modellek [4]). A szóösszekötés feladatának a nehézsége, hogy a fordítási modell segítségével képes az összetartozó szópárok megtalálására, viszont a fordítási modell a szószinten összepárosított korpusz segítségével építhető fel. Ez a paradoxon a közismert expectation-maximization, röviden EM-algoritmus [18] segítségével oldható fel. Az algoritmus első lépésben uniform eloszlást feltételez minden korpuszban lévő szópárra. Ezután két lépést alkalmaz felváltva: első lépésben az aktuális összekötések alapján kiszámolja a fordítási modell súlyait; a második lépésben ezen súlyok alapján újragenerálja a szóösszekötéseket. A két lépést addig ismétli, amíg az állapottér nem konvergál. A módszer egyik implementálása a GIZA++ rendszer [7], melyet munkám során alkalmaztam. Az IBM 1 modell – aminek a feladata az összetartozó kifejezések összepárosítása és ezek segítségével a mondatok fordítása – egy igen jelentős megszorítással rendelkezik, miszerint a modell alapján több forrásnyelvi szót nem lehet ugyanahhoz a célnyelvi szóhoz kötni, azaz a sok-egy reláció nem engedélyezett [4], [17], [2. 86 old.]. Ez a megkötés jelentősen megnehezíti az agglutináló nyelvekre történő fordítást, mivel így nem lehet helyesen összekötni az angol szavakat a megfelelő ragozott magyar alakokkal.

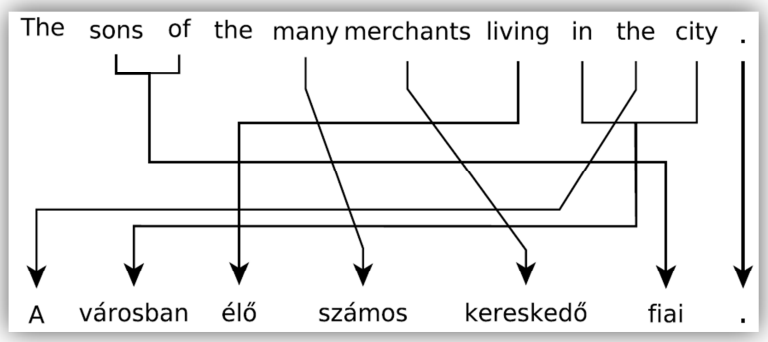
#### *2.2.3.5 Szórendbeli különbséget büntető modell*

A log-lineáris modell által – a fordítási és a nyelvmodell mellett – bevezetett első kiegészítő komponens a szórendbeli különbséget büntető modell (distortion model). A modellnek feladata a fordítás során a szavak helyes sorrendjének a meghatározása, valamint a keresési gráf méretének a csökkentése. A Mosesben használatos átrendezési modell azzal a megközelítéssel él, hogy a célnyelvi mondat szavainak sorrendje hasonlít a forrásnyelvi mondat szavainak sorrendjéhez. Emiatt

DOI:10.15774/PPKE.ITK.2015.004

a modell bünteti a fordítás során jelentkező túl nagy távolságú átrendezéseket, valamint definiálja a dekódolás során lehetséges maximális átrendezés távolságát [2]. A gyakorlatban ez azt jelenti, hogy a fordítórendszer a monoton fordítást támogatja, és ebben az esetben képes a legjobb minőségű eredményt elérni.

A modellben használt közelítés továbbra is megnehezíti a grammatikailag távoli nyelvek fordítását, mivel túl erősen bünteti a nagy távolságú átrendezéseket. Ez problémát okoz például angol-német nyelvpárok közötti fordítás esetén is, ahol a mellékmondat végén szereplő ige fordítását a modell olyan mértékben bünteti, hogy az SMT rendszer nem fordítja le. Hasonló a helyzet az angol-magyar vagy a japán-angol fordításnál is.




**3. ábra: A szórendi különbség szemléltetése egy angol-magyar példán**

A 3. ábra egy angol-magyar példa alapján mutatja be, hogy a *city* szó forrásnyelvi pozíciója 8 egység távolságra van a célnyelvi fordítás (*városban*) pozíciójától. Emiatt, ha a szórendbeli különbséget büntető modell ennél kisebb értékben maximalizálja a lehetséges átrendezések távolságát, a dekóder nem lesz képes helyesen lefordítani ezt a szót.

#### *2.2.3.6 A lexikalizált átrendezési modell*

A lexikalizált átrendezési modell (phrase based lexicalized reordering model) lényege, hogy egy kifejezés fordítása után három úton folytatódhat a fordítás folyamata:

- balról jobbra történő monoton fordítás;
- a soron következő kifejezés átugrása;
- a megelőző kifejezés fordítása.

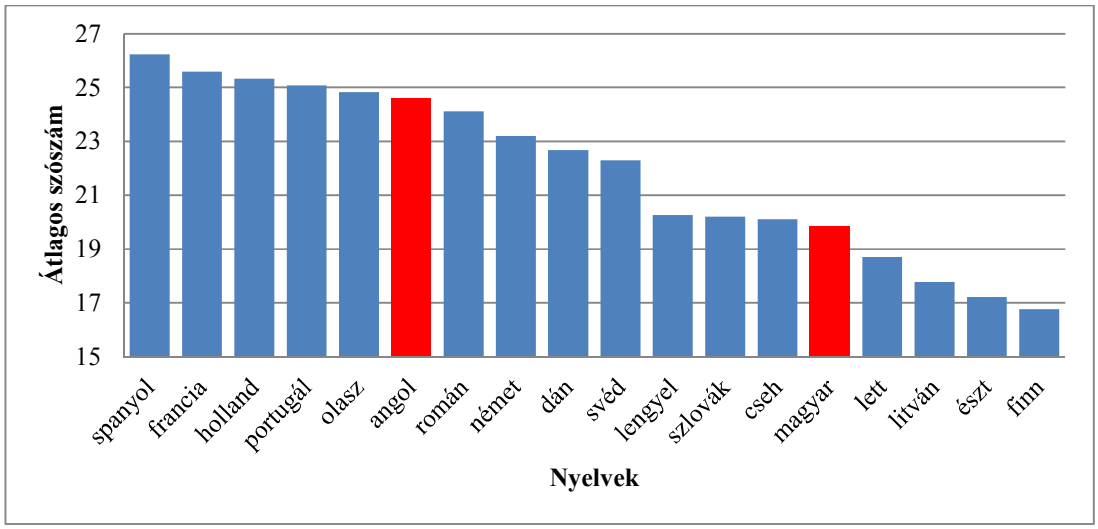
Például az *„I can count the stars visible.”* mondat esetén a *the* szó fordítása után nem a soron következő *stars* szó fordítása következik, hanem az azt következő *visible* szóé, majd ezután visszalép a megelőző szóra. Így kapjuk meg a *„Meg tudom számolni a látható csillagokat.”* fordítást. A

DOI:10.15774/PPKE.ITK.2015.004

lexikalizált átrendezési modellben a Moses rendszer a maximum likelihood becslés segítségével kiszámolja a fordítási modellben megtanult kifejezésekhez a három lehetséges úthoz tartozó valószínűséget. Ez a számítás a párhuzamos kétnyelvű tanítóhalmaz alapján történik.

#### 2.2.3.7 *A mondathossz-harmonizációs modell*

A mondathossz-harmonizációs modell azzal a feltételezéssel él, hogy az eredeti mondat és a lefordított mondat szavainak száma hasonló. Ennek megfelelően a modell feladata, hogy kiszűrje a szószámban jelentősen eltérő fordítási javaslatokat. Könnyen belátható, hogy a modell által használt feltételezés túlságosan erős feltétel. A 4. ábra egy példát mutat be, ahol az Europarl korpusz [20] alapján kiszámoltam, hogy átlagosan hány szóból áll egy mondat a korpuszban szereplő nyelvekben. A diagramból kiolvasható, hogy a különböző típusú nyelvek között jelentős eltérés mutatkozik a mondatok átlagos szószámában, emiatt a mondathossz-harmonizációs modell hangsúlyozott figyelembe vétele ronthatja a fordítórendszer minőségét.




**4. ábra: A mondatok átlagos szószáma**

### 2.2.4 A statisztikai gépi fordítás típusai

Az első SMT rendszerek szó alapon működtek. Mivel azonban a szavak fordítása függ azok környezetétől, bevezetésre kerültek a kifejezésalapú fordítórendszer-megvalósítások, amik a szavak helyett szócsoportokkal dolgoznak. Ebben a kontextusban kifejezésnek tekintünk bármilyen tokensorozatot. A módszer továbbfejlesztett változatai már nemcsak a szóalakokat használják fel, hanem a mondatok szintaktikai jellemzőit is figyelembe veszik.

DOI:10.15774/PPKE.ITK.2015.004

#### *2.2.4.1 Szóalapú gépi fordítás*

A zajoscsatorna-modell alapján egy forrásnyelvi mondat fordítását az összes lehetséges célnyelvi szósorozat vizsgálatával kaphatjuk meg. A legtöbb természetes nyelv esetén egy szó csak néhány célnyelvi szóra fordulhat, ezért felesleges a fordítás során a célnyelvi szótár minden elemét figyelembe venni. A lehetséges fordítások közül a megfelelő megtalálásában a már lefordított szövegrész segíthet. Ahhoz, hogy a fordítórendszer képes legyen megtalálni a releváns szópárokat, a rendszer tanítása során meg kell találni és össze kell párosítani az összetartozó szópárokat. A szóalapú gépi fordítórendszer megoldást nyújt erre a problémára úgy, hogy a zajoscsatorna-modellben bemutatott fordítási modellt kiegészíti egy szóösszekötő modellel. Az új modell feladata meghatározni a párhuzamos mondatokban az összetartozó szavakat. Ennek segítségével a fordítási modell már csak a releváns szópárokat és azok valószínűségét tartalmazza. Ez a modell formálisan az (5) egyenlettel írható le,

$$P(S|T) = \sum_{A \in \alpha} P(S, A|T) = \cdots = \frac{\varepsilon}{(l+1)^k} \prod_{i=1}^{k} \sum_{j=0}^{l} P(s_i|t_j) \tag{5}$$

ahol $A$ a szóösszekötő modell, $S = s_1 \dots s_k$ a forrásnyelvi mondat és $T = t_1 \dots t_l$ a célnyelvi mondat. A képletben szereplő $j = 0$ esetben a forrásnyelvi szónak nincs megfeleltetése a célnyelvi oldalon, melynek jelölésére az úgynevezett NULL token szolgál. Az ilyen fordítási modellel működő fordítórendszert nevezzük szóalapú fordítónak, melynek futási ideje a fordítandó mondat hosszával lineárisan arányos.

A szóalapú fordítási modell egyik megszorítása azonban az, hogy a forrás és célnyelvi szavak között egy-sok relációt feltételez. A természetes nyelvek többségére azonban nem igaz ez a feltevés. Vegyük például az angol-magyar nyelvpárt, ahol egy több szóból álló angol kifejezés (*„in my house”*) fordítása magyarul nagy valószínűséggel egy szó lesz (*„házamban”*). A megszorításnak köszönhetően ebben az esetben az angol kifejezés nagy része a NULL tokennel lesz párosítva, emiatt nem kerül be a fordítási modellbe, tehát a modell építése során információvesztés történik. A szóalapú rendszer másik hiányossága, hogy annak ellenére, hogy a szó környezete jelentősen befolyásolhatja a fordítás minőségét, semmilyen kontextuális információt nem használ fel egy szó fordítása során. Számos olyan eset létezik, amikor egy kifejezés helyes fordítása teljesen eltér a kifejezés szavankénti fordításától. Ezek az úgynevezett idiomatikus vagy nem kompozicionális szerkezetek, mint például a magyar *Hol volt, hol nem volt...* az angol *Once upon a time* kifejezésnek felel meg, vagy a *majd kiugrik a bőréből* szókapcsolat, amit *to be over the moon*-nak fordítunk.

DOI:10.15774/PPKE.ITK.2015.004

#### 2.2.4.2 *A kifejezésalapú gépi fordítás*

A kifejezésalapú vagy frázisalapú fordítási modell [7], [8] lényege, hogy a rendszer által használt fordítási egység nem maga a szó, hanem különböző hosszúságú kifejezések. Ez azért előnyös, mert így a modell a szó környezetéből származó információkat is felhasználja a fordítás során. Ezáltal a fordítás menete a következők szerint alakul: a forrásnyelvi mondatot $(S)$ $I$ darab $S = \bar{s}_i^I$ szegmensre daraboljuk, majd ezeket a szegmenseket, mint önálló $\bar{s}_i$ egységeket lefordítjuk egy $\bar{t}_j$ célnyelvi frázisra. Végül a célnyelvi szegmenseket a megfelelő sorrendbe rendezzük. A kifejezésalapú fordítási modell formálisan a (6) egyenlettel írható le,

$$P(S|T) = P\left(\bar{s}_1^k \middle| \bar{t}_1^l\right) = \prod_{i=1}^{k} \phi(\bar{s}_i|\bar{t}_i) \tag{6}$$

ahol $\phi(\bar{s}_i|\bar{t}_i)$ az $i$. kifejezéspárhoz tartozó valószínűség.

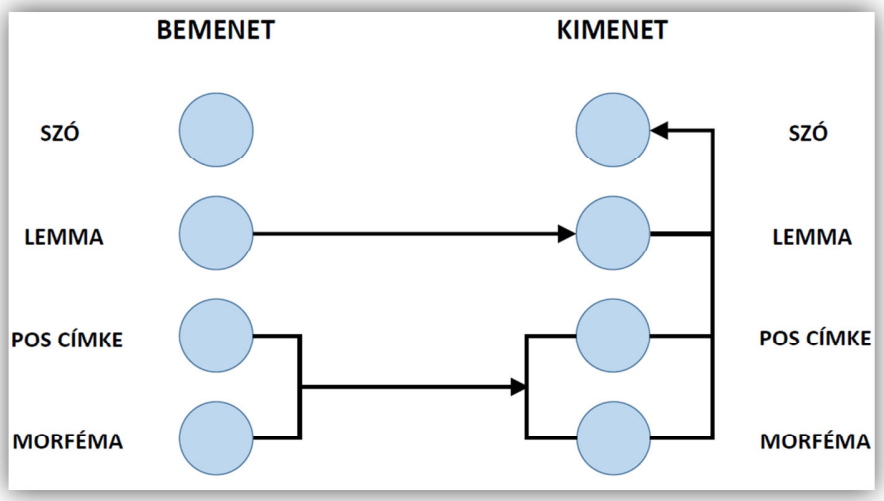
Annak ellenére, hogy ez a modell több információt tartalmaz, tanításának és alkalmazásának komplexitása hasonló a szóalapú modellekhez. A kifejezésalapú fordítás előnye a szóalapú rendszerrel szemben, hogy olyan eseteket is képes kezelni, amikor egy szónak több szóból álló fordítása van, vagy amikor több szó fordítása határozza meg egy célnyelvi szó fordítását. Továbbá a szavak helyett szócsoportok fordítása képes feloldani a fordítás során felmerülő többértelműségeket.

#### 2.2.4.3 *Faktoros gépi fordítás*

A frázisalapú modellek esetén a szavak reprezentálása hiányos, mivel a szóalakokat veszi figyelembe. Ebből kifolyólag a rendszer ugyanannak a szónak a különböző toldalékolt alakjait egymástól független tokenekként kezeli. Így például a *vártam* szó ismerete semmilyen többletinformációval nem segíti a *várok* szó fordítását, annak ellenére, hogy ugyanaz a tövük és számban-személyben is megegyeznek. Viszont a *vár* szó helyes fordításában sokat segítene, ha ismernénk annak az adott kontextusban érvényes szófaját.

A faktoralapú fordítási modell [21] a kifejezésalapú fordítás módszerének egy kiterjesztése, mely szószintű nyelvi és lexikális jellemzőket (mint például szófaji címkék, szótő stb.) integrál a fordítási folyamatba. A faktoros modell célja, hogy csökkentse az adathiányból származó nehézségeket oly módon, hogy külön kezeli a szavak lemmájának és egyéb morfológiai jellemzőinek a fordítását, majd a célnyelvi oldalon történő szóalakgenerálással állítja elő a felszíni alakot a jellemzővektor alapján. Az 5. ábra a faktoros fordítás modelljét ábrázolja, ahol először a mondatban szereplő szavak lemmájának fordítása történik. Ezzel párhuzamosan a morfoszintaktikai címkéket egymástól függetlenül szintén megfelelteti a célnyelvi reprezentációnak, majd ezek segítségével állítja elő a céloldalon a szóalakot.

DOI:10.15774/PPKE.ITK.2015.004




**5. ábra: Faktoros fordítási modell szemléltetése [21]**

A kísérletek azt mutatják, hogy a faktoros modellek jól működnek morfológiailag hasonló nyelvek esetén [21]. Ezzel ellentétben nehézkesen használhatók akkor, ha a nyelvek grammatikailag távol állnak egymástól, és a célnyelv gazdag morfológiájú. Ebben az esetben a rendszer a szótöveket nagy valószínűséggel helyesen fordítja, viszont a statisztikai alapon működő morfológiai generátor képtelen a célnyelvi oldal szóalakjait helyesen előállítani [22]. Ez az adathiányproblémának köszönhető, mivel a tanítóanyagban nagy valószínűséggel nincs benne minden szó minden lehetséges szóalakja.

#### 2.2.4.4 *A faalapú gépi fordítás*

A gépi fordítás eddig felsorolt típusai a mondatokra úgy tekintettek, mint szavak sorozatára. Emiatt nem vették figyelembe a szavak között fennálló szintaktikai viszonyokat. A közvetlen összetevős elemzési fa a szavak közötti relációk egy lehetséges reprezentációs formája. A faalapú fordítórendszer (TbSMT – Tree-Based Statistical Machine Translation) [23]–[25] a szintaktikai elemzésből származó többletinformáció segítségével javít a fordítás minőségén. A TbSMT rendszer sajátossága, hogy az eddig szóalapú komponensekből álló modelleket környezetfüggetlen generáló szabályokkal egészíti ki, ahol a nemterminális szimbólumok a közvetlen összetevős elemzés nemterminális szimbólumainak felelnek meg. A rendszer ezeket a szabályokat automatikusan, egy szintaktikailag elemzett párhuzamos korpuszból tanulja meg. Munkám során nem alkalmaztam ezt a módszert, mivel a magyar nyelvre nincs megfelelően jó minőségű szintaktikai elemzőrendszer, amely biztosítani tudná a pontos elemzést a további feldolgozáshoz.

DOI:10.15774/PPKE.ITK.2015.004

## 2.3 Kiértékelés

A gépi fordítás kiértékelése nagy kihívást jelent, mivel egy mondatnak több valid fordítása lehet, melyek karakter- vagy szószinten nem összehasonlíthatók. Ez abból adódik, hogy ugyanaz a tartalom akár többféle módon is kifejezhető rokon értelmű kifejezések, illetve a szórendi különbségek segítségével. Például *„A vádlott maga alatt vágta a fát, amikor..."* és *„A gyanúsított saját érdekei ellen cselekedett, amikor...."* mondatok ugyanannak az angol mondatnak *(„The suspect acted against his own interests when ..."*) a helyes fordításai, mégis egy szóalapú vagy egy karakteralapú automatikus kiértékelő számára ezek összehasonlíthatatlan fordítások.

Az SMT rendszer fordításának kiértékelésére kézenfekvő megoldás emberi kiértékelők alkalmazása. Megfelelően képzett és nyelvi tudással rendelkező emberi erőforrás segítségével rendkívül pontos kiértékelés érhető el. Koehn és Monz [26] a fordítórendszer kimenetének pontosságát és gördülékenységét több bírálóval pontoztatta 1-5-ig terjedő skála alapján. Az így kapott eredményeket átlagolták, és ily módon határozták meg a fordítás minőségét. Az emberi kiértékelés hátránya azonban, hogy nagyon lassú, költséges és munkaigényes folyamat.

A rendszerek gyors és olcsó elemzéséhez tehát szükség van automatikus kiértékelő módszerekre. Az SMT rendszer automatikus értékelésének alapvető módszere a lefordított mondatnak egy referenciamondathoz való hasonlítása különböző jellemzők mentén. Napjaink legnépszerűbb kiértékelő módszere a BLEU (BiLingual Evaluation Understudy) [27], mely megoldást kínál a szavak sorrendjéből adódó probléma kezelésére is. A módszer hasonlít a PER algoritmushoz (Position independent word Error Rate) [28], ám az utóbbival ellentétben figyelembe veszi a több szóból álló frázisok referenciafordítással való egyezéseit 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 számításának módja formálisan a következőképpen írható le:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} \omega_n \log p_n\right) \tag{7}$$

ahol BP (brevity penalty) a rendszer fedését hivatott értékelni, oly módon, hogy lepontozza a referenciafordításnál sokkal rövidebb fordításokat. $p_n$ a módosított pontosság, $w_n$ az n-gramok súlya (tipikusan 1 értéket vesz fel).

A rendszereket a szóalapú BLEU (továbbiakban w-BLEU) metrika mellett morfémaalapú BLEU-vel (továbbiakban mm-BLEU) is kiértékeltem annak érdekében, hogy minősíteni tudjam a

DOI:10.15774/PPKE.ITK.2015.004

morféma szintű algoritmusaimat. Az mm-BLEU számítása a szóalapú rendszerek esetében a fordítás utólagos morfológiai elemzésével történt. Az mm-BLEU érték számítása során a referenciafordításnak és az SMT kimenetén megjelenő fordításnak megfelelő morfémasorozat összehasonlítása történik. Megjegyzendő, hogy az mm-BLEU különbözik az m-BLEU metrikától [29], mely egy felügyelet nélküli szegmentáló által generált pszeudo-morfémákra van számolva. Az mm-BLEU érték a rendszernek a fordítás során történő helyes morféma-előállítási képességének mérésére szolgál.

A BLEU metrika előnyei ellenére több publikáció ([30]–[32]) is figyelmeztet arra, hogy számos esetben az algoritmus nem korrelál az emberi kiértékeléssel. Például ha két fordítási hipotézis csak a kifejezések sorrendjében tér el, ugyanazt a BLEU pontot kapja. Mivel egy mondat szavainak lehetséges permutációja a mondat hosszával faktoriálisan növekszik, egy hosszabb mondat esetén számtalan fordítási javaslat előállhat, ami hasonló BLEU értéket kap, viszont csak kevés olyan van köztük, ami az emberi kiértékelés számára is elfogadható. A BLEU módszer alulpontozza a szinonimák és a parafrázisok – amikor a referenciamondatban szereplő szó helyett nem egy szót, hanem annak körülírását adja a rendszer fordításként – használatát. Végül a BLEU módszer nem képes a fedés hatékony mérésére. Az erre a célra integrált brevity penalty változó csak nagyvonalakban ad információt a vizsgált mondat fedéséről. Ezen hibák miatt a BLEU pontozás nem ad pontos képet a rendszer által generált fordításról, mivel nem feltétlenül korrelál az emberi kiértékeléssel. Callison-Burch et al. [31] szerint egy fordítórendszer pontos minősítése során mindenképpen szükséges az automatikus módszerek mellett az emberi kiértékelés elvégzése is.

DOI:10.15774/PPKE.ITK.2015.004

# II. A statisztikai gépi fordítórendszer minőségének javítása

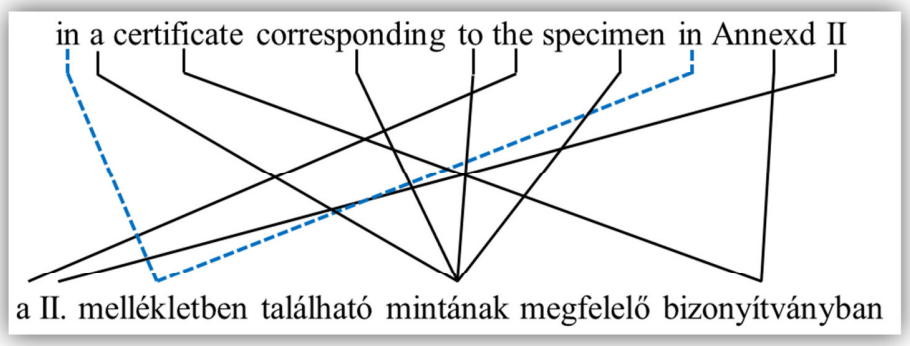
Ebben a fejezetben bemutatom az angol és a magyar nyelvek közti legfontosabb grammatikai különbségeket, illetve az általam kidolgozott megoldási lehetőségeket, melyek elősegítik egy jobb minőségű fordítórendszer létrehozását.

A magyar az agglutináló (ragozó) nyelvek családjába tartozik, ami lehetővé teszi a toldalékok halmozását. Szintén jellemző a többféle alakváltozat mind a szótövek, mind a toldalékok terén. Nyelvünk gazdag esetrendszerrel rendelkezik, megkülönbözteti a határozatlan („alanyi") és a határozott („tárgyas") ragozást, a főnévi igenév pedig ragozható (*látnom, látnod, látnia* stb.). Az angol elsősorban izoláló nyelv, melyben a mondatokat izolált szótövek alkotják, a nyelvtani viszonyokat pedig a funkciószavak és a mondat szavainak sorrendje fejezi ki. Ugyanakkor számos példáját hordozza a flexiónak (a tő megváltoztatásával járó ragozásnak), főleg a rendhagyó esetekben. Ezek alapján a nyelv flektálónak is tekinthető, de lassan tart az izoláló felé; hiszen például a mai angol már nem tartalmaz felszínen megjelenő esetragokat, mivel azok az idők során lekoptak [33]. A magyar a páros szerveket (pl. *kéz, láb, szem, fül*) és a több birtokos egy-egy birtokát is egyes számban mondja (pl. *élik az életüket,* nem pedig *életeiket*), a számneves névszói csoportok pedig alakilag egyes számúak, és így is egyeztetjük őket az igével. Nyelvünkben hiányzik az indoeurópai nyelvekre jellemző birtoklást kifejező ige (*én birtoklok valamit* helyett *nekem van valamim*). Számottevő különbség mutatkozik az igeidők számában; az angol 12-féle igeidejét a magyar hárommal (jelen, múlt, jövő) képes kifejezni. Fontos eltérés a passzív szerkezet alkalmazása, ami a magyarban létezik ugyan (-tatik, -tetik toldalék), ám ez ilyen formában a mindennapos nyelvhasználatból már kikopott. A passzív szerkezetet a magyarban különböző struktúrákkal helyettesítjük.

Ahogy azt már a 2.2. fejezetben kifejtettem, az SMT rendszer minőségét a dekóder által használt modellek milyensége határozza meg. A fordítási modell építéséhez nélkülözhetetlen a szószinten összepárosított kétnyelvű korpusz, melyet a szóösszekötő rendszer segítségével hozunk létre. Az általánosan használt statisztikai szóösszekötőnek azonban a nyelvek közt fennálló különbségek miatt nagyon nehéz dolga van. Egyrészt probléma az agglutináló és az izoláló tulajdonság különbségéből adódó eltérés, miszerint egy angol funkciószó megfelelője általában nem egy

DOI:10.15774/PPKE.ITK.2015.004

magyar szó, hanem egy toldalék. Ezáltal a szóösszekötő nem tudja azt hova kapcsolni, illetve előfordulhat más természetű hiba is, amikor az összes azonos angol funkciószót a magyar mondat egyetlen szavához kapcsolja. Gyakori jelenség továbbá, hogy a magyar szóhoz nem az angol szótövet köti, hanem a főnévi frázis funkciószavát kapcsolja. Ezeket a hibatípusokat a 6. ábra szemlélteti egy példamondat segítségével.




**6. ábra: Példa a szóösszekötő helytelen működésére**

Másrészt gondot okoz, hogy némelyik magyar toldaléknak (például tárgyrag) nincs megfeleltethető angol párja, emiatt ez szószinten nem tanulható meg. Az eddig felsorolt szerkezeti különbségek vezetnek a két nyelv mondataira jellemző **átlagos szószám- és morfémaszám-különbségekhez**. További nehézség a két nyelv közti jelentős **szórendi eltérés.** A magyar mondat szórendje nem kötött, ugyanis azt elsősorban nem szintaktikai szabályok, hanem pragmatikai tényezők határozzák meg. Ugyanakkor semleges mondatok esetében rendkívül összetett szintaktikai megszorítások is fellépnek. Így például az alany-állítmány-tárgy sorrend mellett gyakran előfordul a tárgy-alany-állítmány vagy az állítmány-alany-tárgy szórend is attól függően, hogy a mondat melyik részét szeretnénk hangsúlyozni. A kiemelni kívánt információ rögtön a ragozott ige elé, az ún. fókuszpozícióba helyezendő.

Az angollal ellentétben a magyar mondatban a szórendtől függetlenül a ragozás egyértelműen utal az elemek mondatbeli szerepére. Az angol szórendje azonban sokkal kötöttebb a magyarénál, jellemzően alany-állítmány-tárgy alakú. A hagyományos SMT rendszer amellett, hogy csak kis távolságú, lokális átrendezéseket képes kezelni, a legjobb minőséget monoton fordítás esetén képes elérni. Mivel nem rendelkezik nyelvi tudással, így a szabad szórendű nyelvek nehezen kezelhetők ezzel a módszerrel.

DOI:10.15774/PPKE.ITK.2015.004

Birch [34] szerint a gépi fordítás minőségét a forrás- és célnyelvek megválasztásának függvényében három tulajdonság befolyásolja: a célnyelv nyelvtani összetettsége, a szórendi eltérés mértéke és a két nyelv közötti történeti kapcsolat. Az angol-magyar fordítás esetén a Birch-féle tulajdonságok megnehezítik a jó minőségű fordítás előállítását, hiszen a két nyelv összehasonlítása során már bemutattam, hogy a nyelvpár esetében nemcsak a szórendi eltérés számottevő, hanem a magyar nyelv agglutináló tulajdonságából származó célnyelvi összetettség problémája is fennáll.

Munkám során célom az automatikus gépi fordítás minőségének javítására irányuló módszerek kidolgozása volt, melyeket az angol-magyar nyelvpár esetén teszteltem, és ezeken mutatom be. Ezt egyrészt a szórendi eltérésekből fakadó nehézségek megoldására a forrásnyelvi mondatok szósorrendjének kézzel írt szabályokkal történő megváltoztatásával oldottam meg. Ezt a folyamatot a II. fejezetben mutatom be. Másrészt a magyar mint célnyelv összetettségével a 4. fejezetben foglalkozom, ahol a két nyelv szószáma közti eltérés és a morfológiai különbségek kiküszöbölése volt a célom.

DOI:10.15774/PPKE.ITK.2015.004

# 3 Szórendi különbségek csökkentése szintaxismotivált átrendezési szabályok alkalmazásával

A II. fejezetben elemzett problémák tükrében célom egy olyan hibrid fordítórendszer létrehozása volt, amely a statisztikai gépi fordítás előnyeinek kihasználása mellett igyekszik csökkenteni a szórendi különbségekből és a morfológiai sokszínűségből adódó nehézségeket. A nyelvtanilag távoli nyelvek fordítása esetén szórendi átrendezés segítségével bíztató eredményeket értek el ([35]–[38]). Munkám során a szóösszekötő rendszer javítása és a monoton fordítás elérése érdekében szintaxisorientált szórendi átrendezéseket végeztem általam megfogalmazott szabályok segítségével a forrásnyelvi mondatokon. Emellett a forrásnyelvi mondat szavain összekötéseket alkalmaztam, mellyel közelítettem a célnyelvi mondat morfémáinak sorrendjét.

## 3.1 A forrásnyelvi mondatok szórendi átrendezésének elméleti háttere és megvalósítása

A statisztikai gépi fordítás során a nyelvpárok között fennálló szórendi különbségek nagymértékben megnehezítik a fordítórendszer működését, mind a rendszer betanítása, mind a használata során. Ahogy azt az 2.2. fejezetben kifejtettem, az SMT rendszer tanításának egyik kulcsfontosságú lépése a párhuzamosan lefordított mondatok szavainak összepárosítása. Mivel a statisztikai módszer alapját a kétnyelvű párhuzamos mondatokban szereplő szavak megfeleltetésére épített valószínűségek képezik, ezért a szóösszerendelés minősége alapjaiban meghatározza a végső fordítás minőségét is. A Moses keretrendszerben lévő statisztikai alapú szóösszekötő rendszer számára nehéz feladat összepárosítani a nagy távolságban levő szópárokat, ezenkícül a szórendbeli különbségeket büntető modell miatt a dekóder nem képes nagy távolságú átrendezésekre a fordítás során. A forrásnyelvi mondatok szósorrendjének a célnyelvi mondathoz történő igazítása nagyban megkönnyíti a szóösszekötő munkáját, valamint a dekódolás során fellépő nagy távolságú átrendezési problémákat lokális átrendezési feladattá egyszerűsíti. Következésképpen, a fordítórendszer tanítása során a nyelvpárt az alaprendszerénél jobban reprezentáló statisztikák jönnek létre; ez pedig javítja a végső fordítás minőségét.

Munkám során a fordítási folyamatot kiegészítettem egy előfeldolgozási lépéssel, melyben az általam megfogalmazott, kézzel írt, szintaxismotivált szabályok alkalmazásával végeztem átrendezést a forrásnyelvi szövegeken. A feladat alapja – melyet korántsem egyszerű megvalósítani – az átrendezendő szerkezetek felismerése és azonosítása. Ehhez első lépésként a forrásnyelvi

DOI:10.15774/PPKE.ITK.2015.004

angol mondatokra szófaji egyértelműsítést, közvetlen összetevős elemzést és a függőségi relációk meghatározását is elvégeztem a Stanford Parser [39], [40] segítségével. Bizonyos függőségi relációkról elmondható, hogy a nekik megfelelő magyar szerkezet az angoltól eltérő, vagyis más szórendet követ. Az ilyen egyértelműen definiálható relációkra (pl. elöljárók vs. esetragok/névutók) fogalmaztam meg kézzel írt átrendezési szabályokat. Az alkalmazott szabályok csak azokat a szórendi eltéréseket szüntetik meg, amelyek a két nyelv között szabályszerűen fellépnek; nem volt célom a magyar „szabad szórendjéből" adódó eltérések eltüntetése. Nagyon egyszerű példa az angol *in my house* kifejezés, mely az átrendezés és összevonások után „house_my_in" formára alakult, amely megfelel a magyar *házamban* alaknak. Az ilyen rövid szókapcsolatok során a szabályok alkalmazása nem jelent nagy problémát, azonban hosszabb mondatok esetén az egymáshoz kapcsolódó részek egészen távol is eshetnek, illetve több függőségi kapcsolatban is érintettek lehetnek.

Továbbá olyan morfológiai elemek is beszúrásra kerültek, melyek az eredeti angol mondatban nincsenek explicit módon jelölve, a magyar megfeleltetés miatt azonban szükségesek (pl. tárgyrag). Természetesen figyelembe vettem az összetartozó szerkezeti egységeket; ezeket az átrendezés során is egységként kezelve, egyben helyeztem át. Ezen kívül a magyar nyelv szabályait betartva az esetleges többszörösen átrendezett szerkezetek esetében a toldalékok megfelelő pozícióját is szem előtt tartottam, ami így megfelel a képző-jel-rag sorrendnek.

## 3.2 Felhasznált eszközök és erőforrások

### 3.2.1 A tanító- és a teszt halmazok felépítése

Az elérhető angol-magyar párhuzamos korpuszok többsége nem alkalmas egy általános SMT rendszer betanítására, mivel csupán egy-két terület terminológiáját tartalmazzák. Munkám során az elérhető legnagyobb és témáját tekintve legáltalánosabb párhuzamos korpuszt, a BME Média Oktató és Kutató Központ és az MTA Nyelvtudományi Intézete által készített Hunglish korpuszt [41], [42] használtam. Ez a korpusz egyidejűleg több területről tartalmaz szövegeket: szépirodalom, magazin, jog, filmfeliratok. A rendszer betanítása során nehézséget jelent azonban, hogy az egyes részek minősége meglehetősen változó. Az így létrejött korpusz mérete 1 202 205 párhuzamos mondatpár. A 7. ábra szemlélteti a korpusz előfeldolgozásának egyes lépéseit, melyek a következők:

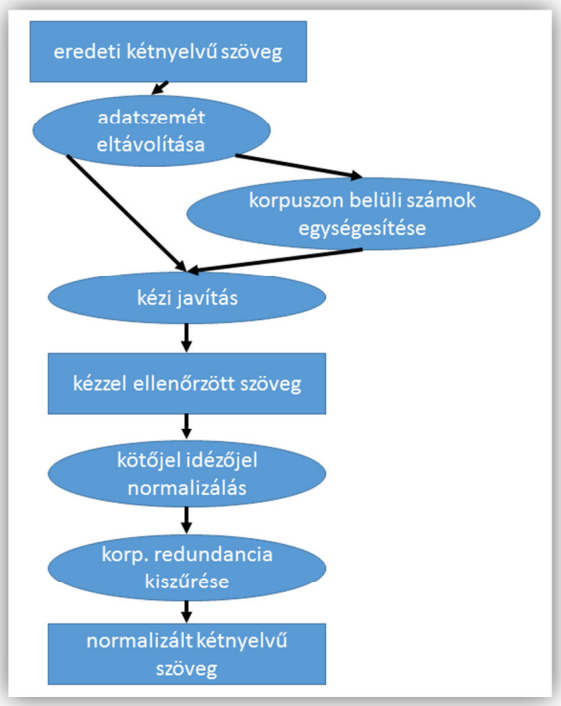
- Első lépésként az eredeti kétnyelvű szövegből eltávolítottam azokat a sorokat, melyek nem tartalmaztak betűket vagy számokat.
- A korpuszban előfordult, hogy az angol és magyar oldalon nem ugyanazok a számok voltak egymás mellett, ezért először automatikus módszerrel javítottam ki ezeket a hibákat, majd

DOI:10.15774/PPKE.ITK.2015.004

kézzel ellenőriztem a változásokat.

- Az így létrejött szövegben a magyar oldalon a kötőjelek, és az angol oldalon ennek megfelelő idézőjelek normalizálását végeztem oly módon, hogy a magyar oldalt az angol szerkezettel azonos formátumúra hoztam. Ezt úgy értem el, hogy mindkét esetben csak egy kötőjelet hagytam az idézet és a narratív szöveg között. Például az *– Ön nem figyelt ide – mondta.* és a *"You were not listening" he said.* mondatpárt az alábbi formára alakítottam át: *Ön nem figyelt ide – mondta.* és *You were not listening – he said.*
- Az elvégzett simítások után a korpuszban létrejött redundáns mondatokat kiszűrtem, így megkaptam a korpusz végső formátumát, melyet munkám során alkalmaztam.

A korpuszból véletlenszerűen kiválasztott 1000 mondatot teszthalmaznak, 1000 mondatot paraméteroptimalizációs halmaznak, a fennmaradó részt pedig tanítóhalmaznak használtam.




**7. ábra: A korpusz előfeldolgozásának folyamata**

### 3.2.2 Morfoszintaktikai elemzőrendszerek

Az előfeldolgozás első lépéséhez szükség volt egy robusztus morfoszintaktikai egyértelműsítőre, közvetlen összetevős és függőségi elemzőre az angol mondatok átrendezéséhez, illetve a morfémaalapú fordítás miatt egy, a magyar nyelv elemzésére alkalmas eszközre.

DOI:10.15774/PPKE.ITK.2015.004

Magyar szövegek elemzésére a PurePos2 [43] nevű rendszert használtam a szavak szótövének megtalálására, valamint a morfoszintaktikai egyértelműsítésre. A PurePos2 rendszer a HUMOR [44], [45] kódrendszert használja. Annak érdekében, hogy a két címkekészletet összehangoljam, és biztosítsam, hogy minden címkének legyen párja a másik nyelven az összetett címkéket felbontottam atomi egységekre. A morfémaalapú fordítórendszer a szóalakok helyett azok *szótő#morfoszintaktikai címkék* alakját használja. Ezeket a rekordokat a fordítás után vissza kell alakítani szóalakokká, amihez a HUMOR [44] morfológiai generátor modulját alkalmaztam.

Az angol nyelvű szövegeken a teljes morfoszintaktikai egyértelműsítés mellett közvetlen összetevős szintaktikai elemzést, valamint függőségi elemzést végeztem a Stanford Parsert [39] rendszer segítségével, amely az egy szabadon hozzáférhető angol szintaktikai elemző. Az elemző hozzáférhető változatát a Wall Street Journal anyagait tartalmazó Penn Treebank [46] egy töredékén tanították. Az elemzés minősége sokkal fontosabb, mint a gyorsasága, hiszen az elemzés és a szóösszekötés offline történik, a tanítás során csak egyszer (illetve a fordítandó szöveget kell még elemezni), ezért az elemző lexikalizált változatát használtam. A lexikalizált környezetfüggetlen nyelvtani (CFG – context free grammar) szabályokat használatával az elemző magasabb pontossággal működik, mint a nem lexikalizált változat. Hátránya azonban, hogy valamivel lassabb és jelentősen nagyobb a működéséhez szükséges modellek mérete [47]. Ám még ezzel a módszerrel is nagyon sok olyan eset fordult elő, melyeket az elemző nem tudott megfelelően kezelni.

A Stanford Parser sorba kapcsolt elemekből álló rendszer, amelynek első szófaji egyértelműsítő komponense önmagában meglehetősen sok hibát generál, amelyet azután minden további komponens továbbiakkal tetéz. A korábbi komponensek hibáit a láncban később következő moduloknak nem áll módjukban javítani. Ezek a szófajilag rosszul elemzett szavak és a rosszul megállapított függőségi relációk a saját rendszeremben is kritikus problémát jelentenek, hiszen ezen relációk alapján történik az átrendezési szabályok alkalmazása. Ez azt jelenti, hogy ha egy eleve rosszul elemzett szöveget rendezünk át, akkor az így kapott hibás átrendezés inkább ront, mint javít a fordítás minőségén.

Az első ilyen hibaforrás a helytelen szófaji címkék használata az elemző számára ismeretlen szavak, vagy az ismeretlen kontextusban megjelenő ismert szavak esetében. A legtipikusabb hiba a főnevek, melléknevek és igék összetévesztése, amely szinte minden esetben végzetes következményekkel jár az elemzés egészére. Mivel mind a szintaktikai, mind a függőségi elemzés ilyen félrevezető információkon alapul, a hiba továbbterjed a rendszerben. Erre látható példa az 1. táblázatban, ahol a *sound* vagy a *cash* szavak helytelen elemzése (ige helyett főnév) miatt elvesztik a mondatok az állítmányukat. Ha egy szó rossz szófaji címkét kap a fordítandó mondat-

DOI:10.15774/PPKE.ITK.2015.004

ban (például sound [NN] címkét kapott a [VB] helyett), akkor az erre épülő szószintű átrendező rendszer nem a mondat fordítását támogató átrendezéseket fogja elvégezni. Például a *100 million sound good to me* mondaton a rendszer a rossz elemzés miatt a helyes *100 millió jól hangzik számomra* szórend helyett a *számomra lévő 100 millió jól hangzás* szerkezetnek megfelelő átrendezéseket végezte el (ennek megfelelően helytelenül került az átrendezett angol mondatba a *lévő*nek megfelelő *xxx/xxx* karaktersorozat is). Emellett előfordulhat az is, hogy egy szónak többféle szófajú fordítása is szerepel a fordítási modellben, melyek közül a szövegkörnyezettől függően több is lehet helyes. Ezért ha az aktuálisan fordítandó mondatban rossz címke szerepel, akkor az annak megfelelő hibás fordítás kerül az eredménybe.

| | |
|---|---|
| Elemzett angol mondat | *100/CD million/CD sound/NN good/JJ to/TO me/PRP ./.* |
| Átrendezett angol mondat | *me/PRP_to/TO xxx/xxx 100/CD million/CD good/JJ sound/NN./.* |
| Elemzett angol mondat | *For/IN airline/NN personnel/NNS ,/, we/PRP cash/NN personal/JJ checks/VBZ up/RP to/TO $/$ 100/CD ./.* |
| Átrendezett angol mondat | *airline/NN personnel/NNS_for/IN ,/, personal/JJ cash/NN up/RP_checks/VBZ_we/PRP 100/CD_$/$_to/TO ./.* |

**1. táblázat: Példa a hibás elemzésre**

## 3.3 A létrehozott átrendezési szabályok

Két fő csoportba sorolható átrendezési szabályokat alkalmaztam, melyeket a következő fejezetekben ismertetek.

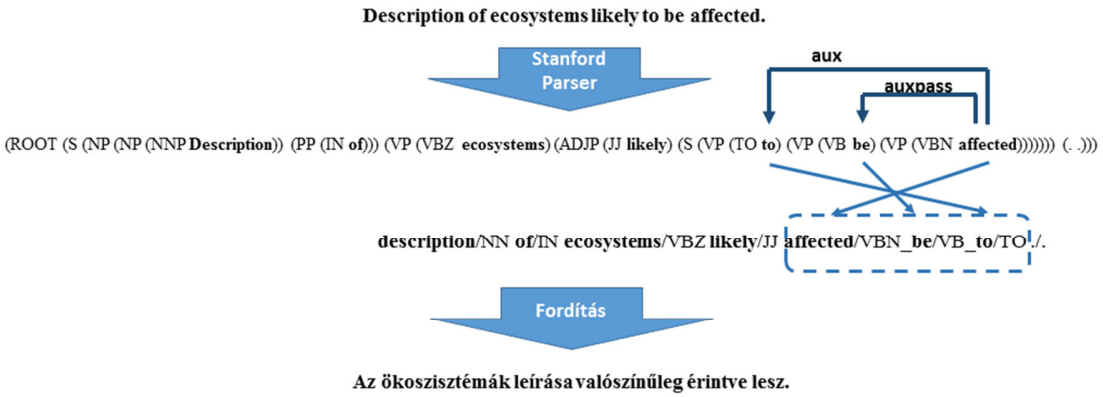
### 3.3.1 Szórendi átrendezést és morféma-összevonást/felbontást tartalmazó szabályok

Ezek a szabályok a függőségi relációk meghatározása után a közvetlen összetevős szerkezetet is figyelembe véve alakítják át a szavak sorrendjét, ezzel egy időben szükség esetén össze is vonják azokat. Az angol mondatokban sok olyan információ nincs jelen, ami a magyar oldalon toldalékként szerepel. Ezeket a Stanford Parser [39] által meghatározott függőségi relációk alapján határoztam meg, és a közvetlen összetevős elemzés segítségével rendeztem át. Előfordulnak továbbá olyan esetek is, amikor az angol különálló szóként jelöli a magyar toldaléknak megfelelő morfémákat, amiket így rácsatolhatunk a megfelelő szóra. Ezek az összevonások kisebb szerkezetek átrendezését jelentik. Olyan szabályok kerülnek végrehajtásra, mint a passzív, a segédigés, az elöljárós és a birtokos szerkezetek átalakítása, az angolban hátravetett módosítók előremozgatása és még néhány, ritkábban előforduló szabály. Fontos az átrendezési szabályok végrehajtásának sor-

DOI:10.15774/PPKE.ITK.2015.004

rendje is, mivel bizonyos függőségi relációkat több szabály esetén is figyelembe veszünk. A szabályok ismertetési sorrendje megegyezik alkalmazásuk sorrendjével.

**Határozói igenév:** A magyar nyelvben nincs konkrét megfelelője az angol *'to be VBN'* passzív formátumnak. Ahhoz, hogy értelemben a lehető legközelebbi fordítást kapjuk átalakítottam *VBN_to_be* sorrendre, mivel ez az alak a magyar *'-va, -ve'* határozóragnak felel meg. A keresett frázist a függőségi elemzés AUX (auxiliary – segédige) és AUXPASS (auxiliary passive – passzív segédige) függőségi relációk alapján határoztam meg. A változtatásokat a 8. ábra szemlélteti.



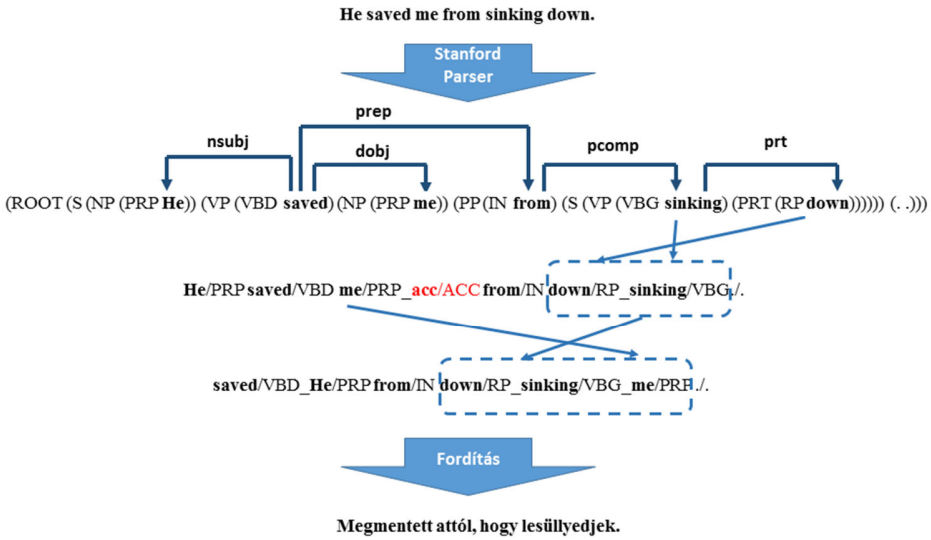

8. ábra: A *„to be VBN"* frázis átrendezése

**Jelzős szerkezetben szereplő 'of':** Az angol nyelvben előfordulnak a *'kind of sth', 'sort of sth', 'lot(s) of sth'* és *'type of sth'* kifejezések, ahol az *'of'* funkciószó nem birtokos szerkezetet jelöl, hanem valaminek a típusát, fajtáját. Ezekben az esetekben az *'of'* helyett az *'xf'* jelölést alkalmaztam, és összekapcsoltam a szerkezetet, hogy megkülönböztethető legyen a birtokos szerkezet *'of'*-jától. Például az elemzett *a/DT different/JJ type/NN of/IN braking/VBG device/NN* mondatból az átrendezés után *a/DT different/JJ type/NN_xf/IN braking/VBG device/NN* lett.

**Igekötők**: Az angol nyelv PRT (phrasal verb particle – igekötő) relációi a magyarban igekötőként jelennek meg, ezért ezeket az ige elé helyezem, és a magyar nyelvtan szabályai szerint ahhoz kapcsolom őket. Például a *you/PRP are/VBP going/VBG to/TO shoot/VB us/PRP down/RP* az átrendezés után *you/PRP are/VBP going/VBG to/TO us/PRP down/RP_shoot/VB* lett.

**Tárgyrag**: A magyar a mondat tárgyát a *'-t'* tárgyraggal jelöli, ám ennek az angolban nincs megfelelője. Ennek következtében a mondat tárgyát a szintaxist nem ismerő SMT rendszer csak azokban az esetekben tudja kezelni, melyek szerepeltek a tanítóanyagban, önmagától viszont képtelen rá. A rendszer támogatására az angol mondat tárgyát a DOBJ (direct object – a mondat tárgya) függőségi reláció alapján keresem meg, és tárgyraggal (acc/ACC) egészítem ki. Például az *Open/VBD the/DT door/NN* elemzett mondat az átalakítás után *Open/VBD the/DT door/NN_acc/ACC* lett.

DOI:10.15774/PPKE.ITK.2015.004



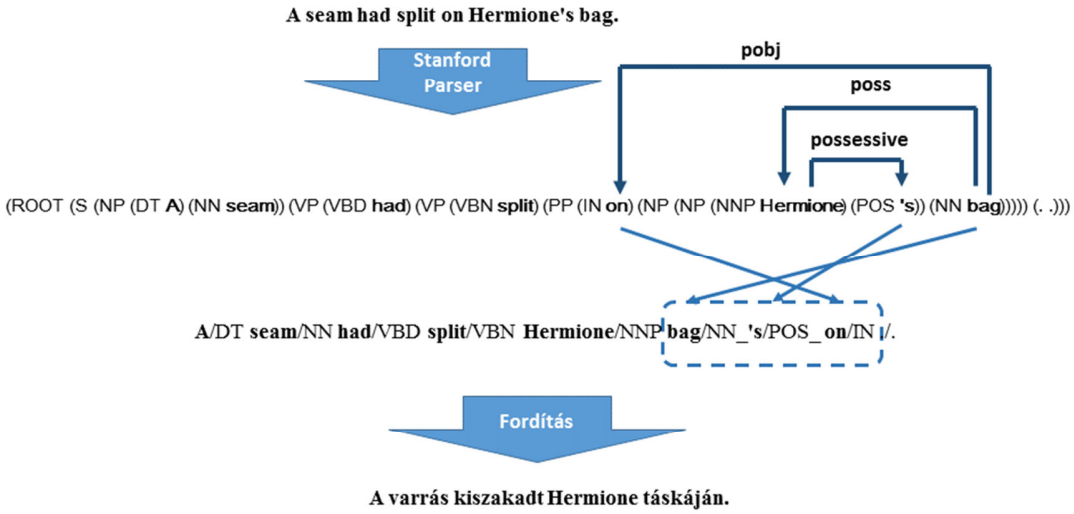

**9. ábra: Az *„attól, hogy”* szerkezet kezelése**

Abban az esetben, ha az eredeti angol mondatot egy *’-ás, -és’* szerkezettel fejeznénk ki, akkor azt a magyarban általában egy alárendelő mellékmondatként fordítjuk. A magyar mondatban a főmondat tárgya rejtett marad. Mivel a mellékmondat igéjének ragozását a tárgyesetű névmás határozza meg, ezért ezt a megfelelő pozícióba mozgatom, és elhagyom az ily módon feleslegessé vált tárgyragot. Az attól, hogy szerkezet a főigétől, valamint a két tagmondatot összekapcsoló prepozícióból következik, amit a fordítórendszernek kell megtanulnia. Az átrendezendő szerkezetet a PCOMP (prepositional complement – elöljárószó kiegészítő) függőségi relációval határoztam meg. A változtatásokat a 9. ábra szemlélteti.

**Részeseset**: Az angolban létezik részestárgy eset, amikor a cselekvés valakinek a részére történik. Ezt az esetet a tárgyraghoz hasonlóan az angolban a mondatbeli pozíción kívül nem jelöli semmi, a magyar viszont a *’-nak, -nek’* raggal fejezi ki. Az elemzett mondat IOBJ (indirect object – indirekt tárgy) relációja alapján állapítottam meg, hogy melyik szóhoz kell a részeseset ragját (dat/DAT) csatolni. Például az *I/PRP give/VBP Peter/NNP a/DT toast/NN* mondat a változtatás után *I/PRP give/VBP Peter/NNP dat/DAT a/DT toast/NN* alakú lett.

**Birtokos személyjel**: Az angol nyelvben megjelenő birtokos névmás a magyar nyelvben nem minden esetben fordítható birtokos névmássá, ilyenkor birtokos személyjelként van a főnévhez csatolva. A függőségi elemzés POSS (possessive modifier – birtokos módosító) relációja alapján rendeztem át a frázis szavainak sorrendjét. Erre példa az *A/DT sad/JJ little/JJ mouse/NN contorted/JJ his/PRP$ mouth/NN* mondatból az átrendezési szabályokat alkalmazva *A/DT sad/JJ little/JJ mouse/NN contorted/JJ mouth/NN his/PRP$* alakot kaptam.

DOI:10.15774/PPKE.ITK.2015.004




**10. ábra: Birtokos személyjel szerkezet kezelése**

Másfelől az angolban *'s*-szel jelölt birtokos szerkezet birtokoshoz van kapcsolva. Ezt úgy rendeztem át (10. ábra), hogy a magyarban ennek megfelelő birtokos személyjel a megfelelő helyre – a birtok után – kerüljön. A frázis szavainak sorrendjét a POSS (possession modifier – birtokhatározó) és POSSESSIVE (possessive modifier – birtokjel) függőségek alapján változtattam meg.

**Számegyeztetés:** Míg az angolban a többes számú főnév (NNS) minden esetben jelölve van, addig a magyarban csak akkor, ha általánosságban beszélünk (pl. *évek*). Ha konkrét esetről van szó, akkor viszont számnév segítségével tudjuk megállapítani a többest (pl. *öt év*). Ahhoz, hogy ezekben az esetekben összhangba hozzam a két nyelvet, szótövesítettem az angol főnevet (a Morpha [48] nevű program segítségével), hogy egyes számú fordítás (NN) jöjjön létre. A számegyezetetést határozott esetben a függőségek NUM (numeric modifier – számhatározó), míg határozatlan (more, many) esetben az AMOD (adjectival modifier – mennyiségjelző) relációja alapján végeztem. Például a *five/CD years/NNS* frázist *five/CD year/NN* formájúra alakítottam, határozatlan esetben pedig a *many/JJ Gods/NNS* szókapcsolatot *many/JJ God/NN* alakúra.

**Az 'of' elöljárószóval kifejezett birtokos szerkezet:** Az egyik legszembetűnőbb különbség az angol és a magyar nyelv szórendjében a birtok és a birtokos sorrendje. Míg az angolban a birtok szerepel elöl, addig a magyarban ez a szerkezet végén található. Két esetet különböztetünk meg, mégpedig az egyszeres és a többszörös birtokos esetét.

DOI:10.15774/PPKE.ITK.2015.004

Egy birtokos esetén megkülönböztetünk rövid és hosszú birtokos szerkezetet. Rövid birtokos (*a birtokos birtoka*) esetén csak az egész frázis elé kerül névelő, míg a hosszú birtokos frázis (*a birtokos**nak a** birtoka*) esetén a birtokoshoz kapcsolódó *-nak, -nek* raggal és a birtok elé kerülő névelővel bővül ki a szerkezet. Mivel mindkét szerkezet nyelvtanilag helyes, munkám során minden esetben a rövid birtokos szerkezetet alkalmaztam. Azért választottam a rövid szerkezetet, mivel így az átrendezési szabályok alkalmazása során nem kerülnek be az angol oldallal párosíthatatlan toldalékok és névelők a szövegbe, melyek aztán megnehezítik a szóösszekötést és a fordítást.

Egy birtokos esetén először megkeresem a birtokot és a birtokost kifejező főnévi frázisokat, melyeket ezután megcserélek. Ezután a megfelelő pozícióba a birtokhoz csatolom az *of* elöljárószót, hogy ez a fordítás során birtokos személyjelként jelenjen meg. Például:

I like to see the sons of the merchants.
like/VBP_I/PRP see/VB_to/TO the/DT merchants/NNS sons/NNS_of/IN_acc/ACC ./.

Hosszabb birtoklási lánc (birtok *of* birtokos$_1$ *of* ... *of* birtokos$_n$) esetén, a névelőt csak az első birtokosnál és a birtoknál hagyom meg, a köztes birtokosok elől pedig kitörlöm. A *-nak, -nek* ragot a második birtokostól kezdődően minden további birtokoshoz csatolom (***THE*** *birtokos$_n$ birtokos$_{n-1}$**_of_NAK*** *... birtokos$_1$**_ of_NAK THE** birtok*). Például:

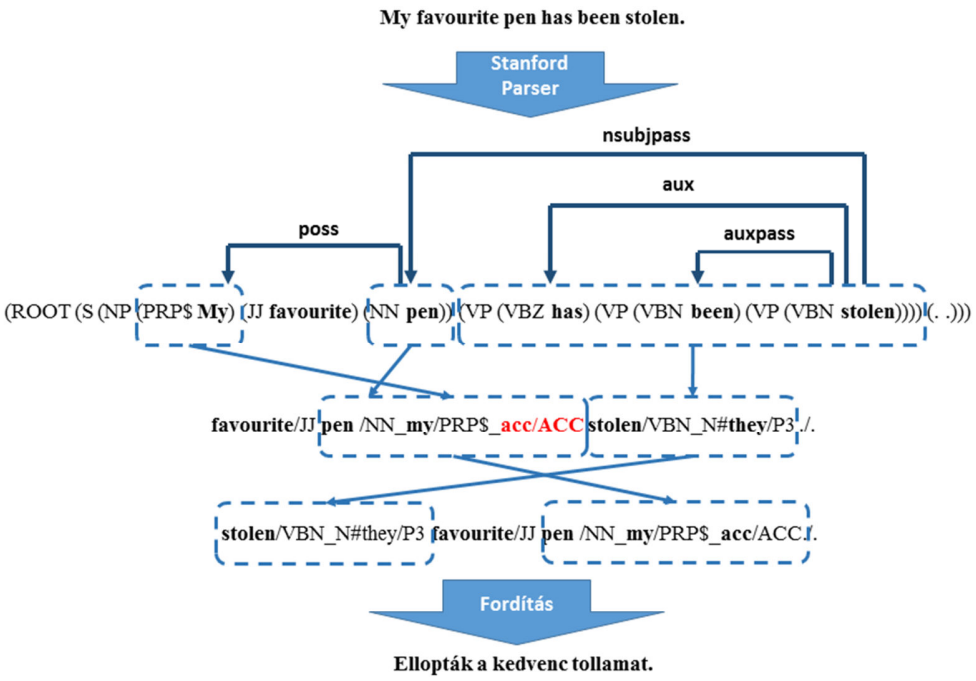
I like to see the black color of the hats of the sons of the merchants.
like/VBP_I/PRP see/VB_to/TO the/DT merchants/NNS sons/NNS_of/IN_nak/NAK the/DT hats/NNS_of/IN_nak/NAK the/DT black/JJ color/NN_of/IN_acc/ACC ./.

**Passzív szerkezet kezelése**: A mai magyar nyelv általában aktív szemléletű, vagyis ha ismerjük az alanyt, akkor általában cselekvő szerkezetet használunk. Emellett azonban létezik az angol passzív szerkezetnek megfeleltethető szenvedő alak is. Ez a szenvedő szerkezetet mára szinte teljesen kikopott nyelvünkből. Azonban jelenleg is vannak olyan nyelvi eszközök, melyekkel az angol passzív szerkezet fordítható. Ilyen például a szenvedőige-képző (*-tatik, -tetik,* például *kéretik*) alkalmazása; a mondat aktív szerkezetűvé alakítása általános alany használatával (például *elvitték a ...*); vagy a létige és határozói igenév együttes alkalmazása (például *A kocka el van vetve.)*. [49]

Munkám során az angol passzív szerkezetet egységesen kezeltem. Célom ugyanis egy helyes fordítás generálása volt, még akkor is, ha az nem egyezik a referencia mondattal. Az átrendezés során az aktív mondat (alany-állítmány-tárgy) szórendjére változtattam a passzív mondatot szórendjét. Ha az angol passzív mondatnak nem ismert az alanya, akkor a mondat fordítása során

DOI:10.15774/PPKE.ITK.2015.004

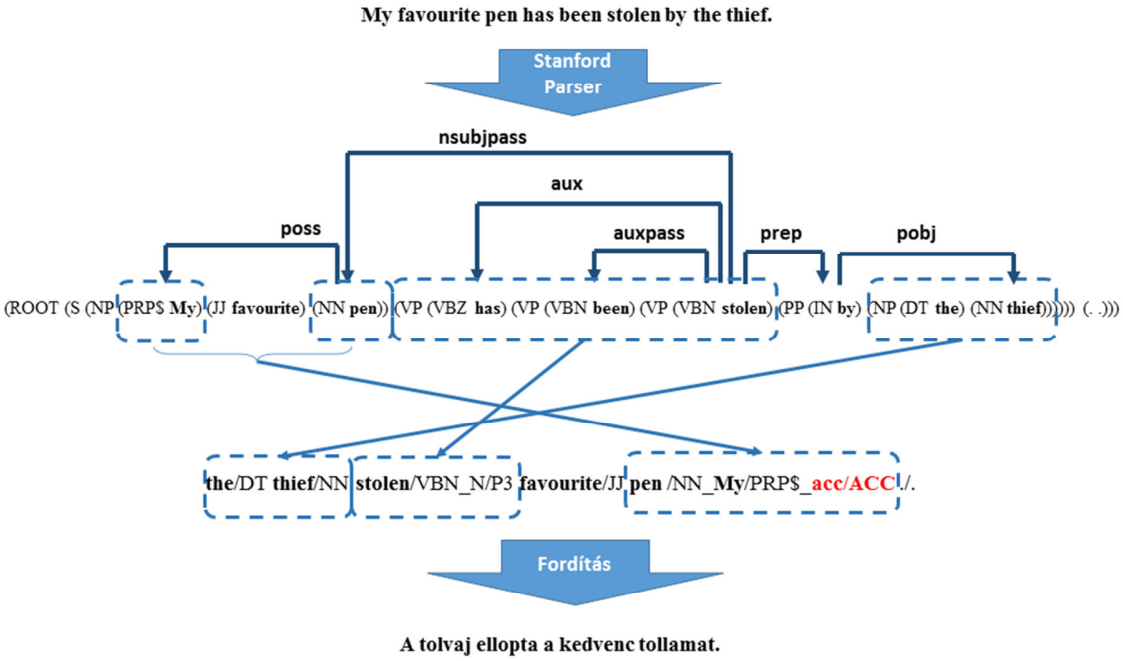
általános alanyú igeragozást alkalmaztam. Az angol segédigékben tárolt igeidőre vonatkozó információt egyeztettem a főigével. A passzív szerkezet alanya az NSUBJPASS (passive nominal subject – a passzív szerkezet alanya) függőségi reláció alapján ismerhető fel. Nemcsak ezt az alanyt, hanem a közvetlen összetevős elemzésből meghatározott egész alanyi csoportot helyeztem át a mondat igéje mögé, továbbá tárgyragot (acc/ACC) is fűztem hozzá. Az átrendezés lépéseit a 11. ábra szemlélteti.




**11. ábra: Passzív szerkezet átrendezése I.**

Ezzel ellentétben ismert alanyt esetén az alanyi csoportot az ige elé helyeztem. A mondat alanyát a *by* elöljárószó és a PREP (prepositional modifier – elöljárószós módosító) és POBJ (object of a preposition – az elöljárószó tárgya) függőségi kapcsolatok alapján határoztam meg. A változtatásokat a 12. ábra mutatja be.

DOI:10.15774/PPKE.ITK.2015.004




**12. ábra: Passzív szerkezet átrendezése II.**

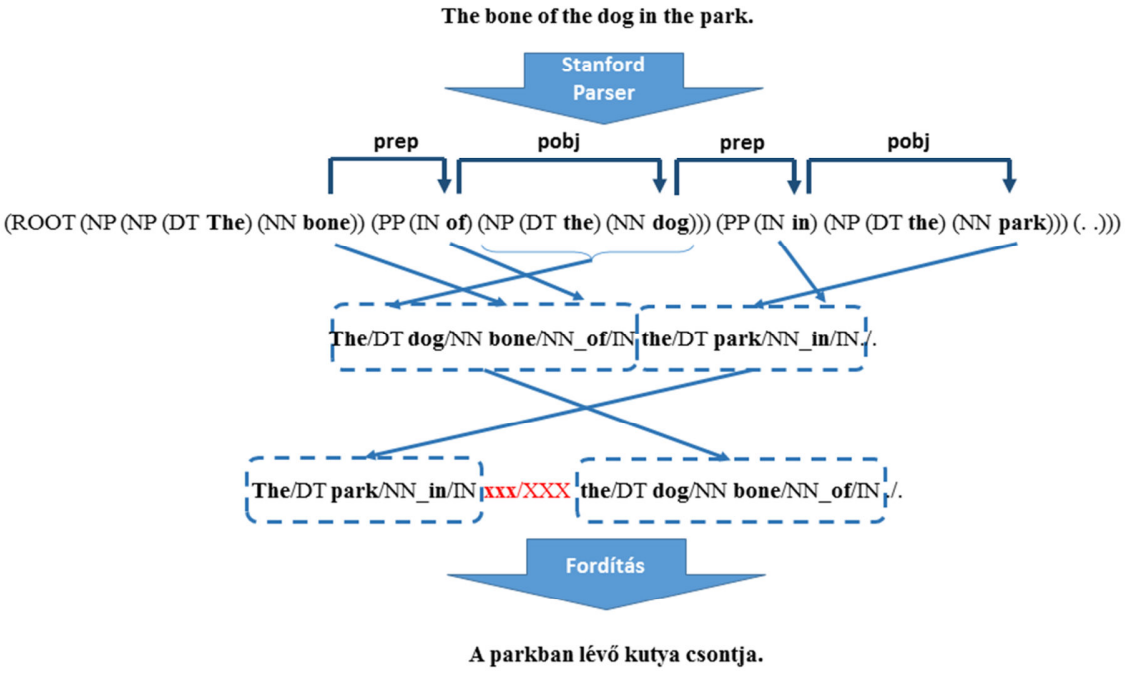
**Az elöljárószavak kezelése:** Az angol és a magyar szórend közti különbség, hogy az angol elöljáró a magyarban többféle módon jelenhet meg; például névutó vagy toldalék formájában, esetleg teljesen más szerkezettel fordítható (pl. *in ten minutes – tíz perc múlva* vagy to my knowledge – tudomásom *szerint*). Ezzel szemben az angolban a prepozíció mindig megelőzi a főnévi frázist, amelyre vonatkozik. A különböző esetekre külön átrendezési szabályokat hoztam létre. A prepozíciót a PREP (prepositional modifier – elöljárószós módosító) függőségi reláció segítségével tudom azonosítani.

Abban az esetben, ha a magyar névutóval fejezi ki az angol elöljárószót, annak mindig a főnévi frázis után kell állnia. Ennek megfelelően a következő szabály alkalmazásával az angol *between, without, before, behind, above, below, under,* valamint *over* esetekben megcseréltem a sorrendet, és ezeket a PREP függőségi relációval meghatározott prepozíciókat a relációban szereplő főnévi frázis végére helyeztem. Például az angol *We/PRP can/MD talk/VB before/IN the/DT lesson/NN* mondat az átrendezés után *can/MD_we/PRP talk/VB the/DT lesson/NN before/IN* alakú lett.

DOI:10.15774/PPKE.ITK.2015.004

Minden további esetben az angol prepozíciót a magyarban toldalékolt szóként kezelem, emiatt a POBJ (object of the preposition – az elöljárószó tárgyvonzata) kapcsolat jobbjához csatoltam az elöljárószót. Abban az esetben, ha a főnévi frázis kötőszót is tartalmaz, akkor a magyarban a kötőszó mind a két oldalán található főnév megkapja a prepozíciónak megfelelő toldalékot. Ennek megfelelően a CONJ (conjunction – összekötendő szavak) relációval összekapcsolt főnévre is rákapcsolom az adott elöljárószót. Például az *I/PRP go/VBP to/TO the/DT cinema/NN with/IN PeterNNP and/CC Paul/NNP* mondatot átrendezve a *go/VBP_I/PRP the/DT cinema/NN_to/TO PeterNNP_with/IN and/CC Paul/NNP_with/IN* formátumot kaptam.

Az angol hátravetett elöljárószós módosító csoportnak a magyarban vagy egy az angol szerkezethez hasonló szórend felel meg, vagy egy fordított szórendű *„lévő”*-s szerkezettel fordítható. Abban az esetben, ha a prepozíció része egy birtokos szerkezetnek, ez a magyarban csak egy *„lévő”*-s szerkezetre fordítható. Ez alapján az elöljárószavas csoportot 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 csoport közé beillesztek egy *„xxx”* sztringet. Ezt a folyamatot a 13. ábra szemlélteti.

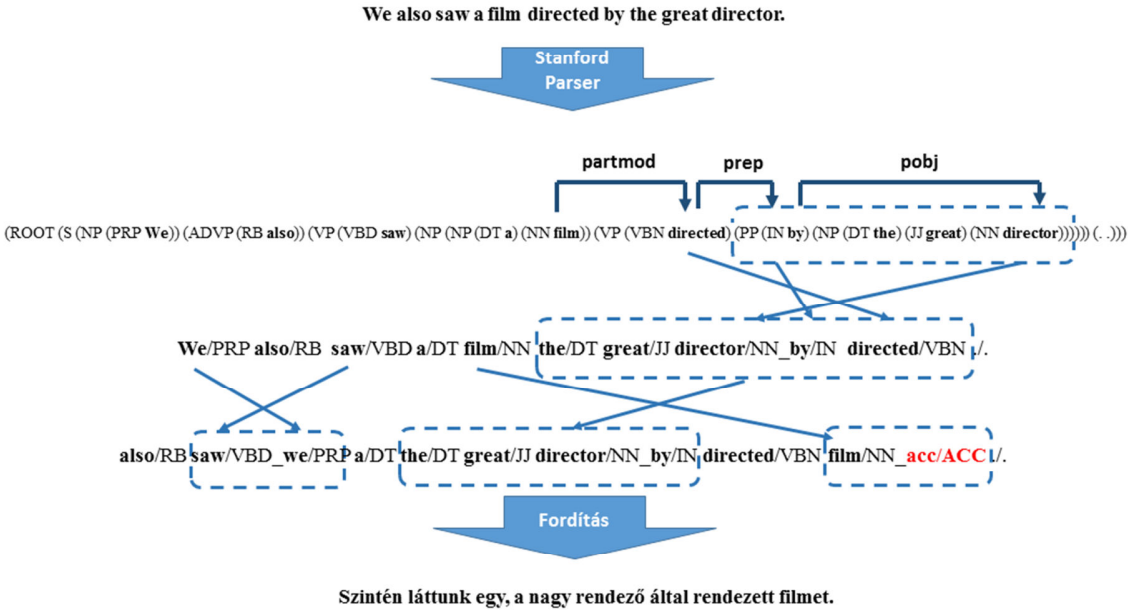



**13. ábra: Elöljárószók kezelése**

DOI:10.15774/PPKE.ITK.2015.004

**Melléknévi igenevek kezelése:** Az AMOD (adjectival modifier – melléknévi módosító) reláció a főnévi frázishoz kapcsolt melléknevet jelöli. A magyar nyelvben ez a szerkezet folyamatos melléknévi igenévnek fordul, ami mindig megelőzi a főnevet, amire vonatkozik. Az angol nyelv esetén a melléknév a magyarhoz hasonlóan általában a főnév előtt áll, de bizonyos esetekben a főnév mögött foglal helyet. Erre az esetre létrehoztam egy szabályt, amelyik abban az esetben, amikor a melléknév hátrább van, mint az AMOD relációban mellette álló főnév, akkor a főnévi frázisban a determináns után helyezem át. Erre példa az *I/PRP want/VBP to/TO find/VB the/DT most/RBS beautiful/JJ star/NN visible/JJ* mondat, melyet átrendeztem *want/VBP_I/PRP find/VB_to/TO the/DT most/RBS beautiful/JJ visible/JJ star/NN* alakúra.

Hasonló a helyzet a magyar befejezett melléknévi igenevekkel is, ami az angolban a past participle-nek felel meg. Megfigyelhető, hogy a past participle-t leginkább jelzőként használjuk vonzatokkal vagy bővítményekkel együtt. Ez a frázis ilyenkor a jelzett szó után áll. A magyarban általában, a jelzett szó előtt helyezkedik el a melléknévi igeneves csoport, melyben a melléknévi igenév a frázis végén található. Ezt a szerkezetet a PARTMOD (participle modifier – melléknévi igenévi módosító) reláció segítségével találom meg. Először a past participle frázis igéjét a szerkezet végére helyezem át, majd az egész szerkezetet a jelölt szóhoz tartozó főnévi frázis elé helyezem. A folyamatot a 14. ábra szemlélteti.




**14. ábra: Befejezett melléknévi igenév kezelése**

DOI:10.15774/PPKE.ITK.2015.004

**Egzisztenciális szerkezet:** Ebben az esetben a *there* és az utána szereplő létige együtt felelnek meg a magyar egzisztenciális *van*-nak. Ez az EXPL (expletive – egzisztenciális szerkezet) reláció alapján található meg. Erre példa a *there/EX is/VBZ a/DT ghost/NN* szerkezet, mely az átrendezés után *is/VBZ_there/EX a/DT ghost/NN* alakú lett, amely a magyar *van egy szellem* szókapcsolatnak felel meg.

**Segédigék kezelése:** Ez a szabály az igei frázisokat alakítja megfelelő formátumra, és egyezteti számban, személyben, időben és módban. Célom volt a magyar nyelvű ige szóalakjának helyes generálása megfelelő igeragozással. Az NSUBJ (nominal subject – a mondat alanya) függőségi relációval meghatároztam a mondat alanyát, melynek számával és személyével az igét egyeztetni kell. Az alanynak megfelelő névmást csatoltam címkeként az ige után, hogy ezáltal a dekóder a megfelelő módon generálja ki a szóalakot. Halmozott alany esetén a megfelelő névmást kézzel írott szabályok alapján határoztam meg, melyhez a CONJ (conjunction – összekötendő szavak) és CC (coordination – kötőszó) relációkat vettem alapul (pl. *I and Peter → we, you and Tom → you*[PL], *Peter and Paul → they* stb.). Másfelől átrendezési szabályokat írtam az angol segédigék alapján, hogy szerkezetükben a lehető legjobban megközelítsem a magyar fordítást. Öt csoportba soroltam a kezelendő angol segédigéket, melyeket különböző módokon kezelek:

- *may, might*
    - Jelen időben egyszerű átrendezést végeztem, mivel a *may* és a *might* a feltételes mód jeleként fordítható. Például: *I may/might eat → eat_I_may/might → megenném*
    - A múlt idejű *may/might have VBN* frázist a *lehet hogy VBN volna* frázisra fordítható. Mivel a *volna* segédigének az angolban nincs megfelelője, ezért az angol szerkezetet kiegészítettem ezzel. Például: *I may/might have been eaten/VBN → may/might_have eaten_I_volna → lehet, hogy megettem volna*
- *would*
    - Jelen időben ugyanúgy kezeltem, mint a may/might esetet.
    - Múlt időben a *would* segédigét elhagytam, helyette a *volna* szót illesztem a szerkezet végére. Például*: I would have eaten → eaten_I volna → megettem volna*
- *must, should*
    - Jelen időben az egyeztetett igéhez csatoltam a névmást, és az igei frázis után rendezem a *should* segédigét. Például: I should eat → eat_I should → ennem kéne

DOI:10.15774/PPKE.ITK.2015.004

  - Múlt időben a *should have VBN* kifejezést *VBN should_have volna*-ra rendezem át. Például: *I should have eaten → eaten_I should_have volna → ennem kellett volna*

- *will*
  - Jelen időben a *will* segédigére csatoltam a toldalékokat, mivel a magyarban a *fog* segédige megfelelően ragozott alakja lesz a fordítás. Például: *I will eat → eat will_I → enni fogok*
  - Befejezett jövő időben: Például: *I will have eaten → eat will_I_have → addigra meg fogom enni*
- *can, could*
  - Jelen időben a *will* jelen idejéhez hasonlóan működik. A *can* segédigéhez csatoltam a toldalékokat. Például: *I can eat → can_I eat → tudok enni*
  - Múlt időben pedig be kellett szúrni egy *volna*-t. Például: *I could have eaten → could_I volna eaten → tudtam volna enni*

Külön átrendezési szabállyal kellett kezelni azokat az eseteket, ha a mondatnak névszói-igei típusú állítmánya van. Ilyenkor az NSUBJ (nominal subject – a mondat alanya) reláció nem az alanyt és az igei állítmányt köti össze, hanem az állítmány névszói részével kapcsolja össze. Ezután azonban meg kellett keresni az állítmány igei részét is az AUX (auxiliary – segédige) relációval. Másrészt ezeknek az eseteknek a fordítása az állítmány névszói részének és a segédige megfelelően toldalékolt alakjának együttese lesz. Ebből adódóan meg kellett cserélni a névszó és az ige sorrendjét, valamint minden toldalékot a *be* segédigéhez kellett csatolni. Ezt az alábbi átrendezési szabályok alkalmazásával értem el:

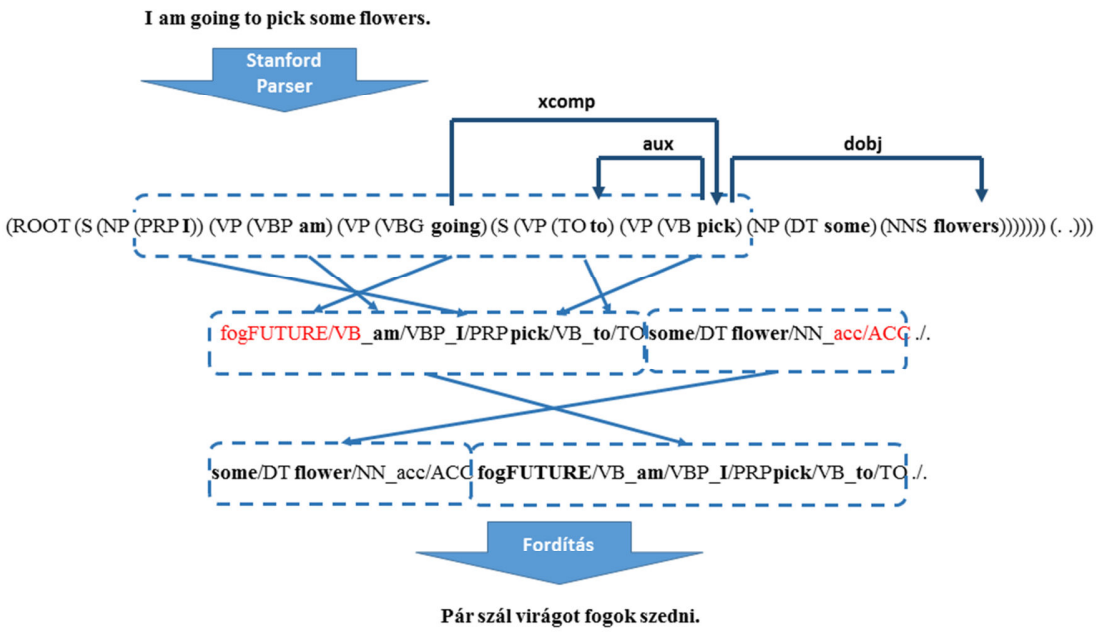
- *will:* Ebben az esetben a fent leírt szabályt alkalmaztam kiegészítés nélkül.
  - *I will be a lion. → a/DT lion/NN be/VB_will/MD_I/PRP ./.*
  - *I will have been a lion. → a/DT lion/NN been/VBN_will/MD_have/VB_I/PRP ./.*
- *may, might, would, can, could:* Ezeknél a segédigéknél múlt idő esetén az igei frázist egy *volna*-val kellett kiegészítenem.
  - *I may be a lion. → a/DT lion/NN be/VB_may/MD_I/PRP ./.*
  - *I may have been a lion. → a/DT lion/NN been/VBN_may/MD_have/VB_I/PRP_volna/WOULD ./.*
  - *I would be a lion. → a/DT lion/NN be/VB_would/MD_I/PRP ./.*
  - *I would have been a lion. → a/DT lion/NN been/VBN_would/MD_have/VB_I/PRP_volna/WOULD ./.*

DOI:10.15774/PPKE.ITK.2015.004

- *I can be a lion.* → *a/DT lion/NN be/VB_can/MD_I/PRP ./.*
- *I could have been a lion.* → *a/DT lion/NN been/VBN_could/MD_have/VB_I/PRP_volna/WOULD ./.*

- *must, should:* Ebben az esetben az állítmány névszói részéhez csatoltam a *-nak* toldalékot.
  - *I should be a lion.* → *a/DT lion/NN_nak/NAK should/MD be/VB_I/PRP ./.*
  - *I should have been a lion.* → *a/DT lion/NN_nak/NAK should/MD_have/VB_volna/WOULD been/VBN_I/PRP ./.*

**Jövő idős szerkezet:** Az angol *'going to'* jövő idős frázis a magyarban a *fog* segédige ragozott alakjával és az igéből képzett főnévi igenév segítségével fejezhető ki. Bevezettem egy új címkét *(fogFUTURE)* a *fog* segédige jelölésére, valamint ehhez kapcsoltam az angol szerkezetből kinyert toldalékokat, mivel magyarban ezt a segédigét egyeztetem számban és személyben. Továbbá az angol *'to'* funkciószót az ige mögé kapcsoltam, hogy az a magyarban főnévi igeneves fráziként jelenjen meg. A frázist az XCOMP (open clausal complement – mellékmondat kiegészítéss) és AUX (auxiliary – segédige) függőségi kapcsolatok alapján azonosítottam. Továbbá meghatároztam a cselekmény tárgyát a DOBJ (direct object – a mondat tárgya) függőségi relációval. A tárgyhoz kapcsoltam a tárgyragot (acc/ACC), hogy a magyar oldalon ez megjelenjen. Végül az így kapott két szerkezetet megcseréltem. A folyamatot a 15. ábra szemlélteti.




**15. ábra: Jövő idős szerkezet kezelése**

DOI:10.15774/PPKE.ITK.2015.004

**Mutató névmások**: A *this, these, that, those* mutató névmások a magyar fordításban mutató névmással kifejezett határozott minőségjelzős szerkezetként jelennek meg. Ezt szerkezetet a DET (determiner – névelő) relációval határoztam meg. A *this, these* szavak jelölésére létrehoztam egy új XXTHIS címkét, ezt egységesen a magyar *ez* névmásra fordítottam. Ezzel párhuzamosan a *that, those* szavak XXTHAT címkét kaptak, és az *az* névmásra lettek fordítva. A magyar *ez, az* mutató névmások megkapják a jelzett szó toldalékait (ezért nincs szükség a többes számot külön kezelni). Mivel a magyarban a mutató névmás és a főnév közé bekerül egy határozott névelő, ezért az átrendezési szabály ezt a hiányt is kezeli. Például az *I/PRP broke/VBD the/DT mirror/NN in/IN this/DT house/NN* angol mondatot *I/PRP broke/VBD the/DT mirror/NN_acc/ACC XXTHIS/DT_in/IN the/DT house/NN_in/IN* formájúra alakítottam át.

### 3.3.2 Redundanciák feloldása, utófeldolgozás

Ezek a szabályok elsősorban az előző csoportba tartozó átrendezések mellékhatásai miatt szükségesek.

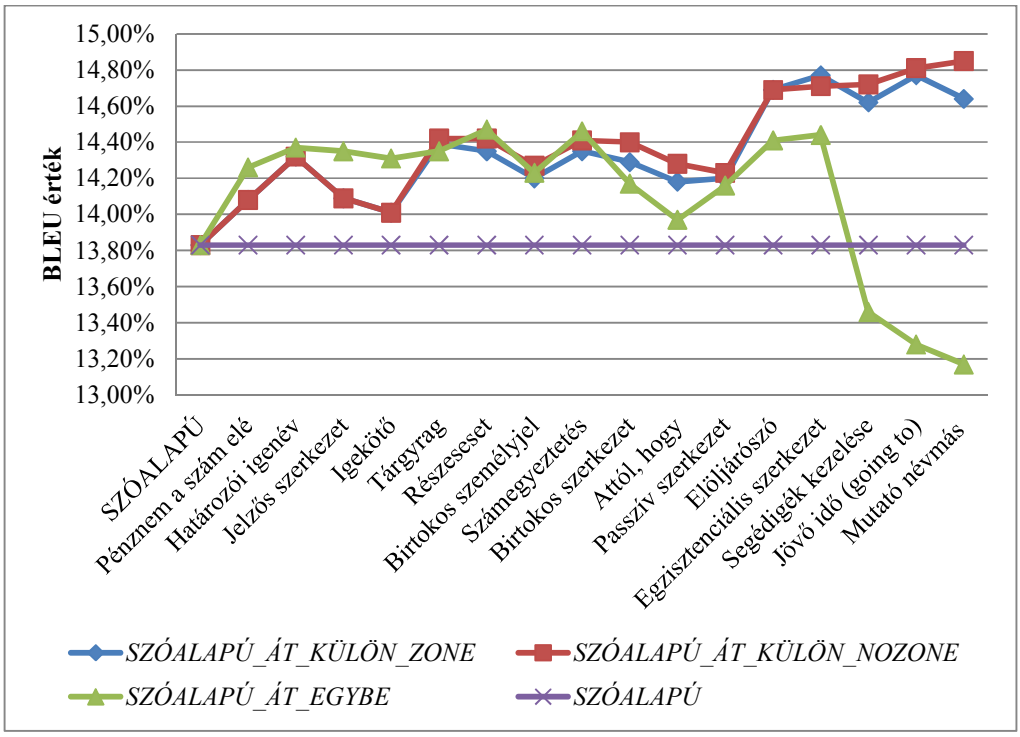
**Számok és pénznemek kezelése:** Magyar nyelvben szokásos formázási technika, hogy a többjegyű számokat három jegyenként szóközökkel elválasztva tagolják (pl. 10 000 000). Elsődleges feladat a tanítóhalmazban a számok detekciója, illetve a széttagolt számok reguláris kifejezések segítségével leírt összefűzése. Ugyanezt a folyamatot kell elvégezni a korpusz angol oldalán azzal a különbséggel, hogy itt a szóközök helyett vesszővel tagolják a számokat (pl. 10,000,000). További eltérés a két nyelv között, hogy az angolban ponttal (xxx.xxx) a magyarban (xxx,xxx) vesszővel választjuk el a tizedesjegyeket. A számok normalizációját követi a pénznemek (dollár: $ és euró: € karakter) kezelése. Az angolban ezek a karakterek az összeg előtt találhatók, míg a magyarban az után írjuk őket; ezért egy átrendezési szabályt alkottam ezen különbség feloldására, ami például a *$/$ 100000/CD is/VBZ not/RB enough/JJ* mondatot *100000/CD_$/$ is/VBZ not/RB enough/JJ* formájúra alakította.

**A redundáns névelők kezelése:** Ha a főnévi frázisok átrendezése során több névelő kerül egymás mellé, akkor az ismétlődéseket törlöm.

DOI:10.15774/PPKE.ITK.2015.004

## 3.4 Az eredmények ismertetése

Ebben a fejezetben ismertetem az átrendezési szabályokkal kiegészített rendszer eredményeit. A rendszert a Hunglish korpuszon [50] tanítottam, melyen az 3.2.1. fejezetben bemutatott módosításokat végeztem. A forrásoldali angol szövegen a Stanford Parser [39] segítségével közvetlen összetevős és függőségi elemzést végeztem, és az így kapott szintaktikailag elemzett angol szövegen végeztem el a 3.3. fejezetben ismertetett átrendezési szabályokat. A szófaji címkéket csak az átrendezésnél használtam, ezután eltávolítottam a szövegből, így megkaptam az átrendezett angol szöveget. Az így létrehozott szóalapú fordítórendszer az átrendezett angol szövegről fordít a változtatás nélküli magyar szövegre, és a fordítás során már nem használja fel a szófaji címkéket. (A szófaji címkéket is felhasználó morfológiailag módosított fordítórendszerek vizsgálatával a 4. fejezetben foglalkozom.)




**16. ábra: A szabályokkal kiegészített rendszerek eredményei az alaprendszerhez képest**

A rendszert fokozatosan építettem fel, vagyis minden lépésben eggyel több szabállyal egészítettem ki a már meglévőket. Így pontosabban lehet vizsgálni a hozzáadott szabályok hatását. Az egyes szabályok hozzáadási sorrendjét, illetve az adott szabállyal kiegészített rendszer eredményét a 16. ábra szemlélteti. Az ábra egy oszlopa bemutatja az aktuális és a már előtte szereplő átrendezési szabályok összhatását az alaprendszerhez (*SZÓALAPÚ*, átrendezést nem tartalmazó rendszerhez) képest, melyet az ábrán lila színnel jelöltem. A Moses rendszer dekóderében lehető-

DOI:10.15774/PPKE.ITK.2015.004

ség van arra, hogy a fordítandó mondatban definiáljunk olyan szócsoportot, amit a dekóder egy egységként kezel; ezt zónának nevezzük. Továbbá az egy egységen belüli szavak fordítását nem befolyásolja azok környezete. Munkám során létrehoztam a teszthalmaz egy olyan változatát, ahol az átrendezett kifejezéseket külön fordítási egységbe csoportosítottam (ez az ábrán kékkel jelölt *ZONE* rendszer). Az ábrán pirossal jelölt *NOZONE* pedig az egyszerű szóalapú fordítórendszert mutatja. Az eredmények vizsgálatából kiderült, hogy a szabályok alkalmazásával mindkét esetben az alaprendszernél jobb minőségű fordítást értem el. Megfigyelhető továbbá, hogy a *NOZONE* rendszerek általában jobban szerepelnek a *ZONE* rendszerekhez képest.

A legjobb eredményt az utolsó szabály hozzáadásával alkotott *SZÓALAPÚ_ÁT_KÜLÖN_NOZONE* rendszer érte el, szám szerint 14,85%-os BLEU értékkel. Ez 7,38%-os relatív javulás a *SZÓALAPÚ* 13,83%-ához képest. A 16. ábra bemutatja az egyes szabályok hatását a fordítás minőségére.

| | |
|---|---|
| Angol ref | For years I 've struggled to rid our kind of any hereditary weaknesses |
| *SZÓALAPÚ* | Az évek , az ilyen meg olyan ősi gyengeség jele |
| Jelzős szerkezettel | Évek óta már küszködött , hogy megszabaduljon a fajta minden ősi gyengeség |
| Magyar ref | Évekig igyekeztem ... megszabadítani a fajtánkat az öröklődő gyengeségektől . |

**2. táblázat: Példamondat a jelzős szerkezet fordítására**

Megfigyelhető, hogy néhány szabály hozzáadásával számottevően csökkent a *SZÓALAPÚ_ÁT_KÜLÖN_ZONE* és *SZÓALAPÚ_ÁT_KÜLÖN_NOZONE* rendszerek BLEU értéke. Az első ilyen pont a jelzős szerkezet kezelése – *kind of, sort of, type of, lots of* –, ahol a szerkezet *of* funkciószavát *xf* karaktersorozattá alakítottam. Az eredmények mélyebb vizsgálata során megállapítottam, hogy a teszthalmaz 1000 mondatából 6 esetben fordulnak elő ezek a kifejezések. Ezekben az esetekben a fordításban megjelent a szerkezetek megfelelő magyar fordítása (*fajta, féle, típusú)*. Erre mutat példát a 2. táblázat. A szabály hatására 234 mondatban változott a BLEU érték; ebből 112 esetben javult és 122 mondatnál csökkent. Ezzel ellentétben a *SZÓALAPÚ_ÁT_EGYBE* rendszer esetében a BLEU érték csak kis mértékben csökken. Ez annak köszönhető, hogy ilyenkor az az *'xf'* a szerkezet fejéhez van csatolva, ezáltal kevésbé van hatással a többi mondat fordítására. Hasonló hatás figyelhető meg az igekötők kezelésénél is.

| | |
|---|---|
| Angol ref | Did you feed him his pills today ? |
| *SZÓALAPÚ* | nem takarmány neki a pirulákat ma ? |
| Birtokos személyjellel | meg KELL odaadni a gyógyszereit ? |
| Magyar ref | Beadtad neki a gyógyszereit ? |

**3. táblázat: Példamondat a birtokos személyjel helyes fordítására**

DOI:10.15774/PPKE.ITK.2015.004

A következő BLEU csökkenés a birtokos szerkezet kezelésére bevezetett szabály esetében figyelhető meg. A teszthalmazban 359 mondatban történt változás: 178 esetben javulás (például 3. táblázat), 181 mondatnál romlás. Az eredmények mélyebb vizsgálatából kiderült, hogy néhány esetben az angol mondat birtokos determinánsának és birtokának megcserélése a magyar oldalon a főnév elhagyását eredményezte (például a 4. táblázat). Ez jelentős BLEU veszteséget okoz.

| | |
|---|---|
| Angol ref | Sir , my team is ready now . |
| *SZÓALAPÚ* | uram , a csapat készen áll . |
| Birtokos személyjellel | uram , a készen áll . |
| Magyar ref | Uram , a csapatom már készen áll . |

**4. táblázat: Példamondat a birtokos személyjel helytelen fordítására**

A passzív szerkezet a *SZÓALAPÚ_ÁT_KÜLÖN_NOZONE* rendszer esetében mutatkozik némi romlás. Ez annak tudható be, hogy a korpusz mondataiban a passzív szerkezetet többféleképpen fordítják. Abban az esetben mikor a teszt halmazban a passzív szerkezet nem az általam javasolt módszerrel fordul, akkor az a BLEU értéket csökkenti függetlenül attól, hogy a lefordított mondat helyes-e.

Azon átrendezési szabályok esetén, amikor a magyar oldalon is csak egy helyes fordítás lehetséges (például az igeragozásnál vagy a birtokos szerkezet kezelésénél), nagymértékű javulás figyelhető meg a szabály hozzáadása után.

| | w-BLEU | mm-BLEU |
|---|---|---|
| *SZÓALAPÚ* | 13,83% | 59,32% |
| *SZÓALAPÚ_ÁT_KÜLÖN_NOZONE* | **14,85%** | 58,06% |
| *SZÓALAPÚ_ÁT_KÜLÖN_ZONE* | 14,64% | 57,94% |
| *SZÓALAPÚ_ÁT_EGYBE* | 13,17% | 57,21% |

**5. táblázat: A legjobb rendszerek eredményei**

Kipróbáltam továbbá egy olyan rendszerarchitektúrát (*SZÓALAPÚ_ÁT_EGYBE*), amikor az átrendezés után a magyarhoz hasonlóan a lemmához kapcsolom a hozzátartozó funkciószókat (ez az ábrán zölddel van jelölve). Ezzel a módszerrel az angol igéket és főneveket agglutináló tulajdonsággal ruháztam fel. Az így létrehozott kísérleti rendszer azonban rosszabb BLEU pontot ért el (13,17%) a *SZÓALAPÚ*-hoz képest, ugyanis az angol oldalon is megjelent az adathiány-probléma. Az 5. táblázat a legjobb minőségű rendszerek eredményeit szemlélteti.

DOI:10.15774/PPKE.ITK.2015.004

## 3.5 Kapcsolódó munkák, előzmények

A gépi fordítórendszereknek megoldást kell nyújtani a nyelvek között fennálló eltérések kezelésére. Az elsők között Brown et al. [4] foglalkoztak statisztikai gépi fordítással. Az általuk létrehozott szóösszekötő rendszerbe integrálták a szórendi átrendezést támogató modult, az úgynevezett magasabb szintű IBM modelleket (IBM 3-5). Azonban ezen modellek nem kezelik hatékonyan az átrendezést, mert kevés strukturális és környezeti információt használnak fel, mivel csupán a szavak cseréjéhez szükséges lépésszámmal dolgoznak, és a párhuzamos korpuszból semmilyen információt nem használnak fel [35].

Az SMT dekóderek a megfelelő szórend elérésére törekednek a fordítás során. Az összes lehetséges átrendezés megvizsgálásáról Knight [13] megállapította, hogy NP-teljes probléma. Annak érdekében, hogy a feladat véges időn belül elvégezhető legyen, megszorításokat kell alkalmazni a dekódolás során. Egyrészt Berger et al. [51], valamint Wu [52] a lehetséges átrendezések számának korlátozásával érték el a feladat komplexitásának csökkentését. Egy másik szemlélet szerint a dekóder a balról jobbra történő fordítás során azzal a feltételezéssel él, hogy a szórend a forrásnyelvi és célnyelvi oldalon nagyjából hasonló, és emiatt a rendszer nem támogatja a nagyobb szórendi eltéréseket. Koehn és munkatársai [53], [54] az átrendezést két modellel valósították meg a dekódolás során. Az első az úgynevezett szórendbeli különbséget büntető modell (2.2.3.5. fejezet), melyben ha egy helyes fordítás nagy eltéréseket tartalmaz a szórendben, azt a rendszer lepontozza. A második egy lexikalizált átrendezési modell (2.2.3.6. fejezet), mely a kifejezéseken belüli átrendezéseket tárolja. Ezt azért fontos megemlíteni, mivel munkám során a Moses nevű [9] SMT implementációval dolgoztam, melynek dekódere szintén használja ezt a modellt. A Koehn-féle módszer előnye, hogy az átrendezési szabályokat grammatikai tudás nélkül, csupán a párhuzamos kétnyelvű korpuszból képes előállítani. Hátránya viszont, hogy a nyelvtani szabályszerűségek helyett csak átrendezési példákat tanul meg a rendszer, ami adathiányproblémához vezet. Emiatt azokban az esetekben, amik nem szerepeltek a tanítóhalmazban, nem képes javaslatot tenni.

A statisztikai módszereken alapuló rendszerek önmagukban nem elegendőek a szórendi átrendezések kezelésére, ezért megoldást jelenthet a PBSMT (kifejezésalapú gépi fordító) dekóderjének megváltoztatása. Az egyik megközelítés a log-lineáris modell alapú dekóder kiegészítése átrendezést támogató modulokkal, mint például Zens et al. [55], Zhang et al. [56], [57] és Feng et al. [58] esetében, akik új nyelvmodellt építettek a dekóderbe, melyet a már előzőleg átrendezett korpuszon tanítottak be. Másrészt a szórendi különbséget büntető modell lecserélése lehet megol-

DOI:10.15774/PPKE.ITK.2015.004

dás, mely helyett Al-Onaizan és Papineni [59] egy szórendi hasonlóságon alapuló, míg Xiang et al. [38] egy szintaktikai tulajdonságokat is használó modellt alkalmaztak. Crego és Marino [36], [60] a hagyományos frázisalapú dekóder helyett, létrehoztak egy egyedi kétnyelvűszópár-alapú fordítási modellt, mely nem azt vizsgálja, hogy egy adott kifejezésnek mi a fordítása, hanem az adott kifejezés fordításánál figyelembe veszi annak környezetét is. Zhang et al. [61] egy új átrendezési modellt hoztak létre, mely az eredeti PBLRM-mel ellentétben a kifejezések egymáshoz viszonyított szomszédsági kapcsolatait és ezek valószínűségeit határozza meg. Li et al. [62] egy módosított dekódert alkalmaztak, amely az eredeti forrásnyelvi mondat mellett annak átrendezett variációit is felhasználja. Az eddig leírt rendszerek hátránya, hogy szabadon nem elérhetők, és az újraimplementálásukba fektetett munka nem arányos az általuk elért fordítások minőségének javulásával. Továbbá olyan előfeldolgozó eszközöket igényelnek, melyek magyar nyelvű megvalósítása még kutatás tárgyát képezi.

A következő módszer a szórendi különbségek kezelésére a forrásnyelvi mondatok előfeldolgozási lépésként történő átrendezése oly módon, hogy az minél jobban közelítse a célnyelvi mondatok szósorrendjét. Azzal, hogy az átrendezés által a forrás- és célnyelvi mondatok szavai közel azonos pozícióba kerültek, az eredeti PBSMT dekóderrel jobb minőségű fordítás érhető el. Az átrendezési szabályok előállításának közkedvelt módja gépi tanuláson alapul, ahol a szabályok automatikusan a párhuzamos korpusz alapján kerülnek meghatározásra. Xia és McCord [63] az átrendezési mintákat angol-francia párhuzamos szövegek közvetlen összetevős elemzései és szóösszekötések segítségével állították elő. Visweswariah et al. [64] hasonló stratégiát alkalmaztak, melyben módosítottak az átrendezés valószínűségének számításán. Costa-Jussá és Fonollosa [65] egy maximum entrópián alapuló rendszerrel egészítették ki a szóösszekötés-alapú szabálymeghatározást. A Rottmann és Vogel [66] által készített fordítórendszer a szóösszekötött korpuszból átrendezési szabályokat tanul, melyeket a fordítás során az eredeti mondat szóhálóján alkalmaz. Niehues és Kloss [67] ezt a módszert kiegészítve nagy hangsúlyt fektet a nagy távolságú átrendezésekre, amelyekhez szófaji egyértelműsítés-alapú szabályokat alkalmaznak. Elming [68], illetve Elming és Habash [69] az eredeti mondatból és annak átrendezett formájából szófát építettek. Az átrendezések valószínűségét nem a forrás-, hanem a célnyelvoldali mondatból számolták. A Jiang et al. által létrehozott szógráfsúlyozó algoritmus [70], valamint a kifejezés-összekötőből származó információ alapján állítják elő az átrendezési szabályokat. Holmqvist et al. [71] egy nyelvfüggetlen, szóösszekötő-alapú átrendezést (alignment-based reordering) készítettek, melyben a forrásnyelvi szöveget a kezdeti szóösszekötések alapján átrendezték, majd egy második szóösszekötést végeztek. Ennek célja, hogy a szavak sorrendje közelebb kerüljön a célnyelvi mondathoz. Végül ezt a második szóösszekötést alkalmazták az eredeti mondaton. Lerner és Petrov [72]

DOI:10.15774/PPKE.ITK.2015.004

az átrendezéseket diszkriminatív osztályozó segítségével állították elő a szintaktikailag elemzett forrásnyelvi szövegből. Huang és Pendus [73] bemutatták, hogy jobb fordítási eredmény érhető el, ha a teljesen lexikalizált átrendezési szabályokat (ahol a nemterminálosok mellett a terminális elemek is fel vannak tüntetve) a belőlük előállított részlegesen, vagy egyáltalán nem lexikalizált szabályokkal (csak nemterminálisokat tartalmaznak) egészítjük ki. Herrmann et al. [74] munkájuk során több átrendező módszert kombináltak. Rendszerükbe integrálták a szószintű lexikalizált átrendezési modellt, a morfológiai szinten történő POS-alapú átrendezéseket, valamint szintaktikai szinten a közvetlen összetevős elemzési fán alapuló módszereket. Az így létrehozott rendszerek hátránya, hogy grammatikailag eltérő nyelvek esetén automatikus módszerekkel nehezen ismerhetők fel a komplex nyelvtani különbségek.

Az automatikus módszer jól teljesít morfológiailag egyszerűbb nyelvek esetében, azonban a morfológiailag gazdag, agglutináló vagy szerkezetileg egymástól távol álló nyelveknél nem elégséges. Olyan nyelvpárok esetén, amelyek szintaktikai struktúrájukban és szórendjükben nagyon különböznek egymástól, a kutatások egyre inkább a szintaxisalapú modellek és hibrid metódusok alkalmazása felé tolódnak; mint például a tisztán statisztikai metódusoknak az előfeldolgozási folyamat során kézzel készített szabályokkal történő kiegészítése. Ilyen megközelítések alkalmazása nagyban elősegíti a fordítás minőségének javulását. Léteznek olyan, jól megfogható szerkezeti különbségek a nyelvpárok között, melyek nagyobb pontossággal azonosíthatók kézzel írt szabályokkal, mint gépi tanulással. Ilyen például a Berger et al. [51] által a francia nyelvben előforduló birtokos szerkezetre (*NOUN1 de NOUN2*) alkalmazott átrendezési szabály. A francia birtokos szerkezetnek az angol fordítása vagy hasonló szerkezetű (*NOUN1 of NOUN2*), vagy fordított szórendű lesz (*NOUN2 NOUN1*). Munkájuk során egy maximumentrópia-alapú modell segítségével döntötték el, hogy érdemes-e a francia birtokos szerkezetet tartalmazó kifejezés szavait felcserélni a fordítás előtt. Másik példa lehet az ige mellékmondatbeli pozíciója német és angol nyelv esetén. Míg a németben az ige a mellékmondat végén van, addig az angolban a szerkezet elején az alany után található. Ezt a jelenséget Collins et al. [75] fogalmazták meg lehetséges átrendezési szabályként német-angol fordítás esetén, Popović és Ney [76] pedig az angol-német irányra implementálták ugyanezt. Ehhez a német mondat közvetlen összetevős analízisét végezték el. Popović és Ney [76] emellett angol és spanyol nyelvpárok esetén a főnév és a melléknév sorrendjének összehangolására írtak átrendezési szabályokat. Wang et al. [77] ugyanezen megfontolásból különböző nyelvcsaládba tartozó nyelvpár (kínai-angol) szórendi különbségeinek áthidalására fogalmaztak meg törvényszerűségeket közvetlen összetevős elemzés alapján.

DOI:10.15774/PPKE.ITK.2015.004

Az automatikus módszerek szóösszekötések alapján keresik meg az átrendezési mintákat. Agglutináló nyelvek esetében (magyar, török, finn stb.) azonban a szóösszekötő rendszer alacsony pontosságú, hiszen például az angol funkciószavaknak (mint az elöljárószók, igekötők stb.) nincs egyértelmű megfelelője (például: *házamban – in my house*), emiatt azokat általában rossz szóhoz rendeli. A nyelvek közti strukturális különbség nemcsak a mondat szavainak számában, hanem gyakran a szórendben is megmutatkozik. Ebből adódóan az automatikus módszerek helyett kézenfekvő megoldásnak tűnik kézzel írt szabályok alkalmazása. Patel et al. [78] angol-hindi nyelvpárra készített félig lexikalizált kézzel írt szabályokat összetevős elemzés alapján. Yeniterzi és Oflazer [37] az angol-török nyelvpárok közti gépi fordításra mutattak olyan megközelítést, melyben egyidejűleg végeztek szintakszisalapú átrendezést a forrásnyelvi oldalon, illetve morfológiai szegmentációt, melynek célja egyben a szóösszekötések minőségének javítása. Ezt oly módon érték el, hogy az egy szóhoz tartozó morfémákat összekapcsolták a fordítási folyamat során (például: *on their economic relations – economic relation_s_their_on*).

Az angol-magyar gépi fordítás során felmerülő nehézségek és nyelvtani különbségek nagymértékben hasonlítanak az angol-török nyelvpár estében tapasztaltakhoz. Munkám során többek közt Yeniterzihez hasonlóan az angol és a magyar mondat szószámkülönbségére kerestem megoldást. Méréseim során bebizonyosodott, hogy a Yeniterzi-féle szóösszekapcsolás alkalmazásával magyar nyelv esetén nem sikerült javulást elérni, ezért más megoldást kerestem a probléma leküzdésére. Ezenkívül Yeniterzivel ellentétben a magyar szóalakjának előállításához nem eredeti PBSMT dekódert használtam, hanem egy magyar nyelven működő morfológiai generátort, a HUMOR-t [44].

DOI:10.15774/PPKE.ITK.2015.004

## 3.6 Összefoglalás

A statisztikai alapú fordítórendszer számára nehézséget okoz az egymástól grammatikailag távol eső nyelvek közti fordítás. Ez elsősorban a nyelvek közt felmerülő jelentős szórendi különbségből fakad. A fejezetben bemutattam, hogy ez a probléma nagymértékben akadályozza többek közt az angol-magyar nyelvpár közti fordítást is. Az általánosságban használt tisztán statisztikai alapon működő dekóderimplementáció csak a lokális szórendi átrendezésekre képes.

Munkám során létrehoztam egy hibrid fordítórendszert, mely a statisztikai módszerek mellett az általam írt komplex szabályrendszer segítségével javít a fordítás minőségén. A kézzel írt átrendezési szabályokat a forrásnyelvi – angol – szövegen alkalmaztam előfeldolgozási lépésként. Az átrendezendő szerkezetek megtalálása a közvetlen összetevős elemzés és a függőségi relációk alapján történt. Célom az angol nyelvű mondatok átalakítása a magyarhoz jobban hasonlító szerkezetekké. Ezekkel a transzformációkkal sikerült olyan fordításokat létrehozni, melyek 1,02% BLEU, azaz 7,38%-os relatív javulást értek el a hagyományos rendszerhez képest. 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 szórendi különbséget okozó jelenséget, ezek finomítása további kutatás témája lehet.

Az általam készített hibrid fordítórendszer egyedülálló megoldást jelent angol-magyar nyelvpár fordítása esetén.

Kapcsolódó tézis:

**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]

DOI:10.15774/PPKE.ITK.2015.004

# 4 Morfológiai különbségek kezelése a jobb minőségű statisztikai gépi fordítás érdekében

Az előző fejezetben bemutatott algoritmus az angol-magyar nyelvpár esetén fennálló szórendi különbségek okozta nehézségeket kezeli az angol oldalon végzett szórendi átrendezés segítségével. Birch et al. elmélete [34] alapján a fordítórendszer minőségét a szórendi különbség mellett a célnyelv összetettsége, valamint a két nyelv közti grammatikai eltérés is nagymértékben befolyásolja. Ez a különbség az angol-magyar nyelvpár esetében a nyelvek tulajdonságai miatt számottevő, ahogy azt a II. fejezet bevezetőjében kifejtettem.

Fontos észrevennünk, hogy az SMT rendszer számára a magyar mint **célnyelv összetettsége** komoly kihívást jelent, hiszen egy statisztikai alapú dekóder nem generálja a szóalakot, csak a tanítóanyagban szereplő szóalakok közül választ. Ez az agglutináló nyelvek esetében nagy hátrányt jelent, mivel hiába szerepel egy szó a tanítóanyagban, és lenne jól fordítható, a rendszer képtelen a szótövet más toldalékkal ellátni. A probléma kiküszöbölésére az SMT fordítási folyamatába egy morfológiai generátort építettem, mely egy szótőből morfoszintaktikai címkék segítségével képes a szóalakot előállítani.

| | | teszthalmaz | tanítóhalmaz |
|---|---|---|---|
| mondatok száma | | 1000 | 1 026 836 |
| átlagos szószám a mondatban | angol | 14,137 | 14,173 |
| | magyar | 11,672 | 11,764 |
| átlagos morfémaszám a mondatban | angol | 16,764 | 16,768 |
| | magyar | 18,391 | 18,429 |

**6. táblázat: Angol és magyar közti szó- és morfémaszám-különbség**

A két nyelv között egy másik, jelentős **különbség** is megfigyelhető a **morfoszintaktikában**, ugyanis a magyar nyelv agglutináló tulajdonságából kifolyólag egy szó ragozott alakjának az angolban egy nagyobb szószámú szerkezet felel meg, ahol a magyar szóhoz csatolt toldalékok különböző funkciószavakkal kerülnek kifejezésre. Emiatt egy angol mondat átlagos szószáma jóval több a magyarénál, viszont ez az arány a morfémák szintjén fordított. A két nyelv mondataira jellemző átlagos szó- és morfémaszám különbséget a Hunglish korpuszon [41], [42] végzett mérések igazolják. A 6. táblázatból kiderül, hogy – a teszt- és a tanítóhalmaz méretétől függetlenül – egy angol mondat átlagosan 14 szóból áll, ám a magyarban ez a szám csak 11. Ezzel ellentétben, a morfémák szintjén a magyar oldalon mérhető a magasabb érték (magyar: ~18, angol: ~16). Megfigyelhető továbbá, hogy a morfémaszinten mért eltérés relatíve kisebb a szavak szintjén mértnél.

DOI:10.15774/PPKE.ITK.2015.004

Ez a különbség az angol-magyar fordítás esetében arra készteti az SMT rendszer szóösszekötő modulját, hogy sok-egy vagy sok-sok típusú kapcsolatokat alakítson ki, vagy esetleg összekötés nélkül hagyja a szavakat. Ahogy azt már a 2.2.3.4. fejezetben kifejtettem, az IBM-1 modelleken alapuló szóösszekötő rendszer a sok-egy kapcsolatot nem tudja kezelni, a nem megfelelő szóösszekötések pedig hiányzó vagy helytelen szavakat eredményeznek a fordításban.

Szószám-harmonizáció segítségével elérhetjük, hogy a szóösszekötőnek ne kelljen sok-egy típusú kapcsolatokat kezelnie. A harmonizáció vagy az angol oldalon történő szóösszevonással, vagy a magyar oldal morfémákra bontásával oldható meg. Munkám során mindkét lehetséges megoldást megvizsgáltam. Először bemutatom a létrehozott rendszerek működését, majd ismertetem az elért eredményeket.

## 4.1 A létrehozott módszerek bemutatása

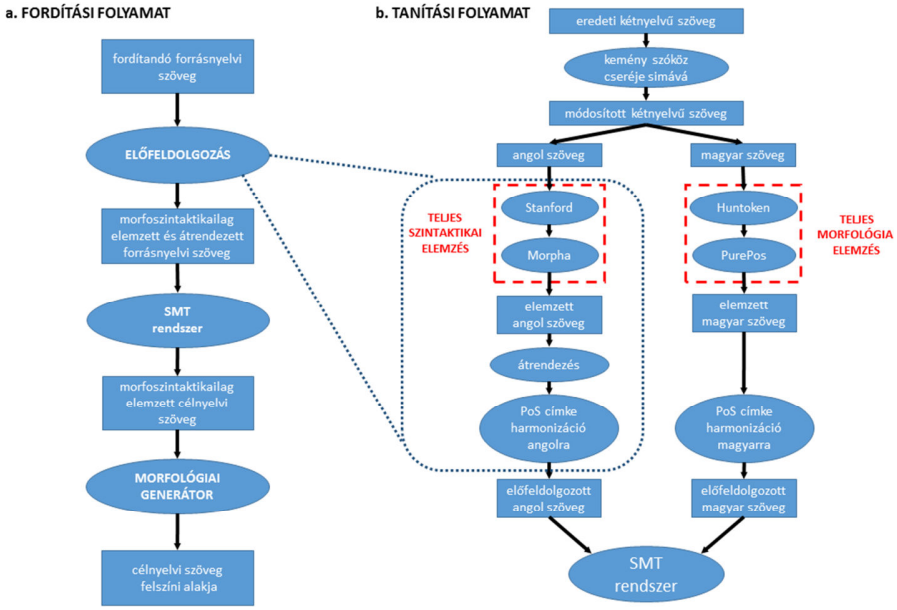
A Moses rendszer által használt statisztikai alapú dekóder (2.2.3.3. fejezet) legnagyobb hiányossága agglutináló nyelvek esetében, hogy csak a tanítóanyagban szereplő szóalakokat tudja bizonyos valószínűséggel ajánlani a fordítás során. Emiatt, ha külön a szótő és külön a toldalék – csak más szótővel – már szerepelt a tanítóanyagban, viszont együtt még nem, akkor a dekóder nem tudja kigenerálni a megfelelő szóalakot. Például a *car* és az *in my house* nyelvi egységek alapján nem képes az *in my car* kifejezésből az *autóban* szót előállítani. Egy másik nehézség, amivel a statisztikai alapú dekódernek meg kell küzdenie, az a szavak fordítása során fellépő többértelműség feloldása. Ez abban nyilvánul meg, hogy a fordítórendszer a tanítás során egy tetszőleges angol főnévhez vagy igéhez számtalan magyar szóalakot lát példaként, emiatt a fordítás során ezek a szóalakok mind felmerülnek mint lehetséges fordítási javaslat. Például a Hunglish korpusz alapján a *house* szó a következő magyar megfelelőkkel párosul:

- *ház* *(you really like this **house**? → tényleg tetszik ez a **ház**?)*
- *házban* *(There is a bath in the **house**. → A **házban** van egy fürdőszoba.)*
- *házat* *(I'll leave the **house**. → Elhagyom a **házat**.)*
- *házba* *(He went into the **house**. → Bement a **házba**.)*
- *házába* *(He went into her **house**. → Bement a **házába**.)*
- *háza* *(This is my girlfriend's **house**. → Ez a barátnőm **háza**.)*
- *házból* *(are you calling from the safe **house**? → az őrzött **házból** hívsz?)*
- ... stb.

DOI:10.15774/PPKE.ITK.2015.004

A fenti példából is látszik, hogy az angol-magyar fordítás nem oldható meg szószintű rendszerrel, hanem a morfémák szintjén kell a feladatot elvégezni. A statisztikai dekóder ezen gyengeségeinek kiküszöbölésére létrehoztam egy hibrid rendszert, mely a fordítást **morfémaalapú statisztikai módszer**rel végzi, valamint a magyar szóalak előállítását **morfológiai generátor** segítségével oldja meg. Rendszeremben a magyar nyelven működő HUMOR nevű eszközt [44] használtam. A morfémaalapú fordítás nagy előnye a szószintűvel szemben, hogy míg szavak szintjén egy szónak akár több száz alakját kéne külön esetként kezelni, addig a morfémaalapú rendszernek elég ismernie a szavak lemmáját és a hozzá kapcsolódó lehetséges toldalékokat. A szabályalapú generátornak köszönhetően lehetőség van olyan szóalakokat is előállítani, amelyek nem szerepelnek a tanítóanyagban.

Munkám során létrehoztam és megvizsgáltam három különböző rendszerarchitektúrát. Az első egy morfológiailag elemzett szavakkal működő szóalapú SMT, melyben a fordítás végén a morfológiai generátor nagy pontossággal állítja elő a szóalakot. A második egy morfológiai alapon működő fordító, ahol minden morféma mint önálló szó szerepel, és tulajdonképpen címkék közötti fordítás történik. Végül az előző két rendszer előnyeit ötvözve létrehoztam egy faktoros alapon fordító SMT rendszert. A következőkben bővebben bemutatom a rendszerek működésének folyamatát, majd ismertetem azok előnyeit és hátrányait.




**17. ábra: Fordítási és tanítási folyamat**

DOI:10.15774/PPKE.ITK.2015.004

A 17. ábra bemutatja az általam módosított fordítási folyamatot, mely a fordítandó angol szöveg teljes szintaktikai elemzésével és lemmatizálásával kezdődik. A Stanford Parser [39] segítségével szófaji egyértelműsítést végeztem, továbbá meghatároztam a közvetlen összetevős elemzést és a függőségi relációkat. A lemmatizálás a Morpha [48] alkalmazás segítségével történt. A következő lépésben az annotált angol szövegen a 3. fejezetben leírt szórendi átrendezést hajtottam végre. Az előfeldolgozás utolsó lépéseként elvégeztem a két nyelv morfoszintaktikai címkekészletének harmonizációját. Ezzel az volt a célom, hogy minden magyar morfémacsoporthoz külön POS címke tartozzon. Ennek megfelelően a több morfoszintaktikai információt tartalmazó címkéket szétbontottam. Az angol esetében a Penn Treebank [46] címkekészlete például egy címkében tárolja az ige szófaját és annak idejét. A szerkezet felbontásával az ige szófaja és annak ideje külön-külön címkével lett jelölve ([VBD]→[VB][Past]; [VBN]→[VB][PPart]; [VBG]→[VB][ING] és [VBZ]→[VB][Z]; [VBP]→[VB][P]). Másfelől a többes számú főneveket jelölő címkéket is felbontottam ([NNS]→[NN][PL] és [NNPS]→[NNP][PL]). Ezzel párhuzamosan a magyar (HUMOR) címkekészleten a következő átalakításokat végeztem:

- többesszám címke szétválasztása (például [PSe1i]→[PL][PSe1])
- múlt idő címkéjének szétválasztása (például [Me3]→[Past][e3])
- felszólító mód szétválasztása (például [Pe3]→[Subj][e3])
- feltételes mód szétválasztása (például [Fe3]→[Cond][e3])
- tárgyas ragozás szétválasztása (például [Te3]→[e3][Def], [Tt3]→[t3][Def], [TMe3]→[Past][e3][Def], [TPe3]→[Subj][e3][Def], [TFe3]→[Cond][e3][Def])
- alanyi ragozás szétválasztása (például [Ie3]→[e3][Obj2], [It3]→[t3][Obj2], [IMe3]→[Past][e3][Obj2]; [IPe3]→[Subj][e3][Obj2]; [IFe3]→[Cond][e3][Obj2]
- felsőfokú szerkezet átalakítása, ahol a felsőfokot reprezentáló címkét a szerkezet végére teszem [FF][MN]→[MN][FF]

Ez azért előnyös formátum, mert így csökkent a címkékben tárolt redundáns információ mennyisége, továbbá a magyar oldal legtöbb morfoszintaktikai címkéjének lett angol megfelelője, ami fordítva is igaz. Az előfeldolgozás során létrehozott, morfoszintaktikailag elemzett és átrendezett szöveg az SMT rendszer bemenete, ami ebből morfoszintaktikailag elemzett magyar mondatot állít elő. Végső lépésként a morfológiai generátor állítja elő a szóalakokat.

DOI:10.15774/PPKE.ITK.2015.004

A rendszer tanítása során a tanítóanyag magyar oldalán is szükséges az angolhoz hasonló előfeldolgozó lépéssorozat alkalmazása (17. ábra). Az angollal ellentétben magyar nyelvre nincsen a generátor címkekészletével kompatibilis szintaktikai elemző, emiatt a célnyelvi oldalon szófaji egyértelműsítést és lemmatizálást végeztem. Ezt a PurePos2 [43] nevű teljes morfoszintaktikai egyértelműsítő rendszerrel oldottam meg, míg a szavak tokenizálását a HunToken [79] alkalmazás hajtotta végre.

A fenti lépések alapján létrehoztam a **morfológiailag elemzett szavakkal működő szóalapú hibrid fordítórendszer**t. Ezt a rendszert a továbbiakban *SZÓALAPÚELEMZETT*-nek fogom hívni. A 7. táblázatban az előző példában szereplő mondatok láthatók, miután átalakítottam azokat a *SZÓALAPÚELEMZETT* rendszer SMT dekóder moduljának ki- és bemeneti formátumára. Megfigyelhető, hogy a módszer segítségével a különböző magyar szóalakoknak megfelelő angol nyelvű szószerkezetek jöttek létre. Az alkalmazott módszernek köszönhetően javult a fordítórendszer tanulásának minősége, mivel jelentősen csökkent a nyelvpárok között megfigyelhető szószámbeli eltérés, ami kedvez a szóösszekötő modul működésének. Az alkalmazott módszerrel lényegében az angol szavakat agglutináló viselkedésűvé alakítottam át. Ennek köszönhetően jelentősen csökkent a szavak fordítási lehetőségeinek száma, mivel a magyar toldalékolt szavaknak az angol oldalon is meglesz a megfelelő párja. Azzal, hogy az angol szavakat ragozhatóvá alakítottam, felerősödött az agglutináló nyelvek esetében tapasztalható adathiány-probléma, mivel a morfémák összekapcsolásával a forrásnyelvi szövegben található típusok száma növekedett meg (Hunglish korpusz alapján a sima angol szövegben 160 449 típus van, míg a *SZÓALAPÚELEMZETT* rendszerben 578 110 különböző szóalak található).

| | |
|---|---|
| be/[VB] [Z] he/[PRP] there/[EX] a/[DT] bath/[NN] the/[DT] **house/[NN] in/[IN]** ./[.] | van/[IGE] [e3] egy/[DET] kád/[FN] a/[DET] **ház/[FN] [INE]** ./[PUNCT] |
| leave/[VB] will/[MD] i/[PRP] the/[DT] **house/[NN] acc/[ACC]** !/[.] | elhagy/[IGE] [e1] [Def] a/[DET] **ház/[FN] [ACC]** !/[PUNCT] |
| go/[VB] [Past] he/[PRP] the/[DT] **house/[NN] into/[IN]** !/[.] | bemegy/[IGE] [Past] [e3] a/[DET] **ház/[FN] [ILL]** !/[PUNCT] |
| go/[VB] [Past] he/[PRP] **house/[NN] her/[PRP$] into/[IN]** !/[.] | bemegy/[IGE] [Past] [e3] a/[DET] **ház/[FN] [PSe3] [ILL]** !/[PUNCT] |
| this/[DT] girlfriend/[NN] my/[PRP$] **house/[NN] 's/[POS]** be/[VB] [Z] he/[PRP] ./[.] | ez/[FN_NM] a/[DET] barátnő/[FN] [PSe1] **ház/[FN] [PSe3]** ./[PUNCT] |

**7. táblázat: Példamondatok a *SZÓALAPÚELEMZETT* rendszer összekötött morfémáira**

DOI:10.15774/PPKE.ITK.2015.004

A felmerülő adathiány-probléma kiküszöbölésére mind a forrás-, mind a célnyelvi oldalon a morfémákat leválasztottam a szótőről, ily módon létrehoztam egy **morfémaalapú hibrid fordítórendszert**. Az SMT fordítás után az egymás után lévő célnyelvi morfoszintaktikai címkéket újraegyesítem a szótővel, amiből a morfológiai generátor előállítja a szóalakokat. Ezt a rendszert *MORFÉMAALAPÚ* rendszernek fogom nevezni. A rendszer a toldalékmorfémák helyett az azoknak megfelelő morfoszintaktikai címkékkel dolgozik. Erre azért van szükség, mivel ezzel kiküszöbölhető a morfémák között fennálló homonímia okozta többértelműség (például a múlt idő jele és a tárgyrag esetén). A szavak morfémákra bontásának segítségével sikerült elérni, hogy az angol és a magyar mondatokban a szavak száma megegyezik (8. táblázat). A *MORFÉMAALAPÚ* rendszer legnagyobb erőssége, hogy megteremti az SMT rendszer működéséhez szükséges ideális körülményeket (monoton fordítás, azonos/hasonló szószámú nyelvpárok) ezáltal jó minőségű morféma szintű fordítás állítható elő. A morfológiai generátor integrálásával pedig sikeresen állítható elő az agglutináló nyelv szóalakja. A módszer másik fontos tulajdonsága, hogy a tanítóhalmazban szereplő szavak esetén képes produktívan előállítani sokmorfémás szerkezetek felszíni alakját még akkor is, ha ez az alak nem szerepelt a korpuszban.

| | |
|---|---|
| be/[VB] [Z] he/[PRP] there/[EX] a/[DT] bath/[NN] the/[DT] **house/[NN] in/[IN]** ./[.] | van/[IGE] [e3] egy/[DET] kád/[FN] a/[DET] **ház/[FN] [INE]** ./[PUNCT] |
| leave/[VB] will/[MD] i/[PRP] the/[DT] **house/[NN] acc/[ACC]** !/[.] | elhagy/[IGE] [e1] [Def] a/[DET] **ház/[FN] [ACC]** !/[PUNCT] |
| go/[VB] [Past] he/[PRP] the/[DT] **house/[NN] into/[IN]** !/[.] | bemegy/[IGE] [Past] [e3] a/[DET] **ház/[FN] [ILL]** !/[PUNCT] |
| go/[VB] [Past] he/[PRP] **house/[NN] her/[PRP$] into/[IN]** !/[.] | bemegy/[IGE] [Past] [e3] a/[DET] **ház/[FN] [PSe3] [ILL]** !/[PUNCT] |
| this/[DT] girlfriend/[NN] my/[PRP$] **house/[NN] 's/[POS]** be/[VB] [Z] he/[PRP] ./[.] | ez/[FN_NM] a/[DET] barátnő/[FN] [PSe1] **ház/[FN] [PSe3]** ./[PUNCT] |

**8. táblázat: Példamondatok a *MORFÉMAALAPÚ* rendszer különálló morfémáira**

A modell hátránya, hogy a Mosesbe integrált EM-alapú szóösszekötő rendszer (GIZA++ [7], 2.2.3.4. fejezet) számára a funkciószavak és toldalékok pontos párosítása nehéz, mivel ugyanaz a funkciószó több különböző szóhoz tartozhat, ami bizonytalanná teszi a szóösszekötő munkáját. Például ha egy mondatban két főnév található és az egyik közülük többes számú, akkor előfordulhat, hogy a többes számot jelölő [PL] címke a másik főnévhez kerül. A nagyon gyakori toldalékok összekötésének nehézségét mutatja a tanítóhalmazon végzett szóösszekötés, miszerint a korpuszban szereplő [PL] címkék 39%-a maradt összekötés nélkül, 13% nem lett főnévhez csatolva a nem monoton módon történő összekötés miatt, míg 1% több morfémához (akár 8-hoz is) lett kötve. Ennek következtében az SMT rendszerben használt algoritmusok zajos frázistáblát építenek.

DOI:10.15774/PPKE.ITK.2015.004

További nehézség, hogy mivel a morfémák külön fordítási egységként vannak kezelve, ezért a rendszer számára egyenrangúak a lemmákkal. A lemmák így elvesztik fölérendeltségüket a toldalékokkal szemben. Emiatt az elválaszthatatlan morfémasorozatok esetében (mint például főnév+többes szám) – melyeknek mindig együtt kéne mozogni – a dekóder szétszórja ezeket a toldalékokat a mondatban. Ezen kívül előfordulnak olyan hibák, hogy a frázishatárokon lévő „morfémazaj" miatt a nyelvmodell gyakran kiszűri a jó fordításokat, és emiatt egy-egy lemma elveszik. Ilyenkor hiába sikerül a rendszernek a toldalékot helyesen lefordítani, a fordításban nem jelenik meg. Ez a hiba a kiértékelés során nagymértékben rontja a fordítás minőségét.

Munkám során olyan módszert kerestem, mely biztosítani tudja, hogy egyetlen szótő se vesszen el, emellett azonban elő tudja állítani a tanítóanyagban nem szereplő toldalékolt alakot. A **faktoros szóalapú fordítási modell**t találtam alkalmasnak arra, hogy megoldást nyújtson a felmerült problémákra. A faktoros modell – továbbiakban *FAKTORALAPÚ* – lényege, hogy a fordítási feladatot több független fordítórendszer kombinálásával valósítja meg. Ahogy azt a 2.2.4.3. fejezetben bemutattam, a faktoralapú rendszer több független fordítórendszert kapcsol össze oly módon, hogy a fordítási kimenetükből generáló lépéssel állítja elő a szóalakokat. A létrehozott *FAKTORALAPÚ* rendszerben két fordítórendszer működik párhuzamosan. Az első a forrásnyelvi szavak lemmáját fordítja le, míg a második a *SZÓALAPÚELEMZETT* rendszerben bemutatott morfémákra bontott szófaji címkék között végez fordítást. A folyamat során a szófaji egyértelműsítés végett a szó lemmájához kapcsolom annak fő POS címkéjét (9. táblázat). Ha a fordítást két külön kifejezésalapú SMT rendszerrel végezném, akkor nem tudnám megoldani, hogy ugyanazon forrásnyelvi mondatnak a fordítása során a két rendszer kimenetén azonos szószámú fordítás szerepeljen. Ezzel ellentétben a faktoralapú SMT képes biztosítani a két párhuzamosan futó rendszer kimenetén a célnyelvi mondatban a szavak számának egyenlőségét. Az általam létrehozott rendszer a gyakorlatban elterjedt megközelítéstől eltér abban, hogy elhagyja a végső szóalakot generáló lépést, és ezt a morfológiai generátorra bízza. Az így felépített *FAKTORALAPÚ* rendszer egyesíti a *MORFÉMAALAPÚ* és a *SZÓALAPÚELEMZETT* rendszerek előnyeit, mivel a módszer képes a tanítóhalmazban nem látott szavak előállítására is, ezzel párhuzamosan megőrzi a lemmát a szerkezet fejeként. Cserébe a *FAKTORALAPÚ* rendszer fordítása nagyon lassú.

A morfológiai elemzés, valamint a szótövesítés bevezetésével az angol mondatban elveszik a segédige számára vonatkozó információja, mivel a *was* és *were* is ugyanazt a szófaji címkét kapja, ami azonban nem tartalmazza azt, hogy egyes vagy többes számról van-e szó. Ennek érdekében módosítottam a szófaji címkéket egyes számban VBZ, többes számban VBP alakra. Így a *was/[VBD]* helyett a *be/[VBZ][PAST]*, a *were/[VBD]* helyett pedig a *be/[VBP][PAST]* címkéket használtam.

DOI:10.15774/PPKE.ITK.2015.004

| | |
|---|---|
| be/[VB]\|[VB][Z]_he/[PRP] there/[EX]\|[EX] a/[DT]\|[DT] bath/[NN]\|[NN] the/[DT]\|[DT] **house/[NN]\|[NN]_in/[IN]** ./[.]\|[.] | van/[IGE]\|[IGE][e3] egy/[DET]\|[DET] kád/[FN]\|[FN] a/[DET]\|[DET] **ház/[FN]\|[FN][INE]**[1] ./[PUNCT]\|[PUNCT] |
| leave/[VB]\|[VB]_will/[MD]_i/[PRP] the/[DT]\|[DT] **house/[NN]\|[NN]_acc/[ACC]** ./[.] | elhagy/[IGE]\|[IGE][e1][Def] a/[DET]\|[DET] **ház/[FN]\|[FN][ACC]** ./[PUNCT]\|[PUNCT] |
| go/[VB] [Past] he/[PRP] the/[DT] **house/[NN]\|[NN]_into/[IN]** !/[.] | bemegy/[IGE]\|[IGE][Past][e3] a/[DET]\|[DET] **ház/[FN]\|[FN][ILL]** !/[PUNCT]\|[PUNCT] |
| go/[VB]\|[VB][Past]_he/[PRP] **house/[NN]\|[NN]_her/[PRP$]_into/[IN]** ./[.]\|[.] | bemegy/[IGE]\|[IGE][Past][e3] a/[DET]\|[DET] **ház/[FN]\|[FN][PSe3]\|[ILL]**[2] ./[PUNCT]\|[PUNCT] |
| this/[DT]\|[DT] girlfriend/[NN]\|[NN]_my/[PRP$] **house/[NN]\|[NN]_'s/[POS]** be/[VB]\|[VB][Z]_he/[PRP] ./[.]\|[.] | ez/[FN_NM]\|[FN_NM] a/[DET]\|[DET] barátnő/[FN]\|[FN][PSe1] **ház/[FN]\|[FN][PSe3]** ./[PUNCT]\|[PUNCT] |

**9. táblázat: Példamondatok a *FAKTORÁLAPÚ* rendszerből;**
**szerkezete:** lemma/[fő POS címke] | lemmához és a toldalékaihoz tartozó POS címkék

## 4.2 Az eredmények ismertetése

A létrehozott rendszerek áttekintése után bemutatom azok eredményességét, illetve összehasonlítom őket más létező rendszerek eredményeivel (Google Translate [80], Bing Translator [81], MetaMorpho [82]). A kiértékelés során a bemutatott rendszereket különféle beállítások mellett vizsgáltam meg. Egyrészt mindegyik rendszert teszteltem szórendi átrendezéssel vagy anélkül (az átrendezés kódja a rendszer nevében *ÁT*). Másrészt az SMT dekóder lehetőségeit vizsgáltam különböző paraméterek mellett (a *T0* kóddal jelölt a monoton dekódolás, azaz amikor nincs dekóder általi átrendezés; és *T6,* amikor maximum 6 szó távolságban rendezhet át a dekóder).

Teszteléseim során a Hunglish korpuszt [41], [42] használtam, amit a 3.2.1. fejezetben bemutatott arányban bontottam fel. A korpusz morfémákra bontott, valamint a 3.2.1. fejezetben bemutatott módon előfeldolgozott változatát használtam. Az összehasonlítást az alaprendszer ismertetésével kezdem (továbbiakban *SZÓALAPÚ*-nak nevezem), ami egy szószintű kifejezésalapú SMT fordítórendszer. Ebben az esetben a korpuszon előfeldolgozási lépésként csak tokenizálást és

[1] Ine=Inessive='in'
[2] 1PxS1=Possessor:1Sg='my'

DOI:10.15774/PPKE.ITK.2015.004

kisbetűsítést végeztem. A frázistáblában a maximális frázishosszt 7-re állítottam, 5-gramokból álló nyelvmodellt építettem, valamint a szórendbeli különbséget büntető modell maximális átrendezési távolságának a 6 értéket adtam. Emellett monoton fordítást alkalmazó dekóderrel is végeztem méréseket. Az előző fejezetben ismertetett *SZÓALAPÚELEMZETT*, *MORFÉMAALAPÚ* és *FAKTORALAPÚ* modellek esetén is létrehoztam azok 4-4 variációját. Az így kapott 16 rendszer és a további 3 kereskedelmi rendszer w-BLEU és mm-BLEU értékeit a 10. táblázatban foglaltam össze.

Az eredményekből látható, hogy az *SZÓALAPÚ_T6* nevű rendszer 13,83% szószintű, valamint 59,32% morfémaszintű BLEU pontosságot ért el. A vizsgálat során az alaprendszeren is végrehajtottam a forrásoldali átrendező szabályaimat. Az így létrejött *SZÓALAPÚ_ÁT_T0* és *SZÓALAPÚ_ÁT_T6* rendszerben a szavakat nem bontottam morfémákra, hanem az eredeti szóalakot használtam (nem történt meg a szavak morfológiai elemzése). Megfigyelhető, hogy a forrás oldali átrendezés hatására javult a w-BLEU érték (14,25% és 14,85%), azonban az mm-BLEU értéke csökkent (58,06% és 57,79%).

A 10. táblázat alapján kijelenthető, hogy minden rendszerváltozat esetén jobb minőségű fordítás érhető el akkor, ha az SMT dekóderének engedélyezzük a fordításidei átrendezést, szemben a monoton dekódolással. Továbbá az is kijelenthető, hogy az általam alkalmazott forrásoldali átrendező szabályok szintén minden rendszer esetén javítottak a fordítás minőségén.

A morfológiailag módosított rendszerek minőségének ismertetése: az első ilyen a *SZÓALAPÚELEMZETT* rendszer, ami önmagában 12,89% w-BLEU értéket ért el, ami az átrendezések után 13,05% w-BLEU értékre növekedett. Habár a BLEU értékekből nem nyilvánvaló, mivel csak 0,51% a minőségcsökkenés az alaprendszerhez képest, de a *SZÓALAPÚELEMZETT* rendszer az emberi kiértékelés (13. táblázat) alapján sokkal rosszabb bármely más rendszerénél, ami a nagyszámú „lefordíthatatlan agglutináló angol szóalaknak” köszönhető. Amiatt, hogy a funkciószavakat toldalék gyanánt hozzákapcsoltam a szerkezet fejeléméhez, rendkívül megnőtt a tanítóanyagban nem szereplő szavak (OOV – out-of-vocabulary) száma a teszthalmazban.

DOI:10.15774/PPKE.ITK.2015.004

| Rendszerek | w-BLEU | mm-BLEU |
|---|---|---|
| *SZÓALAPÚ_T0* | 13,56% | 58,93% |
| *SZÓALAPÚ_T6* | 13,83% | 59,32% |
| *SZÓALAPÚ_ÁT_T0* | 14,25% | 57,79% |
| *SZÓALAPÚ_ÁT_T6* | **14,85%** | 58,06% |
| *SZÓALAPÚELEMZETT_T0* | 12,75% | 56,10% |
| *SZÓALAPÚELEMZETT_T6* | 12,89% | 56,84% |
| *SZÓALAPÚELEMZETT_ÁT_T0* | 13,02% | 57,10% |
| *SZÓALAPÚELEMZETT_ÁT_T6* | 13,05% | 57,21% |
| *MORFÉMAALAPÚ_T0* | 11,69% | 63,18% |
| *MORFÉMAALAPÚ_T6* | 12,19% | 63,87% |
| *MORFÉMAALAPÚ_ÁT_T0* | 12,01% | 64,24% |
| *MORFÉMAALAPÚ_ÁT_T6* | 12,22% | **64,94%** |
| *FAKTORALAPÚ_T0* | 9,70% | 56,01% |
| *FAKTORALAPÚ_T6* | 9,84% | 57,09% |
| *FAKTORALAPÚ_ÁT_T0* | 10,50% | 59,56% |
| *FAKTORALAPÚ_ÁT_T0_FIX* | 10,64% | 60,28% |
| *FAKTORALAPÚ_ÁT_T6* | 10,78% | 59,97% |
| *FAKTORALAPÚ_ÁT_T6_FIX* | 10,88% | 60,83% |
| Google Translate [80] | **15,68%** | 55,86% |
| Bing Translator [81] | 12,18% | 53,05% |
| MetaMorpho [82] | 6,86% | 50,97% |

**10. táblázat: A morfológiai módosításokat tartalmazó fordítórendszerek fordítási eredményei**

Az összes *MORFÉMAALAPÚ* rendszer mm-BLEU tekintetében jobb eredményt ért el bármely *SZÓALAPÚ* rendszerhez képest. A legjobb mm-BLEU pontot, a 64,94%-ot *MORFÉMAALAPÚ_ÁT_T6* implementáció ért el, annak ellenére, hogy w-BLEU értéke alulmaradt az előzőekhez képest (12,22%). A *MORFÉMAALAPÚ* modell esetében megfigyelhető, hogy mivel a fordítás során a fordítandó kifejezések alapegységei a morfémák, ezért ezek előfordulhatnak rossz szó mellé kerülve is, hiszen ugyanaz a toldalékmorféma egy mondaton belül többször is előfordulhat. A dekóder fordítási modellje pedig több, az adott mondatban akár nem megfelelő szóhoz is hozzákapcsolhatja ezeket. Így a generálás során a toldalékok nem feltétlenül kerülnek a megfelelő szóra, illetve a kívánt helyen nem jelennek meg. A morfémaalapú fordítás alapvető problémát jelent már a tanítóanyagban szereplő szóösszerendelések (illetve a mi esetünkben morfémaösszerendelések) számára is, amelyek alapján a fordítóban használt frázistábla készül, ugyanis a hosszabb mondatokban ugyanaz a funkcionális morféma számos példányban előfordulhat, és a rendszerben használt Giza++ szópárosító algoritmus (2.2.3.4. fejezet) ezeket hibásan párosítja össze.

Ezt a hibajelenséget javítani tudtam azzal, ha a monoton dekódolás helyett megengedtem a dekódolás alatti átrendezést, így 1,75% relatív (0,21%-os w-BLEU érték) javulást értem el. A

DOI:10.15774/PPKE.ITK.2015.004

monoton dekódolás megakadályozta a dekódert abban, hogy a magyar mondatokban a komment ige előtti részében helyes szórendet hozzon létre. A mondatok ezen részére ugyanis szigorú szórendi szabályok vonatkoznak, ellentétben a topik és a komment ige utáni részében lehetséges sokkal szabadabb szórenddel. Az általam készített átrendezési szabályok nem tartalmazzák ennek a jelenségnek a kezelését, mivel ezek csak hajszálnyi különbségek alapján ismerhetők fel, és melyeket az eredeti angol mondatból nem lehet megbízhatóan megállapítani. A feladat megoldására a dekóder belső lexikalizált átrendezési modellje próbálja meg – bizonyos mértékben sikeresen – kezelni a problémát.

Munkám során a létrehozott rendszeremet összehasonlítottam elérhető kereskedelmi alkalmazásokkal. A tesztanyagot az SMT-alapú Google Translate [80] és a Bing Translator [81] gépi fordítókkal, valamint a szabályalapú MetaMorpho [82] nevű fordítórendszerrel fordítottam le. A rendszerek összehasonlítása nehéz, mivel az SMT-alapú rendszerek felépítése és működése nem publikus. A 10. táblázatban olvasható eredményekből látható, hogy az általam létrehozott rendszerek morféma szinten jobban teljesítenek az automatikus kiértékeléssel, mint a kereskedelmi társaik. Ez esetlegesen annak tudható be, hogy rendszerem használ morfoszintaktikai előfeldolgozást. A w-BLEU alapján a különböző *SZÓALAPÚ* és *MORFÉMAALAPÚ* rendszerek felülmúlják a Bing Translator és a MetaMorpho rendszerek teljesítményét, valamint a *SZÓALAPÚ_ÁT_T6* rendszer csak 0,83%-kal marad el a Google Translate eredményétől. Eredményeimet pozitívnak tartom a fordításhoz általuk és általam használt erőforrások nagyságrendbeli méretkülönbségeinek tükrében. A MetaMorpho rendszer kirívóan alacsony w-BLEU értéke jól szemlélteti a BLEU metrika azon gyengeségét, hogy a szabályalapú rendszereket alulpontozza.

A 11. táblázatban látható, hogy a bemutatott példa esetében a *MORFÉMAALAPÚ_ÁT_T6* rendszer fordítása mind gördülékenység, mind az eredeti jelentéstartalom megőrzése szempontjából javulást ért el *SZÓALAPÚ_T6* rendszerrel szemben. Emellett megfigyelhető, hogy a referenciafordítás nem a szó szerinti tükörfordítás, emiatt viszont az általam készített rendszereket alulpontozza az automatikus kiértékelés. Az ilyen mondatok miatt a rendszer helytelen fordítási modellt épít a tanítás során.

DOI:10.15774/PPKE.ITK.2015.004

| | |
|---|---|
| eredeti angol mondat | After you were picked up at sea, our listening post in Malta intercepted that fax. |
| *MORFÉMAALAPÚ_ÁT_T6* | Miután felvették magát a tengeren, hallgatta a helyünk, hogy Málta állta ezt a faxot. |
| *SZÓALAPÚ_T6* | Azután, hogy felvette a tengeren, a máltai hallgatta az emelkedő, hogy fax. |
| magyar referencia mondat | Miután önt kihalászták, ezt fogták el egy máltai postán. |

**11. táblázat: Egy példamondat a vizsgált rendszerek fordításából I.**

A harmadik a *FAKTORALAPÚ* rendszer még szórendi átrendezést is alkalmazó változata a *FAKTORALAPÚ_ÁT_T6*, mely w-BLEU és mm-BLEU szempontból is alulmúlta mindkét előző rendszercsoport teljesítményét. A *FAKTORALAPÚ* rendszer elméletben jó megoldásnak tűnik az adathiány-probléma megoldására, de a lexikális és grammatikai faktorok fordítása veszélybe kerülhet a Moses rendszer faktoros modellimplementációja miatt. Ez abban nyilvánul meg, hogy ha egy több faktorból álló szó fordítása során valamely faktor fordítása sikertelen, akkor az egész szót ismeretlen szóként kezeli, függetlenül a többi faktor fordításának sikerétől. Például hiába tudja lefordítani a szó lemmáját helyesen, ha a toldalékot nem, a célnyelvi lemma nem kerül a fordításba, helyette a forrásnyelvi szó marad az SMT kimenetén, ami nagymértékben ront a fordítás BLEU értékén. Egy másik felmerülő probléma, hogy mivel külön egységként fordítja a lemmát és a toldalékokat, azok nem kapcsolódnak egymáshoz, így különböző célnyelvi szósorrend generálódik a két faktor fordítása során. Ennek köszönhetően, ha egy szótövekből álló kifejezés szósorrendje lemma szinten [Det N V], addig a morfoszintaktikai címkék szintjén akár [V Det N] is lehet. Emiatt helytelen struktúrák jönnek létre, például főnév kapja az igei toldalékokat, vagy fordítva. Ebből viszont az következik, hogy hiába vannak külön-külön helyesen lefordítva a lemma és a morfológiai jellemzők. Az így létrejött inkonzisztens szerkezetek megakadályozzák a helyes szóalak generálását. Ez a jelenség a magyar nyelv szórendi sajátosságai miatt elég gyakori; a tesztalmaz mondatainak 21%-át érinti.

A fordítás minőségének javítása érdekében utófeldolgozási lépést iktattam a rendszerbe, amelynek célja a szótövek és morfológiai címkék helyes sorrendjének felállítása a faktoros tanítás végén. A szóösszekötés helyességében bízva minden morfoszintaktikai címke megfelelő pozíciója megtalálható. A fordítás során a Moses rendszer képes megmondani, hogy egy adott forrásnyelvi kifejezésnek mely célnyelvi szókapcsolat felel meg. Azok a szavak, amik egy kifejezésen belül vannak, egy fordítási egységnek tekinthetők. Abban az esetben, amikor egy fordítási egységen belül eltér a szórend a szótő és a toldalékok fordítása között, akkor egy általam készített transzfor-

DOI:10.15774/PPKE.ITK.2015.004

mációs lépés segítségével átrendezem a toldalékok címkéit, hogy a megfelelő szótő mellé kerüljenek. A két faktor újrapárosítása után a rendszer egyesíti azokat, és a morfológiai generátor segítségével megadja a végső szóalakokat. Az utófeldolgozó lépéssel rendelkező faktoros rendszereket a *FIX* névvel látom el (10. táblázat).

| | |
|---|---|
| eredeti angol mondat | at my request the ceremony was postponed for a year . |
| *MORFÉMAALAPÚ_ÁT_T6* | *kérésemre halasztották a szertartást .* |
| *SZÓALAPÚ_T6* | *az én kérésemre a szertartás volt .* |
| *FAKTORALAPÚ* | *kérésemre elhalasztották a szertartást egy évre .* |
| magyar referencia mondat | a szertartást kérésemre egy esztendővel elhalasztották. |

**12. táblázat: Egy példamondat a vizsgált rendszerek fordításából II.**

A 12. táblázatban látható példamondat az angol passzív szerkezet fordítását szemlélteti. Láthatjuk, hogy ebben az esetben a *FAKTORALAPÚ* rendszer majdnem tökéletes fordítást adott a *SZÓALAPÚ* és a *MORFÉMAALAPÚ* rendszerekkel ellentétben.

A példamondat alapján belátható, hogy az egyszerű BLEU értékek alapján történő rangsorolás nem feltétlenül felel meg az emberi kiértékelésnek. Ez volt az oka annak, hogy az automatikus kiértékelést sokáig nem alkalmazták a különböző fordítórendszerek a Workshops on Statistical Machine Translation (WMT) [83] által történő hivatalos rangsorolásánál. A WMT egy olyan workshop, ahol a különböző fordítórendszerek versenye történik, adott korpusz alapján. Emiatt munkám során az egyszerű BLEU-érték mellett az WMT által alkalmazott rangsorolási sémával is kiértékeltem az egyes rendszerek eredményeit.

A teszthalmazból 300 véletlenszerűen kiválasztott mondat került emberi kiértékelésre. Öt annotátor rangsorolta a fent leírt rendszer által generált fordításokat, összehasonlítva azokat az eredeti referenciafordítással olvashatóság, gördülékenység és tartalomhűség szempontjából. Minden mondat esetében minden annotátornak öt rendszer fordítását kellett véletlenszerűen megjelenített sorrendben értékelnie. A rendszerek egy normalizált érték alapján lettek rangsorolva, amit egy adott rendszer többi rendszerhez viszonyított szereplése alapján számoltam. A rendszer szegmensenként annyi pontot kap, ahány rendszernél jobb az összehasonlítás során. Az emberi kiértékelés összesített eredményét a 13. táblázat mutatja be, melyből látható, hogy a *MORFÉMAALAPÚ_ÁT_T6* rendszer a tesztelt rendszerek 55,60%-ánál ért el jobb eredményt.

A fordítórendszerek kimenetének manuális vizsgálata során kiderült, hogy a morfológiai és szintaktikai információkkal is dolgozó rendszer jobban fel tudja térképezni az eredeti szöveg nyelvtani összefüggéseit, és sikerrel alkalmazza ezeket a megfelelő szóalak előállítására a fordítás

DOI:10.15774/PPKE.ITK.2015.004

során. A szabályalapú átrendezés javulást ért el nyelvészetileg gazdagabb modellek alkalmazása esetén is. Az alaprendszernél rosszabban teljesítő modelleknek a szóalapú átrendezéses megoldások bizonyultak, legfőképpen azok, amelyek az angol nyelvet agglutináló jellemzőkkel ruházták fel; ez a rossz teljesítmény azonban nem volt meglepő. Azt, hogy a BLEU érték által felállított rangsor mennyire nem felel meg az emberi kiértékelésnek, alátámasztják a következő esetek: egyrészt rendszereim közül BLEU pontozás szempontjából legjobb eredményt a *SZÓALAPÚ_ÁT_T6* rendszer ért el, habár az annotátorok ezt a morfológiai változtatásokat is alkalmazó rendszerek után sorolták. Másrészt pedig az olvasó számára legjobb eredményt elérő kereskedelmi MetaMorpho rendszer kapta a legalacsonyabb BLEU értéket.

| **Rendszer neve** | **Emberi kiértékelés** | **w-BLEU** | **mm-BLEU** |
|---|---|---|---|
| referenciafordítás | 88,33% | | |
| MetaMorpho [82] | 76,30% | 6,86% | 50,97% |
| Google Translate [80] | 72,80% | 15,68% | 55,86% |
| Bing Translator [81] | 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% |

**13. táblázat: A morfológiai módosításokat tartalmazó fordítórendszerek emberi kiértékelése**

A 13. táblázat bemutatja a kereskedelmi rendszerek fordításainak annotátorok által értékelt minőségét. Legjobb eredményt a MetaMorpho érte el, hiszen ez egy magyar nyelvre optimalizált fordítórendszer, ami nyelvtani szabályok alapján nagy pontossággal generálja a megfelelően toldalékolt szóalakokat. Ez a szubjektív értékelők számára nagyban javítja a fordítás olvashatóságát. A legjobb rendszerem minősége is csak megközelíteni tudja a statisztikai alapú kereskedelmi rendszerek eredményét. Munkám során az elérendő reális cél nem a kereskedelmi rendszerek felülmúlása volt, hanem a meglévő erőforrások melletti minőségjavulás elérése az alaprendszerhez képest. Ezt a célt sikerült teljesítenem az automatikus és emberi kiértékelések alapján is.

Az eredmények kiértékelése során megfigyelhető továbbá, hogy a BLEU pontozáshoz használt referenciafordítás az emberi kiértékelés szerint 21,67%-ban rosszabb a rendszer által készített fordításoknál. A jelenség alaposabb vizsgálatánál kiderült, hogy ez annak köszönhető, hogy a párhuzamos korpusz angol és magyar oldala nem minden esetben helyes fordítása egymásnak. Ez nemcsak amiatt jelent problémát, hogy bizonyos kifejezéseket hibásan tanul meg, hanem az au-

DOI:10.15774/PPKE.ITK.2015.004

tomatikus kiértékelés során is sokszor hibás referenciafordításhoz végzi a hasonlítást. Ezért bár az eredeti mondat fordításának megfelel a létrejött fordítás is, ezekben az esetekben semmiképpen nem hasonlítható a referenciához. Levonható az a következtetés, hogy kisméretű, nem megfelelő minőségű korpusz használatával nem lehet jó minőségű fordítórendszert összeállítani. Érdemes még megjegyezni azt is, hogy az emberi kiértékelés esetében a rendszerek minősítése jelenős szórást mutatott. Ezt az is tükrözi, hogy ugyanazon rendszer különböző személyek rangsorolásában eltérő helyen szerepelt (volt rá példa, hogy míg az egyik 92,98%-ra értékelte, a másiknál csak 75,29%-os eredményt ért el).

## 4.3 Kapcsolódó munkák, előzmények

Bisazza és Federico [84] a morfológiailag gazdag török nyelvről fordítottak a morfológiailag egyszerűbb angolra. A forrásnyelvi oldalon előfeldolgozást végeztek, mégpedig a török szavakat morfológiailag elemezték, így létrehozva belőlük a lemmákat és a hozzájuk tartozó toldalékokat. Ezzel elérték, hogy az angol oldalhoz hasonló szószámú mondatot kaptak, amivel megkönnyítették a Giza++ [7] működését. Bisazza és Federico [84] a morfológiailag gazdagabb agglutináló nyelvről fordítottak egyszerűbbre, ezért reguláris kifejezések segítségével egyszerűsítették a forrásnyelvi oldalt. Hasonlóképpen jártak el Mermer et al. [85] is, akik a török és az arab nyelvekről fordítottak angolra. Munkájuk során felügyelet nélküli gépi tanulási módszerrel végeztek morfológiai elemzést a forrásoldalon. Hasonló kutatást végzett héber-angol nyelvpárra Singh és Habash [86], akik különböző technikák segítségével morfológiailag elemezték a héber szavakat, mint például reguláris kifejezéssel, gépi tanulási módszerrel és szabályalapú morfológiai elemzővel. Az említett munkák azonban nem foglalkoztak a célnyelvi oldal szóalakjának előállításával. Emiatt nem volt szükségük az angol oldal módosítására, nem alkalmaztak átrendezési szabályokat, illetve nem foglalkoztak a két nyelv közötti szó- és morfémaszintű különbségekkel sem.

Ramasamy et al. [87] angol és tamil nyelvek között készítettek gépi fordítórendszert. Munkájuk során a célnyelvi szavakról leválasztották a toldalékokat, annak érdekében, hogy az angol funkciószavaknak meglegyen a megfelelő fordítása. Yeniterzi és Oflazer [37] az angol-török fordítót fejlesztettek. A forrásnyelvi oldalon szófaji egyértelműsítést, függőségi elemzést és szórendi átrendezést végeztek, míg a célnyelvi oldalon csak morfológiai elemzést alkalmaztak. Rendszerük a faktoralapú dekódolás során először a forrásnyelvi szóalak fordítását végzi, azonban ha ez nem sikerül, akkor a lemma és a címkék alapján próbálja előállítani a végleges szóalakot. A módszer gyengesége, hogy a statisztikai alapú dekóder nem képes azoknak a szavaknak a szóalakját előállítani, amelyekre nem látott példát a tanítóanyagban. Oflazer és Durgar El-Kahlout [88], [89] is az

DOI:10.15774/PPKE.ITK.2015.004

angol-török fordítással foglalkozott. Mindkét oldalon morfoszintaktikai elemzést végeztek, és a toldalékok helyett azok szófaji kategóriáját használták fel a fordítás során. A morfémaalapú fordítás legjobb kimeneti javaslatait egy szóalapú nyelvmodellel újrarangsorolták, viszont a pontos szóalakot nem generálták ki; a szótövet és a morfoszintaktikai címkéket hagyták meg a rendszer kimenetén. Luong et al. [90] az angol és a finn nyelv közti fordítást próbálták tökéletesíteni. Ehhez a forrás- és célnyelvi oldalon is lemmára és morfémákra bontották a szóalakot. A morfémákat leválasztották a szótőről, illetve prefix és szuffix jelöléssel látták el azokat. Az így létrehozott elemzett szövegekből párhuzamosan tanítottak be szó- és morfémaalapú fordítási modellt, melyekből létrehoztak egy kombinált – csak grammatikus összetételekből álló – morfémaalapú fordításhoz felhasználható modellt. A fenti megvalósítással ellentétben a toldalékmorfémák helyett azok morfoszintaktikai címkéit alkalmaztam a fordítás során, ezért sokkal kisebb a többértelműség esélye (például a *–t* a múlt idő jeleként vagy tárgyragként szerepel). Munkám során a kombinált fordítási modell helyett a faktoros tanítással értem el, hogy a lemma és toldalékegységek helyesen őrződjenek meg. Yeniterzi és Oflazer [37] munkájával ellentétben az ismeretlen szavak alakját nem egy statisztikai dekóderrel állítottam elő, hanem morfológiai generátorral.

Clifton és Sarkar [29] az angol és finn nyelvek közt végeztek fordítást oly módon, hogy felügyelet nélküli gépi tanulási módszerrel morfológiailag elemezték a finn szöveget; míg az angol oldalon semmilyen előfeldolgozást nem végeztek. Rendszerükbe utófeldolgozó modulként a tanítóhalmazon tanított környezetfüggetlen nyelvtanon alapuló morfológiai generátort integráltak, melynek segítségével a célnyelvi oldal szóalakjait állították elő. Nem fektettek hangsúlyt a két nyelv szórendkülönbségéből fakadó problémák kezelésére, továbbá azokra az esetekre, melyeket az angol nyelv nem jelöl külön (ilyen például az accusativus vagy dativus esetek). Módszerük hátránya, hogy az általuk használt felügyelet nélküli tanulással tanított generátor korpuszfüggő, míg – az általam is alkalmazott – szabályalapú generátor a szóalakot pontosabban képes előállítani.

## 4.4 Összefoglalás

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.

Ebben a fejezetben bemutattam egy olyan hibrid statisztikai gépi fordítórendszert, amely morfológiai generátor segítségével állítja elő a célnyelvi szavak ragozott alakját. A morfológiai generátor a statisztikai alapú dekóderrel szemben nagy pontossággal képes előállítani olyan szóalakokat is, amelyek nem szerepeltek a tanítóhalmazban. Az általam létrehozott rendszer a szóala-

DOI:10.15774/PPKE.ITK.2015.004

pú fordítással ellentétben morfológiailag elemzett forrás- és célnyelvi szövegeken dolgozik. A homonímia kezelése érdekében a toldalékmorfémák helyett az azoknak megfelelő morfoszintaktikai címkesorozatot használtam. Több morfémaalapú fordítás (szó-, morféma- és faktoralapú fordítási modellek) segítségével megoldottam az angol és a magyar mondatok között jelentkező szószámkülönbségből 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 rosszabb minőségű fordítást. Ezzel bebizonyosodott, hogy a szóharmonizáció hatására az emberi kiértékelés számára jobb minőségű rendszereket hoztam létre a tisztán statisztikai alapon működőkkel szemben.

A munkám során létrehozott hibrid fordítórendszer nemcsak hazai, hanem nemzetközi viszonylatban is egyedülálló, mivel a szóalakot morfológiai generátor segítségével állítja elő, és ezzel javít az eddig létező rendszerek eredményességén. Továbbá egyedi megoldásnak számít, hogy a faktoros fordítás során nem egy kimeneti faktor jelenik meg, hanem egy jellemzővektor, amely a lemmából és a toldalékokból áll. Ezzel biztosítom, hogy a szótövek a toldalékoktól függetlenül, de azokkal összehangolva kerüljenek fordításra.

Kapcsolódó tézisek:

**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.**

**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 tézisekhez kapcsolódó publikációk: [Laki_1], [Laki_4], [Laki_8]

DOI:10.15774/PPKE.ITK.2015.004

# 5 Statisztikai gépi fordítórendszer minőségének javítása pontosan fordított rövid kifejezések segítségével

Az eddig ismertetett fordítórendszerek kiértékelésénél megfigyelhető, 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. Ahogy azt a II. fejezetben a 6. ábra bemutattam, előfordulhat, hogy a morfémaalapú rendszer nem a megfelelő főnévi frázishoz köti a funkciószavakat. Erre egy másik példa a *„The dogs living in the house eat the bones from the fridge in the kitchen”* mondat fordítása *„A házban élő kutyák megették a hűtőből a csontokat a konyhában*”, ahol a szóösszekötő számára nehéz feladat eldönteni, hogy melyik *in* funkciószót kösse a *ház* és melyiket a *konyha* szavakhoz. A rossz minőségű szóösszekötő hatása a fordítási folyamat további modelljeiben is megjelenik, és ront a végső fordítás eredményességén.

A szóösszekötő gyengeségeinek (2.2.3.4. fejezet) kiküszöbölése érdekében a tanítóhalmazt rövid, pontos fordítású kifejezéspárokkal egészítettem ki. A kiinduló feltevés szerint, az így kiegészített korpuszban a kifejezések pontos fordítása nemcsak segít a pontosabb szóösszekötéseket létrehozni a mondatban, hanem csökkenti a lefordítatlan szavak számát is.

## 5.1 Felhasznált erőforrások

A feladat megoldásához egy egyszerű angol–magyar szótárat használtam [91], melyet először átalakítottam oly módon, hogy egy kifejezésnek csak egyetlen megfelelője legyen. Így 344 924 darab kifejezéspárt kaptam. A fordítórendszerhez szükséges tanítóhalmazt pedig a Hunglish korpusz [50] két aldomainjéből építettem fel, a Literature és a Magazines nevű részekből (a továbbiakban LitMag). A LitMag korpusz 654 939 mondatot és 9 425 911 szót tartalmaz.

## 5.2 Az eredmények bemutatása

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 romlott a nyelvi modell minősége. Ennek érdekében meg kellett találni azt a mértéket, hogy hányszor éri meg a szótárat hozzáfűzni a kor-

DOI:10.15774/PPKE.ITK.2015.004

puszhoz. Ezt a küszöbértéket empirikus úton határoztam meg oly módon, hogy az eredeti korpuszhoz egyszer, kétszer, háromszor, négyszer és ötször hozzáadtam a kétnyelvű szótárat. A rendszerek eredményeit a következő táblázat (14. táblázat) 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: | 10,86% |

**14. táblázat: Különböző rendszerek BLEU-eredményei**

A 14. táblázatból látszik, hogy az *ALAPRENDSZER* (10,85% BLEU) értékéhez képest az *1XSZÓTÁR* behelyezésével 3,04%-os relatív – 0,33% BLEU – javulás figyelhető meg, mely mértéke a behelyezett szótárak számától függően folyamatosan csökken. A BLEU érték azért az első esetben a legjobb, mert a szótár mérete összemérhető az eredeti korpusz méretével (fele az eredeti korpusznak), emiatt annak ismétlése viszonylag hamar eltolja a súlyokat. A teszthalmazból kiválasztott példamondat fordításait a 15. táblázat tartalmazza. Az első sorban az eredeti angol mondat olvasható, a másodikban ennek a referenciafordítása; továbbiakban pedig az alaprendszer, illetve a több szótár integrálásával kiegészített SMT rendszerek fordításai találhatók.

Rögtön az első kifejezés elemzésénél feltűnik az *i wonder* szókapcsolat fordításában észlelhető eltérés. Mind az *ALAPRENDSZER*, mind a legjobb eredményt nyújtó első rendszer *(ALAP+1XSZÓTÁR) csak tudnám*-ra, míg a többi a *kíváncsi vagyok*-ra fordítja. Annak ellenére, hogy mind a két fordítás helyes, az automatikus kiértékelővel mégis más eredményt kaptam, 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 fordításhoz képest. A legnagyobb probléma itt is az, hogy mivel ez sem egyezik meg a referenciafordítással, ezért nem kap nagyobb BLEU-értéket.

DOI:10.15774/PPKE.ITK.2015.004

| | |
|---|---|
| Angol referenciafordítás: | " 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 . |

**15. táblázat: A különböző mennyiségű szótár integrálásával készített rendszerek eredményei**

A *said* fordításánál hasonló jelenség figyelhető meg. Az *ALAPRENDSZER kérdezte*, míg a szótáras módszerek a *szólt* fordítást adták. A különbség oka, hogy a hozzáadott szótárban ez volt a megfeleltetése. A példamondat második felének vizsgálatánál látható, hogy az *ALAPRENDSZER* eredménye viszonylag gyenge (*mikor ő az* .). Ez a hibajelenség abból ered, hogy a szóösszekötő a hosszabb mondatok második felét gyakran hozzákapcsolja valamelyik szóhoz, így viszont torzul a fordítási modell. Ebből kifolyólag a dekóder sem tud megbirkózni a hasonló szövegrészekkel, ezért fordulhat elő, hogy a program „összecsapja" a fordítandó mondatok végét. Ezzel szemben a szótáras esetekben megfigyelhető változások bizonyítják a szóösszekötő minőségének javulását. Az *ALAP+1XSZÓTÁR* esetben a rendszer a második tagmondatra jobb fordítást ad, *ALAP+2XSZÓTÁR* esetben megjelenik a *zsibongó tömeg*, *ALAP+3XSZÓTÁR* után pedig a *mikor elindult az összeverődött tömegen* kifejezés lett a rendszer szerinti legjobb fordítás.

A 15. táblázatban szereplő példa a statisztikai gépi fordítórendszerek azon hiányosságát tükrözi, melyet a II. fejezetben már többször említettem; mégpedig 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, 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 köszönhető az, hogy az *into* az első három esetben mintha nem is jelenne meg a fordításban (*tömeg*), a *ALAP+3XSZÓTÁR-as* rendszertől már látható *a tömegen*, ami már ragozott alak ugyan, csak a megfelelő igekötő *(át)* hiányzik róla.

DOI:10.15774/PPKE.ITK.2015.004

Megvizsgáltam a különböző rendszerek 1-9-gramos kifejezésekre vonatkozó BLEU értékeit is (16. táblázat). Megfigyelhető ugyanis, hogy az *ALAPRENDSZER*hez képest 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 *ALAP+1XSZÓTÁR*as rendszer eredménye szinte az összes esetben jobb lett, mint az *ALAPRENDSZER*, tehát ekkor közelítette 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 *ALAP+5XSZÓTÁR* esetben, hogy az 1-gram értéke sokkal magasabb még az *ALAP+1XSZÓRÁR*as rendszerénél is, de már 2-gram esetén alacsonyabb lesz nála, míg 5-gram esetén már az *ALAPRENDSZER*nél is.

| | 1-gram | 2-gram | 3-gram | 4-gram | 5-gram | 6-gram | 7-gram | 8-gram | 9-gram |
|---|---|---|---|---|---|---|---|---|---|
| *ALAPRENDSZER* | 47,05 | 16,29 | 7,07 | 3,54 | 1,94 | 1,14 | 0,74 | 0,57 | 0,46 |
| ***ALAP+1XSZÓTÁR*** | **47,60** | **16,62** | **7,35** | **3,78** | **2,09** | **1,25** | **0,81** | **0,60** | **0,46** |
| *ALAP+2XSZÓTÁR* | 47,55 | 16,46 | 7,25 | 3,75 | 2,02 | 1,19 | 0,75 | 0,57 | 0,43 |
| *ALAP+3XSZÓTÁR* | 47,32 | 16,33 | 7,09 | 3,64 | 1,94 | 1,09 | 0,68 | 0,47 | 0,33 |
| *ALAP+4XSZÓTÁR* | 47,24 | 16,32 | 7,10 | 3,63 | 1,93 | 1,08 | 0,68 | 0,47 | 0,31 |

**16. táblázat: A különböző rendszerek BLEU értékei különböző hosszú kifejezések esetén**

## 5.3 Kapcsolódó munkák

Durgar és Oflazer [89] munkájukban az általuk használt korpuszon betanítottak egy fordítási modellt, melyből a frázistábla legnagyobb valószínűségű kifejezéspárjait visszahelyezték a tanító korpuszba. Ezzel sikerült javítaniuk az angol-török fordítás minőségét. Holmqvist et al. [92] angol és német nyelvpáron végeztek gépi fordítást. Munkájukban a szóösszekötő rendszer elemzése során egészítették ki a korpuszt szószinten összekötött tanítóanyaggal, amit a fordítási modell építése előtt eltávolítottak a rendszerből. Ezzel a módszerrel nem sikerült javulást elérniük. A fenti rendszerekhez képest az általam bemutatott módszer előnye, hogy mivel a tanítóanyaghoz egy szótárt integráltam nemcsak a szóösszekötő rendszer lett pontosabb, hanem az OOV szavak aránya is nagymértékben csökkent. Habash [93] munkája során az ismeretlen szavak fordítását egy szótár segítségével oldotta meg, ahol az OOV szavakhoz tartozó szótárbejegyzésekkel kiegészítette a fordítási modellt. Minden bejegyzéshez egy viszonylag kicsi valószínűséget rendelt, hogy ne rontsa el a frázistábla súlyozását. Azzal, hogy a tanítóhalmazt kétnyelvű szótárral egészítettem ki, a fordítórendszer képes megtanulni a szótárban található kifejezésekhez tartozó helyes súlyozást. Okuma et al. [94] bemutattak egy olyan fordítórendszert, ami a fordítás során a fordítandó mon-

DOI:10.15774/PPKE.ITK.2015.004

datban az ismeretlen szavakat kicseréli egy szófajilag megegyező, tanítóhalmazban szereplő ismert szóra. Az átalakított mondaton elvégzi a fordítást, majd visszahelyezi az ismeretlen szó fordítását – amit a szótár segítségével határoz meg – a célnyelvi mondatba. Az általuk bemutatott rendszer hátránya, hogy agglutináló nyelvek esetén elő kell állítani az ismeretlen szó megfelelően toldalékolt szóalakját – melyet a szótár segítségével nem lehet megállapítani –, így ezekben az esetekben a módszer rendkívül nehezen alkalmazható. Vogel és Monson [95] létrehoztak egy kézi szótárból épített fordítási modellt. A létrehozott fordítási modellt kiegészítették a szótárban szereplő szótövekből automatikusan generált ragozott szóalakokkal. A szótár bejegyzéseihez tartozó valószínűségeket egy párhuzamos korpusz segítségével állapították meg. Végül bemutattak egy olyan rendszert is, ahol a szótárat kiegészítették párhuzamos korpusszal. Velem ellentétben, munkájuk során nem vizsgálták meg, hogy milyen hatással van a fordítás minőségére, ha a szótárt többször is hozzáadják a párhuzamos korpuszhoz.

## 5.4 Összefoglalás

A szóösszekötő a fordítás során sokszor nehezen párosítja az összetartozó kifejezéseket. Ez főleg akkor fordul elő, ha a kifejezések nyelvtanilag különböző szerkezet miatt távol állnak egymástól, vagy nagyon különbözők. A túl hosszú mondatok is nehézséget okoznak a szóösszekötőnek. A probléma megoldására a tanítóhalmazba integráltam egy rövid, pontos fordítású kifejezéspárokból álló szótárat. A rendszer egyedisége, hogy nemcsak az egyszeri hozzáadást vizsgáltam, hanem a rendszert a szótár többszöri integrálásával is teszteltem. A legjobb esetben sikerült 11,18%-os relatív javulást elérni a fordítás minőségében. A szótár többszöri hozzáadása miatt folyamatosan csökkent a BLEU érték. Ennek oka az eredeti szótár relevanciájának csökkenése, illetve a fordítási és nyelvi modellek deformációja. Ezzel ellentétben az emberi kiértékelés számára a hosszabb mondatok fordítása jelentősen javult.

Kapcsolódó tézis:

**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.**

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

DOI:10.15774/PPKE.ITK.2015.004

# III. Statisztikai gépi fordítás alkalmazása teljes morfoszintaktikai egyértelműsítésre

Egy általános szövegfeldolgozási folyamat általában egy bottom-up típusú moduláris felépítésű architektúra, amiben az elemzési lánc a morfológiai elemzéssel kezdődik, és a szemantikai, valamint a diskurzuselemzéssel zárul. A feldolgozási folyamat egyik első eleme a teljes morfoszintaktikai egyértelműsítő rendszer, aminek a feladata, hogy egyértelműen meghatározza a szavak szótövét, és megállapítsa azok morfoszintaktikai címkéit (POS – Part-of-Speech). A morfoszintaktikai, vagy más néven szófaji címkék meghatározása olyan komplex probléma, melyet a különböző, egyre jobban teljesítő implementációk ellenére még mindig nem sikerült tökéletesen megoldani. Az első, erre a célra létrehozott eszközök rendre külön-külön végezték a szófaji címkézést és a szótövesítést. Mivel ezek a kezdeti megvalósítások képezik az újabb rendszerek alapját, így ezek is külön kezelik a két feladatot. Ennek köszönhetően kevés olyan eszköz létezik, amely nagy pontosságú teljes morfoszintaktikai egyértelműsítésre képes annak ellenére, hogy ez a feladat kulcsfontosságú a morfológiailag gazdag nyelvek elemzése során. Továbbá, csak néhány olyan módszer létezik, amely grammatikailag nagyon különböző nyelvek esetében is egyformán magas pontossággal működik. A létező nyelvspecifikus alkalmazások a viszonylag magas pontosságot különböző nyelvfüggő eszközök (lexikon, szabályok, morfológiai elemző) integrálásával érik el, emiatt viszont rendkívül nehezen alkalmazhatók más nyelvekre. A legelterjedtebb ilyen eszközök gépi tanulási módszereken alapulnak. Annak ellenére, hogy ezek a rendszerek nem felügyelt módszerekkel tanulják meg a címkézéshez használt szabályokat, a működésükhöz szükséges nyelvi jellemzőket továbbra is emberi közreműködéssel kell megállapítani. A nehézséget az okozza, hogy az ideális nyelvi jellemzők megtalálása szintén nehéz feladat. Ezzel ellentétben, a statisztikai gépi fordításon alapuló rendszerek előzetes nyelvi tudás nélkül alkalmasak az alapvető fordítási szabályok automatikus felismerésére [96].

Napjaink legjobb minőségű címkézői a legtöbb nyelv esetén elérik a 96-97% pontosságot [97, o. 342], [98]–[101]. A 90% közeli pontosság egyszerű módszerekkel (mindig a tanítóhalmazban leggyakoribb annotációt rendeli a szóhoz) is megvalósítható, ám a maradék 10% eléréséhez

DOI:10.15774/PPKE.ITK.2015.004

már bonyolult metódusok rendszerbe építése szükséges. Elmondható továbbá, hogy a szófaji egyértelműsítés pontossága messze túlszárnyalja a természetesnyelv-feldolgozás többi területén elért eredményeket. Ebből adódóan a szófaji egyértelműsítés feladatát sokan már megoldott problémának tekintik, mivel az elérhető javulás már csak nagyon elenyésző [102], [103]. Ám ha olyan szemszögből közelítjük meg a problémát, hogy ez a lépés csak előfeldolgozás a magasabb szintű nyelvfeldolgozási folyamat számára, akkor az itt elért pár tized százalékos javulás nagyban elősegíti a feldolgozási lánc további moduljainak munkáját.

## 6.1 A teljes morfoszintaktikai egyértelműsítés feladata és nehézségei

A teljes morfoszintaktikai egyértelműsítés feladata, hogy a mondat szavairól egyértelműen meghatározza azok helyes szótövét és morfoszintaktikai elemzését. Gyakorlatilag ez a folyamat a morfológiai elemző egyszerűsített változata, mivel itt nem a szó teljes belső szerkezete térképeződik fel (abszolút tő, toldalékok), csupán az aktuális szövegkörnyezetben esedékes elemzését állapítja meg.

A **szótövesítés**, vagy más néven lemmatizálás, az az algoritmikus folyamat, amelyik meghatározza egy szó szótári alakját. A szótövesítő megvalósítása nehéz feladat agglutináló nyelvek esetén, mivel egy szónak rengeteg szóalakja lehetséges, így a szótövesítés folyamata csak olyan komplex feladatok leküzdésével valósítható meg, mint például a kontextus megértése és a szófajok meghatározása. Léteznek olyan esetek is, amik statisztikai módszerek segítségével nem kezelhetők. Ilyen például a magyar nyelvben az ikes igék szótövesítése, mivel ezek az igék csak az egyes szám 3. személyű ragozott alakjukban térnek el a nem ikes igéktől. Emiatt az ikes igék listájának ismerete nélkül egy statisztikai rendszer képtelen egy adott igéről eldönteni, hogy az egy ikes, vagy egy nem ikes ige.

A morfoszintaktikai egyértelműsítő másik komponense a **szófaji egyértelműsítés**. A szófaji egyértelműsítés az a folyamat, amelyik a szövegben található szavak szófaji többértelműségét feloldja, valamint ellátja a szavakat a megfelelő morfoszintaktikai címkével. Mivel egy szónak több szófaja is lehet, a mondatban elfoglalt szerepe, valamint az általános lexikai jelentése alapján dönthető el, hogy mi a megfelelő szófaji címkéje. Emiatt az egyértelműsítés feladata nem oldható meg lexikonból történő kiválasztással.

A szófaji egyértelműsítő rendszer a szavakat címkékkel látja el, amelyek kódolt formában tárolják a nyelvekben megfigyelhető szófaji osztályokat. A nyelvészek több szempont alapján különböztetnek meg elsődleges szófaji típusokat (úgymint a főnév, az ige és a melléknév), valamint

DOI:10.15774/PPKE.ITK.2015.004

további alkategóriákat. Ezek képezik a lexikális kategorizáció alapját, ám ezen felül a nyelvészeti modellek további másodlagos kategóriákat (elöljárók, határozók stb.), illetve az elsődleges és másodlagos típusokon belül további alkategóriákat is használnak [104], [105]. Az alkategóriák magukba foglalják a morfoszintaktikai különbségeket (mint például a szám, személy, idő, mód) valamint arra szolgálnak, hogy megadják a különböző szintaktikai és szemantikai viselkedési mintákat (főnevek esetén például: konkrét vagy elvont, egyéni vagy kollektív, megszámlálható vagy megszámlálhatatlan stb). Ha a szövegfeldolgozási lánc magasabb szintű rendszerei úgy kívánják, a természetes nyelvi szövegfeldolgozás során használt osztályozás eltér a szigorúan vett nyelvészet-orientált szófaji kategóriáktól.

A teljes morfoszintaktikai egyértelműsítés sokkal nehezebb és bonyolultabb feladat a gazdag morfológiával rendelkező és az agglutináló nyelvek esetén, mint a morfológiailag egyszerűbb nyelveknél [106]. Ez a probléma elsősorban a nagyfokú adathiányból ered, hiszen a nyelv ragozó tulajdonságából adódóan egy szónak rengeteg szóalakja van, amelyek közül nehezebb a megfelelőt kiválasztani. Míg például egy angol szónak körülbelül 4-6 különböző szóalakja lehetséges, addig az agglutináló nyelvek esetében ez a szám elérheti akár a több százat is. Ha például egy angol és egy magyar nyelven működő statisztikai rendszert szeretnénk egyforma számosságú szófaji címkekészlettel betanítani, akkor ehhez magyar nyelv esetén az angolnál 8-10-szer nagyobb tanítókorpuszra van szükség [107].

Az egyik legtöbbet vitatott kérdés a címkekészlet mérete. Amíg a nem agglutináló nyelvek esetében a címkekészlet nagysága általában 50 és 200 között mozog, addig az agglutináló és erősen ragozó nyelveknél ez a szám nagyságrendekkel magasabb [108]. Ebből adódóan sokkal nagyobb tanítóhalmazokra lenne szükség, hogy a statisztikai modellek építéséhez megfelelő mennyiségű információ álljon rendelkezésre. Ezen nyelvek esetén azonban nincsenek ilyen méretű adathalmazok (annotált nyelvi források).

A másik, gyakran vitatott kérdés ragozó nyelvek esetén a morfológiai rendszerben rejlő többértelműség problémája, ami többek között a morfológiai címkekészlet méretével van összefüggésben. Az, hogy egy adott szó melyik, vagy éppen hány morfológiai osztályhoz van rendelve, a szófaji egyértelműsítő rendszer számára nem mindig eldönthető, mivel a kérdés olyan tényezőktől is függhet, amik a szófaji egyértelműsítés szintjén nem elérhetők. Például a *Pistinek van egy Lacinak adott tolla* mondatban nehezen dönthető el, hogy a *Pistinek* szó birtokos, vagy részes esetű főnév.

A gyakorlatban bebizonyosodott [107], hogy a hibás elemzések nagy részét nem a fenti problémák okozzák. Magyar nyelv esetén ugyanakkora korpuszban előforduló szavak fajtája sok-

DOI:10.15774/PPKE.ITK.2015.004

szorosa az angolban levőknek (például a Hunglish [50] korpuszban 1,5-ször több szóalak van a magyar szövegben, mint az angolban), ami az adathiány-problémához vezet. Ebből következik, hogy magyar nyelv esetén egy adott szó lehetséges címkevalószínűség-eloszlásának becslése egy lexikalizált elemző számára sokkal nehezebb. Továbbá könnyen belátható, hogy egy magyar korpuszban sokszorosa a ritka szavak mennyisége és ezzel párhuzamosan a tanítóanyagban nem szereplő szavak (OOV – out-of-vocabulary) aránya is egy ugyanakkora angol nyelvű korpuszhoz képest (17. táblázat). Egy nagy pontosságú egyértelműsítő rendszernek rendelkeznie kell megfelelő stratégiával az OOV szavak kezelésére, mivel a hibásan elemzett szavak nehézséget jelentenek a további szövegfeldolgozás során.

| | szegmens | token | szóalak | OOV | OOV/token |
|---|---|---|---|---|---|
| angol | 1000 | 14137 | 3562 | 77 | 0,54% |
| magyar | 1000 | 11672 | 4842 | 323 | 2,78% |

**17. táblázat: OOV szavak aránya azonos méretű korpusz esetén (Hunglish korpusz)**

Végül a teljes morfoszintaktikai egyértelműsítő rendszer sebességét és minőségét nagymértékben befolyásolja az a modell, aminek a segítségével egy szó a környezetéből származó információt kinyeri. Az elemzendő szó szófaját a mondatban körülötte szereplő összes szó befolyásolja, viszont így a rendszer állapottere exponenciálisan növekszik a mondat hosszának arányában. Általánosan elterjedt megoldás ezen hatás kezelésére, ha a vizsgált szó csak egy szűkebb környezetét vesszük figyelembe, ám ez információveszteséget okoz.

## 6.2 A teljes morfoszintaktikai egyértelműsítés, mint gépi fordítási feladat

A teljes morfoszintaktikai egyértelműsítés feladata megfogalmazható úgy, hogy a POS-címkék sorozatából ($T = t_1, \dots, t_m$) keressük azt a címkesorozatot ($\hat{T}$), amelyik a vizsgált mondat ($W = w_1, \dots, w_n$) szavainak helyes elemzését tartalmazza. Statisztikai megközelítés alapján a keresett címkesorozat a $P(T|W)$ valószínűségek közül a legmagasabb értékkel rendelkező lesz. Formálisan ez a következőképpen írható le:

$$\hat{T} = \underset{T}{\operatorname{argmax}} P(T|W) = \underset{T}{\operatorname{argmax}} \frac{P(W|T)P(T)}{P(W)} = \underset{T}{\operatorname{argmax}} P(W|T)P(T) \qquad (8)$$

A 2.2.1. fejezetben bemutatott Bayes-tételt alkalmazva az egyértelműsítés feladata is felbontható két modell kombinációjára, név szerint a lexikai valószínűség modellre $\big(P(W|T)\big)$, valamint a címkeátmenet-valószínűség modellre $\big(P(T)\big)$. Ezt a megközelítést használja a rejtett Markov modell (HMM – hidden Markov model) alapú szófaji egyértelműsítő rendszer.

DOI:10.15774/PPKE.ITK.2015.004

Összevetve a (2) és (8) egyenleteket láthatjuk, hogy az SMT feladata izomorfnak tekinthető a HMM-alapú szófaji egyértelműsítéssel, ahol az SMT $P(T)$ nyelvi modellje a címkeátmenet-valószínűség modellnek, míg a $P(S|T)$ fordítási modell a lexikai-valószínűség modellnek feleltethető meg. Az egyértelműsítőként működő fordítórendszer esetén a forrásnyelvnek tehát az eredeti szöveg, célnyelvnek pedig az ehhez tartozó annotáció felel meg. A legnagyobb különbség az SMT dekóderrel implementált modell és a HMM-alapú egyértelműsítő között az a módszer, amivel azok a két valószínűségi modellt létrehozzák [109].

A teljes morfoszintaktikai egyértelműsítés feladata lényegesen egyszerűbb, mint a természetes nyelvek közötti fordítás, mivel az elemzés során nincsen szórendi, valamint szószámbeli eltérés a forrás- és a célnyelvi oldalak között. Következésképp a kifejezésalapú SMT rendszerben nincs szükség az átrendezési és a mondathossz-harmonizációs modellekre. Mivel az egyértelműsítő rendszer esetén a forrás- és a célnyelv szavai között egy-egyértelmű megfeleltetés van, ezért a rendszer tanítása során nincs szükség a párhuzamos korpusz szavai közti statisztikai módszerrel történő szóösszekötésre. Ennek köszönhetően az egyértelműsítő rendszer tanítása lényegesen gyorsabb és pontosabb a természetes nyelvek közti fordítórendszer betanításához képest. Ezek alapján az SMT-alapú egyértelműsítő rendszer a (9) képlet alapján a következőképpen írható le:

$$\hat{T} = \underset{T}{\operatorname{argmax}} P(W|T)P(T) = \underset{T}{\operatorname{argmax}} \prod_{i=1}^{|T|} \phi(\bar{s}_i|\bar{t}_i) P_{LM}(t_i|t_{i-l}^{i-1}) \tag{9}$$

ahol $\phi(\bar{s}_i|\bar{t}_i)$ a kifejezés lexikai valószínűség modell, ahol a $\bar{s}_i$-k több szóból álló kifejezések, a $\bar{t}_i$-k pedig a hozzájuk tartozó szófaji címkék sorozata. A $P_{LM}\left(t_i\middle|t_{i-l}^{i-1}\right)$ egy n-gram alapú címkeátmenet-valószínűség modell. Fontos megemlíteni, hogy a kifejezésalapú modellek erőssége, hogy az egyértelműsítés során képesek a szavak környezeti információit is felhasználni.

## 6.3 Az SMT-alapú teljes morfoszintaktikai egyértelműsítő rendszer felépítése

Munkám során létrehoztam egy SMT-alapú teljes morfoszintaktikai egyértelműsítő rendszert. Elsődleges célom volt, hogy a létrehozott módszer nyelvfüggetlen legyen, az alkalmazott komponensek nagy része így került megvalósításra. Ezek mellett néhány nyelvfüggő komponenst is integráltam a rendszerbe (magyar nyelvű morfológiai elemző, névelők kezelése a magyar nyelvben), de ezek használata csupán opcionális, így a létrehozott rendszer nyelvfüggetlennek tekinthető. A nyelvfüggő komponensekre azért volt szükség, mert a kiértékelés során célom volt a magyar nyelvű korpuszon a lehető legjobb eredmény elérése. Ebben a fejezetben bemutatom az általam létre-

DOI:10.15774/PPKE.ITK.2015.004

hozott rendszert, valamint ismertetem működését.

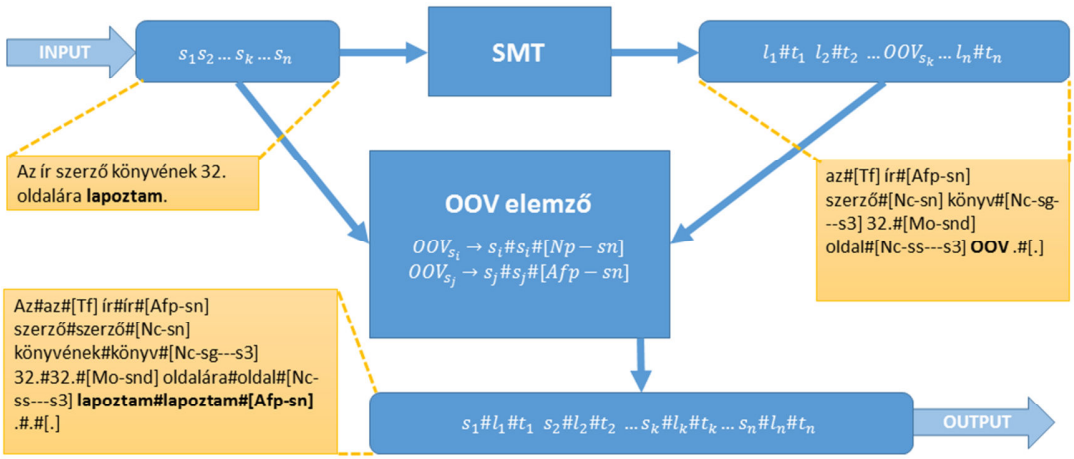
### 6.3.1 Az SMT-n alapuló egyértelműsítő alaprendszer

Ahogy azt a 2.2. fejezetben bemutattam, a statisztikai gépi fordítórendszer betanítható, hogy elvégezze a teljes morfoszintaktikai egyértelműsítés feladatát. Ehhez nem kell más, mint létrehozni egy olyan párhuzamos, kétnyelvű korpuszt, ahol a forrásnyelvi oldalon maga az elemzendő nyers szöveg áll, míg a célnyelvi oldalon a szavak lemmájából és POS címkéjéből álló egységek (*„szótő#POS”)*. Az OOV szavak ebben az esetben mindig a tanítóanyagban legnagyobb gyakorisággal előforduló címkét kapják. Mivel a kódrendszerek többsége különbséget tesz a tulajdonnevek és köznevek közt, ezért én is külön kezelem a kis és a nagy kezdőbetűs szavakat. Ennek megfelelően a kisbetűs szavakhoz tartozó leggyakoribb címke a melléknév ([Afp-sn][3]) lesz, míg a nagybetűs szavak a főnév-tulajdonnév ([Np-sn][4]) címkét kapják. Ezt a rendszert kiindulási változatként tekintem és a továbbiakban *ALAP*-nak fogom nevezni. Az *ALAP* rendszer működését a 18. ábra szemlélteti.

Megfigyelhető, hogy ez a rendszer a különböző részfeladatok (OOV szavak kezelése, lemmatizálás kezelése) elvégzésére az elérhető legegyszerűbb algoritmusokat használja. Emiatt a komplexebb algoritmusok és modulok integrációjával jelentősen javítható az egyértelműsítő rendszer minősége. A következő fejezetekben bemutatom azokat a módszereket, amikkel sikerült javítanom a rendszerem minőségén. A rendszerekhez tartozó eredményeket az 6.4. fejezetben részletezem.

[3][Afp-sn]: melléknév, alapfok, egyes szám, alanyeset
[4][Np-sn]: főnév, tulajdonnév, egyes szám, alanyeset

DOI:10.15774/PPKE.ITK.2015.004




**18. ábra: Az *ALAP* rendszer folyamatábrája és a lépések bemutatása egy példamondaton**

### 6.3.2 Mondatkezdő és mondatzáró szimbólumok

Bizonyos szavak szófaja függ a mondaton belüli helyétől. Az elemzés során különösen fontos megőrizni azt a tudást, hogy a szó a mondat első vagy utolsó pozíciójában áll, mivel ez az információ nem jelenik meg az egyértelműsítő rendszer fordítási modelljében. Ennek érdekében mind tanítás, mint tesztelés esetén a mondatokat kiegészítettem egy mondatkezdő és egy mondatvég szimbólummal. Ennek köszönhetően az SMT rendszer modelljeibe bekerül az az információ is, hogy az adott kifejezés mondatkezdő, a mondat közepén található, vagy éppen mondatvégi pozícióban van. Ezt a rendszert továbbiakban *SZIMB*-nek fogom nevezni.

### 6.3.3 A számjegyek, az azonosítók, a százalékok és a római számok kezelése

A korpuszban szereplő, számjegyeket tartalmazó szavak esetén a feladat nehézségét az adathiányprobléma jelenti, mivel a tanítóanyagban a lehetséges összes számnak szerepelnie kellene ahhoz, hogy megfelelően címkézni lehessen ezeket. Az ebből fakadó hibák kezelésére a számjegyeket tartalmazó szavakat reguláris kifejezések segítségével egységes osztályokba (szám, százalék, id) soroltam. A számjegyek különböző osztályokra való bontására azért volt szükség, mert a kódrendszerek többsége különbséget tesz a számok, százalékok, valamint az azonosítók elemzése között. Az arab számjegyeken kívül szükséges volt a római számokat is azonosítani, mivel bizonyos kódkészletek esetén külön POS címke tartozik hozzájuk, valamint előfordul, hogy a szótövük a megfelelő arab szám. A számjegyek kezelését tartalmazó rendszert a továbbiakban a *SZÁM* névvel fogom ellátni.

DOI:10.15774/PPKE.ITK.2015.004

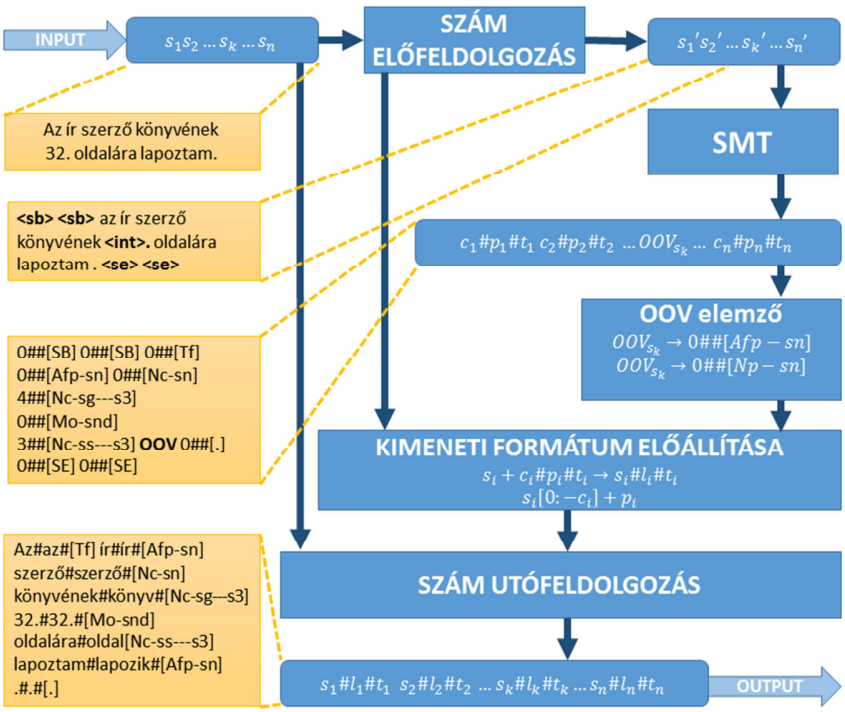
### 6.3.4 A célnyelvi címkekészlet méretének csökkentése

Az *ALAP_SZÁM* rendszer az elemzendő szavakat szótő és POS címke párosra (*„szótő#POS”*) fordítja. Mivel a tanítóanyag célnyelvi oldalán az összes forrásnyelvi szó szótöve megtalálható, ezért a célnyelvi szótárban előforduló különböző szavak számossága legalább akkora, mint a korpuszban előforduló különböző szavak számossága. 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. Mivel a gyakorlatban minden szóhoz önálló címke tartozik, ezért az egyértelműsítés feladata a szó-címke párokból álló lexikonból való kiolvasáshoz közelít. Ezért gyengül a környezetből származó információk fontossága.

Erre a 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 elmondható, hogy a nagymértékű információveszteség ellenére viszonylag kis mértékű volt az elemző rendszer minőségének javulása. A kapott eredményeket a 6.4.2. fejezetben részletesen bemutatom.

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 [43] 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. Például a *„hazám”* szó fordítása *„2#a#[Nc-sn---s1]”* lesz. A továbbiakban ezt a reprezentációt fogom használni és ezeket a rendszereket a *TÖRÖLCSATOL* néven fogom hívni. A rendszer működését a 19. ábra szemlélteti.

DOI:10.15774/PPKE.ITK.2015.004




**19. ábra: A *TÖRÖLCSATOL_SZÁM* rendszer folyamatábrája**

### 6.3.5 A prefixek kezelése

A *TÖRÖLCSATOL* rendszerben a szótövek tárolására használt reprezentáció kifejezetten alkalmas a szuffixumok kezelésére, emiatt jól alkalmazható ragozó nyelvek esetén. Az architektúrából következik, hogy nehézséget jelent, ha a vizsgált szó prefixet tartalmaz. Erre példa magyar esetén a melléknevek felsőfokú ragozása vagy az igekötős igék. A probléma, hogy *<törö#,csatol>* rekord használata során az egész szóalakot ki kell törölni, és magát a lemmát kell hozzáilleszteni az üres sztringhez. Erre egy példa a *legpirosabb* szó, amit a *<11#piros>* rekord kódol. A lemmakezelő algoritmus ezen gyengesége miatt elveszti működésének értelmét, mivel így minden prefixet tartalmazó szó önálló annotációs címkével rendelkezik, és ez nem csökkenti a célnyelvi címkekészlet méretét. Ennek ellenére a módszer képes kezelni ezeket az eseteket is, bár ezt nem a leghatékonyabb módon teszi.

A prefixek kezelésére létrehoztam egy nyelvfüggő modult, amelyik reguláris kifejezés segítségével a felsőfok jelét szuffixum pozícióba helyezi át. Ez a gyakorlatban azt jelenti, hogy a felsőfok *leg-* prefixét a szó elejéről a *-bb* szuffixum elé rendeztem *(...legbb)*. Így az előző példát folytatva a *legpirosabb* szóból *„pirosalegbb”* lett, így ennek a sztringnek a *<6#’’>* rekord lett a reprezentációja. Ennek a módszernek köszönhetően a lemmák szuffixumalapú meghatározása nagyobb pontossággal működik.

DOI:10.15774/PPKE.ITK.2015.004

### 6.3.6 Az ismeretlen szavak kezelése osztályozási módszerrel

Oravecz és Dienes [107] szerint az egyértelműsítés során jelentkező leggyakoribb hiba a tanítóhalmazban nem szereplő szavak helytelen elemzéséből adódik. A probléma súlyossága a nyelvek morfoszintaktikai tulajdonsága alapján jelentősen eltérhet, mivel egy agglutináló nyelvű korpuszban sokkal több ismeretlen szó szerepel, mint egy ugyanakkora méretű izoláló nyelvű szövegben (17. táblázat).

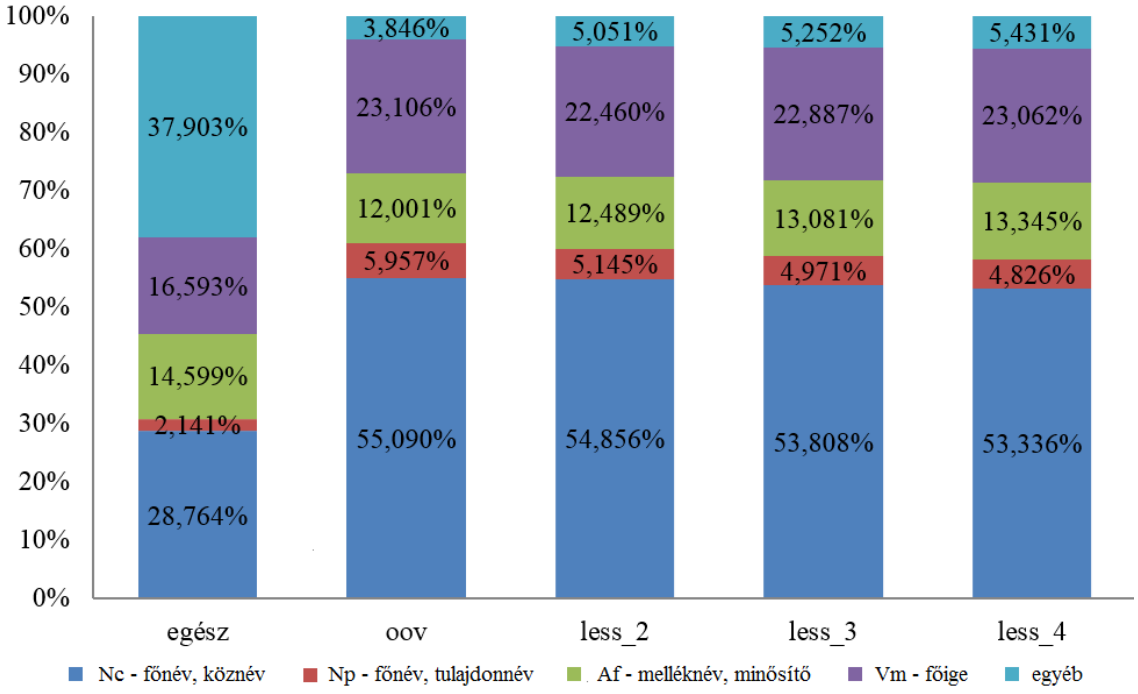
Az OOV szavak elemzése azért jelent problémát, mert az egyértelműsítő rendszernek semmilyen előzetes ismerete nincs ezekről a szavakról. Ezt az eddig bemutatott rendszereim úgy kezelték, hogy az ismeretlen szavakhoz a tanítóanyagban szereplő leggyakoribb címkét rendelték. Ezzel a módszerrel olyan arányú javulás érhető el, amilyen a tanítóanyagban a leggyakoribb címke aránya az OOV szavak között.

Az egyértelműsítő rendszer jobb minőségű működéséhez szükséges az ismeretlen szavak viselkedésének pontosabb modellezése. Dermatas és Kokkinakis [110] azzal a feltételezéssel éltek, hogy megfelelően nagy korpusz esetén az OOV szavak a ritka szavakhoz hasonlóan viselkednek. Munkájuk során ezt a feltevést empirikus módon bizonyították. Belátható, hogy ez a feltételezés morfoszintaktikailag távolabbi nyelvek esetén is igaz. Ezt egy magyar nyelvű példa segítségével szemléltetem, amiben az adatot a Szeged Korpusz 2 [111] szolgáltatta (a korpuszt részletesen az 3.2.1. fejezetben mutattam be). A korpuszt két részre osztottam: egy tanítóanyagra, ami 846 562 tokenből és 98 595 különböző szóalakból (type) állt; valamint egy teszthalmazra, ami 103 931 tokent és 23 337 szóalakot tartalmazott. Az így létrehozott teszthalmaz szavai közül 7321 token (7,044%) nem szerepelt a tanítóhalmazban. A tanítóhalmaz szavaiból csoportokat hoztam létre az előfordulási gyakoriságuk alapján, így ezek a csoportok szimulálják a ritka szavak viselkedését. A csoportokat a bennük szereplő szavak előfordulási gyakorisága alapján neveztem el, így például a *LESS_2* csoport azokat a szavakat tartalmazza, amelyek kevesebb, mint kétszer szerepelnek a tanítóanyagban. Az így létrehozott szócsoportok segítségével három különböző aspektusból vizsgáltam meg az OOV szavak és a ritka szavak kapcsolatát.

Elsőként megvizsgáltam a leggyakoribb szófaji címkék (köznév, tulajdonnév, melléknév, ige és az összes többi szófaji kategória) arányát a különböző adatcsoportokban. Az eredményeket a 20. ábra mutatja. A diagramon jól látható, hogy az OOV szavak típusmegoszlásának aránya nagyon hasonlít az eltérő küszöbbel mért ritka szavak típusmegoszlásához. Ugyanakkor a teljes korpuszban már jelentősen eltérő eloszlás figyelhető meg. Az ábrán látható, hogy amíg a teljes kor-

DOI:10.15774/PPKE.ITK.2015.004

pusz alig több mint negyede köznév, addig az OOV és a ritka szavak valamivel több, mint a fele tartozik ebbe a kategóriába. Ezzel szemben arányában jóval kevesebb ige van a teljes korpuszban, mint a többi csoportban. A melléknevek tekintetében figyelhető meg a legkisebb mértékű eltérés. Érdemes megfigyelni továbbá, hogy az egyéb szófaji kategóriákba tartozó szavak közül csak nagyon kevés van, amelyik egyáltalán, vagy csak nagyon kis számban fordul elő a ritka szavakból álló halmazban. Ez annak tudható be, hogy a korpusz egyéb kategóriájába a zárt szóosztályok (például névelő, névmások, indulatszó, kötőszó stb.) szavai kerülnek, melyek nagy gyakorisággal fordulnak elő a korpuszban. A ritka szavak egyéb kategóriájába főként a határozószók tartoznak.



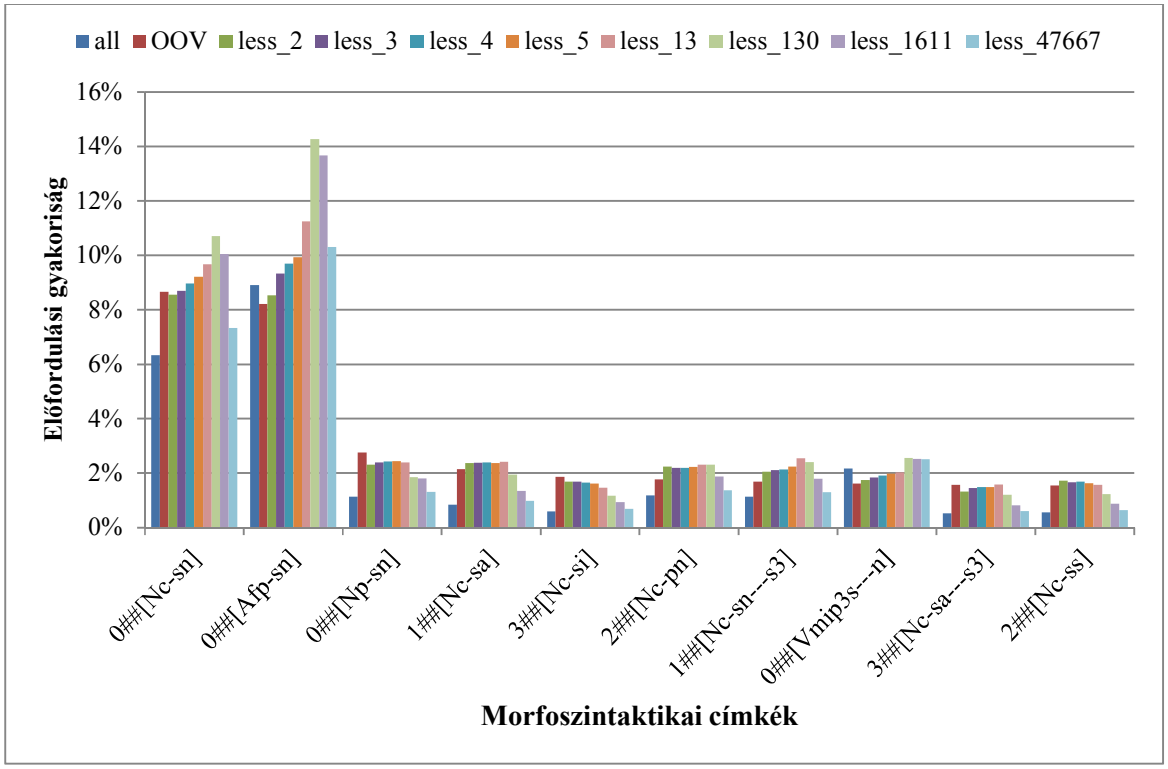

**20. ábra: A leggyakoribb szófaji címkék aránya a különböző adatcsoportokban**

A szófaji eloszlások mellett a *TÖRÖLCSATOL* rendszerben bemutatott teljes morfoszintaktikai címkék (lemmatranszformáció + POS címke) arányát is megvizsgáltam. A 21. ábra az OOV korpuszban szereplő tíz leggyakoribb címke előfordulásának arányát mutatja be a különböző adatcsoportokban. Az ábrán jól látszik, hogy a különböző címkék előfordulási aránya jelentősen eltér az OOV szavak és a tanítóhalmaz esetén. Ezzel szemben jelentős a hasonlóság az OOV korpusz és a ritka szavakat tartalmazó csoportok között. A diagram alapján megfigyelhető, hogy az ismeretlen szavak között a leggyakoribb címkekategória a toldalék nélküli köznév, valamint az OOV korpusz szavai között gyakran fordulnak elő ismert szavak ragozott alakjai.

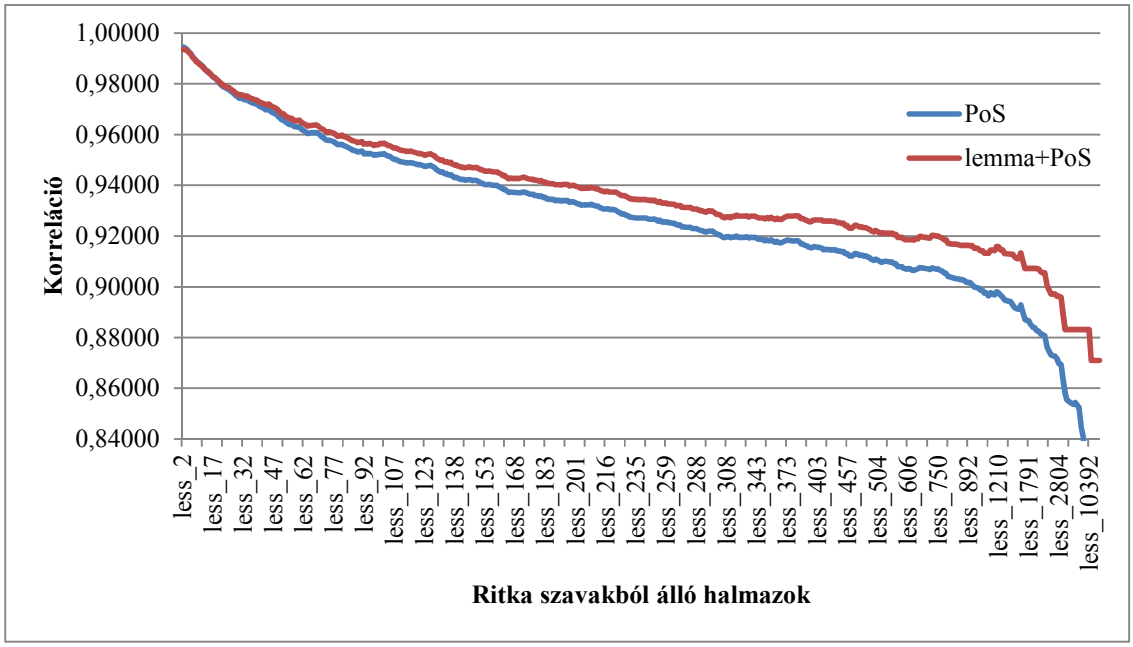
DOI:10.15774/PPKE.ITK.2015.004

Végül meghatároztam az OOV korpusz valamint a különböző korpuszcsoportok egymáshoz viszonyított korrelációját, mind POS címke, mind teljes morfológiai címke tekintetében. Elsőként mindegyik korpuszcsoportból létrehoztam egy címke gyakorisági vektort, majd ezen vektorok között határoztam meg ezek Pearson-korrelációját [112]. (A halmazok közti diszjunkt címkéket 0 súllyal vettem figyelembe.) Az összes adatot ábrázoló grafikont a 22. ábra szemlélteti. Jól látható az a tendencia, hogy a legmagasabb korrelációja az ismeretlen szavaknak a tanítókorpuszban egyszer szereplő szavakból (*LESS_2*) álló részkorpusszal van, minél több szót veszünk hozzá ehhez a korpuszhoz a korreláció annál alacsonyabb lesz.

DOI:10.15774/PPKE.ITK.2015.004



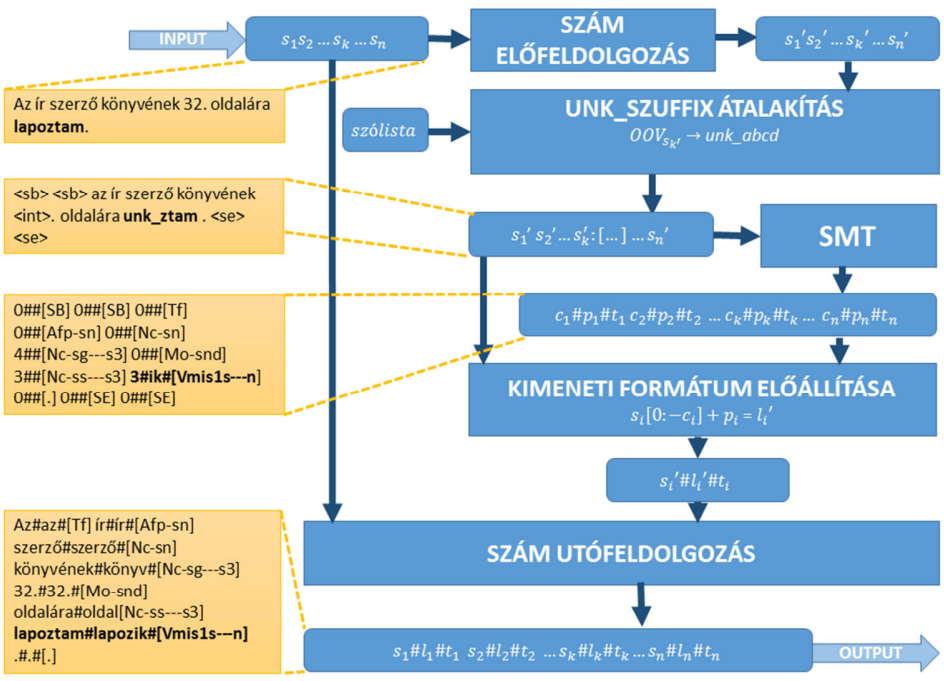

**21. ábra: Az OOV korpuszban szereplő tíz leggyakoribb címke előfordulásának aránya különböző adatcsoportokban**




**22. ábra: Az ismeretlen szavak és a ritka szavak közti korreláció**

DOI:10.15774/PPKE.ITK.2015.004

A fent bemutatott mérések alapján megállapítható, hogy az ismeretlen szavak jól modellezhetők a tanítóhalmazban szereplő ritka szavak segítségével, emiatt több modult is megvizsgáltam, amelyek a ritka szavak segítségével igyekeznek az ismeretlen szavakat egyértelműsíteni. 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 ragozások 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 karaktersorozat optimális hosszát a 6.4.2. fejezetben bizonyítom. Annak érdekében, hogy különbséget tudjak tenni a kis- és a nagybetűs OOV szavak között az egyértelműsítés során, a nagybetűs ismeretlen szavakat kapitalizált „*UNK_ABCD”*-re cseréltem. 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. Ezt a rendszeremet a továbbiakban *UNKSZUFFIX* -nek fogom hívni. A rendszer működését a 23. ábra mutatja be.



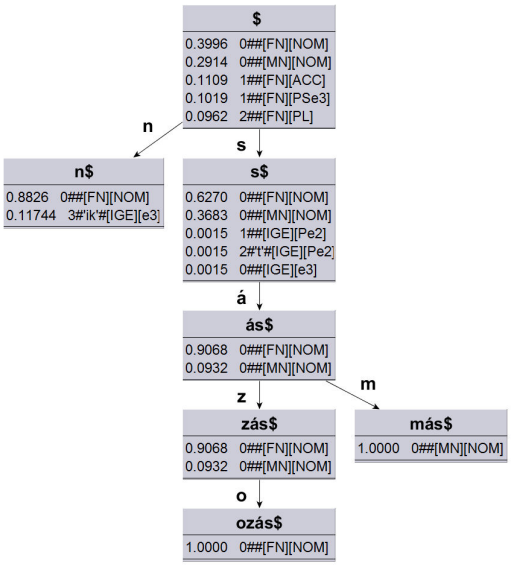

**23. ábra: A *TÖRÖLCSATOL_SZÁM_UNKSZUFFIX* rendszer folyamatábrája és a lépések bemutatása egy példamondaton**

DOI:10.15774/PPKE.ITK.2015.004

### 6.3.7 Szóvégalapú teljes morfoszintaktikai ajánlórendszer integrálása

Az *UNKSZUFFIX* rendszer gyengesége, hogy az OOV szavak fix hosszúságú szuffixumát veszi figyelembe az egyértelműsítés során. Ez a megszorítás problémát jelenthet, mivel a különböző toldalékok mérete nem fix, hanem rugalmasan változhat. Például a magyar nyelv esetében, egy szóhoz egyidejűleg több toldalék is kapcsolódhat, vagy bizonyos toldalékok csak egy karakterből állnak. Emiatt érdemes olyan módszert is megvizsgálni, amelyik nem azonos súllyal veszi figyelembe a szóvégi karaktereket. Ennek a feltételnek felel meg az úgynevezett végződésfa-alapú ajánlórendszer (suffix guesser). A guesser feladata, hogy egy ismeretlen szóról megbecsülje, hogy mekkora valószínűséggel tartozik egy morfológiai címkéhez. Hasonló kutatások megmutatták [43], [113], [114], hogy az egyértelműsítő rendszerek minősége javítható guesser integrálásával.

Munkám során az SMT-alapú teljes morfoszintaktikai egyértelműsítési láncba előfeldolgozó lépésként egy suffix 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 suffix 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.




$$P(0\#\#[FN][NOM]|facebookozás) = \theta_0 P(0\#\#[FN][NOM]) \times \theta_1 P(0\#\#[FN][NOM]|\text{”s”}) \times \theta_2 P(0\#\#[FN][NOM]|\text{”ás”}) \times \theta_3 P(0\#\#[FN][NOM]|\text{”zás”}) \times \theta_4 P(0\#\#[FN][NOM]|\text{"ozás"})$$

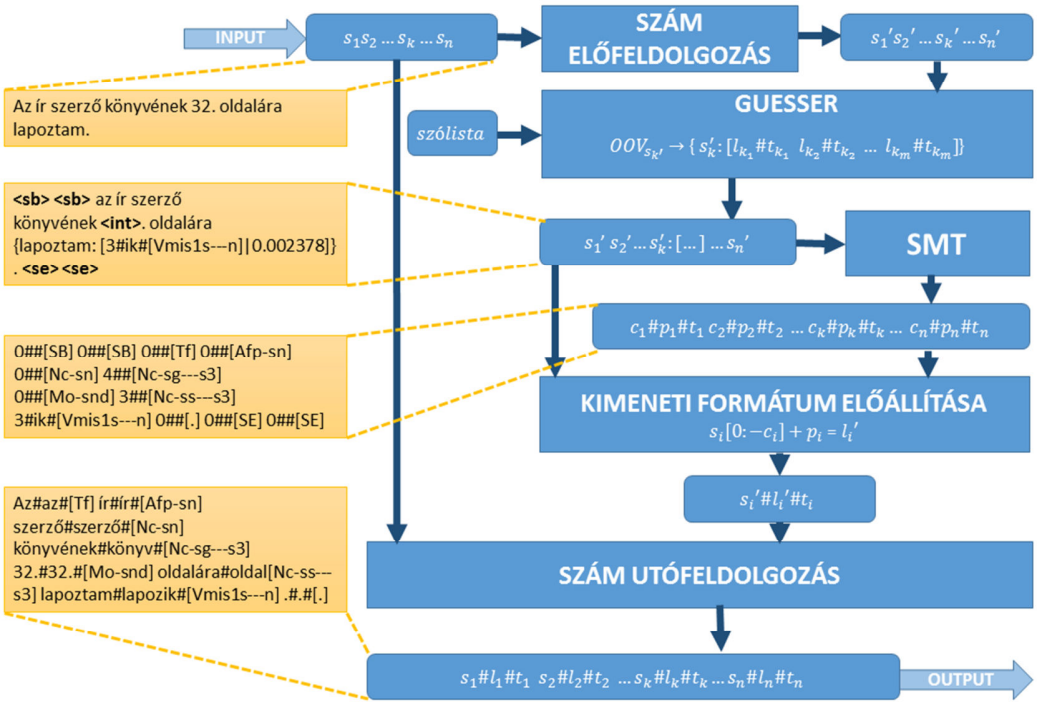
**24. ábra: A végződésfában való keresés folyamatának bemutatása egy példa segítségével**

DOI:10.15774/PPKE.ITK.2015.004

A program egy új szó elemzése során a szó karakterei alapján végigmegy a végződésfa megfelelő csúcsain. A gráf éppen aktuális csúcsaiban található valószínűségek súlyozott szorzata alapján számolja ki a szóhoz tartozó címkék valószínűségét. Ezt szemlélteti a 24. ábra, ahol a számára ismeretlen *facebookozás* szóra keresi az elemzési javaslatokat. Az elemzés során végigmegy a végződésfa megfelelő csúcsain, majd ezek segítségével minden egyes címkére kiszámolja, hogy mekkora valószínűséggel rendelhető a vizsgált szóhoz. Ez formálisan a következőképpen írható le:

$$P(T|W) = \prod_{i=0}^{n} \theta_i \times P(T|\,[w_{n+1-i} \dots w_{n+1}])$$
$$n = [1; |W|] \tag{10}$$
$$\theta_0 > \theta_1 > \cdots > \theta_n$$

ahol $T$ a morfoszintaktikai címke, $W = [w_0, w_1, \dots, w_n]$ a vizsgált szó és $\theta_i$ a súly paraméter. Továbbiakban ezt a rendszert *GUESSER*-nek fogom nevezni. Mivel a nagybetűs szavak máshogy viselkednek, mint a kisbetűsek, ezért ebben az esetben is célszerű külön kezelni őket. Ennek megfelelően külön guessert használok a kis- és a nagybetűs szavakra. A 25. ábra a rendszer működését szemlélteti. Látható, hogy a guesser előfeldolgozó lépésként működik. Az ismeretlen szavakat egy szólista alapján találja meg, amit a tanítóanyag alapján állítok össze.




**25. ábra: Az *TÖRÖLCSATOL_SZÁM_GUESSER* rendszer folyamatábrája és a lépések bemutatása egy példamondaton**

DOI:10.15774/PPKE.ITK.2015.004

### 6.3.8 Morfológiai elemző integrálása

A 6.3.7.fejezetben bemutatott guesser hátránya, hogy egy ismeretlen szó elemzése során semmilyen nyelvi tudást nem vesz figyelembe, helyette tisztán statisztikai számítások alapján minden lehetséges szófaji címkéhez rendel egy valószínűség-értéket, végül ezen valószínűségek legjobbjait javasolja elemzésnek. Ezzel szemben a morfológiai elemző nyelvtani tudás segítségével meg tudja határozni egy tetszőleges szó lehetséges lemmáit – ha az szerepel az elemző szótárában – és a hozzájuk tartozó POS címkéket. Hátránya viszont, hogy nem rendel valószínűség-értéket a lehetséges címkékhez, nem nyelvfüggetlen, és függ a morfoszintaktikai kódrendszertől is. Ehhez nagy háttértudásra, beépített szótárra van szüksége, hiszen ismernie kell az összes lehetséges szótövet.

A morfológiai elemzőt a guesserrel kombinálva integráltam rendszerembe; így a morfológiai elemző által adott javaslatokhoz valószínűség-értékeket tudtam rendelni. A rendszeremben a következő módon működik együtt a két rendszer a morfológiai elemző által adott kimenet alapján:

- ha egy találat van, akkor a valószínűség 1;
- ha több a találat, akkor a guesserből rendeli hozzá a valószínűség-értékeket;
- ha nem ad találatot, akkor a guesser legvalószínűbb javaslatait használja.

A morfológiai elemző segítségével tulajdonképpen a guesser által adott ajánlásokat szűkítem, és tartom meg ezekből a nyelvtanilag lehetséges találatokat.

## 6.4 Az SMT-alapú egyértelműsítő rendszer minőségének bemutatása

Ebben a fejezetben egy magyar nyelvű példarendszer segítségével bemutatom az általam létrehozott SMT-alapú teljes morfoszintaktikai egyértelműsítő rendszer eredményeit, valamint a 6.3. fejezetben leírt modulok és algoritmusok integrálásának hatását az eredmények minőségére nézve. Elsőként bemutatom a rendszer betanításául szolgáló korpuszt, majd ismertettem a tanítóanyagon elért eredményeimet. Jelen fejezetben a célom az adott tanítóanyagon elérhető legmagasabb pontosságú egyértelműsítő rendszer megalkotása volt, így itt a nyelvfüggetlen módszerek mellett nyelvfüggő modulokat is használtam. Rendszeremet HuLaPos2 rendszernek neveztem el, a továbbiakban így fogok hivatkozni rá.

### 6.4.1 A felhasznált erőforrás

Munkám során a Szeged Korpusz 2-t [111] használtam, ami a legnagyobb méretű magyar nyelvű kézzel annotált korpusz. A Szeged Korpusz 2 morfoszintaktikailag elemzett és kézzel egyértelműsített természetes nyelvi szöveges adatbázis. A korpusz tartalmazza az eredeti szóalakokat, a sza-

DOI:10.15774/PPKE.ITK.2015.004

vak szótövét, valamint az egyes szavakhoz tartozó MSD kódrendszerben tárolt morfoszintaktikai kódokat.

Az MSD (Morpho-Syntactic Description) kódrendszer [115] egy nemzetközi kódrendszer, amely morfoszintaktikai szemszögből közelíti meg a korpusz elemzését. Ez a kódrendszer majdnem minden európai nyelv különböző jellemzőinek kódolására, de legfőképp morfológiai elemzésére alkalmazható. A szavak morfoszintaktikai jellemzőit (mint például szófaj, eset, szám, személy, igeidő, igemód stb.) különböző karaktersorozatokkal reprezentálja. Azokban az esetekben, amikor egy adott „helyiértéken" álló attribútum hiányzik, vagy nem értelmezhető az adott nyelven, a '-' karaktert használja. Az első helyeken a főbb POS kategóriák kódjai szerepelnek.

A korpusz egyik előnye, hogy kézzel ellenőrzött, ezért egészen pontos adathalmaznak tekinthető. Másik előnye pedig, hogy általános szövegeket tartalmaz, és emiatt nem témaspecifikus. Mérete azonban csupán kicsivel több, mint 82 000 fordítási egység, ami 1,2 millió tokent (154 238 különböző szóalak) tartalmaz, így viszonylag kis méretű korpusznak számít.

A tanítóanyag építése során az eredeti korpusz néhány sorát elhagytam. A Szeged Korpusz 2-ben léteznek hibás mondatok, szavak (helyesírási hibák, elütések, idegen szavak stb.), melyeket a korpuszépítés során hibakóddal láttak el. Ezeket a mondatokat eltávolítottam a korpuszból azért, hogy a rendszerem a tanulás során ne találkozzon ilyen jellegű hibákkal. A Szeged Korpusz 2 egy másik tulajdonsága, hogy a több szóból álló kifejezéseket, amik általában névelemek (named entity) vagy különféle igei frázisok, egy közös címkével látták el. Például a *Magyar Nemzeti Bank* szóhármast egy kifejezésként annotálták, egyetlen tulajdonnévként. A szófaji egyértelműsítőként működő szóalapú gépi fordítórendszer működését nagymértékben megnehezítené, ha a rendszernek a névelem-felismerést is el kellene végeznie. Mivel nem volt célom, hogy az általam létrehozott rendszer névelem-felismerést is végezzen, ezért elhagytam a korpuszból azokat a mondatokat, amelyek több szóból álló kifejezéseket tartalmaznak. A fentiek alapján létrehozott csökkentett méretű korpusz 64 395 mondatból, 1 042 546 tokenből és 112 100 különböző szóalakból áll.

A rendszer minőségét az egyértelműsítés, a szótövesítés, valamint a teljes morfoszintaktikai egyértelműsítés pontosságával mértem. A pontosságot szavak és mondatok szintjén egyaránt kiszámítottam. Munkám során a rendszeremet a szószintű teljes morfoszintaktikai egyértelműsítés pontosságára optimalizáltam. A korpuszt 10%-10%-80%-ban osztottam fel fejlesztői, teszt- és tanítóhalmazokra. A rendszer paramétereit a fejlesztői halmaz segítségével állítottam be, míg a végső kiértékelést a teszthalmazon végeztem. A rendszereim eredményeit Wilcoxon-féle előjeles rangszám próba [116] segítségével vizsgáltam meg, hogy azok statisztikailag szignifikánsak-e.

DOI:10.15774/PPKE.ITK.2015.004

### 6.4.2 Az eredmények ismertetése

Elsőként létrehoztam az **SMT-n alapuló egyértelműsítő alaprendszert** (6.3.1. fejezet), amit *ALAP*-nak nevezek. A rendszer pontossága 91,281%-os POS címkézés esetén, és 94,303% a szótövesítés során, valamint 91,257% a teljes morfoszintaktikai egyértelműsítés tekintetében (18. táblázat).

A **számjegyeket tartalmazó mondatok** (6.3.3. fejezet), valamint a **mondatkezdő és mondatvégi szimbólumok** (6.3.2. fejezet) kezelésével sikerült az *ALAP* rendszer POS taggelésének pontosságát 0,196%-kal, a szótövesítést 0,03%-kal valamint a teljes morfoszintaktikai egyértelműsítést 0,193%-kal javítani. (*ALAP_SZIMB_SZÁM)*

| | Szószintű | | | Mondatszintű | |
|---|---|---|---|---|---|
| | POS | Lemma | Összes | POS | Összes |
| *ALAP* | 91,281% | 94,303% | 91,257% | 35,371% | 35,294% |
| *ALAP_SZIMB* | 91,352% | 94,326% | 91,389% | 36,016% | 35,968% |
| *ALAP_SZIMB_SZÁM* | 91,477% | 94,333% | 91,450% | 36,591% | 36,514% |

**18. táblázat: A alaprendszerek eredményei**

A 6.3.4. fejezetben kifejtettem azt a kérdést, hogy miként befolyásolja az egyértelműsítő rendszer minőségét a **célnyelvi címkekészlet méretének csökkentése**. Az *ALAP* rendszer architektúrájából következik, hogy a célnyelvi szótár mérete lényegében megegyezik a forrásnyelvi szótár méretével. Ez a Szeged Korpusz 2-ben a 91 178 különböző szóalak (type) mellett a célnyelvi szótár 94 175 különböző *„szótő#POS”* (6.3.1. fejezet) egységből áll. Az első megközelítésem a célnyelvi szótár méretének a csökkentésére a címkékben tárolt információ redukálása volt. A lemmatizálás elhagyása esetén (a rendszer csak szófaji egyértelműsítést végez) a célnyelvi címkekészlet mérete megegyezik a korpuszban szereplő címkék számosságával, ami a mi esetünkben 995 elem. Ez a címkekészlet közel százszoros csökkentésének felel meg. Az így kapott rendszert *ALAP_SZIMB_SZÁM_CSAKPOS*-nak nevezem. A szótövesítés elhagyásával a POS címkézés 0,064%-al javult az *ALAP_SZIMB_SZÁM* rendszerhez képest, ami a hibák 0,754%-a.

A címkekészlet további csökkentését a POS címkekészlet redukálásával értem el. A teljes címkekészlet helyett csupán a 34 darab fő POS címkére fordítottam, ami további harmincszoros méretbeli változást jelent az *ALAP_SZIMB_SZÁM_CSAKPOS* rendszerhez képest. Az így létrehozott

DOI:10.15774/PPKE.ITK.2015.004

rendszer 95,471% pontosságú, ami a hibák további 46,504%-os javulásának felel meg. Ezt a rendszert *ALAP_SZIMB_SZÁM_FŐPOS*-nak nevezem.

A fenti eredményekből következik, hogy nem éri meg a tárolt információ mennyiségének a csökkentése, mivel így csak kismértékű az elérhető minőségjavulás. Ezért választottam a 6.3.4. fejezetben bemutatott módszert, aminek a segítségével az információ mennyiségének a megőrzése mellett redukálható a célnyelvi szótár mérete. Az így működő rendszer (*TÖRÖLCSATOL_SZIMB_SZÁM*) a szótövek helyett egy rekordot tárol *(<töröl#csatol#POS>)*, amiben a lemmák előállításához szükséges transzformációk vannak leírva. Az ilyen típusú célnyelvi címkéből 3554 darab van, ami harmincszor kisebb, mint az eredeti *ALAP_SZIMB_SZÁM* rendszer szótára. Az így felépített rendszer minősége a lemmatizálás megtartása mellett 91,447%-os pontosságot, vagyis közel azonos eredményt ért el az *Alap_SZIMB_SZÁM* rendszerhez képest. A 19. táblázatból kiderül, hogy a szófaji egyértelműsítés pontosságának növekedése ellenére, ez a szótövesítés minőségének csökkenését eredményezte, valamint a teljes morfoszintaktikai egyértelműsítésben is kisebb mértékű csökkenés figyelhető meg. A kapott eredmények kiértékelése során kiderült, hogy az eredménycsökkenés azoknak a szavaknak az elemzéséből adódik, melyek az egyes toldalékok előtt megváltoztatják alapformájukat (pl. *apa - apám, haza - hazát, ...*). Ezekben az esetekben a szóvégi karakterek törlése sikeres, viszont a csonka-szó kiegészítése nem.

| | 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% |

**19. táblázat: A célnyelvi címkekészlet csökkentésével felépített rendszerek eredményei**

A rendszer kimenetének további elemzése során megvizsgáltam a keletkező hibák összetételét. A teszthalmaz 103 930 tokenből áll, amelyek közül az egyértelműsítő rendszer 95 039-et helyesen elemzett, míg a maradék 8891 szó annotációját elrontotta. A helytelen elemzések közül kiemelkedik az OOV szavak egyértelműsítéséből fakadó hiba, ami a rossz elemzések 70,01%-át adja. Ez alapján a minőségben a legnagyobb változás az ismeretlen szavak elemzésének javításával érhető el.

DOI:10.15774/PPKE.ITK.2015.004

A *TÖRÖLCSATOL_SZIMB_SZÁM* rendszer az ismeretlen szavakhoz a leggyakoribb annotációs címkét rendeli hozzá, így csak az ebbe a címkébe tartozó szavakat sikerül helyesen elemeznie. Mivel a teszthalmazban szereplő 7321 OOV szó közül csak 1096 ragozatlan, kis kezdőbetűs melléknév (0##[Af-sn]) és nagy kezdőbetűs tulajdonnév (0##[Np-sn]) szerepel, ezért ezzel a módszerrel az ismeretlen szavak 14,97 %-át egyértelműsítette helyesen a rendszer.

Az első módszer, amit megvizsgáltam az ismeretlen szavak egyértelműsítése érdekében az **ismeretlen szavak osztályozása** volt (6.3.6. fejezet). A módszer lényege, hogy az egyértelműsítés során a tanítóhalmazban nem szereplő szavak a szuffixumuk alapján közös csoportokba kerültek oly módon, hogy az OOV szót egy *unk_abcd* sztringre cserélem, ahol az *abcd* a szó utolsó négy karakterét jelenti.

Ehhez elsőként megvizsgáltam, hogy a magyar nyelvű korpusz esetén mekkora az az ideális karakterszám, amit érdemes megtartani az ismeretlen szavak végéről. A mérések eredményét a 20. táblázat tartalmazza, melyből kiolvasható, hogy 0 és 7 hosszúságú szóvég közötti intervallumon mértem meg az egyértelműsítő rendszer minőségét (*unk_0*-tól *unk_7*-ig). A legmagasabb pontosságú eredményt abban az esetben értem el, amikor a szavak utolsó négy karakterét hagytam meg (*unk_4*). Ennek értelmében a további mérések során ezt a paraméterbeállítást használom.

| Rendszernév | Szószintű | | | Mondatszintű | |
|---|---|---|---|---|---|
| | POS | Lemma | Összes | POS | Összes |
| *unk_0* | 91,494% | 94,329% | 91,445% | 36,995% | 36,703% |
| *unk_1* | 92,367% | 94,567% | 91,460% | 39,645% | 35,924% |
| *unk_2* | 94,413% | 96,056% | 93,377% | 48,906% | 42,958% |
| *unk_3* | 95,599% | 96,975% | 94,562% | 56,117% | 49,106% |
| ***unk_4*** | **96,027%** | **97,819%** | **95,362%** | **58,660%** | **54,129%** |
| *unk_5* | 95,581% | 97,627% | 95,088% | 55,763% | 52,188% |
| *unk_6* | 94,595% | 96,951% | 94,288% | 49,599% | 47,442% |
| *unk_7* | 93,553% | 96,049% | 93,308% | 44,715% | 42,804% |

**20. táblázat: Az OOV szavak végén különböző számú karakter megtartásával készített rendszerek eredményei**

DOI:10.15774/PPKE.ITK.2015.004

A 6.3.6. fejezetben bemutattam, hogy az OOV szavak viselkedése a tanítóanyagban ritkán előforduló szavak viselkedésével modellezhető, ahol ritka szónak számít a tanítóanyagban egy bizonyos küszöbérték alatti gyakorisággal előforduló szó. Például ha ez a küszöb 2, akkor a tanítókorpuszban kevesebb, mint 2-szer szereplő szavakat cserélem le (*LESS_2* nevű rendszer). Ezt az *UNKSZUFFIX* rendszer a modellek tanítása során használja ki oly módon, hogy a tanítóanyagból a ritka szavakat cseréli le az *unk_szuffix* sztringre. A kérdés ebben az esetben az, hogy mekkora értéknél válasszuk meg a küszöbértéket ahhoz, hogy a legjobb eredményt érjük el. A mérési eredményeket a 21. táblázat tartalmazza. A táblázatból kiolvasható, hogy a rendszer minőségét a 2-től 5-ig terjedő intervallum mellett vizsgáltam meg és a legmagasabb pontosságot akkor értem el, amikor a küszöbérték 2 volt, vagyis csak a hapaxokat cseréltem ki *unk_szuffix* sztingre. Ez azzal magyarázható, hogy ha túl sok szót cserélek *unk_szuffix* szóra, akkor a létrehozott csoportok már nem egyértelműen kapcsolódnak egy-egy szófaji címkéhez, hanem átfedések jönnek létre. Továbbá, a megfigyelt eredmények azzal magyarázhatók, hogy az OOV szavaknak a hapaxokból álló korpusszal a legmagasabb a korrelációja (22. ábra), ami a küszöbérték emelésével folyamatosan csökken.

| Rendszernév | Szószintű | | | Mondatszintű | |
|---|---|---|---|---|---|
| | POS | Lemma | Összes | POS | Összes |
| *LESS_2* | 96,027% | 97,819% | **95,362%** | 58,660% | 54,129% |
| *LESS_3* | 95,957% | 97,660% | 95,197% | 58,244% | 53,082% |
| *LESS_4* | 95,913% | 97,549% | 95,086% | 57,858% | 52,265% |
| *LESS_5* | 95,866% | 97,452% | 94,974% | 57,442% | 51,387% |

**21. táblázat: Az *UNKSZUFFIX* rendszer tanításához felhasznált ritka szavak küszöbértékének meghatározása**

Annak érdekében, hogy kihasználjam az SMT-alapú egyértelműsítő rendszer azon tulajdonságát, miszerint a belső modelljeiben tárolt kifejezések hossza változtatható, illetve ezek helyes megválasztása nagyban befolyásolja a fordítórendszer minőségét, megvizsgáltam, hogy mely paraméterbeállítások mellett teljesít legjobban a rendszer. Az eredményeket a 22. táblázat szemlélteti, melyből kitűnik, hogy a legjobb eredmény akkor érhető el, ha a nyelvmodellben 3, a fordítási modellben pedig 6 az aktuálisan elemzendő szó környezetének ablakmérete.

DOI:10.15774/PPKE.ITK.2015.004

| A FORDÍTÁSI MODELL ABLAKMÉRETE | A NYELVMODELL ABLAKMÉRETE | | | | | |
|---|---|---|---|---|---|---|
| | *2* | *3* | *4* | *5* | *6* | *7* |
| *2* | 95,296% | 95,362% | 95,328% | 95,324% | 95,341% | 95,342% |
| *3* | 95,312% | 95,375% | 95,344% | 95,331% | 95,338% | 95,349% |
| *4* | 95,309% | 95,377% | 95,362% | 95,337% | 95,345% | 95,351% |
| *5* | 95,305% | 95,382% | 95,355% | 95,321% | 95,333% | 95,338% |
| *6* | 95,298% | **95,383%** | 95,348% | 95,320% | 95,333% | 95,337% |
| *7* | 95,293% | 95,380% | 95,346% | 95,319% | 95,333% | 95,337% |

**22. táblázat: Az unkSzuffix rendszer eredménye a fordítási és nyelvmodellek függvényében**

A 23. táblázat az *UNKSZUFFIX* rendszerrel elért legjobb eredményt mutatja. Látható, hogy az eddigi legjobb eredményhez képest jelentősnek mondható, 3,9 %-os javulást értem el. A rendszer kiértékelése megmutatta, hogy a javulás az OOV szavak pontosabb elemzésének köszönhető, mivel a szócsoportosításon alapuló módszer segítségével az ismeretlen szavak 73,064%-a helyesen lett elemezve szemben a *TÖRÖLCSATOL_SZIMB_SZÁM* rendszer 14,971%-os pontosságával. Ez azt jelenti, hogy a teszthalmazban szereplő OOV szavak egyértelműsítéséből származó hibák közel harmadát (31,679 %) sikerült javítani. Érdemes megemlíteni, hogy amíg az OOV szavak egyértelműsítése javult, az ismert szavak elemzése 6 %-os romlást mutatott. Ez annak köszönhető, hogy a *UNKSZUFFIX* rendszer tanítóhalmazából eltávolításra kerültek a hapaxok, így ezek is ismeretlen szavak lettek a rendszer számára.

| Rendszernév | Szószintű | | | Mondatszintű | |
|---|---|---|---|---|---|
| | POS | Lemma | Összes | POS | Összes |
| *TÖRÖLCSATOL_SZIMB_SZÁM* | 91,496% | 94,330% | 91,447% | 36,977% | 36,684% |
| *TÖRÖLCSATOL_SZIMB_SZÁM_UNK SZUFFIX* | 96,025% | 97,828% | **95,383%** | 58,752% | 54,284% |

**23. táblázat: Rendszerek eredményei III.**

DOI:10.15774/PPKE.ITK.2015.004

| | Szószintű | | | Mondatszintű | |
|---|---|---|---|---|---|
| | POS | Lemma | Összes | POS | Összes |
| *LESS_2* | 95,496% | 97,248% | 94,941% | 55,116% | 51,233% |
| *LESS_3* | 95,503% | 97,248% | 94,943% | 55,316% | 51,371% |
| *LESS_4* | 95,513% | 97,267% | 94,959% | 55,362% | 51,510% |
| *LESS_5* | 95,516% | 97,273% | 94,967% | 55,331% | 51,525% |
| *LESS_6* | 95,511% | 97,275% | 94,968% | 55,270% | 51,510% |
| *LESS_7* | 95,510% | 97,277% | 94,968% | 55,285% | 51,541% |
| *LESS_8* | 95,506% | 97,272% | 94,962% | 55,254% | 51,510% |
| *LESS_9* | 95,515% | 97,273% | 94,968% | 55,285% | 51,525% |
| *LESS_10* | 95,511% | 97,273% | 94,966% | 55,239% | 51,495% |
| *LESS_11* | 95,514% | 97,279% | 94,969% | 55,239% | 51,495% |
| *LESS_12* | 95,514% | 97,278% | 94,970% | 55,208% | 51,464% |
| *LESS_13* | 95,516% | 97,280% | 94,973% | 55,223% | 51,495% |
| *LESS_14* | **95,516%** | **97,281%** | **94,976%** | **55,239%** | **51,525%** |
| *LESS_15* | 95,515% | 97,280% | 94,974% | 55,223% | 51,525% |
| *LESS_16* | 95,514% | 97,282% | 94,975% | 55,223% | 51,525% |
| *LESS_17* | 95,511% | 97,280% | 94,973% | 55,193% | 51,495% |
| *LESS_18* | 95,510% | 97,276% | 94,970% | 55,177% | 51,464% |
| *LESS_19* | 95,515% | 97,276% | 94,971% | 55,177% | 51,448% |
| *LESS_20* | 95,515% | 97,277% | 94,971% | 55,193% | 51,464% |

**24. táblázat: A guesser tanítóanyag-méretének meghatározása szógyakoriság alapján**

Ahogy azt a 6.3.7. fejezetben bemutattam, az OOV szavak pontosabb egyértelműsítése érdekében rendszeremhez **szóvégalapú teljes morfoszintaktikai ajánlórendszert** integráltam (*GUESSER*). A suffix guesser előfeldolgozó lépésként működik a rendszerben, és az *UNKSZUFFIX* rendszerhez hasonlóan a tanítóhalmazban szereplő ritka szavakkal tanítottam be. Az előző rendszerhez hasonlóan itt is kulcskérdés a megfelelő szógyakorisági küszöbérték megtalálása, melynek segítségével a guesser tanítóhalmazának mérete határozható meg. A megfelelő küszöbérték megtalálása érdekében a rendszert 2 és 20 közötti intervallumon vizsgáltam (*LESS_2*-től *LESS_20*-ig), ami azt mutatja, hogy a tanítóhalmazban mekkora gyakorisággal előforduló szavakat teszem bele

DOI:10.15774/PPKE.ITK.2015.004

a guesser tanítóanyagába. Az eredményeket a 24. táblázat tartalmazza. A táblázatokból kiolvasható, hogy a legjobb eredményt akkor értem el, amikor a guessert a tanítóhalmazban kevesebb, mint tizennégyszer előforduló szavakkal tanítottam.

Az előző rendszerekhez hasonlóan megvizsgáltam, hogy a rendszer minősége miként változik a modellekben használt kifejezések méretének függvényében. Méréseim eredményeit a 25. táblázat szemlélteti, mely megmutatja, hogy a legjobb paraméterbeállítás a legfeljebb két szó hosszú kifejezésekből álló fordítási modell és a 10 szó hosszú nyelvmodell esetén érhető el.

| | | *A FORDÍTÁSI MODELL KIFEJEZÉSHOSSZA* | | | | |
|---|---|---|---|---|---|---|
| | | *2* | *3* | *4* | *5* | *6* |
| *A NYELVMODELL KIFEJEZÉSHOSSZA* | *2* | 94,971% | 94,964% | 94,958% | 94,932% | 94,737% |
| | *3* | 94,971% | 94,963% | 94,955% | 94,929% | 94,737% |
| | *4* | 94,971% | 94,963% | 94,955% | 94,922% | 94,737% |
| | *5* | 94,969% | 94,962% | 94,954% | 94,921% | 94,737% |
| | *6* | 94,969% | 94,962% | 94,954% | 94,742% | 94,737% |
| | *7* | 94,968% | 94,962% | 94,953% | 94,742% | 94,721% |
| | *8* | 94,968% | 94,961% | 94,951% | 94,742% | 94,721% |
| | *9* | 94,966% | 94,961% | 94,950% | 94,740% | 94,721% |
| | *10* | **94,976%** | 94,966% | 94,960% | 94,943% | 94,740% |
| | *11* | 94,974% | 94,966% | 94,959% | 94,938% | 94,740% |
| | *12* | 94,972% | 94,964% | 94,959% | 94,932% | 94,737% |

**25. táblázat: A *GUESSER* rendszer eredményei a fordítási- és nyelvmodellekben alkalmazott kifejezések hosszának függvényében**

További vizsgálataimmal arra kerestem a választ, hogy az egyértelműsítő rendszer minőségét hogyan befolyásolja az ismeretlen szavakhoz rendelt elemzési javaslatok száma. Ez a mennyiség attól függ, hogy a guesser kimenetén lévő legvalószínűbb címkék közül hányat rendelünk hozzá az OOV szóhoz, és ezáltal a dekódernek hány javaslat közül kell kiválasztania a megfelelő elemzést. A rendszer alapbeállításként a 10 legvalószínűbb címkét adja át a dekóder számára. A 26. táblázat bemutatja az 1 és 6 közötti értékek esetén elért fordítási eredményeket. A legjobb fordítás akkor született, ha ez az érték 1, vagyis a guesser csak az OOV szóhoz rendelt legnagyobb valószínűségű címkét adja át az elemzőnek.

DOI:10.15774/PPKE.ITK.2015.004

A 27. táblázatban felsorolt eredményekből látható, hogy a morfológiai guesserrel kiegészített *TÖRÖLCSATOL_SZIMB_SZÁM_GUESSER* rendszer 96,177%-os fordítási pontosságot ér el, ami közel 0,8%-os javulás a *TÖRÖLCSATOL_SZIMB_SZÁM_UNKSZUFFIX* rendszerhez képest. Az eredmények mélyebb elemzése során kiderült, hogy míg az utóbbi rendszer 1972 ismeretlen szót elemez helytelenül, ez a szám a guesserrel kiegészített rendszer esetében csak 1441 darab. Emellett azonban nemcsak az OOV szavak elemzésében történt javulás, ugyanis az ismert (nem OOV) szavaknál a *TÖRÖLCSATOL_SZIMB_SZÁM_GUESSER* rendszer csak 2532 szó esetében ront elemzést. A *TÖRÖLCSATOL_SZIMB_SZÁM_UNKSZUFFIX* rendszernél ez az érték 2826, mivel itt a ritka szavak is OOV szavak lesznek; ez pedig gyengébb eredményt jelent az eddigi legjobb a *TÖRÖLCSATOL_SZIMB_SZÁM* rendszer 2666 darab rossz elemzéséhez képest is. Következésképp a *TÖRÖLCSATOL_SZIMB_SZÁM_GUESSER* rendszer nemcsak a fordítás szempontjából ért el jobb eredményt, hanem az ismeretlen szavak kezelésében is jobban teljesít az előző rendszerekhez viszonyítva.

| | | Szószintű | | | Mondatszintű | |
|---|---|---|---|---|---|---|
| | | POS | Lemma | Összes | POS | Összes |
| *AZ ELEMZÉSI JAVASLATOK SZÁMA* | 1 | 96,511% | 98,595% | **96,177%** | 62,465% | 59,692% |
| | 2 | 95,940% | 97,670% | 95,320% | 57,951% | 53,513% |
| | 3 | 95,737% | 97,440% | 95,129% | 56,595% | 52,327% |
| | 4 | 95,641% | 97,361% | 95,053% | 55,978% | 51,988% |
| | 5 | 95,588% | 97,321% | 95,013% | 55,609% | 51,695% |
| | 6 | 95,557% | 97,304% | 94,993% | 55,470% | 51,603% |

**26. táblázat: Az OOV szavakhoz rendelt elemzési javaslatok számának változtatása**

A morfológiai guesser hátránya, hogy nem rendelkezik semmiféle grammatikai háttérismerettel, pusztán csak statisztikai számítások alapján a ritka szavakból következtet a lehetséges címkékre. Emiatt olyan címkéket is figyelembe vesz lehetőségként, melyek nyelvtanilag helytelenek. Ezen a problémán tud segíteni, ha a szuffix-guessert **morfológiai elemzővel** kombinálom (6.3.8. fejezet). Annak ellenére, hogy a morfológiai elemző integrálásával a rendszer elveszíti a nyelvfüggetlen tulajdonságát, ez bizonyul logikus lépésnek az adott korpuszon történő legjobb egyértelműsítés elérése érdekében.

DOI:10.15774/PPKE.ITK.2015.004

A morfológiai elemző rendszer integrálásának legnagyobb nehézsége, hogy előállítása erőforrásigényes, mivel minden nyelv és címkekészlet esetén külön-külön létre kell hozni a kézzel írott szabályokat. Ez az általam használt Szeged Korpusz 2 [111] esetén is problémát jelentett, hiszen nem létezik az MSD kódrendszer ezen verziójával kompatibilis morfológiai elemző. A morfológiai elemző helyettesítésére egy a korpuszból felépített morfológiai lexikont alkalmaztam. A lexikon alkalmazása nem modellezi tökéletesen az elemző működését, mivel így a teszthalmaz nem tartalmaz olyan szót, amely nincs benne a lexikonban és ne ismernénk az elemzését. A modell ezen gyengesége azonban kiküszöbölhető oly módon, hogy csak azokhoz a szavakhoz használom a lexikont, amelyeket egy másik kódkészlettel működő morfológiai elemző is ismer. Ehhez a HUMOR kódrendszert használó HUMOR [44] elemzőt integráltam a rendszerembe.

| | 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% |

**27. táblázat: A különböző felépítésű *TÖRÖLCSATOL* rendszerek eredményei**

A 28. táblázat mutatja az eddig bemutatott rendszerek és a morfológiai lexikonnal kiegészített rendszer eredményei közti különbséget. Az adatok számszerűsített elemzéséből megállapítottam, hogy a 103 930 tokenből csak 915 esetén hibás a lemmatizálás. Az OOV szavak tekintetében csak 1111 esetben lett az elemzés hibás. Így az OOV szavak elemzése 84,82%-ban helyes, ami a guesserrel kiegészített rendszerhez képest további 23%-os javulást eredményez. A nem OOV szavak elemzése érdemben nem változott (csak 3 szóval javult a helyes elemzések száma a *TÖRÖLCSATOL_SZIMB_SZÁM_GUESSER* rendszerhez képest).

DOI:10.15774/PPKE.ITK.2015.004

| | Szószintű | | | Mondatszintű | |
|---|---|---|---|---|---|
| | POS | Lemma | Összes | POS | Összes |
| *TÖRÖLCSATOL_SZIMB_SZÁM* | 91,496% | 94,330% | 91,447% | 36,977% | 36,684% |
| *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% |

**28. táblázat: A morfológiai guesser és lexikon integrálásával felépített rendszerek eredményei**

### 6.4.3 Az SMT-alapú egyértelműsítő rendszer összehasonlítása más magyar nyelvű rendszerekkel

A 6.4.2. fejezetben azt mutattam be, hogy az általam felépített SMT-alapú teljes morfoszintaktikai egyértelműsítő rendszer mekkora pontossággal képes működni egy magyar nyelvű adathalmazon. A továbbiakban a létrehozott rendszereim eredményét fogom összehasonlítani különböző, magyar nyelvre is alkalmazható rendszerekkel.

Az összehasonlítást a Szegedi Tudományegyetem Informatikai Tanszékcsoportjának **magyarlanc** [117] nevű elemzőjének vizsgálatával kezdtem. Ezt az elemzőt azonban azért nem alkalmaztam munkám során, mivel nem ad lehetőséget a belső modellek egyéni tanítására, tanítóhalmaza tartalmazza a teszthalmaz mondatait is, ami gátolja az összehasonlítás helyességét, továbbá egyszerűsített MSD kódrendszere nem kompatibilis a Szeged Korpusz 2 [111] által használttal.

Az egyik összehasonlításra használt rendszer a nyelvfüggetlen HMM-alapú **HunPos** [113] nevű szófaji egyértelműsítő. A HunPos rendszer legfontosabb újítása az eredeti HMM algoritmussal szemben, hogy a vizsgált szó elemzése közben figyelembe veszi a megelőző szó elemzését is. Méréseik alapján ezzel a technikával jelentősen növelhető a rendszer pontossága. Mivel a HunPos csak szófaji egyértelműsítést végez, ezért szükséges volt a **CST lemmatizáló**val [118] való kiegészítése annak érdekében, hogy a teljes morfoszintaktikai egyértelműsítésben elért eredményét tudjam a saját rendszeremével összehasonlítani.

A másik ilyen rendszer az Orosz és Novák által a HunPos alapján létrehozott nyílt forráskódú HMM-alapú teljes hibrid morfológiai annotáló eszköz, a **PurePos2** [43], amely a szótövesítést és morfoszintaktikai címkézést egyidejűleg végzi. A PurePos2 egyik előnye, hogy igény szerint működtethető morfológiai egyértelműsítővel is, mely nélkül viszont képes a nyelvfüggetlen viselkedés produkálására is.

DOI:10.15774/PPKE.ITK.2015.004

Az összehasonlításhoz először összefoglalom a HMM- és az SMT-alapú morfológiai elemzők közti legjelentősebb különbségeket, valamint bemutatom a gépi fordító módszerből származó előnyöket, melyek a következőképpen fogalmazhatók meg:

- Míg a HMM-alapú rendszerek többségénél a címkeátmenet-valószínűség modell trigram-alapú, addig az SMT rendszer a címkék kiválasztásához az összes megelőző szó címkéjét fel tudja használni. Ez azt jelenti, hogy a fordítórendszer az elemzés során sokkal több környezeti információval rendelkezik.
- Jelentős különbség a két módszer által használt simító algoritmus, mivel a Moses toolkit a Kneser-Ney simítást [119] alkalmazza, ami egy rendkívül hatásos módszer a HMM-alapú rendszer lineáris interpolációjával szemben.
- A létező HMM-alapú megoldások a szavak lehetséges címkéinek a megállapítása során semmilyen környezeti információt nem használnak, csupán az aktuális szó lehetséges morfoszintaktikai címkéit nézik. Ezzel szemben a kifejezésalapú gépi fordítórendszer fordítási modellje képes több szóból álló kifejezéseket egy fordítási egységként kezelni. Ennek köszönhetően az SMT-alapú szófaji egyértelműsítő rendszer felhasználja a szavak kontextuális információit is.
- Fontos különbség a dekódoló algoritmus, valamint a dekódolás sorrendje. A HMM-alapú rendszerek leggyakrabban Viterbi algoitmust használnak, ami szigorúan balról jobbra irányítottsággal végzi a szövegek elemzését. Ezzel szemben az SMT esetében a dekódoló metódus a nyalábolt keresési algoritmus egy hatékony és gyors változatát, az úgynevezett verem dekódolást alkalmazza, mely képes tetszőleges sorrendű működésre. Ennek köszönhetően a fordításalapú módszer a problémás kifejezések egyértelműsítése érdekében segítségül tudja hívni a mondatban tőlük hátrébb elhelyezkedő egyértelműen felcímkézhető szavak elemzését.

Az általam készített nyelvfüggő (*TÖRÖLCSATOL_SZIMB_SZÁM_MORFLEXIKON*) és nyelvfüggetlen (*TÖRÖLCSATOL_SZIMB_SZÁM_UNKSZUFFIX, TÖRÖLCSATOL_SZIMB_SZÁM_GUESSER*) SMT-alapú egyértelműsítő rendszerek lemmatizálás és teljes morfoszintaktikai egyértelműsítés tekintetében jobb eredményt értek a *HUNPOS+CST_SZÓTÖVESÍTŐ* és a *PUREPOS2* rendszerekhez képest (29. táblázat). Szófaji egyértelműsítés tekintetében a guessert és a morfológiai lexikont tartalmazó rendszerek teljesítenek a legjobban (96,511% és 96,624%). A Wilcoxon-próba alapján rendszereim statisztikailag szignifikánsan jobbak, mind egymáshoz viszonyítva, mind a többi rendszerhez viszonyítva.

DOI:10.15774/PPKE.ITK.2015.004

| | Szószintű | | | Mondatszintű | |
|---|---|---|---|---|---|
| | POS | Lemma | Összes | POS | Összes |
| *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% |

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

A 29. táblázat a különböző nyelvfüggetlen egyértelműsítő rendszerek eredményeinek összehasonlítását mutatja be. A HunPos és PurePos2 rendszerek mellett a harmadik vizsgált rendszer a **Morfette** [120] – moduláris adatvezérelt statisztikai rendszer –, ami teljes morfoszintaktikai egyértelműsítést végez. Szófaji egyértelműsítés tekintetében ez a legjobb teljesítményű rendszer a vizsgált rendszerek közül (0,133%-kal magasabb pontosságot ért el a legjobb rendszeremhez képest); viszont lemmatizálásban 3,071%-kal, teljes morfoszintaktikai egyértelműsítés esetén 2,722%-kal gyengébb a *TÖRÖLCSATOL_SZIMB_SZÁM_MORFLEXIKON* rendszernél. Ezzel egyidejűleg ezekben a mérőszámokban a nyelvfüggetlen *TÖRÖLCSATOL_SZIMB_SZÁM_GUESSER* rendszer is jobban teljesít a Morfette-nél.

Az utolsó vizsgált rendszer az **NLTK** eszközgyűjtemény [121] maximumentrópia-alapú szófaji egyértelműsítője [122], mely minden tekintetben a leggyengébb eredményt ért el a vizsgált rendszerek között. A HunPos-hoz hasonlóan ennél a rendszernél is a CST lemmatizálót integráltam szótövesítés céljából. A CST lemmatizáló egy olyan szabad forráskódú gépi tanuláson alapuló szótövesítő rendszer, amely a prefixumok, infixumok és szuffixumok egyidejű kezelésével határozza meg egy szóalak lemmáját.

## 6.5 Az SMT-alapú teljes morfoszintaktikai egyértelműsítő rendszer nyelvfüggetlen viselkedése

Az eddigiekben rendszeremet a magyar nyelvű MSD kódrendszerrel működő Szeged Korpusz 2-re optimalizáltam. Mivel a HuLaPos2 a morfológiai elemző kivételével nyelvfüggetlen, megvizsgáltam, mennyire alkalmazható különböző morfológiai gazdagságú nyelvek esetén.

DOI:10.15774/PPKE.ITK.2015.004

A HuLaPos2 rendszert öt különböző nyelven (szerb, horvát, bolgár, portugál és angol) elérhető és egy magyar (de más kódkészlettel működő) szófaji egyértelműsítő rendszer eredményeivel hasonlítottam össze. Az összehasonlításhoz olyan eszközöket választottam, amelyeknél elérhető a felhasznált tanítóanyag, valamint a korpusz pontos felosztása is reprodukálható. A rendszerek pontosságának részletes összehasonlítását a 30. és 31. táblázatokban foglaltam össze, ahol az első táblázatba azok a rendszerek kerültek, amelyek teljes morfoszintaktikai egyértelműsítést hajtanak végre (magyar, horvát, szerb), míg a második táblázatban szereplők csak szófaji egyértelműsítést végeznek, lemmatizációt azonban nem (angol, portugál, bolgár).

Szerb és horvát nyelvre Agić et al. [123] készítettek szófaji címkéző és szótövesítő alkalmazást 2013-ban. A rendszert a HunPos és a CST szótövesítő [118] kombinációjából építették fel. Munkájuk során a SETimes.HR [123] (a Southeast European Times horvát nyelvű újság szövegei) korpuszt használták, ami egy szerb és horvát nyelvű szövegeket és függőségi viszonyokat tartalmazó treebank. A szerb és a horvát korpuszrészek egyenként körülbelül 4000 kézzel lemmatizált és morfoszintaktikailag elemzett mondatot tartalmaznak. Rendszeremet 100-100 mondatból álló teszthalmazzal értékeltem ki. A 30. táblázat eredményeiből látható, hogy POS címkézés esetén a HuLaPos2 teljesítménye szignifikánsan meghaladja Agićék rendszerét, míg a szótövesítésben elért eredmény is közelít annak eredményességéhez. A különbség a javasló algoritmus működéséből ered: a CST rendszerben a szótő-transzformációk nemcsak szuffixumok lehetnek, hanem a tetszőleges helyű változások is. Ezzel szemben a HuLaPos2 által használt guesser csak a szóvégi változást képes kezelni.

Az általam készített HuLaPos2 rendszert a 6.3. és 6.4. fejezetekben leírt módon az MSD kódrendszerű Szeged Korpuszt 2-n fejlesztettem és teszteltem. Ugyanennek a korpusznak létezik egy HUMOR kóddal [44] címkézett változata, ahol az eredeti kódokat automatikus módszerrel konvertálták át. A HuLaPos2 rendszert a PurePos2 rendszer morfológiai elemzőt használó, valamint anélkül működő (tehát nyelvfüggetlen) változataival hasonlítottam össze. A 30. táblázatban szemléltetett eredmények megmutatták, hogy rendszerem az összes mért esetben jobb eredményt ért el a PurePos morfológiai elemző nélküli változatával szemben, és szófaji címkézés esetén pontossága megközelíti a PurePos morfológiai elemzős (PurePos+MA) változatát.

DOI:10.15774/PPKE.ITK.2015.004

| Nyelv | Rendszer | Szószintű pontosság | | |
|---|---|---|---|---|
| | | címkézés | szótövesítés | teljes |
| magyar (HUMOR) | **HuLaPos2** | 96,70% | 98,23% | 97,62% |
| | PurePos | 96,50% | 96,27% | 94,53% |
| | 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%** | - |

**30. táblázat: Különböző nyelvű teljes morfoszintaktikai egyértelműsítő rendszerek eredményeinek összehasonlítása.**

A következő három nyelv (bolgár, portugál és angol) esetében csak a POS címkézés eredményességét tudtam összehasonlítani (31. táblázat) rendszeremmel, mivel az elérhető korpuszok nem tartalmazták a szavak lemmáit.

Georgi Georgiev et al. [101] létrehoztak egy morfológiai lexikonnal és nyelvtani szabályokkal kiegészített irányított tanuláson alapuló szófaji egyértelműsítő rendszert bolgár nyelvre. Eszközüket a BulTreeBank korpuszon [124] tanították és tesztelték. A 31. táblázat eredményeiből látható, hogy a HuLaPos2 teljesítménye nagymértékben meghaladja a nyelvtani tudással nem rendelkező, tisztán statisztikai módszereket használó rendszerek minőségét. Annak ellenére, hogy a HuLaPos2 semmilyen nyelvspecifikus eszközzel nincs támogatva, jobban teljesít, mint a morfológiai lexikont használó eszköz, valamint pontossága megközelíti a Georgiev által készített legjobb rendszerét (irányított tanulás + lexikon + szabályok).

Portugál nyelvre a Maia és Xexéo [125] által 2011-ben készített HMM-alapú rendszert vettem összehasonlítási alapul. Módszerük az eredeti HMM módszertől annyiban tér el, hogy a címkeátmenet-valószínűségi modell nem szó-, hanem karakter-n-gram alapú. Ez az eszköz a Floresta Sintá(c)tica Treebank anyagán [126] lett tanítva, melyből az első 10% volt a teszthalmaz, a fennmaradó 90% pedig a tanítóhalmaz. Ugyanezekkel a beállításokkal a HuLaPos2 pontossága 1,2%-kal felülmúlta a portugál címkéző eredményeit.

Ami az angol nyelvet illeti, a Penn Treebank [46] WSJ korpuszát használtam. Az összehasonlíthatóság érdekében a korpuszt Collins alapján [127] szokás felosztani 18-3-3 arányban (41 111 mondat tanításra, 7201 optimalizálásra, 6272 mondat tesztelésre).

A 31. táblázat a HuLaPos2 és a másik négy rendszer által elért eredményeket mutatja. Megfigyelhető, hogy a HuLaPos2 pontossága meghaladja a TnT, továbbá a Mora és Peiró-féle [109] rendszerek által elért értékeket. A Stanford Parser 2.0 [39] valamint az SCCN [100] rendszerek nyelvfüggő nyelvi jellemzőket használnak. Az eredmények egyedülállóak abban a tekintet-

DOI:10.15774/PPKE.ITK.2015.004

ben, hogy az általam felépített rendszer a tanítóanyagon kívül semmilyen más lexikai adatbázist, vagy előzetes tudást nem használ.

| Nyelv | Rendszer | A címkézés pontossága |
|---|---|---|
| bolgár | TnT [128] | 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 [128] | 96,46% |
| | kifejezésalapú fordító [109] | 96,97% |
| | **HuLaPos2** | 97,08% |
| | Stanford Parser 2.0 [39] | 97,32% |
| | SCCN [100] | 97,50% |

**31. táblázat: Különböző nyelvű szófaji egyértelműsítő rendszerek eredményeinek összehasonlítása**

## 6.6 Kapcsolódó munkák, előzmények

Többféle, különböző módszereken alapuló megoldás létezik morfológiai elemzés megvalósítására. Ez a fejezet azokat az implementációkat mutatja be, amelyek a munkám szempontjából relevánsnak bizonyultak.

Az első megoldások előre definiált **kézzel írott szabály**rendszereket alkalmaztak szófaji egyértelműsítésre ([129]–[131]). A szabályalapú rendszerek legnagyobb gyengesége, hogy megalkotásuk rendkívül sok emberi ráfordítást vesz igénybe, valamint komoly nyelvészeti tudást feltételez. Ráadásul ezek a rendszerek nem bizonyultak túlságosan robosztusnak, főleg ha új domainen vagy egy új nyelven alkalmazták őket. A gépi tanulási módszerek alkalmazása éppen ezeket a problémákat hivatottak orvosolni, valamint adatforrások széles tárházát képesek kiaknázni, mint például lexikonok, nagy egynyelvű szövegek, párhuzamos kétnyelvű korpuszok stb.

Az első úttörő megoldással Brill [132] állt elő, aki a számítógépre bízta a szabályok megalkotását. Módszerének lényege, hogy felügyelet nélküli gépi tanulási módszerrel egy tetszőleges címkéző rendszer kimenete és a referenciaelemzés segítségével javító szabályokat definiál. A tanítás során a rendszer előre definiált sablonok alapján találja meg a szükséges transzformációk helyét. Ezután a létrehozott javító szabályokat a vizsgált egyértelműsítő rendszer kimenetén, mint utófeldolgozó lépést alkalmazva javítható a címkézés pontossága. Emiatt ezt a megközelítést a szakirodalomban **transzformáció-alapú** gépi tanulási módszernek nevezik (TBL – transformation

DOI:10.15774/PPKE.ITK.2015.004

based learning). A transzformáció-alapú módszer célja nem a közvetlen szófaji egyértelműsítés, hanem egy létező egyértelműsítő rendszer minőségének javítása. A létrehozott rendszer rugalmasan képes a különböző sablonokat kezelni, és hatásosan képes a belső modelljébe integrálni őket. Ellenben a módszer rendkívül lassú, ami igencsak megnehezíti az alkalmazását, valamint hajlamos a túltanulásra.

Yarowsky et al. [133] munkájuk során erőforrásokban szegény nyelvek számára készítettek nyelvtechnológiai eszközöket. Ehhez a gyakran kutatott nyelvek létező eszközeit használták fel, ahol már rendelkezésre álltak jó minőségű nyelvfeldolgozó eszközök, mint például az angol vagy japán nyelvek. Egy párhuzamos kétnyelvű korpusz angol oldalát a Brill-tagger [132] segítségével címkézték, majd szóösszekötést alkalmaztak. Ily módon próbálták meg „átvetíteni" a címkéket a korpusz másik nyelvének szavaira. A megoldás hiányossága, hogy csak a Penn TreeBank [46] fő szófaji címkéit alkalmazza, másrészt pedig a szóösszekötő teljesítménye nagyban befolyásolja a rendszer minőségét. Fossum és Abney [134] ezt a módszert terjesztette ki oly módon, hogy nemcsak egy „forrásnyelvet" vesz figyelembe, hanem több nyelvből érkező információkat is felhasznál az elemzés során.

A teljes egészében statisztikai módszereket használó szófaji egyértelműsítő rendszerek közül a két legjobban elterjedt típus a **maximumentrópia-alapú**, valamint a HMM-alapú megvalósítások. A MaxEnt módszer lényege, hogy a rendszer egy mondat címkézését előre definiált jellemzők súlyozott szorzatából állítja elő. A módszer rugalmassága abból fakad, hogy bármilyen – általunk fontosnak tartott – jellemzőt felhasználhatunk, legyen az akármilyen egyszerű vagy összetett. Továbbá nem szükséges, hogy a tulajdonságok egymástól függetlenek legyenek, amiből következik, hogy a rendszer felhasználhat az átfedésekből és az egymástól kölcsönösen függő jellemzőkből adódó információkat. A MaxEnt-alapú szófaji egyértelműsítő létrehozása Ratnaparkhi [135] nevéhez fűződik. A módszert 2000-ben Toutanova és Manning fejlesztették tovább [39], [40] új jellemzők integrálásával, melyek a szótárban nem szereplő szavak, a ritka szavak és a funkciószók elemzését hivatottak javítani. Az általuk létrehozott rendszer az úgynevezett Stanford Parser. A Stanford Parser egy magyar nyelvre készített adaptációja az úgynevezett magyarlanc [117], ami egy magyar nyelvre specializálódott modulokat tartalmazó NLP alkalmazás. A Malecha és Smith [122] által készített, az NLTK toolkit [121] programgyűjteményhez integrált, szófaji egyértelműsítő rendszer is MaxEnt-alapú. Chrupała et al. [120] létrehozták a Morfette nevű kifejezésalapú teljes szintaktikai egyértelműsítő rendszert, melyben MaxEnt-alapú osztályozó modellekkel végzik a lemmatizálást és a címkézést. Finch és Sumita [136] egy MaxEnt-alapú POS taggert készített, melybe feature-ként integrálták a gépi fordítórendszer frázistábláját. Ennek segít-

DOI:10.15774/PPKE.ITK.2015.004

ségével fel tudták használni az SMT rendszernek a szövegkörnyezetből származó információ kinyerését, ami eltér a MaxEnt módszerétől. A legjobb eredményt a két módszer kombinációjából felépített rendszerrel sikerült elérniük, mellyel mindegyik teszthalmazon jobb eredményt értek el az eredeti MaxEnt módszerhez képest.

A MaxEnt módszer előnye, hogy hatékonyan kihasználja a környezetből származó információkat azáltal, hogy a címkézés során nem tekinti a szavakat egymástól függetlennek. Ezt azon az áron teszi, hogy a MaxEnt-alapú rendszerek lassúak mind a tanítás, mind a dekódolás tekintetében. További nehézség rejlik a releváns jellemzők megfogalmazásában, hiszen ezek legalább olyan nehezen definiálhatók, mint a szabályalapú rendszer szabályai. Nagyon nehéz egy olyan, mindent lefedő jellemzővektort kialakítani, amely hatékonyan jellemzi a vizsgált problémát.

A szófaji egyértelműsítés feladatára létrehozott számos módszer közül a legelterjedtebbek a **rejtett Markov-modellen** alapulók. A HMM-alapú megközelítés a címkézés feladatát a zajoscsatorna-modell segítségével közelíti, ezáltal egy szó helyes szófaját két valószínűségi változó (a lexikai-valószínűség modell és a címkeátmenet-valószínűség modell) maximalizált szorzatával határozza meg. Habár nem egyértelmű, ki készítette az első Markov-modellen alapuló POS taggert, a szakirodalomban 1976-ban elsőként Bahl és Mercer [137] foglalkozott a témával. Egy másik korai munka, amely a statisztikai alapú címkézést népszerűsítette, Church nevéhez fűződik [138], aki standard Markov-modellt használt egy egyszerű simító algoritmussal. Ezek után több munka foglalkozott a Markov-modellen alapuló szófaji címkézőkkel [139]–[141].

Az egyik legpontosabb és elérhetőségének köszönhetően legtöbbet idézett HMM-alapú POS tagger a Brants által 2000-ben készített TnT tagger [128]. Sikerének titka az alapötlethez képest az alkalmazott simító algoritmus, valamint a szótárban nem szereplő szavak kezelése. A simítást egy környezetfüggetlen lineáris interpolációval végzi, ami formálisan a következőképpen írható le:

$$P(t_i|t_{i-1}t_{i-2}) \cong \lambda_1 P(t_i) + \lambda_2 P(t_i|t_{i-1}) + \lambda_3 P(t_i|t_{i-1}t_{i-2}) \quad (11)$$

ahol $\lambda_i \in [0;1]$ és $\lambda_1 + \lambda_2 + \lambda_3 = 1$. A rendszer általában a $\lambda_i$ értékeket a – tanító- és tesztkorpusztól eltérő – optimalizációs korpuszból tanulja. Az ismeretlen szavak eloszlásának becslését egy végződésfa-alapú ajánlórendszer segítségével oldja meg, amit a tanítóhalmazban kevesebb, mint 10-szer szereplő szavakon tanítottak. A TnT tagger másik érdekes jellemzője a kapitalizáció figyelembe vétele a címkekészlet kialakítása során; megfigyelhető ugyanis, hogy a címkék valószínűség-eloszlása a nagy- és kis kezdőbetűs szavak környékén egymástól eltér. A kapitalizációt a következőképp építették be a modellbe:

DOI:10.15774/PPKE.ITK.2015.004

$$P(t_3|t_2,t_1) \rightarrow P(t_3,c_3|t_2,c_2,t_1,c_1) \quad (12)$$

A tagger hatékonyságának növelése érdekében a mondatok beolvasásánál a Viterbi algoritmussal párhuzamosan a beam search algoritmust is alkalmazza. Az így felépített TnT tagger 97%-os pontosságot ért el a Penn Treebank korpuszon [46].

Az első, magyar nyelvre készült statisztikai módszereken alapuló POS taggert Oravecz és Dienes [107] készítette. Rendszerük alapja a TnT tagger, melyhez morfológiai lexikont és suffix guessert integráltak. Munkájukban bebizonyították, hogy az általuk vizsgált rendszerek így sokkal nagyobb pontosságot értek el a magyar mint ragozó nyelv esetében. Az eredmények javulása főleg a szótárban nem szereplő szavak pontosabb elemzésének köszönhető.

Halácsy et al. [142] kipróbálta a létező POS tagger architektúrákat magyar nyelvre. Kísérleteket végeztek HMM- és MaxEnt-alapú rendszerekkel is, amikhez a hunmorph rendszert [143] mint morfológiai elemzőt integrálták. Megmutatták, hogy a sztochasztikus komponensek és a szimbolikus morfológiai elemzők hatásosan kombinálhatók egymással. Legjobb rendszerük 98,17%-os pontosságot ért el, és elég robusztusnak bizonyult az OOV szavak területén is.

Egy másik munkájukban Halácsy et al. HunPos [113] néven reimplementálták a TnT taggert [128]. A szótárban nem szereplő szavak kezelésére egy morfológiai lexikont alkalmaztak, ami a bemenet minden szavához tartalmazza a lehetséges morfoszintaktikai címkéket. A lexikon találatait a TnT-hez hasonló végződésfa-alapú ajánlórendszer segítségével súlyozzák. A kis- és nagybetűs szavakat külön guesserrel kezelik. Ez a módszer nagyban javítja a címkézés pontosságát a morfológiailag gazdag nyelvek – mint például a magyar – esetében. A HunPos rendszer legfontosabb újítása az eredeti HMM algoritmussal szemben, hogy a vizsgált szó elemzése közben figyelembe veszi a megelőző szó elemzését is, ami formálisan a (13) egyenlettel írható le.

$$P(w_i|t_i) \rightarrow P(w_i|t_{i-1},t_i) \quad (13)$$

Méréseik alapján ezzel a technikával jelentősen növelhető a rendszer pontossága. A rendszer MSD kódokat [115] használ, és 98,24%-os pontosságot ért el [142].

Orosz és Novák 2013-ban létrehoztak egy nyílt forráskódú, HMM-alapú teljes hibrid morfológiai annotáló eszközt, a PurePos2-t [43], [114], amely a szótövesítést és morfoszintaktikai címkézést egyidejűleg végzi. A HunPos és TnT rendszerekhez hasonlóan a speciális tokenek kezelésére egy lexikális modellt tartalmaz, valamint a szótárban nem szereplő szavak elemzésére suffix-trie guessert használ. Ezekkel a rendszerekkel ellentétben azonban a PurePos egy morfológiai elemzőt foglal magába (HUMOR). A HUMOR elemző [45], [144] a HUMOR kódrendszeren

DOI:10.15774/PPKE.ITK.2015.004

alapul. A PurePos2 a TnT-hez hasonlóan Viterbi-dekódert alkalmaz, ám ezenkícül alternatív megoldásként beam search dekódoló algoritmust is használ. A rendszer gyorsan tanítható, és egyszerűen integrálhatók bele különböző szabályalapú komponensek. A PurePos rendszer gazdag morfológiájú nyelvek esetén alkalmazható hatékonyan, továbbá abban az esetben, ha csak kis méretű tanítóanyag áll rendelkezésre.

Ahogy azt a 6.4.3. fejezetben bemutattam, az SMT-alapú szófaji egyértelműsítő rendszer izomorfnak tekinthető a HMM-alapú megközelítéssel. A legnagyobb különbség a két módszer között a belső modellek kifejtésében rejlik. A fenti alkalmazásokhoz képest a legnagyobb eltérés a rendszer által használt ablak mérete, ami azt jelenti, hogy egy szó elemzéséhez sokkal több környezeti információt használ fel.

Mora és Peiró [71] munkájuk során statisztikai gépi fordítót alkalmaztak szófaji egyértelműsítésre. A rendszert az angol nyelv morfológiai egyértelműsítésére tervezték, de lemmatizálásra nem. Munkájukban a tanítóanyagban nem szereplő szavak kezelésére egy szógyakoriságon alapuló modellt és egy 11 elemből álló szuffixum listát alkalmaztak. A tanulmány arra az eredményre jutott, hogy a legjobb eredmények angol nyelvre úgy érhetők el, ha a fordítandó frázisok maximális hosszát és a nyelvi modell rendjét is 3-ra állítják be. Az általuk bemutatott beállítások nem alkalmazhatók ragozó nyelvek esetén, mivel ebben az esetben olyan sok toldalék van, ami nem sorolható fel egy egyszerű toldaléklistában.

## 6.7 Összegzés

A szófaji egyértelműsítés feladatának megoldására számos alkalmazás létezik, viszont rendkívül ritka az olyan, amelyik teljes morfoszintaktikai egyértelműsítést (párhuzamosan lemmatizálást is) végez. A jó minőségű teljes morfoszintaktikai egyértelműsítés kulcsfontosságú az agglutináló nyelvek feldolgozása során.

Munkám során egy új megközelítést alkalmaztam a teljes morfoszintaktikai egyértelműsítés feladatának megoldására: létrehoztam egy statisztikai gépi fordításon alapuló nyelvfüggetlen rendszert, ami egyidőben végez lemmatizálást és szófaji egyértelműsítést. A célnyelvi szótár méretének csökkentése érdekében a szótöveket egy szuffixum-alapú reprezentációban tároltam. Az ismeretlen szavak hatékony kezelésének céljából az elemzési folyamatba egy végződésfa-alapú ajánlórendszert 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 a módszer hatékonyságát. Az eredmények vizsgálatából megálla-

DOI:10.15774/PPKE.ITK.2015.004

pítható, 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.

Kapcsolódó tézisek:

**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.**

**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.**

**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.**

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

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

DOI:10.15774/PPKE.ITK.2015.004

# IV. Záró fejezetek

## 7 Összefoglalás: 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 morfoszintaktikai 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 szóalak statisztikai módszerrel történő előállítása. Nehézséget okoz továbbá az egymástól nyelvtanilag távol álló nyelvek közti fordítás, mivel gyakran jelentős szórendi és szószámbeli különbség mutatkozik köztük. 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 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 a jelentős szórendkülönbséggel rendelkező nyelvpárok fordításának megoldására. Emiatt létrehoztam egy olyan hibrid fordítórendszert, amely általam megfogalmazott szintaxismotivált szabályokat alkalmaz előfeldolgozásként a forrásnyelvi angol szövegen, hogy a szórendből adódó különbségeket feloldja. Ezzel célom a két nyelv (angol-magyar) szórendjének közelítése volt, ami megkönnyíti a fordítórendszer – eredetileg csak lokális átrendezésekre képes – dekóderének munkáját. Az angol-magyar nyelvpárra alkotott szabályok segítségével az alap fordítórendszer eredményeihez képest javulást értem el.

DOI:10.15774/PPKE.ITK.2015.004

**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 morfológiailag bonyolult nyelvek szóalakjának előállítása nagy nehézséget jelent a fordítórendszer dekódere számára az adathiány-problémából kifolyólag, 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 szövegen, a szóalak pedig morfológiai generátor segítségével kerül előállításra. Az adathiány és a homonímia csökkentése érdekében a szavak toldalékmorfémái helyett az azoknak megfelelő morfoszintaktikai címkéket alkalmaztam. Az általam felépített morfológiai generátort alkalmazó architektúrák az emberi kiértékelés számára könnyebben érthető, folyamatosabb fordítás előállítására voltak képesek a statisztikai dekódert használó fordítórendszerekkel szemben.

**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, mellyel az agglutináló és izoláló nyelvek mondatpárjaiban megfigyelhető szószámkülönbségre 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

DOI:10.15774/PPKE.ITK.2015.004

kapcsolódó szófaji címkékből álló rekordot, melyből a 2. tézisben bemutatott morfológiai generátor állítja elő a feszíni szóalakot.

*A fordítás minőségének javításában az igazi áttörést az 1.-3. tézisekben leírt rendszerek együttes alkalmazása jelentette.*

**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 kiegyensúlyozza a hosszabb szegmenseket tartalmazó tanítóhalmazból számított statisztikát, robosztusabbá téve a fordítási modellt.**

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

A szóösszekötő a fordítás során sokszor nehezen párosítja az összetartozó kifejezéseket. Ez főleg akkor fordul elő, ha a kifejezések nyelvtanilag különböző szerkezet miatt távol állnak egymástól, vagy nagyon különbözők. A túl hosszú mondatok is nehézséget okoznak a szóösszekötőnek. A probléma megoldására a tanítóhalmazba integráltam egy rövid, pontos fordítású kifejezéspárokból álló szótárat. A rendszer egyedisége, hogy nemcsak az egyszeri hozzáadást vizsgáltam, hanem a rendszert a szótár többszöri integrálásával is teszteltem. A legjobb esetben sikerült 11,18%-os relatív javulást elérni a fordítás minőségében. A szótár többszöri hozzáadása miatt folyamatosan csökkent a BLEU érték. Ennek oka az eredeti szótár relevanciájának csökkenése, illetve a fordítási és nyelvi modellek deformációja. Ezzel ellentétben az emberi kiértékelés számára a hosszabb mondatok fordítása jelentősen javult.

DOI:10.15774/PPKE.ITK.2015.004

# II. TÉZISCSOPORT

Dolgozatom második felében a teljes morfoszintaktikai egyértelműsítés egy teljesen új megközelítését mutattam be azáltal, hogy a feladat megoldására statisztikai gépi fordítórendszert alkalmaztam. 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 megközelíti az egyes nyelvfüggő rendszerek által elért eredményeket is.

**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 morfoszintaktikai egyértelműsítő rendszert, mely szimultán végez lemmatizálást és szófaji egyértelműsítést. Bebizonyítottam, hogy a célnyelvi címkekészlet komplexitásának csökkentésével javítható az egyértelműsítő rendszer teljesítménye. Rendszeremben a lemmákat egy szuffixum-alapú reprezentációban tároltam, mellyel a minőségjavulás mellett képes voltam csökkenteni a célnyelvi címkekészlet elemszámát.

**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 elemzése. Ez különösen igaz az agglutináló nyelvek esetében, hiszen egy szótőnek akár több száz szóalakja is

DOI:10.15774/PPKE.ITK.2015.004

lehet, ám ezek közül nem mind szerepel a tanítóhalmazban, így az egyértelműsítő rendszernek semmilyen előzetes ismerete nincs ezekről a szavakról. Az ismeretlen szavak egyértelműsítésének javítása érdekében egy végződésfa-alapú morfológiai ajánlórendszert integráltam az elemzési láncba. Ennek köszönhetően nagymértékben sikerült javítani az OOV szavak egyértelműsítésének pontosságát.

**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 morfoszintaktikai 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.

**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 morfoszintaktikai egyértelműsítő nyelvfüggő morfológiai elemzővel kiegészítve további minőségjavulást eredményezett. Ezzel a módszerrel létrehoztam egy nagy pontosságú rendszert magyar nyelvre, mely a lemmatizálást 99,12% pontossággal végzi, a tanítóanyagban nem szereplő szavak 84,82%-át helyesen elemzi, a teljes morfoszintaktikai egyértelműsítés tekintetében pedig 96,50% pontosságú.

DOI:10.15774/PPKE.ITK.2015.004

# 8 Az eredmények alkalmazási területei

A disszertációmban leírt munkák olyan feladatok megoldására irányultak, melyek elősegítik egyrészt a nyelvek közti fordítás minőségének, másrészt a teljes morfoszintaktikai egyértelműsítés pontosságának javulását. A hibrid gépi fordítással kapcsolatos eredményeim sikeresen integrálhatóak tetszőleges SMT architektúrába. Az elért eredmények alátámasztották, hogy a morfológiai információ fordítási láncba való beépítése 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. bemutatta [Orosz_1, Orosz_2], az SMT-alapú egyértelműsítő rendszer kifejezetten alkalmas arra, hogy különböző elveken működő egyértelműsítő rendszerek kombinációjával jelentősen javítsa azok pontosságát.

DOI:10.15774/PPKE.ITK.2015.004

# 9 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.

DOI:10.15774/PPKE.ITK.2015.004

[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 fordító rendszer 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.

DOI:10.15774/PPKE.ITK.2015.004

# 10 Irodalomjegyzék


[1] N. Indurkhya and F. J. Damerau, *Handbook of Natural Language Processing*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2010.

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

[3] J. Hutchins, "Towards a Definition of Example-Based Machine Translation," in *Proceedings of Workshop on Example-Based Machine Translation, MT Summit X*, Phuket, Thailand, 2005, pp. 63–70.

[4] 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.

[5] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, Jul. 1948.

[6] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, Oct. 1948.

[7] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[8] F. J. Och and H. Ney, "The Alignment Template Approach to Statistical Machine Translation," *Comput Linguist*, vol. 30, no. 4, pp. 417–449, Dec. 2004.

[9] 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.

[10] A. Stolcke, "SRILM-an extensible language modeling toolkit.," in *7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, USA, 2002, vol. 2, pp. 901–904.

[11] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an open source toolkit for handling large scale language models," in *9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, 2008, pp. 1618–1621.

[12] A. Levenberg and M. Osborne, "Stream-based randomised language models for SMT," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, Singapore, 2009, pp. 756–764.

[13] K. Knight, "Decoding Complexity in Word-replacement Translation Models," *Comput. Linguist.*, vol. 25, no. 4, pp. 607–615, 1999.

[14] P. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models," in *Machine translation: From real users to research*, Springer, 2004, pp. 115–124.

[15] C. Tillmann, S. Vogel, H. Ney, and A. Zubiaga, "A DP based search using monotone alignments in statistical translation," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, 1997, pp. 289–296.

[16] C. Tillmann and H. Ney, "Word reordering and a dynamic programming beam search algorithm for statistical machine translation," *Comput. Linguist.*, vol. 29, no. 1, pp. 97–133, 2003.

[17] R. C. Moore and C. Quirk, "Faster Beam-Search Decoding for Phrasal Statistical Machine Translation," in *Proceedings of MT Summit XI*, Copenhagen, Denmark, 2007, pp. 321–327.

DOI:10.15774/PPKE.ITK.2015.004

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Stat. Soc. Ser. B Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[19] F. J. Och and H. Ney, “Improved Statistical Alignment Models,” in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Hongkong, China, 2000, pp. 440–447.

[20] P. Koehn, “Europarl: A Parallel Corpus for Statistical Machine Translation,” in *Conference Proceedings: the tenth Machine Translation Summit*, Phuket, Thailand, 2005, vol. 5, pp. 79–86.

[21] P. Koehn and H. Hoang, “Factored Translation Models,” in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, 2007, pp. 868–876.

[22] A. F. Gelbukh, Ed., *Computational Linguistics and Intelligent Text Processing - 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II*, vol. 7817. Springer, 2013.

[23] C. Quirk, A. Menezes, and C. Cherry, “Dependency Treelet Translation: Syntactically Informed Phrasal SMT,” in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, Ann Arbor, USA, 2005, pp. 271–279.

[24] Y. Liu, Q. Liu, and S. Lin, “Tree-to-string alignment template for statistical machine translation,” in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006, pp. 609–616.

[25] T. P. Nguyen, A. Shimazu, T.-B. Ho, M. Le Nguyen, and V. Van Nguyen, “A tree-to-string phrase-based model for statistical machine translation,” in *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, Manchester, UK, 2008, pp. 143–150.

[26] P. Koehn and C. Monz, “Manual and Automatic Evaluation of Machine Translation Between European Languages,” in *Proceedings of the Workshop on Statistical Machine Translation*, Stroudsburg, PA, USA, 2006, pp. 102–121.

[27] 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.

[28] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf, “Accelerated DP based search for statistical translation,” in *Fifth European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997, pp. 2667–2670.

[29] A. Clifton and A. Sarkar, “Combining morpheme-based machine translation with post-processing morpheme prediction,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Portland, USA, 2011, pp. 32–42.

[30] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, USA, 2005, pp. 65–72.

[31] C. Callison-Burch and M. Osborne, “Re-evaluating the role of BLEU in machine translation research,” in *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006, pp. 249–256.

[32] K. O. A. Cuneyd Tantug and I. D. El-Kahlout, “BLEU+: a Tool for Fine-Grained BLEU Computation,” in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco, 2008.

DOI:10.15774/PPKE.ITK.2015.004

[33] B. Siklósi and G. Prószéky, “Statisztikai gépi fordítás eredményének javítása morfológiai elemzés alkalmazásával,” Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar, Budapest, 2009.

[34] A. Birch, M. Osborne, and P. Koehn, “Predicting Success in Machine Translation,” in *EMNLP2008, Proceedings of the Conference, 25-27*, Honolulu , Hawaii, USA, 2008, pp. 745–754.

[35] D. Xiong, Q. Liu, and S. Lin, “Maximum entropy based phrase reordering model for statistical machine translation,” in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006, pp. 521–528.

[36] J. M. Crego and J. B. Marino, “Reordering Experiments for n-Gram-Based SMT,” in *IEEE ACL Spoken Language Technology Workshop*, Palm Beach, Aruba, 2006, vol. 6, pp. 242–245.

[37] 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.

[38] B. Xiang, N. Ge, and A. Ittycheriah, “Improving reordering for statistical machine translation with smoothed priors and syntactic features,” in *Proceedings of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Portland, USA, 2011, pp. 61–69.

[39] 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.

[40] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network,” in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada, 2003, pp. 173–180.

[41] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy, “Parallel corpora for medium density languages,” in *Recent Advances in Natural Language Processing (RANLP 2005)*, Borovets, Bulgaria, 2005, pp. 590–596.

[42] D. Varga, P. Halácsy, A. Kornai, V. Nagy, L. Németh, and V. Trón, “Parallel corpora for medium density languages,” *Amst. Stud. THEORY Hist. Linguist. Sci. Ser. 4*, vol. 292, pp. 247–258, 2007.

[43] 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.

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

[45] G. Prószéky and A. Novák, “Computational morphologies for small Uralic languages,” *Inq. Words Constraints Contexts Festschr. Honour Kimmo Koskenniemi His 60th Birthd.*, pp. 116–125, 2005.

[46] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, “Building a Large Annotated Corpus of English: The Penn Treebank,” *Comput. Linguist.*, vol. 19, no. 2, pp. 313–330, 1993.

[47] D. Klein and C. D. Manning, “Fast exact inference with a factored model for natural language parsing,” in *Advances in neural information processing systems*, 2002, pp. 3–10.

[48] G. Minnen, J. Carrol, and D. Pearce, “Applied morphological processing of English,” *Nat. Lang. Eng.*, vol. 7, no. 3, pp. 207–223, 2001.

DOI:10.15774/PPKE.ITK.2015.004

[49] K. E. Kiss, F. Kiefer, and P. Siptár, *Új magyar nyelvtan*, 3. kiadás. Budapest: Osiris Kiadó, 2003.
[50] P. Halácsy, A. Kornai, L. Németh, B. Sass, D. Varga, T. Váradi, and A. Vonyó, "A Hunglish korpusz és szótár.," in *III. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, 2005, pp. 134–142.
[51] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, no. 1, pp. 39–71, 1996.
[52] D. Wu, "A polynomial-time algorithm for statistical machine translation," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, Santa Cruz, California, USA, 1996, pp. 152–158.
[53] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Edmonton, Canada, 2003, pp. 48–54.
[54] P. Koehn, A. Axelrod, A. Birch, C. Callison-Burch, M. Osborne, D. Talbot, and M. White, "Edinburgh system description for the 2005 IWSLT speech translation evaluation.," in *IWSLT 2005*, Pittsburgh, USA, 2005, pp. 68–75.
[55] R. Zens, H. Ney, T. Watanabe, and E. Sumita, "Reordering Constraints for Phrase-based Statistical Machine Translation," in *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, 2004, pp. 205–211.
[56] Y. Zhang, R. Zens, and H. Ney, "Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation," in *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, Rochester, USA, 2007, pp. 1–8.
[57] Y. Zhang, R. Zens, H. Ney, and L. F. Informatik, "Improved chunk-level reordering for statistical machine translation.," in *Proceedings of International Workshop on Spoken Language Translation*, Trento, Italy, 2007, pp. 21–28.
[58] M. Feng, A. Mauser, and H. Ney, "A source-side decoding sequence model for statistical machine translation," in *Conference of the Association for Machine Translation in the Americas*, Denver, USA, 2010.
[59] Y. Al-Onaizan and K. Papineni, "Distortion models for statistical machine translation," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006, pp. 529–536.
[60] J. M. Crego and J. B. Marino, "Syntax-enhanced N-gram-based SMT," in *Proceedings of the Machine Translation Summit*, Copenhagen, Denmark, 2007, pp. 111–118.
[61] D. Zhang, M. Li, C.-H. Li, and M. Zhou, "Phrase Reordering Model Integrating Syntactic Knowledge for SMT.," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, 2007, pp. 533–540.
[62] C. H. Li, M. Li, D. Zhang, M. Li, M. Zhou, and Y. Guan, "A Probabilistic Approach to Syntax-based Reordering for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, 2007, vol. 45, pp. 720–727.
[63] F. Xia and M. McCord, "Improving a statistical MT system with automatically learned rewrite patterns," in *Proceedings of the 20th international conference on Computational Linguistics*, Geneva, Switzerland, 2004, pp. 508–514.
[64] K. Visweswariah, J. Navratil, J. Sorensen, V. Chenthamarakshan, and N. Kambhatla, "Syntax based reordering with automatically derived rules for improved statistical machine translation," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 2010, pp. 1119–1127.

DOI:10.15774/PPKE.ITK.2015.004

[65] M. R. Costa-Jussà and J. A. Fonollosa, “Statistical machine reordering,” in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 2006, pp. 70–76.
[66] K. Rottmann and S. Vogel, “Word reordering in statistical machine translation with a POS-based distortion model,” in *Proceedings of the 11th International Conference onTheoretical and Methodological Issues in MachineTranslation*, Skövde, Sweden, 2007, pp. 171–180.
[67] J. Niehues and M. Kolss, “A POS-based model for long-range reorderings in SMT,” in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, 2009, pp. 206–214.
[68] J. Elming, “Syntactic reordering integrated with phrase-based SMT,” in *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, UK, 2008, pp. 209–216.
[69] J. Elming and N. Habash, “Syntactic reordering for English-Arabic phrase-based machine translation,” in *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece, 2009, pp. 69–77.
[70] J. Jiang, J. Du, and A. Way, “Source-side Syntactic Reordering Patterns with Functional Words for Improved Phrase-based SMT,” in *Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation*, Beijing, China, 2010, pp. 19–27.
[71] M. Holmqvist, S. Stymne, L. Ahrenberg, and M. Merkel, “Alignment-based reordering for SMT,” in *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2012, pp. 3437–3440.
[72] U. Lerner and S. Petrov, “Source-Side Classifier Preordering for Machine Translation.,” in *Proceedings of the EMNLP 2013*, Seattle, USA, 2013, pp. 513–523.
[73] F. Huang and C. Pendus, “Generalized Reordering Rules for Improved SMT,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013, vol. 2, pp. 387–392.
[74] T. Herrmann, J. Niehues, and A. Waibel, “Combining Word Reordering Methods on different Linguistic Abstraction Levels for Statistical Machine Translation,” in *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, Altanta, USA, 2013, pp. 39–47.
[75] M. Collins, P. Koehn, and I. Kučerová, “Clause restructuring for statistical machine translation,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, USA, 2005, pp. 531–540.
[76] M. Popovic and H. Ney, “POS-based word reorderings for statistical machine translation,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy, 2006, pp. 1278–1283.
[77] C. Wang, M. Collins, and P. Koehn, “Chinese Syntactic Reordering for Statistical Machine Translation.,” in *EMNLP-CoNLL*, Prague, Czech Republic, 2007, pp. 737–745.
[78] R. N. Patel, R. Gupta, P. B. Pimpale, and S. M, “Reordering rules for English-Hindi SMT,” in *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, Sofia, Bulgaria, 2013, pp. 34–41.
[79] L. Németh and A. Zséder, *huntoken*. Budapest: Budapesti Műszaki és Gazdaságtudományi Egyetem, 2003.
[80] *Google translate*. Google.
[81] *Bing translator*. Microsoft Translation.
[82] A. Novák, L. Tihanyi, and G. Prószéky, “The MetaMorpho translation system,” in *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, 2008, pp. 111–114.

DOI:10.15774/PPKE.ITK.2015.004

[83] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(Meta-) evaluation of machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, 2007, pp. 136–158.

[84] A. Bisazza and M. Federico, "Morphological pre-processing for Turkish to English statistical machine translation," in *International Workshop on Spoken Language Translation*, Tokyo, Japan, 2009, pp. 129–135.

[85] C. Mermer and H. Kaya, "The TÜbİTAK-UEKAE statistical machine translation system for IWSLT 2007," in *4th International Workshop on Spoken Language Translation 2007*, Trento, Italy, 2007, pp. 144–148.

[86] N. Singh and N. Habash, "Hebrew Morphological Preprocessing for Statistical Machine Translation," in *Proceedings of the Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, 2012, pp. 43–50.

[87] L. Ramasamy, O. Bojar, and Z. Žabokrtský, "Morphological Processing for English-Tamil Statistical Machine Translation," in *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, Mumbai, India, 2012, pp. 113–122.

[88] K. Oflazer and I. D. El-Kahlout, "Exploring different representational units in English-to-Turkish statistical machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, 2007, pp. 25–32.

[89] 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.

[90] M.-T. Luong, P. Nakov, and M.-Y. Kan, "A hybrid morpheme-word representation for machine translation of morphologically rich languages," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Massachusetts, USA., 2010, pp. 148–157.

[91] A. Vonyó, *A mindenki által keresett ingyenes angol–magyar magyar–angol köznapi, műszaki és szlengszótár*. 1999.

[92] M. Holmqvist, S. Stymne, J. Foo, and L. Ahrenberg, "Improving Alignment for SMT by Reordering and Augmenting the Training Corpus," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, 2009, pp. 120–124.

[93] N. Habash, "Four Techniques for Online Handling of Out-of-vocabulary Words in Arabic-English Statistical Machine Translation," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, Columbus, Ohio, USA, 2008, pp. 57–60.

[94] H. Okuma, H. Yamamoto, and E. Sumita, "Introducing a Translation Dictionary into Phrase-Based SMT.," *IEICE Trans.*, vol. 91-D, no. 7, pp. 2051–2057, Sep. 2008.

[95] S. Vogel and C. Monson, "Augmenting Manual Dictionaries for Statistical Machine Translation Systems," in *Fourth International Conference on Language Resources and Evaluation, LREC'04*, Lisbon, Portugal, 2004, pp. 1593–1596.

[96] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, Pearson Education International, 2009.

[97] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. Cambridge, USA: The MIT Press, 1999.

[98] L. Shen, G. Satta, and A. K. Joshi, "Guided Learning for Bidirectional Sequence Classification," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 2007, pp. 760–767.

[99] J. Hajič, J. Raab, M. Spousta, and others, "Semi-supervised training for the averaged perceptron POS tagger," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, 2009, pp. 763–771.

DOI:10.15774/PPKE.ITK.2015.004

[100] 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.
[101] G. Georgiev, V. Zhikov, K. I. Simov, P. Osenova, and P. Nakov, “Feature-Rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian.,” in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 2012, pp. 492–502.
[102] E. Giesbrecht and S. Evert, “Is part-of-speech tagging a solved task? an evaluation of pos taggers for the German Web as Corpus,” in *Proceedings of the Fifth Web as Corpus Workshop*, San Sebastian, Spain, 2009, pp. 27–35.
[103] C. D. Manning, “Part-of-speech tagging from 97% to 100%: is it time for some linguistics?,” in *Computational Linguistics and Intelligent Text Processing*, Springer, 2011, pp. 171–189.
[104] J. M. Anderson, *A Notional Theory of Syntactic Categories*. New York, USA: Cambridge University Press, 1997.
[105] J. R. Taylor, *Linguistic Categorization*. Oxford University Press, 2003.
[106] J. Hajič, “Morphological tagging: Data vs. dictionaries,” in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, Seattle, Washington, 2000, pp. 94–101.
[107] C. Oravecz and P. Dienes, “Efficient Stochastic Part-of-Speech Tagging for Hungarian,” in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’02)*, Las Palmas, Spain, 2002, pp. 710–717.
[108] H. Papageorgiou, P. Prokopidis, V. Giouli, and S. Piperidis, “A Unified POS Tagging Architecture and its Application to Greek,” in *LREC*, Athens, Greece, 2000, pp. 1455–1462.
[109] 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.
[110] E. Dermatas and G. Kokkinakis, “Automatic Stochastic Tagging of Natural Language Texts,” *Comput. Linguist.*, vol. 21, no. 2, pp. 137–163, Jun. 1995.
[111] D. Csendes, J. Csirik, and T. Gyimóthy, “The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus,” in *Text, Speech and Dialogue*, vol. 3206, P. Sojka, I. Kopecek, and K. Pala, Eds. Springer Berlin / Heidelberg, 2004, pp. 41–47.
[112] K. Pearson, “Note on regression and inheritance in the case of two parents,” *Proc. R. Soc. Lond.*, vol. 58, no. 347–352, pp. 240–242, 1895.
[113] 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.
[114] G. Orosz and A. Novák, “PurePos – an open source morphological disambiguator,” in *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*, Wroclaw, Poland, 2012.
[115] T. Erjavec, “MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora,” in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Paris, France, 2004, pp. 1535–1538.
[116] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biom. Bull.*, vol. 1, no. 6, pp. 80–83, Dec. 1945.
[117] J. Zsibrita, V. Vincze, and F. Richárd, “magyarlanc : A Toolkit for Morphological and Dependency Parsing of Hungarian,” in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*, Hissar, Bulgaria, 2013, pp. 763–771.

DOI:10.15774/PPKE.ITK.2015.004

[118] B. Jongejan and H. Dalianis, “Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike,” in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapore, 2009, pp. 145–153.
[119] F. James, “Modified Kneser-Ney Smoothing of n-gram Models,” Research Institute for Advanced Computer Science (RIACS), 2000.
[120] G. Chrupała, G. Dinu, and J. Van Genabith, “Learning Morphology with Morfette,” in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco, 2008, vol. 8, pp. 2362–2367.
[121] S. Bird, “NLTK: the natural language toolkit,” in *Proceedings of the COLING/ACL on Interactive presentation sessions*, Sydney, Australia, 2006, pp. 69–72.
[122] G. Malecha and I. Smith, “Maximum Entropy Part-of-Speech Tagging in NLTK,” *Unpubl. Course-Relat. Rep. Httpwww People Fas Harv. Edugmalecha*, 2010.
[123] Ž. Agić, N. Ljubešić, and D. Merkler, “Lemmatization and Morphosyntactic Tagging of Croatian and Serbian,” in *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, Sofia, Bulgaria, 2013, pp. 48–57.
[124] A. Chanev, K. Simov, P. Osenova, and S. Marinov, “Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007,” vol. 309, N. Nicolov, G. Angelova, and R. Mitkov, Eds. Amsterdam & Philadelphia: John Benjamins, 2007, pp. 321–330.
[125] M. R. de H. Maia and G. B. Xexéo, “Part-of-Speech Tagging of Portuguese Using Hidden Markov Models with Character Language Model Emissions,” in *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, Cuiabá, Brazil, 2011, pp. 159–163.
[126] C. Freitas, P. Rocha, and E. Bick, “Floresta Sintá(c)tica: Bigger, Thicker and Easier,” in *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*, Berlin, Heidelberg, 2008, pp. 216–219.
[127] M. Collins, “Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms,” in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, Philadelphia, USA, 2002, pp. 1–8.
[128] 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.
[129] Z. Harris, “String Analysis of Language Structure,” *Int. J. Am. Linguist.*, vol. 30, no. 4, pp. 415–420, 1964.
[130] S. Klein and R. F. Simmons, “A computational approach to grammatical coding of English words,” *J. ACM JACM*, vol. 10, no. 3, pp. 334–347, 1963.
[131] B. B. Greene and G. M. Rubin, “Automatic Grammatical Tagging of English,” Department of Linguistics, Brown University, Providence, Rhode Island, USA, Technical Report, 1971.
[132] E. Brill, “Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging,” *Comput. Linguist.*, vol. 21, no. 4, pp. 543–565, Dec. 1995.
[133] D. Yarowsky, G. Ngai, and R. Wicentowski, “Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora,” in *Proceedings of the First International Conference on Human Language Technology Research*, San Diego, USA, 2001, pp. 1–8.
[134] V. Fossum and S. Abney, “Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora,” in *Natural Language Processing–IJCNLP 2005*, vol. LNAI 3651, Berlin, Heidelberg: Springer-Verlag, 2005, pp. 862–873.

DOI:10.15774/PPKE.ITK.2015.004

[135] A. Ratnaparkhi, “A Maximum Entropy Model for Part-Of-Speech Tagging,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, USA, 1996, pp. 133–142.
[136] A. Finch and E. Sumita, “Transliteration using a phrase-based statistical machine translation system to re-score the output of a joint multigram model,” in *Proceedings of the 2010 Named Entities Workshop*, Uppsala, Sweden, 2010, pp. 48–52.
[137] L. R. Bahl and R. L. Mercer, “Part of speech assignment by a statistical decision algorithm,” in *IEEE International Symposium on Information Theory*, Ronneby, Sweden, 1976, pp. 88–89.
[138] K. W. Church, “A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text,” in *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, USA, 1988, pp. 136–143.
[139] S. J. DeRose, “Grammatical category disambiguation by statistical optimization,” *Comput. Linguist.*, vol. 14, no. 1, pp. 31–39, 1988.
[140] R. Garside, G. Sampson, and G. Leech, *The computational analysis of English: A corpus-based approach*, vol. 57. Longman, 1988.
[141] D. Hindle, “Acquiring disambiguation rules from text,” in *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, Vancouver, Canada, 1989, pp. 118–125.
[142] P. Halácsy, A. Kornai, C. Oravecz, V. Trón, and D. Varga, “Using a morphological analyzer in high precision POS tagging of Hungarian,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy, 2006, pp. 2245–2248.
[143] V. Trón, P. Halácsy, P. Rebrus, P. V. András Rung, and E. Simon, “Morphdb.hu: Hungarian lexical database and morphological grammar,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy, 2006, pp. 1670–1673.
[144] G. Prószéky, “Industrial applications of unification morphology,” in *Proceedings of the Fourth Conference on Applied Natural Language Processing*, Stuttgart, Germany, 1994, pp. 213–214.