EGY MEGSZORÍTÁSALAPÚ SZÁMÍTÓGÉPES MORFOLÓGIAI MODELL ÉS ALKALMAZÁSA URÁLI NYELVEKRE

Doktori értekezés tézisei

Novák Attila

Roska Tamás Műszaki és Természettudományi Doktori Iskola

Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar

Témavezető:
Dr. Prószéky Gábor

2015

1

Bevezetés

A magyarhoz hasonlóan bonyolult morfológiájú nyelvek számítógépes feldolgozása során a nyelvben előforduló lehetséges szóalakok igen magas száma miatt a morfológiai elemzés alkalmazása gyakorlatilag elkerülhetetlen. Disszertációm az ilyen **bonyolult morfológiájú nyelvekre alkalmazható számítógépes morfológiákkal** foglalkozik. A magyar és más „nagy" nyelvek mellett több uráli kisebbségi nyelv számítógépes morfológiájának kidolgozásával is foglalkoztam.

Az utóbbi nyelveknél a morfológia összetettsége mellett az elektronikus formában rendelkezésre álló nyelvészeti erőforrások hiánya miatt sem lehetséges adatvezérelt megközelítést alkalmazni. Továbbá, a meglévő nyelvészeti adatok és leírások is sok esetben hiányosak, vagy ellentmondásosak voltak, ezért a számítógépes modellek ismételt revíziójára volt szükség ezekben az esetekben.

Magyar nyelvű szövegek morfológiai elemzésére Magyarországon hosszú időn keresztül leginkább a MorphoLogic Kft. által kifejlesztett Humor ('High speed Unification MORphology') programot alkalmazták (Prószéky and Kis, 1999). Kutatásom kiindulásaként ezt a programot használtam. Bár maga a program hatékony eszköznek bizonyult, a Humor adatbázisának formátumával problémák voltak a karbantarthatóság, az olvashatóság, a javíthatóság és a bővíthetőség szempontjából. Ezt a problémát az elemzőprogram módosítása nélkül a nyelvi adatbázis többszintűvé tételével sikerült orvosolni.

Az elemző első változatának tőtára az eredeti Humor adatbázison alapult. Ebből eltávolítottam az összes redundáns (megjósolható) tulajdonságot, a szavak megjósolhatatlan tulajdonságait (pl. kategória címke) kézzel ellenőriztem és javítottam, a hiányzó tulajdonságokat pedig pótoltam (pl. morfémahatárok). Jelenleg a Humor adatbázisának mérete így az eredetinek többszöröse, és több speciális szakterület szókincsét is tartalmazza (pl. orvosi vagy üzleti területek szavai). A nyelvtani modell pedig jóval pontosabb az eredetinél.

A magyar mellett létrehoztam spanyol és francia morfológiát is készítettem az elkészült keretrendszer segítségével, továbbá a Humor rendszerre adaptáltam létező holland, olasz és román morfológiákat. Továbbá, közreműködtem számos veszélyeztetett finnugor nyelv morfológiájának létrehozásában (komi, udmurt, mari, mansi)[i] Ugyanebben a projektben két északi szamojéd nyelv, a nganaszan és a tundrai nyenyec morfológiáját is elkészítettem. Egy későbbi OTKA projektum keretében a hanti két nyelvjárására is készült elemző.

Az eredeti Humor elemzőimplementáció nem alkalmas arra, hogy a szó végződése alapján olyan szavak lehetséges elemzéseit előállítsa, amelyeknek a töve nem szerepel az adatbázisában. Nem is lenne egyszerű az algoritmust úgy módosítani, hogy képes legyen ennek a feladatnak a

[i] Ez az NKFP 5/135/2001 projekt keretein belül történt, Fejes Lászlóval közösen.

megoldására. Egy ilyen ismeretlenszó-elemző integrálása az elemzőbe ugyanakkor igen hasznos eszköz lenne, hiszen minden szöveg sok olyan szóalakot tartalmaz, amelynek a töve nem szerepel az elemző szótárában. Emellett nem lehetséges a morfológiai modellek súlyozása vagy gyakorisági információval való ellátása sem, amelyre szükség lenne ahhoz, hogy a morfológia közvetlenül alkalmas legyen adatvezérelt szövegnormalizálási feladatok (pl. automatikus helyesírás-javítás vagy beszédfelismerés) támogatására. Szintén hasznos lenne a modellek súlyozhatósága az ismeretlenszó-elemző által generált javaslatok sorrendezéséhez. Ezek mellett a Humor hátrányaként a morfológiaielemző-szoftver zárt licensze említhető, amely nem teszi lehetővé ezeknek a nyelvi erőforrásoknak a szélesebb körben való hozzáférhetővé tételét.

A fenti problémákat a Humor formátumú morfológiai leírások forrásának véges állapotú leírássá alakításával oldottam meg. Ezek kompilálására és használatára nyílt forráskódú eszközök is rendelkezésre állnak. A véges állapotú reprezentáció használható végződésalapú ismeretlenszó-elemzésre, természetes megoldást kínál gyakorisági információ hozzáadására a modellhez, és lehetővé teszi a súlyozott hibamodellekkel való kompozíciót.

A legszélesebb körben használt véges állapotú morfológiai eszközkészlet a Xerox *xfst-lookup* párosa (Beesley and Karttunen, 2003). Az *xfst* compilerrel különböző formalizmusok alkalmazásával lehet számítógépes morfológiákat leíró véges állapotú transzducereket létrehozni, amelyek morfológiai elemzőként vagy generátorként való működtetésére a *lookup* program szolgál.

Bár a Xerox eszközeit kutatási célra hozzáférhetővé tették 2003-ban Beesley és Karttunen könyvének (Beesley and Karttunen, 2003) publikálásakor, ezek két szempontból nem különböznek lényegesen a Humor elemzőtől: zárt forráskódúak és nem használhatóak súlyozott modellek létrehozására. Ugyanakkor az ismeretlenszó-elemzés problémájának megoldására alkalmasak. Szerencsére néhány évvel ezelőtt létrejöttek az xfst és a lookup nyílt forráskódú alternatívái. Ezen nyílt forráskódú eszközök egyike, a Foma (Huldén, 2009) alkalmas az xfst-lexc formátumú morfológiai leírások kompilálására és működtetésére. Ez tehát lehetővé teszi a zárt forráskód okozta problémák kiküszöbölését. Ezen kívül a szintén nyílt forráskódú HFST-eszközkészlet (Lindén et al., 2011) segítségével a Foma formátumú transzducerek OpenFST (Allauzen et al., 2007) formátumúakká konvertálhatók, amely implementáció viszont lehetővé tette a súlyozott véges állapotú modellek létrehozását.

A szabályalapú morfológiai nyelvtanok létrehozása többféle kompetenciát igényel: ismerni kell a formalizmust, az adott nyelv morfológiáját, helyesírását, morfofonológiáját, és kiterjedt lexikai ismeretekre van szükség. Ezért az egyik további érdekes feladat a morfológia automatikus tanulhatósága szöveges korpuszból vagy lexikai adatbázisokból.

Sok számítógépes morfológiai adatbázis nem tartalmaz külön szabálykomponenst. Ezeket az adatbázisokat általában valamilyen ragozási szótárban szereplő információ konverziójával hozzák létre. A szavak lemmája (és esetleg ettől eltérő töve) mellett valamilyen a szó ragozási paradigmáját leíró információt tartalmaznak (gyakran valamiféle paradigmaazonosító címke formájában), ezt esetleg még valamiféle egyéb lexikai–szintaktikai–szemantikai információval kiegészítve. Szabályok híján azonban az ilyen erőforrások új szavakkal való kiegészítése nem olyan egyszerű, mint a szabályalapú morfológiák bővítése. A gépi tanulás alkalmazása azonban lehetővé teheti, hogy a más morfológiákban a szabálykomponensben leírt tudást

magából az adatbázisból kinyerve azt új szavak ragozási paradigmájának azonosításához használjuk.

Az ismeretlen szavak kezelésére kidolgozott módszerem a tő különböző hosszúságú végződéseit és egyéb lexikai jellemzőit használja jellemzőkként a megfelelő ragozási paradigma kiválasztásához. Általában a leghosszabb illeszkedő végződésre leginkább jellemző morfológiai viselkedést veszi a legnagyobb súllyal figyelembe. Működését egy nyílt forráskódú orosz morfológiai lexikonon mutattam be és értékeltem ki.

A „nyelvek” egy másik csoportja, amikre a morfológiai elemző adaptálását bemutatom, a magyar nyelv különböző változatai. Bemutatom, hogy hogyan tettem lehetővé azon ó-, és középmagyar történeti szövegekben található morfológiai konstrukciók, toldalékmorfémák és allomorfok, illetve tövek és paradigmák elemzését, amelyek a mai magyar nyelvben már nem használatosak. Létrehoztam egy egyértelműsítő rendszert is, ami a szövegek morfoszintaktikai annotációjának kézi és automatikus egyértelműsítésére használható, továbbá egy korpuszkezelő rendszert, aminek segítségével az annotált korpusz kereshető és karbantartható. A másik nyelvváltozat, amire a morfológiai elemzőt kiterjesztettem, a kórházi esetdokumentálás során használt magyar orvosi nyelv. Ez számos tekintetben különbözik a köznapi magyar nyelvtől (rövidítések, idegen szavak, sajátos szóalakok, stb.), aminek kezelésére az elemző lexikonját félautomatikus módszerekkel bővítettem.

2

Az új tudományos eredmények összefoglalása

A számítógépes nyelvi eszközök minőségét alapvetően befolyásolja, hogy az adatbázisuk mennyire jól modellezi az adott nyelvet. A nyelvi modell részei közül kulcsszerepet játszik a morfológiai komponens, ami az adott nyelv szavainak elemzéséért és generálásáért felelős. Jelen kutatásban az összetett morfológiájú nyelvek nyelvészeti szempontból is helyes számítógépes morfológiai modelljeinek több lehetséges megvalósítását tekintettem át.

Létrehoztam egy modellt a morfológia leírására, amit én és kollégáim számos nyelvre alkalmaztunk. A nyelvi leírást és a morfológiai modell használatával készült eszközöket számos kereskedelmi eszközbe is integrálták (helyesírás-ellenőrző programok, tövesítők, szótárprogramok, gépi fordítók), és számos nyelvészeti kutatási projektben felhasználták.

2.1 Morfológiaiadatbázis-leíró keretrendszer

Az agglutináló nyelvek esetén a morfémák lehetséges kombinációinak száma gyakorlatilag végtelen, ezért ilyen nyelvekre komoly kihívást jelent morfológiai elemzőt készíteni. Azok a módszerek, amiket az angolhoz hasonló izoláló nyelvekre használnak, nem használhatóak a bonyolult morfológiájú nyelvek esetén. Magyar nyelvű szövegek morfológiai elemzésére Magyarországon hosszú időn keresztül leginkább a MorphoLogic Kft. által kifejlesztett Humor('High speed Unification MORphology') programot alkalmazták (Prószéky and Kis, 1999). Az elemző modellje a szomszédos morfok közötti megszorításokon alapul, klasszikus 'item-and-arrangement'-típusú elemzést hajt végre. Kutatásom kiindulásaként ezt a programot használtam.

A Humor az elemzendő szót morfok sorozataként elemzi. Minden egyes morf egy morféma egy adott realizációja (allomorfja). Az elemző a szóalakot olyan részekre bontja fel amelyeknek van felszíni alakja (ami a bemeneten megjelenik, ez maga a morf); egy lexikai alakja (a morféma szótári alakja') és egy (opcionálisan strukturált) kategóriacímkéje.

Az elemző mélységi keresést végez a beadott szóalakon a lehetséges elemzések után. Elemzés közben a program kétféle ellenőrzést hajt végre. Egyrészt lokális kompatibilitás-ellenőrzést végez az egymás mellett álló morfok között. Másrészt a szó szintjén is ellenőrzi, hogy a lokálisan kompatibilis morfjelöltek helyes szóstruktúrát alkotnak-e. Ezt az ellenőrzést egy a lehetséges szókonstrukciókat leíró kiterjesztett determinisztikus véges állapotú automata (EFSA) bejárásával valósítja meg. A Humor lexikai adatbázisa a morfémaallomorfok leírását

tartalmazó lexikonból, a szónyelvtant leíró automatából és a szomszédos morfok lokális kompatibilitását ellenőrző kétféle adatszerkezetből áll. Ezek közül az egyiket a folytatási osztályok és a folytatási osztályok kompatibilitását leíró bináris kompatibilitási mátrixok alkotják. A másik adatszerkezet bináris tulajdonság-, és megszorításvektorokból áll. Minden morfnak mind a jobb, mind a bal oldalához tartozik egy-egy folytatásiosztály-azonosító, és van egy jobb oldali bináris tulajdonságvektora, valamint egy bal oldali bináris megszorításvektora.

Az elemző által használt formátumban nagyon nehéz az adatbázist létrehozni, illetve karbantartani, mivel az redundáns és alacsony szintű adatszerkezetekre épül. Ezt a problémát egy olyan morfológiainyelvtan-fejlesztő környezet megtervezésével és implementálásával oldottam meg, amely magas szintű, olvasható, redundanciamentes morfémaalapú nyelvtani és lexikai reprezentációból hozza létre automatikusan a Humor elemző által használt lexikai erőforrásokat. Ez a reprezentáció könnyen karbantartható, bővíthető, és a morfémáknak csak az megjósolhatatlan tulajdonságait tartalmazza. A komplex lexikai egységek (elsősorban az összetett szavak) konzisztens és gazdaságos leírásának elősegítésére beépítettem a rendszerbe egy egyszerű öröklési mechanizmust, amelynek segítségével az összetett lexikai egységek alapesetben az utótagjuktól öröklik a tulajdonságaikat. A magas szintű leírásból a szabályok által leírt műveletek végrehajtásával a rendszer létrehoz egy redundáns, de még könnyen olvasható köztes reprezentációt amely az allomorfokat létrehozó szabályok helyességének könnyebb ellenőrzését is lehetővé teszi. A rendszer biztosítja a létrehozott lexikai erőforrások konzisztenciáját, és automatikusan ellenőrzi a forrásleírást szintaktikai szempontból, illetve figyelmeztet a leírásban szereplő esetleges ellentmondásokra is (pl. hogy egy morf megszorításait egyetlen másik sem elégíti ki, vagy hogy nincs olyan allomorf-generáló szabály, amely az adott morfémaleírásra illeszkedne).

Ezután a köztes reprezentációt a jegyek és megszorítások kódolását megadó leírás alapján a rendszer az elemző által használt redundáns gépközeli formára alakítja át. Bizonyos tulajdonságok bináris tulajdonságokká képződnek le, a többi a folytatási mátrixokat határozza meg, amiket a rendszer automatikusan hoz létre. A keretrendszer nem kötődik egy konkrét nyelvhez, a program módosítása nélkül is hatékonyan használható különböző nyelvek morfológiájának leírására.

A fejlesztőkörnyezet létrehozása után elkészített Humor morfológiák ennek a magas szintű formalizmusnak a használatával íródtak.

1. TÉZIS:

Megterveztem és implementáltam egy morfológiainyelvtan-fejlesztő környezetet, amely egy csak a morfémák megjósolhatatlan tulajdonságait tartalmazó, jól olvasható, magas szintű, redundanciamentes, és ezért könnyen karbantartható és kiegészíthető morfémaalapú nyelvtani és lexikai leírásból automatikusan előállítja a Humor morfológiai elemző által használt lexikai erőforrásokat.

Kapcsolódó publikációk: 13, 14, 54, 55, 56, 61, 63

2.2 A modell alkalmazása különböző nyelvekre

A fejlesztőkörnyezet több nyelv számítógépes morfológiájának megvalósítására használtam. A *magyar, spanyol* és *francia* nyelvekre teljes morfológiai leírásokat készítettem, emellett több uráli nyelv, a *komi*, az *udmurt*, a *mari*, az *északi manysi* és több *hanti* dialektus Humor-alapú számítógépes morfológiájának létrehozásában működtem közre. Továbbá más morfológiai leírások alapján Humor-kompatibilis morfológiát készítettem *hollandra*, *olaszra*, *románra* és *oroszra*, amely nyelvleírások lefedését is bővítettem a konvertálás során. Két északi szamojéd nyelv, a nganaszan és a tundrai nyenyec morfológiáját is elkészítettem. Bár ezek az agglutináló nyelvek szintén az uráli nyelvcsaládhoz tartoznak, a rendkívül bonyolult morfofonológiájuk leírása nehéznek bizonyult a Humor megszorítás-alapú formalizmusával. Ezért ezeket a morfológiákat a Xerox véges állapotú formalizmuson alapuló *lexc* és *xfst* eszközeinek felhasználásával készítettem el.

2. TÉZIS

Számos nyelv számítógépes morfológiáját készítettem el önállóan, illetve más kutatókkal együttműködve.

2a TÉZIS:

A következő nyelvekre az általam definiált formalizmus felhasználásával készítettem vagy adaptáltam számítógépes morfológiát (egyedüli vagy társszerzőként), a formalizmus leíróerejét demonstrálva: magyar, spanyol, francia, holland, olasz, román, komi, udmurt, mari, északi manysi, szinjai és kazimi hanti.

2b TÉZIS:

Két súlyosan veszélyeztetett északi szamojéd nyelv, a nganaszan és a tudrai nyenyec számítógépes morfológiáját a Xerox véges állapotú formalizmusának felhasználásával készítettem el.

Kapcsolódó publikációk: 22, 63, 59, 61, 14, 58, 55, 56, 54, 48, 49, 50

2.3 A magyar morfológia adaptálása sajátos nyelvváltozatokra

Egy nyelvi eszköz alkalmazhatóságát alapvetően befolyásolhatja a konkrét nyelvváltozat, amelyből a feldolgozandó szöveg származik. Ilyenkor az eszközt adaptálni kell az adott nyelvváltozathoz. Két példát vizsgáltam meg, amiken keresztül bemutattam, hogy a létrehozott fejlesztői környezet és formalizmus segítségével a morfológiai elemző könnyen adaptálható speciális nyelvezetű szövegeken való használatra. Az első esetben ó-, és középmagyar történeti szövegeket vizsgáltam, ahol a probléma a már kihalt alaktani szerkezetek, illetve a lejegyzések rendkívüli változatosságának kezelése volt. A másik példa a klinikai szövegek elemzése, ahol a szaknyelvi és sok idegen szót tartalmazó szókincs kezelésére tettem alkalmassá a meglévő magyar morfológiai elemzőt.

3. TÉZIS:

Az általam definiált formalizmus használatával elkészített morfológiák adaptálhatóságát az általam készített magyar morfológiai leírás kiterjesztésével demonstráltam.

3a TÉZIS:

A mai magyarra készített Humor morfológiai elemzőt alkalmassá tettem az ó- és középmagyarban még létező, de mára kihalt alaktani szerkezeteket, toldalékallomorfokat és -morfémákat, paradigmákat és töveket tartalmazó szavak kezelésére.

Kapcsolódó publikációk: 38, 42

3b TÉZIS:

Félautomatikus módszert dolgoztam ki a morfológiai elemző adatbázisának kibővítésére, amely így képessé vált az orvosi terminológia és az idegen szavakat és rövidítéseket nagy mennyiségben tartalmazó klinikai szövegekre jellemző szóalakok elemzésére. Módszer dolgoztam ki az idegen szavak automatikus felismerésére, azok szófajának és kiejtésének automatikus megállapítására. Az így létrehozott elemző jó lefedést biztosít a magyar nyelvű klinikai orvosi szövegeken.

Kapcsolódó publikációk: 2, 37, 65, 36, 7, 1

2.4 A megszorításalapú morfológiák véges állapotú implementációja

A Humor elemzőszoftver zárt licensze nem tette lehetővé az ehhez készült nyelvi erőforrások szabad terjesztését. Ugyanakkor a Humor elemző implementációja nem teszi lehetővé az ismeretlen szavak elemzését (morphological guessing), valamint azt sem, hogy az egyes szavakhoz gyakorisági információt rendeljünk vagy a modellt másképp súlyozzuk. Ezeket a problémákat úgy oldottam meg, hogy a Humor morfológiai erőforrásait olyan véges állapotú leírássá konvertáltam, amely lehetővé teszi a fenti igények kielégítését és rendelkezik nyílt forráskódú implementációval is.

4. TÉZIS:

Implementáltam egy olyan algoritmust, amely a Humor adatbázisokat olyan véges állapotú reprezentációra konvertálja, amely lehetővé teszi az ismeretlen szavak elemzését, a súlyozott modellek létrehozását, és rendelkezik nyílt forráskódú implementációval is.

Kapcsolódó publikációk: 30, 31

2.5 Szótáralapú morfológiák bővítése nyelvtanírás nélkül

Sok számítógépes morfológia adatbázis nem tartalmaz külön szabálykomponenst. Ezeket az adatbázisokat általában valamilyen ragozási szótárban szereplő információ konverziójával hozzák létre. A szavak lemmája (és esetleg ettől eltérő töve) mellett valamilyen a szó ragozási paradigmáját leíró információt tartalmaznak (gyakran valamiféle paradigmaazonosító címke formájában), ezt esetleg még valamiféle egyéb lexikai–szintaktikai–szemantikai információval kiegészítve. Szabályok híján az ilyen erőforrások új szavakkal való kiegészítése nem olyan egyszerű, mint a szabályalapú morfológiák bővítése. A gépi tanulás alkalmazása azonban lehetővé teheti, hogy a más morfológiákban a szabálykomponensben leírt tudást magából az adatbázisból kinyerve azt új szavak ragozási paradigmájának azonosításához használjuk. Módszerem a tő különböző hosszúságú végződéseit és egyéb lexikai jellemzőit használja jellemzőkként a megfelelő ragozási paradigma kiválasztásához. Általában a leghosszabb illeszkedő végződésre leginkább jellemző morfológiai viselkedést veszi a legnagyobb súllyal figyelembe. Működését egy nyílt forráskódú orosz morfológiai lexikonon mutattam be és értékeltem ki, azonban minimális adaptáció után az eszköz bármilyen más nyelvre alkalmazható, amihez rendelkezésre áll a megfelelő morfológiai erőforrás. Emellett éltem azzal a feltételezéssel is, hogy a morfológia mellett létezik olyan elérhető szótár, amiben bizonyos lexikai tulajdonságok is megtalálhatóak, így ezeket a paradigmajavaslatok egyértelműsítése során figyelembe lehet venni. Módszerem jól teljesít minden teszteset során (90% körüli pontossággal), legjobban azonban a ritkán előforduló szavak esetén működik. Éppen ezek hiányoznak az eredeti lexikonból a legnagyobb valószínűséggel.

Az eredményekből az is kiviláglik, hogy a hosszabb szóvégződések előnyben részesítése a rövidebbekkel szemben lényegesen jobb teljesítményhez vezet. Ez akkor is világosan látszik, ha pusztán azt tekintjük, hogy milyen gyakran találja el az algoritmus pontosan a helyes paradigmát. Az elkövetett hibák elemzése azonban még inkább rávilágít a javasolt megoldás erősségeire. Míg a kontrollként használt toldalékguesser algoritmus az adott szóra egyáltalán nem alkalmazható paradigmákat javasol, mikor hibázik, addig az általam bemutatott módszer csupán a szemantikai ismeret hiányából fakadó tévesztéseket követ el. Ezek azonban olyan hibák, amiket emberek ugyanígy elkövetnének.

5. TÉZIS:

Implementáltam egy algoritmust, amely a szótárban nem szereplő szavak helyes ragozási paradigmájának megjósolásával lehetővé teszi, hogy könnyen új lexikai tételeket vegyünk fel szabálykomponenst nem tartalmazó lexikonalapú számítógépes morfológiákba is.

Kapcsolódó publikációk: 27, 5

2.6 Rugalmasan használható szóalak-generáló és lemmatizálóalkalmazások

Az eredeti Humor rendszer nem volt képes szóalak-generálásra, és a létező lemmatizálóalgoritmus nem működött helyesen számos nem triviális szókonstrukcióra. Ezeket

a problémákat úgy oldottam meg, hogy szóalak-generáló lexikon készítésére és használatára szolgáló algoritmusokkal bővítettem a rendszert, és jobb lemmatizáló algoritmusokat terveztem.

A szóalak-generátor mindazokat a szóalakokat előállítja, amelyek a nyelvtana szerint egy adott morfémasorozat lehetséges realizációi lehetnek. A generátor számára az input alapesetben a lemmából és azoknak a kategóriacímkéknek a sorozatából áll, amelyek megadják azokat a morfoszintaktikai jegyeket, amelyeket az adott szóalak megtestesít.

A Humor generátor nem egyszerű inverze a neki megfelelő elemzőnek. Mindazoknak a tetszőlegesen bonyolult képzett és/vagy összetett töveknek a ragozott és képzett alakjait elő tudja állítani anélkül, hogy az inputban bármilyen explicit formában szerepelnének a tövön belüli összetételi határok vagy képzők, amelyeket az elemző meg tud elemzeni, akkor is, ha az egész komplex tő nem szerepel a lexikonában. Ez hasznos lehetőség olyan nyelvek esetében, ahol a morfológiailag komplex tövek gyakoriak. Nem kell amiatt aggódni, hogy a komplex tő egyben szerepel-e a tőadatbázisban, elég, ha az elemző bármilyen helyes elemzést tud rá adni.

A rendszer lehetőséget ad rá, hogy azokban az esetekben, amikor egy adott morfoszintaktikaijegy-halmaz többféle szóalakként is realizálódhat, megadjuk, hogy melyik alakok jelöltebbek/ritkábbak, és melyikek a preferált változatok. Ez hasznos lehet például, amikor a rendszert egy szabályalapú fordítórendszer részeként a fordított szöveg előállítására használjuk, ahol a legkevésbé jelölt szóalakváltozat előállítása kívánatos. A magyar morfológiai leírást ezért kiegészítettem a jelöltségre vonatkozó információval. Az ingadozó alakok közül a jelölteket a rendszer automatikusan kihagyja az adatbázisból, amikor olyan generátor-adatbázist kompilálunk, amely például a gépi fordítóba való integrálás céljából készül.

Mivel az általam vizsgált nyelvek egy részében (pl. a Komiban) a morfémák sorrendje is ingadozást mutat, elkészítettem az elemző egy olyan változatát is, amely egyszerűen sorrendezetlen halmaznak tekinti a kifejezendő morfoszintaktikai jegyeket, és helyesen elő tudja állítani a szóalakot akkor is, ha azok nem abban a sorrendben jelennek meg az inputban, mint ahogy a tényleges szóalakban szerepelniük kell. A Humor generáltort használó gépifordító-rendszer implementálásánál a generátornak ezt a tulajdonságát ki is használták.

A lemmatizáló eszköz, amely az elemző köré épül nem csak a szóalakok lemmáját (szótári alakját) azonosítja, hanem a szófajt és az összes morfoszintaktikai jegyet is visszaadja. Az elemző részletes elemzéseivel ellentétben a lemmatizáló nem tárja fel a szó belső szerkezetét, a kimenetében nem szerepel minden összetételi tag és képző külön elemként. A lemmatizáló elemzései jól használhatóak olyan a morfológiai elemzésre épülő feladatok megoldásához, mint a szófaji egyértelműsítés, a keresőrendszerekben való használathoz szükséges indexelés vagy a szintaktikai elemzés.

A lemmatizálónak két új implementációja készült. Mindkettő helyesen állapítja meg a lemmát, illetve kiszűri az esetleges hibás elemzéseket a különleges szóalaktani konstrukciók, például azon szavak esetében is, ahol a szó nem (csak) a végén toldalékolódik. Így a korábbinál pontosabb eredményt adó lemmatizálóalgoritmusok jöttek létre.

6. TÉZIS:

Új szóalakgeneráló- és lemmatizálóalgoritmusokat dolgoztam ki a Humor rendszer számára.

6a TÉZIS:

Olyan szóalak-generáló Humor modult fejlesztettem ki, amely mindazoknak a tetszőlegesen bonyolult képzett és/vagy összetett töveknek a ragozott és képzett alakjait elő tudja állítani anélkül, hogy az inputban bármilyen explicit formában szerepelnének a tövön belüli összetételi határok vagy képzők, amelyeket az elemző meg tud elemzeni, akkor is, ha az egész komplex tő nem szerepel a lexikonában. A szóalakgenerálót kereskedelmi forgalomban is kapható információ-visszakereső, illetve gépi fordító rendszerbe, illetve kutatási célú prototípusrendszerekbe is beépítettük.

6b TÉZIS:

Olyan Humor alapú lemmatizálóalgortimusokat hoztam létre, amelyek helyesen helyesen állapítják meg a lemmát, illetve kiszűrik az esetleges hibás elemzéseket a különleges szóalaktani konstrukciók, például azon a szavak esetében is, ahol a szó nem (csak) a végén toldalékolódik. A lemmatizálót számos korpuszannotációs projektumban felhasználtuk, és különböző információ-visszakereső és gépi fordító rendszerekbe integráltuk.

Kapcsolódó publikációk (az eszközök alkalmazásairól): 3, 35, 41, 53, 51, 52, 11, 49

2.7 Szöveges korpuszok annotálására, keresésére és javítására alkalmas eszköz

Az annotált korpuszok kézi ellenőrzésének és egyértelműsítésének támogatására olyan webes felületet hoztam létre, amellyel az egyértelműsítési és normalizálási hibák hatékonyan javíthatóak. A rendszer az adott dokumentumot könnyen olvasható többsoros (interlináris) annotációs formában jeleníti meg, és lehetőséget biztosít jelentésglosszák, normalizált alakok, fordítás, illetve egyéb információk kezelésére.

A szövegekben való keresést támogató korpuszkezelő nemcsak azt teszi lehetővé, hogy különböző grammatikai szerkezetekre keressünk a szövegekben példákat, hanem azt is, hogy a kereső találataiban is azonnal kijavíthassuk az esetlegesen még az annotációban vagy a szövegben maradt hibákat, amely javítások ilyenkor az adatbázisba azonnal visszakerülnek. A kereső utóbbi változata csak a megfelelő szakértelemmel és jogosultságokkal rendelkező felhasználók számára elérhető.

A hibakeresés és –javítás egyik hatékony módja, amikor a korpuszban kifejezetten olyan szerkezeteket keresünk, amelyek valószínűleg hibásak, és a valóban hibás találatokat azonnal javítjuk. A javított korpuszt ezután exportálni lehet, és a rendszerbe integrált morfológiai egyértelműsítőt a javított korpusszal újratanítani.

7. TÉZIS:

Olyan egyértelműsítő-rendszert hoztam létre, amely alkalmas a szövegek morfoszintaktikai annotációjának és jelentésglosszáinak automatikus és manuális egyértelműsítésére. Emellett olyan korpuszkezelő alkalmazást hoztam létre, amellyel az annotált korpuszok hatékonyan kereshetőek és karbantarthatóak.

Kapcsolódó publikációk: 42, 38, 49, 50

3

A szerző publikációi

Folyóiratpublikációk

1 Borbála Siklósi, **Attila Novák**, Gábor Prószéky (2016): Context-aware correction of spelling errors in Hungarian medical documents, In: *Computer Speech & Language*, Vol. 35, pp. 219-233, ISSN 0885-2308, http://dx.doi.org/10.1016/j.csl.2014.09.001.

2 György Orosz, **Attila Novák**, Gábor Prószéky (2014): Lessons learned from tagging clinical Hungarian. In: *International Journal of Computational Linguistics and Applications*, Vol. 5 no. 2. ISSN 0976-0962

3 László János Laki, **Attila Novák**, Borbála Siklósi, György Orosz (2013): Syntax-based reordering in phrase-based English-Hungarian statistical machine translation. In: *International Journal of Computational Linguistics and Applications*, Vol. 4 no. 2. pp. 63–78. ISSN 0976-0962

4 István Endrédy, **Attila Novák** (2013): More effective boilerplate removal – the GoldMiner algorithm. In: *Polibits 48.* pp. 79–83. ISSN 1870-9044

Könyvfejezetek

5 **Attila Novák** (2015): Making morphologies the 'easy' way, In: A. Gelbukh (ed.) *Lecture Notes in Computer Science Volume 9041: Computational Linguistics and Intelligent Text Processing* Springer International Publishing, Berlin–Heidelberg. Part I pp. 127–138. ISBN 978-3-319-18110-3

6 Borbála Siklósi, **Attila Novák** (2014): Identifying and Clustering Relevant Terms in Clinical Records Using Unsupervised Methods. In: Besacier, L.; Dediu, A.-H. and Martín-Vide, C. (eds.) *Lecture Notes in Computer Science Volume 8791: Statistical Language and Speech Processing* Springer International Publishing, Berlin–Heidelberg. pp. 233–243 ISBN 978-3-319-11396-8

7 Borbála Siklósi, **Attila Novák**, Gábor Prószéky (2013): Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. In: Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov, Bianca Truthe (eds.) *Lecture Notes in Computer Science Volume 7978: Statistical Language and Speech Processing, First International Conference, SLSP 2013.* Springer, Berlin Heidelberg. pp. 248–259 ISBN 978-3-642-39592-5

8 György Orosz, László János Laki, **Attila Novák**, Borbála Siklósi (2013): Improved Hungarian Morphological Disambiguation with Tagger Combination. In: Habernal, Ivan; Matousek, Vaclav (eds.) *Lecture Notes in Computer Science, Vol. 8082: Text, Speech, and Dialogue, 16th International Conference, TSD 2013.* Pilsen, Czech Republic. Springer, Berlin–Heidelberg. pp. 280–287. ISBN: 978-3-642-40584-6

9 Nóra Wenszky, **Attila Novák** (2013): The hypercorrect key witness. In: Péter Szigetvári (ed.) *VLlxx: Papers presented to Varga László on his 70th birthday.* Department of English Linguistics, Eötvös Loránd University. ISBN 978-963-284-315-5

10 Borbála Siklósi, **Attila Novák** (2013): Detection and Expansion of Abbreviations in Hungarian Clinical Notes. In: F. Castro, A. Gelbukh, M.G. Mendoza (eds.) *Lecture Notes in Computer Science, Vol. 8265: Advances in Artificial Intelligence and Its Applications.* Springer, Berlin Heidelberg. pp. 318–328. ISBN 978-3-642-45114-0

11 László János Laki, György Orosz, **Attila Novák** (2013): HuLaPos 2.0 – Decoding morphology. In: F. Castro, A. Gelbukh, M.G. Mendoza (eds.) *Lecture Notes in Computer Science, Vol. 8265: Advances in Artificial Intelligence and Its Applications.* Springer, Berlin–Heidelberg. pp. 294–305. ISBN 978-3-642-45114-0

12 György Orosz, **Attila Novák**, Gábor Prószéky (2013): Hybrid text segmentation for Hungarian clinical records. In: F. Castro, A. Gelbukh, M.G. Mendoza (eds.) *Lecture Notes in Computer Science, Vol. 8265: Advances in Artificial Intelligence and Its Applications.* Springer, Berlin–Heidelberg. pp. 306–317. ISBN 978-3-642-45114-0

13 **Novák Attila**, Wenszky Nóra (2007): Mire jó és hogyan készül egy számítógépes morfológia. In: Alberti Gábor, Fóris Ágota (eds.) *A mai magyar formális nyelvtudomány műhelyei.* Nemzeti Tankönyvkiadó, Budapest. 157–169.

14 Gábor Prószéky, **Attila Novák** (2005): Computational Morphologies for Small Uralic Languages. In: A. Arppe, L. Carlson, K. Lindén, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund, A. Yli-Jyrä (eds.) *Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday.* Gummerus Printing, Saarijärvi/CSLI Publications, Stanford. pp. 116–125.

15 **Novák Attila** (2002): Többértelmű vagy homályos? In: Kálmán László, Trón Viktor, Varasdi Károly (eds.) *Lexikalista elméletek a nyelvészetben.* Tinta Könyvkiadó, Budapest. (Segédkönyvek a nyelvészet tanulmányozásához 13.) pp. 277–287.

16 **Novák Attila** (2002): HPSG fonológia. In: Kálmán László, Trón Viktor, Varasdi Károly (eds.) *Lexikalista elméletek a nyelvészetben.* Tinta Könyvkiadó, Budapest. (Segédkönyvek a nyelvészet tanulmányozásához 13.) pp. 99–128.

17 Kálmán László, **Novák Attila** (2001): A magyar egyszerű mondat fajtái. In: Kálmán László (ed.): *Magyar leíró nyelvtan*, Mondattan I. Tinta Könyvkiadó, Budapest, 2001. pp. 10–23.

18 Gyuris Bea, **Novák Attila** (2001): A topik és a kontrasztív topik. In: Kálmán László (ed.): *Magyar leíró nyelvtan, Mondattan I.* Tinta Könyvkiadó, Budapest, 2001. pp. 24–53.

19 **Novák Attila**, Dudás Kálmán, Kálmán László (2001): Igevivők. In: Kálmán László (ed.): *Magyar leíró nyelvtan, Mondattan I.* Tinta Könyvkiadó, Budapest, 2001. pp. 54–75.

20 **Novák Attila** (2001): A kommentelőzmények. In: Kálmán László (ed.): *Magyar leíró nyelvtan, Mondattan I.* Tinta Könyvkiadó, Budapest, 2001. pp. 76–91.

21 **Novák Attila** (2001): A hatókör felszíni egyértelműsítése. In: Kálmán László (eds.) *Magyar leíró nyelvtan, Mondattan I.* Tinta Könyvkiadó, Budapest, 2001. pp. 92–97.

22 **Novák Attila** (1999): Inflectional paradigms in Hungarian – The conditioning of suffix- and stem-alternations (Ragozási paradigmák a magyarban – A toldalék- és tőalternációkat kiváltó tényezők), Szakdolgozat, ELTE Elméleti Nyelvészet Szak, Budapest.

23 **Attila Novák** (1998): HPSG Phonology. In: *Lexicon Matters.* ELTE Theoretical Linguistics Programme, Budapest, 1998. pp. 33–48

24 **Attila Novák** (1998): Ambiguity and Vagueness. In: *Lexicon Matters.* ELTE Theoretical Linguistics Programme, Budapest, 1998. 115–120

Konferenciapublikációk

25 **Novák Attila**, Siklósi Borbála (2015): Automatic Diacritics Restoration for Hungarian. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics. pp. 2286–91.

26 Siklósi Borbála, **Novák Attila** (2015): Restoring the intended structure of Hungarian ophthalmology documents. In: *Proceedings of the BioNLP 2015 Workshop at the* 53^{rd} *Annual Meeting of the Association for Computational Linguistics, ACL 2015.* Beijing, China. pp. 152–157

27 **Novák Attila** (2015): „Olcsó” morfológia In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *XI. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 145–157

28 Siklósi Borbála, **Novák Attila** (2015): Nem felügyelt módszerek alkalmazása releváns kifejezések azonosítására és csoportosítására klinikai dokumentumokban. In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *XI. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 237–248

29 Borbála Siklósi, **Attila Novák**, Gábor Prószéky (2014): Resolving Abbreviations in Clinical Texts Without Pre-existing Structured Resources. In: *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014)*. Reykjavík. pp. 69–75

30 **Attila Novák** (2014): A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavík. pp. 1068–1073

31 **Novák Attila** (2014): A Humor új Fo(r)mája. In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *X. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 303–308. ISBN 978-963-306-246-3

32 Siklósi Borbála, **Novák Attila** (2014): Rec. et exp. aut. Abbr. mnyelv. KLIN. szövben – rövidítések automatikus felismerése és feloldása magyar nyelvű klinikai szövegekben. In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *X. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 167–176. ISBN 978-963-306-246-3

33 Siklósi Borbála, **Novák Attila** (2014): A magyar beteg. In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *X. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 188–198. ISBN 978-963-306-246-3

34 Orosz György, **Novák Attila** (2014): PurePos 2.0: egy hibrid morfológiai egyértelműsítő rendszer. In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *X. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 373–377. ISBN 978-963-306-246-3

35 Laki László, **Novák Attila**, Siklósi Borbála (2013): Hunglish mondattan – átrendezésalapú angol-magyar statisztikai gépifordító-rendszer. In: Tanács Attila; Vincze Veronika (eds.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*. SZTE, Szeged. pp. 71–82 ISBN 978-963-306-189-3

36 Siklósi Borbála, **Novák Attila**, Prószéky Gábor (2013): Helyesírási hibák automatikus javítása orvosi szövegekben a szövegkörnyezet figyelembevételével. In: Tanács Attila; Vincze Veronika (eds.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*. SZTE, Szeged. pp. 148–158 ISBN 978-963-306-189-3

37 Orosz György, **Novák Attila**, Prószéky Gábor (2013): Magyar nyelvű klinikai rekordok morfológiai egyértelműsítése. In: Tanács Attila; Vincze Veronika (eds.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*. SZTE, Szeged. pp. 159–169 ISBN 978-963-306-189-3

38 **Novák Attila**, Wenszky Nóra (2013): O & kọzepmagar zoalactany elemzọ. In: Tanács Attila; Vincze Veronika (eds.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*. SZTE, Szeged. pp. 170–181 ISBN 978-963-306-189-3

39 Endrédy István, **Novák Attila** (2013): Egy hatékonyabb webes sablonszűrő algoritmus – avagy miként lehet a cumisüveg potenciális veszélyforrás Obamára nézve. In: Tanács Attila; Vincze Veronika (eds.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* SZTE, Szeged. pp. 297–301 ISBN 978-963-306-189-3

40 György Orosz, László János Laki, **Attila Novák**, Borbála Siklósi (2013): Combining Language-Independent Part-of-Speech Tagging Tools. In: J. P. Leal, R. Rocha, and A. Simoes (eds.) *2nd Symposium on Languages, Applications and Technologies.* Porto: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. pp. 249–257 ISBN 978-3-939897-52-1

41 László János Laki, **Attila Novák**, Borbála Siklósi (2013): English-to-Hungarian Morpheme-based Statistical Machine Translation System with Reordering Rules. In: Marta R. Costa-jussa, Reinhard Rapp, Patrik Lambert, Kurt Eberle, Rafael E. Banchs, Bogdan Babych (eds.) *Proceedings of the Second Workshop on Hybrid Approaches to Machine Translation (HyTra).* Association for Computational Linguistics. pp. 42–50

42 **Attila Novák**, György Orosz, Nóra Wenszky (2013): Morphological annotation of Old and Middle Hungarian corpora. In: Piroska Lendvai, Kalliopi Zervanou (eds.) *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.* Association for Computational Linguistics. pp. 43–48

43 György Orosz, **Attila Novák** (2013): Purepos 2.0: a hybrid tool for morphological disambiguation. In: Galia Angelova, Kalina Bontcheva, Ruslan Mitkov (eds.) *Proceedings of the international conference Recent Advances In Natural Language Processing RANLP 2013.* Hissar, Bulgaria. pp. 539–545 ISSN 1313-8502

44 György Orosz, **Attila Novák** (2012): PurePos – an open source morphological disambiguator. In: Bernadette Sharp, Michael Zock (eds.) *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science.* Wrocław, Poland. pp. 53–63

45 Borbála Siklósi, György Orosz, **Attila Novák**, Gábor Prószéky (2012): Automatic structuring and correction suggestion system for Hungarian clinical records. In: *LREC-2012: SALTMIL-AfLaT Workshop on „Language technology for normalisation of less-resourced languages".* Istanbul, Turkey, 2012. pp. 29–34

46 Siklósi Borbála, Orosz György, **Novák Attila** (2011): Magyar nyelvű klinikai dokumentumok előfeldolgozása. In: *VIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2011).* Szegedi Tudományegyetem, pp. 143–340

47 **Novák Attila**, Orosz György, Indig Balázs (2011): Javában taggelünk. In: *VIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2011).* Szegedi Tudományegyetem, pp. 336–340.

48 Fejes László, **Novák Attila** (2010): Obi-ugor morfológiai elemzők és korpuszok. In: *VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2010).* Szegedi Tudományegyetem, pp. 284–291

49 Bakró-Nagy Marianne, Endrédy István, Fejes László, **Novák Attila**, Oszkó Beatrix, Prószéky Gábor, Szeverényi Sándor, Várnai Zsuzsa, Wagner-Nagy Beáta (2010): Online morfológiai elemzők és szóalakgenerátorok kisebb uráli nyelvekhez. In: *VII. Magyar*

Számítógépes Nyelvészeti Konferencia (MSZNY 2010). Szegedi Tudományegyetem, pp. 345–348

50 István Endrédy, László Fejes, **Attila Novák** , Beatrix Oszkó, Gábor Prószéky, Sándor Szeverényi, Zsuzsa Várnai, Beáta Wágner-Nagy (2010): Nganasan – Computational Resources of a Language on the Verge of Extinction. In: *Creation and Use of Basic Lexical Resources for Less-Resourced Languages: 7th SaLTMiL Workshop (LREC-2010)*. La Valletta, Malta, pp. 41–44

51 **Novák Attila**, Prószéky Gábor (2009): Kísérletek statisztikai és hibrid magyar–angol és angol– magyar fordítórendszerek megvalósítására. In: **VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2009)**. Szegedi Tudományegyetem, pp. 25–34

52 **Attila Novák** (2009): MorphoLogic's submission for the WMT 2009 Shared Task. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation at EACL 2009.* Athens, Greece. pp. 155–159

53 **Attila Novák**, László Tihanyi, Gábor Prószéky (2008): The MetaMorpho translation system. In: *Proceedings of the Third Workshop on Statistical Machine Translation at ACL 2008.* Columbus, Ohio. pp. 111–114

54 **Attila Novák** (2008): Language resources for Uralic minority languages. In: *Proceedings of the SALTMIL Workshop at LREC-2008: Collaboration: interoperability between people in the creation of language resources for less-resourced languages.* Marrakech, pp. 27–32

55 **Novák Attila**, M. Pintér Tibor (2006): Milyen a még jobb Humor. In: *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2006)*. Szegedi Tudományegyetem, pp. 60–69

56 **Attila Novák** (2006): Morphological Tools for Six Small Uralic Languages. In: *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, pp. 925–930

57 **Novák Attila**, Endrédy István (2005): Automatikus ë-jelölő program. In: *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005)*. Szegedi Tudományegyetem, pp. 453–454

58 **Novák Attila**, Wenszky Nóra (2005): Tundrai nyenyec morfológiai elemző és generátor. In: *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005)*. Szegedi Tudományegyetem, pp. 200–208

59 **Novák Attila** (2004): Az első nganaszan szóalaktani elemző. In: *II. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2004)*. Szegedi Tudományegyetem, pp. 195–202

60 **Attila Novák**, Viktor Nagy, Csaba Oravecz (2004): Combining symbolic and statistical methods in morphological analysis and unknown word guessing. In: *Proceedings of The Fourth International Conference on Language Resources and Evaluation (LREC-2004)*. Lisbon, pp. 1255–1258

61 **Attila Novák** (2004): Creating a Morphological Analyzer and Generator for the Komi language. In: *Proceedings of the SALTMIL Workshop at LREC-2004: First Steps in Language Documentation for Minority Languages.* Lisbon, pp. 64–67.

62 **Novák Attila**, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése. In: *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003)*. Szegedi Tudományegyetem, 45–57

63 **Novák Attila** (2003): Milyen a jó Humor? In: *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003)*. Szegedi Tudományegyetem, pp. 138–145

64 **Attila Novák**, Viktor Nagy, Csaba Oravecz (2003): Corpus assisted development of a Hungarian morphological analyser and guesser. In: Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) *Proceedings of the Corpus Linguistics 2003 conference.* UCREL technical paper number 16. UCREL, Lancaster University, pp. 583–590

Kutatási beszámolók

65 Borbála Siklósi, **Attila Novák**, György Orosz, Gábor Prószéky (2014): Processing noisy texts in Hungarian: a showcase from the clinical domain, In: Péter Szolgay (ed.), *Jedlik Laboratories Reports*, Vol. II, no. 3, pp. 5–62 ISSN 2064-3942

Irodalomjegyzék

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFst: A General and Efficient Weighted Finite-State Transducer Library. In Holub, J. and Zdárek, J., editors, *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*, volume 4783 of *Lecture Notes in Computer Science*, pages 11–23. Springer.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI Publications, Ventura Hall.

Huldén, M. (2009). Foma: a Finite-State Compiler and Library. In Lascarides, A., Gardent, C., and Nivre, J., editors, *Proceedings of EACL 2009*, pages 29–32, Athens, Greece. The Association for Computer Linguistics.

Lindén, K., Silfverberg, M., Axelson, E., Hardwick, S., and Pirinen, T. (2011). HFST—Framework for Compiling and Applying Morphologies. In Mahlow, C. and Pietrowski, M., editors, *Systems and Frameworks for Computational Morphology*, volume Vol. 100 of *Communications in Computer and Information Science*, pages 67–85.

Prószéky, G. and Kis, B. (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 261–268, College Park, Maryland. Association for Computational Linguistics.