

TDK dolgozat

Szántó Zsolt

Szegedi Tudományegyetem
Természettudományi és Informatikai Kar
Számítógépes Algoritmusok és Mesterséges Intelligencia
Tanszék

Konstituensalapú szintaktikai elemzés
morfológiailag gazdag nyelvekre

Készítette:
Szántó Zsolt
programtervező informatikus MSc,
II évfolyam

Témavezető:
Dr. Farkas Richárd
adjunktus

Tartalomjegyzék

Bevezetés	4
1. Szintaxis a természetes nyelvekben	6
1.1. Szintaktikai reprezentációk	6
1.2. Szeged Treebank	7
2. Statisztikai konstituenselemzés	9
2.1. Valószínűségi környezetfüggetlen nyelvtanok	9
2.2. Elemzők fejlődése	10
2.3. Berkeley Parser	11
2.3.1. Rejtett alállapotokra épülő szabályok valószínűségének tanulása .	12
2.3.2. Rejtett alállapotok szétbontása	14
2.3.3. Rejtett alállapotok összevonása	14
3. Kapcsolódó munkák	15
4. Módszerek morfológiailag gazdag nyelvek elemzésére	17
4.1. Korpusz és kiértékelési metrikák	17
4.2. Kiterjesztett lexikai modellek	18
4.3. Morfológiai kódok automatikus összevonása	20
4.3.1. Eljárás morfológiai jellemzők értékeinek összevonására	20
4.3.2. Baseline preterminális halmazok létrehozása	23
4.3.3. Eredmények különböző preterminális halmazokkal	23
4.4. Morfológián alapuló jellemzők használata az eredmények újrarangsorolására	25
4.5. A teljes rendszer teljesítménye	27
4.5.1. Korpuszok	27
4.5.2. Eredmények	27
5. Összegzés	29
Köszönetnyilvánítás	30
Irodalomjegyzék	31

Bevezetés

Szintaktikai elemzés szempontjából a világ nyelvei csoportosíthatóak a morfológiai gazdagságuk szintje szerint, ami fordítottan arányos a konfigurációs szintjükkel. A skála egyik végén található az angol, egy erősen konfiguratív nyelv, míg a másik oldalon ott a magyar a maga gazdag morfológiájával és szabad szórendjével [1]. A szintaktikai elemzők általában az angol nyelvet figyelembe véve lettek kifejlesztve, viszont a világ nyelveinek jó része alapjaiban különbözik az angoltól. Különösképpen igaz ez a morfológiailag gazdag nyelvekre, melyek a legtöbb mondat szintű szintaktikai információt a morfológia (azaz a szavak) szintjén, és nem a szórenddel fejezik ki. Ezen különbségek miatt a morfológiailag gazdag nyelvek elemzése olyan technikákat igényel, melyek különböznek az angol nyelvre kifejlesztett módszerektől (vagy kiterjesztik azokat) [2]. Dolgozatomban először egy általános bevezetőt adok a konstituenselemzésről, megmutatom az ehhez használt eljárásokat, erőforrásokat. Ezek után a konstituenselemzés tökéletesítésének érdekében három olyan technikát mutatok be, amelyeket speciálisan a morfológiailag gazdag nyelvek kihívásainak kezelésére dolgoztam ki.

Az utóbbi két évtizedben jelentős mértékben fejlődtek a konstituenselemzők [3, 4, 5, 6], ami elsősorban a Penn Treebank jelenlétének köszönhető [7]. Ám amíg angol nyelven folyamatos fejlődés volt tapasztalható, a morfológiailag gazdag nyelvek kevesebb figyelmet kaptak.

Magyar nyelvre a Szeged Treebank [8], egy – nemzetközi viszonylatban is – nagyméretű, kézzel annotált konstituenskorpusz már közel 10 éve rendelkezésre áll. Annak ellenére, hogy ez kiváló alapanyagul szolgálhatna statisztikai konstituenselemzők fejlesztéséhez, néhány korai kísérletet leszámítva [9], legjobb tudomásom szerint ez idáig senki sem kísérelte ezt meg.

Ebben a tanulmányban a morfológiailag gazdag nyelvek két fő problémájának meg-

oldására keresem a választ. Ezen problémák az optimális preterminálisok (morfológiai kódok) halmazának megtalálása és a szóalakok nagy számának kezelése.

A sztenderd valószínűségi környezetfüggetlen nyelvtanokra épülő konstituenselemzők a preterminálisokat egy-egy struktúra nélküli címkének tekintik. Ezen címkék optimális halmazának meghatározása nagyon kritikus az elemzés hatékonyságára nézve. A két legkézenfekvőbb megoldás, hogy vagy csak a fő szófaji kódokat vagy a teljes morfológiai leírást használjuk címkének, például egy $N\#subpos = c|num = s$ morfológiai címke esetén – ahol az N (főnév) a fő szófajt jelöli, a $subpos = c$ (köznév) és $num = s$ (egyesszám) morfológiai jellemző-érték párok – a fő szófaji kód az N , a teljes morfológiai leírás pedig az egész címke. Előbbi kódolással sok információt veszítünk, míg utóbbi esetén a preterminálisok magas száma miatt az elemzés lassú lehet, a tanulás során az optimalizálási feladat pedig kezelhetetlenné válik. Ezen problémák kezelésére kidolgoztam egy új, teljesen automatikus módszert a morfológiai kódkészlet csökkentésére.

A másik probléma, hogy a toldalékolásnak köszönhetően a morfológiailag gazdag nyelvekben rengeteg eltérő szóalak található (ellentétben az angollal). Ennek következtében elemzéskor a tanítóhalmazban nem vagy csak kevésszer előforduló szavak (úgynevezett ismeretlen vagy ritkán látott szavak) száma nagyon magas, ami negatív hatással van a konstituenselemzők hatékonyságára. Goldberg és Elhadad [10] gondolatait követve kiegészítettem a lexikai modellt külső lexikonok használatával. Megvizsgáltam, hogy egy teljesen felügyelt szófaji egyértelműsítő mennyire alkalmazható az általam javasolt felügyelet nélkülivel szemben a külső lexikonok elkészítésére.

Végül bemutatom, hogy milyen hatással van egy külső újrarangsoroló rendszer a konstituensfák elemzésére, majd az újrarangsoroló rendszer angolra kifejlesztett jellemzőkészletét morfológiai információk felhasználásával bővítem. Ez a megoldás mindkét fentebb említett problémához kapcsolódik. A három technika hatását a magyar mellett másik négy morfológiailag gazdag nyelven, a baszkon, a francián, a héberén és a németen mutatom be.

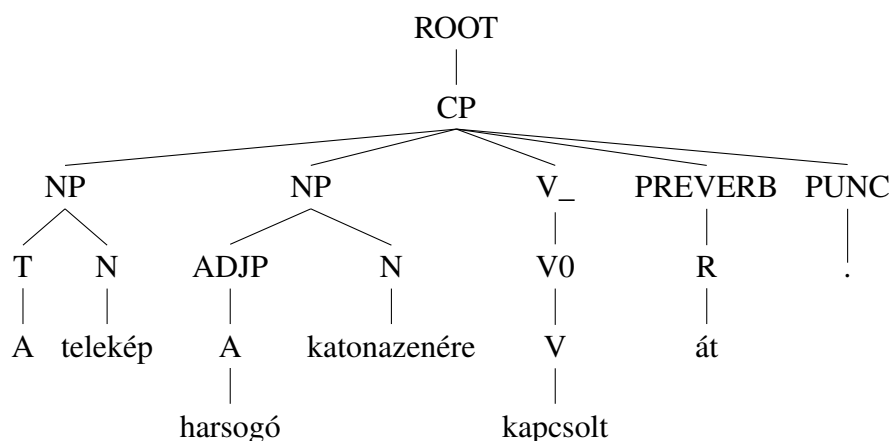
1. fejezet

Szintaxis a természetes nyelvekben

1.1. Szintaktikai reprezentációk

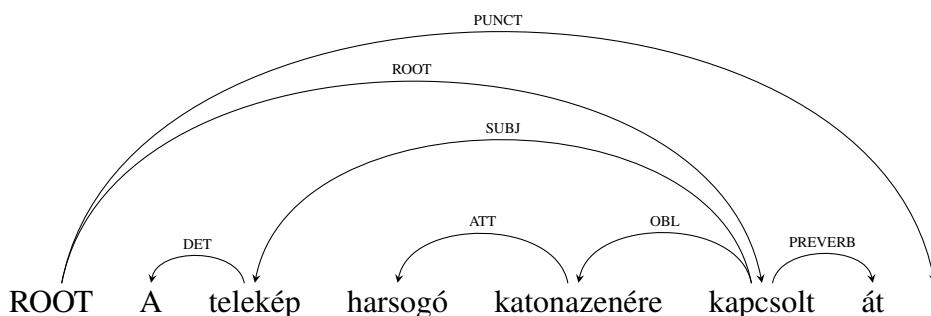
A szintaxis célja a mondatban rejlő nyelvtani kapcsolatok leírása. Az ilyen kapcsolatok megadására több eltérő reprezentáció is létezik. Ilyen reprezentációval (mondatrészek aláhúzása) már az általános iskolában is találkozhattunk. A számítógépes nyelvészetben az úgynevezett konstituens nyelvtanok és függőségi nyelvtanok használata a legelterjedtebb. Egy mondat szerkezetét mindkét esetben egy fa reprezentálja, viszont ezen fák nagyban eltérnek egymástól.

1.1. ábra. Konstituensfa.



Az 1.1. ábra és az 1.2. ábra nagyon jól mutatja a különbséget a két módszer között. A konstituens nyelvtan esetén (1.1. ábra) a szavak csoportjai frázisokat, absztrakt cso-

1.2. ábra. Függőségi fa.



mópontokat alkotnak és ezek írják le a mondat szerkezetét. Az előbbiekkal szemben a függőségi nyelvtanok (1.2. ábra) esetén a szavak egymáshoz kapcsolódnak, a közöttük futó élek írják le a nyelvtani kapcsolatokat.

Az általános vélekedés szerint a szabad szórenddel rendelkező nyelvek esetén, mint például a magyar is, a függőségi fák használata az előnyösebb. Ennek következtében a szabad szórendű nyelvek esetén a konstituensalapú kutatásokra kisebb hangsúly került az elmúlt években. A Szegedi Tudományegyetemen megvalósult kutatások és a 2013-as SPMRL (Statistical Parsing of Morphologically Rich Languages) [11] verseny eredményei némileg cáfolják ezen állításokat. Ezek arra a következtetésre jutottak, hogy a konstituens reprezentáció használatával hasonló, bizonyos esetekben jobb eredmények érhetők el, mint függőségi elemzés esetén. Mindenesetre megállapíthatjuk, hogy abszolút nem bizonyított a függőségi nyelvtanok használatának előnye a morfológiailag gazdag nyelveken, valamint mindkét reprezentációnak vannak előnyei és hátrányai. Ezért a konstituenselemzők vizsgálata egy fontos, de ritkán kutatott téma.

1.2. Szeged Treebank

Konstituensalapú erőforrások tekintetében a magyar nyelv esetén kivételesen jó helyzetben vagyunk, hiszen közel 10 éve rendelkezésre áll egy – nemzetközi viszonylatban is – nagyméretű, kézzel annotált konstituenskorpusz, a SzegedTreebank [8]. A kicsivel több mint 82 ezer mondatot magába foglaló szöveggyűjtemény alapjául a morfológiai infor-

mációkat is tartalmazó Szeged Korpusz [12] szolgált, ebből adódóan a Szeged Treebank is körülbelül 1,2 millió szót és 250 ezer írásjelet tartalmaz. A magyar nyelvű eredmények empirikus kiértékeléséhez a Szeged Treebanket, azaz pontosabban annak újságcikkekből álló részkorpuszát használtam.

A korpusz eredetileg TEI XML formátumban tartalmazza a mondatokat, melyben a hozzájuk tartozó konstituensfa mellett megtalálható az egyes szavak szótöve, szófaja és morfológiai jellemzői, illetve minden szóhoz hozzá van társítva annak minden lehetséges morfológiai elemzése is.

A korpuszban összesen 13 szintaktikai címkét alkalmaznak, ezek közül talán a legfontosabbak a CP, ami az egyes tagmondatokat jelöli, az NP a főnévi csoportokat határolja és a V_ az igét tartalmazza. Ezen felül a korpuszban az ige, a határozói igenév és a főnévi igenév esetén jelölve vannak az adott elemhez tartozó bővítmények is, melyek a tagmondat egy másik szavára nézve olyan szintaktikai szerepeket jelölnék, mint például az alany vagy a tárgy. Ezen bővítményekkel a dolgozat későbbi szakaszaiban nem foglalkozom, ennek több indoka is van. A gyakorlatban ezeket valamilyen utófeldolgozó eljárás keretein belül szokták hozzáadni az egyes konstituensfákhoz.

2. fejezet

Statisztikai konstituenselemzés

A számítógépes szintaktikai elemzés történhet szabályalapú vagy statisztikai módszerekkel. Az utóbbi esetben a rendszer az elemzéshez szükséges információkat egy kézíleg annotált treebankból tanulja. Ezen statisztikai megközelítésre egy általánosan elterjedt módszer a valószínűségi környezetfüggetlen nyelvek használata. Erre épül a későbbiekben részletesebben bemutatásra kerülő Berkeley Parser [5] is, mely jelenleg a legelterjedtebb és talán leghatékonyabb konstituenselemző.

2.1. Valószínűségi környezetfüggetlen nyelvtanok

Környezetfüggetlen nyelvtan alatt olyan formális nyelvtanról beszélhetünk, melyben minden szabály

$$A \rightarrow \alpha$$

alakú, ahol A egy nemterminális, míg α terminálisok és nemterminálisok egy sorozata.

Természetes nyelvek esetén ilyen szabályok lehetnek például az alábbiak:

- $CP \rightarrow NP\ NP\ VP$
- $NP \rightarrow N\ NP$
- $N \rightarrow \textit{paralelepipedon}$

ahol az első szabály azt fejezi ki, hogy a tagmondat állhat egy főnévi csoport, főnévi csoport, igei csoport hármából. A második szabály alapján a főnévi csoportból következhet

főnév, főnévi csoport páros. Míg a harmadik szabály azt írja le, hogy a főnévből mint szófajból következhet a *paralelepipedon* szó.

A nyelvtanból generálható konstituensfákban a szófajok mindig közvetlen a szavak fölött találhatók, ezért a szófajokhoz tartozó nemterminálisokat szokás preterminálisnak is nevezni.

Az egyes szabályok csoportosíthatók aszerint, hogy szintaktikai vagy lexikai szabályról van-e szó. Akkor beszélünk szintaktikai szabályról, ha a szabály baloldalán egy szintaktikai frázis áll, a jobboldalán pedig valamilyen nemterminálisok. Lexikai szabályok esetén pedig a baloldalon minden esetben egy preterminális (szófaji címke) áll, amely valamilyen szót generál. Ez alapján jól látható, hogy a fenti példában az első két szabály szintaktikai, míg a harmadik lexikai.

Valószínűségi környezetfüggetlen nyelvtan esetén a szabályokhoz hozzárendelünk egy valószínűségi értéket, ami azt fejezi ki, hogy a szabály baloldalán található nemterminálisból milyen valószínűséggel vezethető le a szabály jobb oldala, azaz egy $A \rightarrow \alpha$ szabály esetén milyen valószínűséggel következik A -ból α . Ebből következik, hogy az azonos nemterminálisból kiinduló szabályok valószínűségének az összege 1 lesz, azaz minden egyes szimbólumhoz tartozik egy eloszlás, ami megmondja, hogy a szimbólumból az egyes levezetések milyen arányban következnek.

A konstituensfák automatikus elemzése során először – egy kézíleg annotált erőforrás alapján – meg kell határoznunk a fent említett eloszlásokat, ezt a folyamatot hívjuk tanulásnak. Majd az így kapott modell segítségével meghatározhatjuk, hogy egy adott mondatot mely konstituensfa generál a legnagyobb valószínűséggel.

2.2. Elemzők fejlődése

A hagyományos környezetfüggetlen megközelítésnek több hiányossága is van. Ezen hiányosságok egyike éppen a környezetfüggetlenségből fakad, mivel a lehetséges szabályok halmaza erősen korlátozott. A másik nagy probléma, hogy a nyelvészek által használt nemterminálisok csak nagyon korlátozott mennyiségű információ tárolására alkalmasak, mélyebb nyelvtani összefüggések nehezen írhatók le a segítségükkel.

Ezen problémák orvosolására az eredeti környezetfüggetlen leírásnak számos kibő-

vítése született. Ezek általában a nemterminálisok halmazának kibővítésére épültek. Az ilyen módszerek a legtöbb esetben konzisztensek maradnak a konstituenskorpuszal, megtartják annak struktúráját, viszont az elemzés idejére az egyes nemterminálisokat szétbontják több, részletesebb információk tárolására alkalmas új szimbólumra, melyekből bármikor visszanyerhetők az eredeti szimbólumok.

Az egyik első ilyen megvalósítás [13] szerint az egyes nemterminális csomópontokon jelölve lett a szülő csomópont is. Így olyan új szabályok keletkezhetek, mint például a $NP_VP \rightarrow NP_NP\ PP_NP$. A megoldás előnye, hogy teljesen nyelvfüggetlen, egyszerű transzformációkkal bármelyik treebank ilyen formájúra alakítható, illetve az új szabályok több információ kinyerését teszik lehetővé.

Klein és Manning 2003-as munkájukban [14] az egyes frázisokat nyelvészeti szempontok alapján kézzel bontották fel, aminek a segítségével nagyon tömör nyelvtan mellett jónak mondható eredményeket értek el. A készítőik olyan új nemterminálisokat vettek fel, mint például az alanyi és a tárgyas főnévi csoport, aminek következtében a rendszer teljesen nyelvfüggővé vált.

Az utóbbival teljesen szembemenő elképzelés a nemterminális szimbólumok automatikus szétbontása [15]. Ennek a teljesen nyelvfüggetlen megvalósításnak az alapja, hogy az egyes nemterminális szimbólumokat k darab új rejtett szimbólumra (alállapotra) bontjuk. Ezen rejtett alállapotokra épülő szabályok valószínűségeit pedig automatikusan próbáljuk megtanulni. Az alállapotok segítségével a rendszer képes mély nyelvtani összefüggések megtanulására. A módszer hátránya, hogy a keletkező nyelvtan viszonylag kis k esetén is kezelhetetlen méreteket érhet el, valamint miközben egyes nemterminálisokat túlzottan szétbonthat, más esetekben pedig a szükségesnél kevesebb új alállapot keletkezhet. Ezen felül az egyes szabályok nyelvészeti nem, vagy csak nehezen interpretálhatók.

2.3. Berkeley Parser

Az automatikus összevonás és a nyelvészeti szempontokon alapuló módszer előnyeinek egyesítésére épít a jelenleg talán legelterjedtebb és egyben egyik leghatékonyabb konstituenselemző rendszer, a Berkeley Parser [5]. A Berkeley Parser alapja, hogy az egyes

nemterminálisokat különböző számú új állapotra bontja szét és csak azokon a helyeken használ sok állapotot, ahol ez szükséges. Ezzel egy olyan nyelvfüggetlen modellt kapunk, melynek segítségével – az eredeti automatikus szétbontási koncepciónál – sokkal tömörebb nyelvtant állíthatunk elő.

A rendszer az egyes nemterminálisokhoz tartozó állapotokat számának beállításához egy iteratív algoritmust használ. Az algoritmus arra épül, hogy minden lépés elején szétbontja a meglévő állapotokat, az újonnan kapott állapotokra kiszámolja az egyes szabályok valószínűségeit, majd az elemzés számára haszontalannak bizonyuló állapotokat végül összevonja.

A következő fejezetekben részletesen bemutatom a Berkeley Parser tanulási folyamatát, ami alapjául szolgált az általam fejlesztett technikáknak is.

2.3.1. Rejtett állapotokra épülő szabályok valószínűségének tanulása

A Berkeley Parser az egyes lépésekben a rejtett állapotokhoz tartozó szabályok valószínűségének meghatározására a külső-belső (inside-outside) algoritmus egy módosított változatát használja. Az eredeti algoritmust James K. Baker [16] írta le először, a rejtett Markov-modellnél alkalmazott előre-hátra algoritmus környezetfüggetlen nyelvekre való általánosításaként. Ennek következtében a külső-belső algoritmus tekinthető az Expectation-Maximization egy speciális esetének.

Az Expectation-Maximization egy iteratív optimalizálási eljárás, melyet kifejezetten a rejtett változókon alapuló modellek paramétereinek megbecslésére találtak ki. Az algoritmus az Expectation és Maximization lépések felváltva történő ismétlésére épül. Az Expectation lépésben kiszámítja, hogy az egyes szabályok milyen valószínűséggel generálják a mondat egyes részeit. A Maximization részben pedig ezek alapján meghatározza az egyes szabályok valószínűségét.

Ahhoz, hogy ezeket a lépéseket részletesebben is meg tudjuk vizsgálni vezessük be az alábbiakat. Jelölje β a rejtett állapotokra épülő szabályok valószínűségét, amikre optimalizálni szeretnénk. Legyen w egy mondat és T a hozzá tartozó fa, A egy nemterminális, mely az (r, t) indexek közötti szakaszt feszíti ki. Legyenek B és C az A gyermekei, melyek az (r, s) és (s, t) szakaszokat feszítik ki. Ezen felül jelölje A_x , B_y és C_z rendre az A , B és C egy-egy állapotát. Ezek segítségével megadhatjuk az úgynevezett belső

és külső valószínűségeket a rejtett állapotokra nézve. A belső valószínűség $P_{in}(r, t, A_x)$ azt adja meg, hogy az A_x állapotból milyen valószínűséggel generálódik a mondat r és t indexek közötti szakasza ($P(w_{r:t}|A_x)$). A külső valószínűség $P_{out}(r, t, A_x)$ pedig annak a valószínűséget adja meg, hogy a levezetés A_x -ben tart, és az általa generált szakaszon kívül pontosan a mondat többi része szerepel (azaz 1, r indexek közötti szavak, illetve a t és az n indexek közötti szavak, ahol n a mondat hossza) ($P(w_{1:r}A_xw_{t:n})$). Ezek az értékek az alábbi képletek alapján rekurzívan számíthatók:

$$P_{in}(r, t, A_x) = \sum_{x,y} \beta(A_x \rightarrow B_y C_z) P_{in}(r, s, B_y) P_{in}(s, t, C_z) \quad (2.1)$$

$$P_{out}(r, s, B_y) = \sum_{x,y} \beta(A_x \rightarrow B_y C_z) P_{out}(r, t, A_x) P_{in}(s, t, C_z) \quad (2.2)$$

$$P_{out}(s, t, C_z) = \sum_{x,y} \beta(A_x \rightarrow B_y C_z) P_{in}(r, s, B_y) P_{out}(r, t, A_x) \quad (2.3)$$

A belső és külső valószínűségek ismeretében most már felírhatjuk a korábban ismertetett Expectation és Maximization lépéseket. Az előbbi esetben minden egyes T tanító fa minden egyes (nemterminális) csúcsához tartozó minden egyes szabályra ki kell számítanunk az alábbi valószínűséget:

$$P((r, s, t, A_x \rightarrow B_y C_z) | w, T) \propto P_{out}(r, t, A_x) \beta(A_x \rightarrow B_y C_z) P_{in}(r, s, B_y) P_{in}(s, t, C_z) \quad (2.4)$$

A Maximization lépésben az egyes szabályelőfordulásokra (az előző lépésben) kapott értékek alapján újraszámolhatjuk a β szabály valószínűségét az alábbi módon:

$$\beta(A_x \rightarrow B_y C_z) = \frac{\#\{A_x \rightarrow B_y C_z\}}{\sum_{x',y'} \#\{A_x \rightarrow B_{y'} C_{z'}\}} \quad (2.5)$$

Ilyen módon kapunk egy lokális optimalizálási eljárást, mely iteratívan frissíti a rejtett állapotokra vonatkozó szabályok valószínűségét.

2.3.2. Rejtett alállapotok szétbontása

A Berkeley Parser az alállapotokra bontást úgy végzi, hogy először szétbontja az aktuális lépésben meglévő alállapotokat, majd az újonnan tanult alállapotokra újratanulja a modellt. Ezután a haszontalannak bizonyult szétbontásokat összevonja. Szétbontáskor a meglévő nyelvtan minden alállapotát két új alállapotra bontjuk, az új alállapotokra épülő szabályok valószínűségét egyenletes eloszlással kapjuk meg az előző szabályokból minimális véletlen hozzáadása mellett. A véletlen arra szolgál, hogy az Expectation-Maximization számára megtörje a szimmetriát. Végül az így kapott új szabályok segítségével újratanítjuk a modellt.

2.3.3. Rejtett alállapotok összevonása

A rejtett állapotok összevonásának metodikája különösen fontos a dolgozat szempontjából, hiszen ez a módszer képezte az alapját a későbbiekben bemutatásra kerülő morfológiai kódok automatikus összevonását végző eljárásnak.

Az összevonás alapja, hogy az egyes nemterminálisokat csak akkor osztjuk szét több alállapotra, ha ez szükséges. Ehhez a rendszer azt vizsgálja, hogy az egyes szétbontások során mennyi információt nyerünk. Azon szétbontásokat szeretnénk visszacsinálni, ahol ez a nyereség kicsi, azaz, ahol az összevonás kicsi információvesztéssel jár. A rendszer a veszteségek becsléséhez a nemterminális csúcsokat lokálisan vizsgálja, a csúcsokban a veszteséget a csúcs összevonás utáni, illetve összevonás előtti valószínűségének a hányadosából számítja. Egy szétbontás veszteségének a meghatározásához ezeket a veszteségeket minden fa minden csúcsára összeszorozza a rendszer. A Berkeley Parser ezeket a veszteségeket minden egyes szétbontás esetén megvizsgálja, és azokban az esetekben, ahol ez a veszteség kicsi volt, végrehajtja az összevonásokat. A rendszer készítői úgy találták, hogy ha az esetek 50%-ában összevonnak, az drámaian csökkenti a nyelvtan méretét, viszont semmit nem ront a hatékonyságon.

3. fejezet

Kapcsolódó munkák

Angolra a konstituenselemzés egy erősen kutatott terület, amelyet napjainkban elsősorban az előbbieken bemutatott korpusz alapú statisztikai megközelítések urálnak. Ezzel szemben az eddigiekben a morfológiailag gazdag nyelveken történő konstituenselemzésre lényegesen kevesebb figyelmet fordítottak. Számos ilyen nyelven létezik konstituens-korpusz, miközben ezekről publikált (automatikus elemzési) eredmények csak nagyon korlátozott számban találhatók. Slav Petrov [17] futtatta a Berkeley Parsert jó eredményeket elérve arab, bolgár, francia, német és olasz nyelveken, ezen felül korábbi munkák találhatók további nyelveken, mint például: héber [10], koreai [18] vagy spanyol [19]. A 2013-ban megrendezett SPMRL (Statistical Parsing of Morphologically Rich Languages) [11] versenyen a cél 9 morfológiailag gazdag nyelv függőségi és konstituenselemzés volt, amelyhez jól használható korpuszokat is biztosítottak.

A dolgozat egyik fő témája a morfológiai kódok automatikus összevonását végző algoritmus, melyhez egyrészt a preterminális halmaz kézi finomítását tartalmazó munkák tartoznak [20, 19]. De az általam fejlesztett megoldáshoz a legközelebb Dehdari [21] munkája áll, ahol metaheurisztikákat határoztak meg morfológiai jellemzők iteratív törlésére és beszúrására. Az ő megközelítésükben az elemző egy fekete doboz kiértékelőként szerepelt, amit minden egyes részeredmény esetén nulláról újra kellett tanítani, majd leelemzettetni vele a mondatokat. Ezzel szemben az én megvalósításom a Berkeley Parser egy moduljaként nem igényli a többszöri újratanítást, ezzel egy sokkal gyorsabb megoldást adva.

A konstituenselemzésre készített leghatékonyabb rendszerek egy része tartalmaz egy

gazdag jellemzőtérrel rendelkező, újrarangsoroló második lépést [4, 6, 22]. Ezen esetekben a valószínűségi környezetfüggetlen nyelvtanokra épülő elemzőket a lehetséges megoldások kinyerésére alkalmazzák. Az ilyen eljárások nagyméretű jellemzőkészlet (rendszerint néhány millió jellemző) segítségével adnak pontosabb becslést a nyelvtan által meghatározott 50-100 elemzés [4] valószínűségére. Ezek a jellemzőkészletek kifejezetten angolra lettek kifejlesztve. Dolgozatban olyan morfológiai információkra épülő jellemzősablonokat is bemutatok, melyekkel javítható az elemző hatékonysága morfológiailag gazdag nyelvek esetén.

4. fejezet

Módszerek morfológiailag gazdag nyelvek elemzésére

A következő alfejezetekben három olyan technikát fogok bemutatni, amelyekkel javítható a konstituenselemzés hatékonysága morfológiailag gazdag nyelveken [23].

4.1. Korpusz és kiértékelési metrikák

A vizsgálatokhoz a Szeged Treebank [8] újságcikkekből álló alkorpuszát használtam, ami az SPMRL versenyen is alkalmazott tanító/validációs/teszt vágás szerint volt szétbontva. A fejlesztés során az eredmények kiértékelésére a validációs halmazt alkalmaztam, majd a végső összegzésnél a teszhalmazon végeztem el a méréseket. A tanító halmazban összesen 8146 mondat található, míg ugyanez az érték a validációs halmazban 1051. Az egyes mondatokban átlagosan 21,76 token található. Összesen 680 morfológiai címkét tartalmaz a korpusz, ami 16 fő szófaji kód köré csoportosul. A validációs halmazon az ismeretlen szavak aránya 19,94%.

Kiértékeléskor a PARSEVAL [24] metrikát használtam, ami egy F-mérték a szintaktikai fa, mint zárójelezés felett, azaz egy belső csúcs akkor helyes, ha a lefedett szavak és a nemterminális címke is megegyezik a gold standard treebankbeli fával. Maguk a preterminálisok nem számítanak a kiértékelési metrikába, de mint látni fogjuk, a preterminálisok halmazának megválasztása igen fontos.

Ezen felül megvizsgáltam a hibátlanul leelemzett mondatok arányát is.

4.2. Kiterjesztett lexikai modellek

Elsőként egy megoldást ismertetek az ismeretlen szavak problémájára, mely kritikus fontosságú lehet a morfológiailag gazdag nyelvekben a toldalékolás következtében létrejövő rengeteg szóalak miatt. Goldberg és Elhadad [10] ajánlását követve, kiterjesztettem a lexikai modellt a tokenek lehetséges morfológiai elemzéseinek gyakorisági információival.

Minden egyes t címkére és w szóra az alábbi képlet alapján becsültem a $P(t|w)$ valószínűséget:

$$P(t|w) = \begin{cases} P_{tb}(t|w), & \text{ha } c(w) \geq K \\ P_{ex}(t|w), & \text{ha } c(w) = 0 \\ \frac{c(w)P_{tb}(t|w) + P_{ex}(t|w)}{1 + c(w)}, & \text{különben} \end{cases} \quad (4.1)$$

ahol a $c(w)$ a w tanító halmazon vett előfordulásainak a száma, a K egy előre definiált konstans, a $P_{tb}(t|w)$ a treebank alapján számolt valószínűség, míg a $P_{ex}(t|w)$ valószínűségek kiszámítására egy külső lexikont vettem alapul. A konstituenselemző számára szükséges $P(w|t)$ emissziós valószínűségek megkaphatók a $P(t|w)$ valószínűségekből a Bayes-szabály felhasználásával.

A kulcskérdés itt az, hogy hogyan érdemes elkészíteni a külső gyakorisági lexikont, amely $P_{ex}(t|w)$ becslésére szolgál. Goldberg és Elhadad [10] javaslata alapján baseline-nak egy olyan lexikont használtam, melyben az adott szavak lehetséges morfológiai elemzéseit egy morfológiai elemző segítségével határoztam meg, és ezekre a valószínűségeket egyenletes eloszlás alapján számítottam ki. Például az *ír* szó főnévként jelenthet etnikai csoportot vagy igeiként az írás folyamatát. Ebben a modellben mindkét esetben azonos valószínűségi értékkel szerepel, pedig sokkal többször használjuk az *ír* szót igeiként, mint főnévként. Ebből is látszik, hogy részletesebb becslésre van szükség a külső lexikon valószínűségeit illetően. Erre egy megoldás lehet, ha valamilyen eljárás segítségével előállítunk egy jelölt szöveghalmazt, és az eloszlásokat az alapján számítjuk ki.

Goldberg és Elhadad [10] jelentős javulásról számolt be héber nyelvre, amikor az egyenletes eloszlást használó baseline helyett a gyakoriságokat egy olyan nagyméretű korpuszon számolták le, amelyet felügyelet nélküli szófaji egyértelműsítő rendszer [25]

használatával automatikusan annotáltak.

Megmutatom, hogy felügyelt szófaji egyértelműsítéssel ugyanolyan mértékű javulás érhető el. Elsősorban az motiválta a felügyelt egyértelműsítő használatát, hogy – a felügyelet nélküli modellel szemben – nem igényel morfológiai elemzőt, amely meg tudná adni egy szóra a lehetséges morfológiai címkéket. Bár magyar nyelvre rendelkezésre áll morfológiai elemző, de ezen elemzők teljesen nyelvfüggőek, ráadásul az sem garantált, hogy kompatibilisek az adott treebankkel, így közel sem biztos, hogy egy ezekre építő módszer általánosan használható bármely morfológiailag gazdag nyelv esetén. Ezzel szemben bármikor felépíthető egy elfogadható felügyelt morfológiai egyértelműsítő rendszer az adott treebank tanító halmazán.

A címkézetlen szövegekben a morfológiai egyértelműsítés folyamatára a feltételes véletlen mezőkre (CRF) építő MarMot [26] szófaji egyértelműsítő rendszert alkalmaztam. Ez a tisztán statisztikai elemző 97,6%-os pontosságot ért el a validációs halmazon, amely versenyképes a nyelvfüggő szabályokat is alkalmazó magyar nyelvre használt szófaji egyértelműsítőkkal (például a magyarlanccal [27]).

4.1. táblázat. PARSEVAL eredmények és a hibátlanul elemzett mondatok aránya (EX) különböző külső lexikonok használata mellett.

	PARSEVAL	EX
Berkeley Parser	87.22	12.75
egyenletes eloszlás	87.31	14.78
validációs	88.29	15.22
validációs + MNSz	89.27	16.97

A 4.1. táblázat megmutatja az eltérő $P_{ex}(t|w)$ becslések eredményeit a teszhalmazon. Az első sorban az általam abszolút baseline-ként használt ‘Berkeley Parser’ található, ami az elemző eredeti implementációja [5]. Az egyenletes eloszlással készített lexikonhoz a magyarlanc [27] morfológiai elemzőjét használtam. Az utolsó két sor a szófaji egyértelműsítés felhasználásával kapott eredményeket mutatja be. Ehhez a MarMot-ot az újsághírek tanító halmazán tanítottam, és ennek segítségével leelemeztettem a validációs halmazt, illetve – hogy tényleg nagyméretű korpuszal tudjunk dolgozni – 10 millió címkézetlen mondatot a Magyar Nemzeti Szövegtárból [28]. Az eredmények között külön beszámolok a validációs halmazon (validációs) elért eredményekről, illetve a validációs

halmazon és a nagyméretű korpuszon együttesen számolt (validációs + MNSz) gyakoriságok mellett.

Néhány előzetes kísérlet után a K értékét 7-re állítottam be.

A 4.1. táblázatból látható, hogy az ‘egyenletes eloszlás’ mellett, habár a PARSEVAL értékben nem sokat javul, a tökéletesen elemzett mondatok aránya jelentősen megnő. A ‘validációs’ konstrukció tekintélyes növekedést mutatott az ‘egyenletes eloszlással’ szemben is, ami összhangban van a Goldberg és Elhadad által megállapítottakkal. Emellett láthatjuk azt is, hogy a nagyméretű címkézetlen korpusz használata szintén jelentős javulást hozott az eredményekben. A későbbi eredmények vizsgálatához a validációs halmazra építő külső lexikont tartalmazó megvalósítást használtam. Ennek az az oka, hogy több morfológiailag gazdag nyelven is teszteltem, és a Magyar Nemzeti Szövegtárhoz hasonló, nagyméretű korpusz nem áll rendelkezésemre minden nyelven.

4.3. Morfológiai kódok automatikus összevonása

A preterminális címkék halmazának optimális megadása kritikus lehet bármely valószínűségi környezetfüggetlen nyelvtant használó konstituenselemző számára. Morfológiai jellemzők törlésével csökkenthetjük a feladat bonyolultságát, de a szintaxis számára hasznos információkat is elveszíthetünk. Például vegyük a főnevet mint szófajt, a főnév a száma szerint (morfológiai jellemző) kétféle lehet, egyes számú vagy többes számú (a jellemző értékei). A kérdés az lesz, hogy a szintaxis szempontjából az egyes számú és többes számú főnevek eltérően viselkednek-e, ha nem, akkor nem szükséges a szintaktikai elemzés során megkülönböztetnünk ezeket az eseteket.

Ebben a fejezetben bemutatok egy eljárást a preterminálisok optimális halmazának automatikus megadására és empirikus eredmények alapján, különböző baseline-okkal összehasonlítva megmutatom a hatékonyságát.

4.3.1. Eljárás morfológiai jellemzők értékeinek összevonására

A múltban már jelentek meg publikációk a morfológiai kódok számának automatikus csökkentésével kapcsolatban. Ezek egyikében Dehdari [21] bemutatott egy rendszert, melyben az egyes morfológiai jellemzőket egységként kezelte, és ezen egységek iteratív

kerültek törlésre, majd az így kapott új kódkészletet úgy értékelte ki, hogy a tanítástól kezdve újrafuttatta a konstituenselemzőt. Ezzel kapcsolatban két probléma is felmerül. Az első, hogy véleményem szerint a morfológiai jellemzőket nem szabad egységként kezelni, hiszen egy adott jellemző eltérő értékei viselkedhetnek különbözően. Vegyük például a fokot a melléknevek esetén. Itt az alap- és felsőfok hasonlóan viselkedik (vagyis összevonhatóak), amíg az előbbi két érték megkülönböztetése a középfoktól hasznos lehet a szintaktikai elemző számára, mert a középfokú melléknevek általában rendelkeznek egy vonzattal (például: *Kati szebb, mint Zsófi*), míg az alap- és felsőfok nem. A második, hogy az előbbi cikkben az egyes morfológiai jellemzők kerültek törlésre függetlenül attól, hogy milyen szófajhoz tartoznak. Így, ha az eset (Cas) jellemző törlődött, akkor törlődött a főnevek, illetve a melléknevek jellemzői közül is, pedig előfordulhat, hogy az egyes jellemzők egy adott szófaj esetén hasznosak, de más szófaj esetén törölhetők.

Ezekre a megfigyelésekre alapozva terveztem egy új módszert, ami a fő szófaji kódokból kiindulva a jellemzőket külön-külön hozzáadva iteratíván összevonja az egyes morfológiai jellemzők értékeit, miközben az eltérő szófajokhoz tartozó (azonos) jellemzőket külön kezeli. A folyamat eredménye egy csoportosítás az egyes morfológiai jellemzők lehetséges értékei felett. A megközelítésnek egy speciális esete lesz az, amikor egy morfológiai jellemzőkitörlik. Ez akkor fordulhat elő, ha az összevonás során egy morfológiai jellemző minden lehetséges értéke egy csoportba kerül, ekkor a kérdéses jellemzőnek nem lesz többé megkülönböztető szerepe. Ennek következtében erre a munkára tekinthetünk úgy, mint az előbbi módszer egy általánosítására.

Ezen általános megközelítés jelentősen megnöveli a lehetséges preterminális halmazok számát, melyek egyenkénti kiértékelése megvalósíthatatlan lenne egy külső elemző folyamatos újratanításával (a Berkeley Parserrel egy átlagos méretű korpuszon a tanítás és elemzés több mint 1 órát vesz igénybe). Javaslatom szerint nem szükséges az elemző újratanítása minden egyes preterminális halmazra. Globális célom, hogy a konstituenselemzésbeli hasznosságuk alapján lehessen válogatni az egyes halmazok között. Ez megegyezik a Berkeley Parser rejtett állapotokat összevonó eljárásának motivációjával. A Berkeley Parser miután véletlenszerűen szétbontotta a nemterminális alállapotokat, újratanítja a nyelvtant, majd minden egyes szétbontásra kiszámítja, hogy mekkora veszteséggel jár az egyes szétbontott alállapotok összevonása. Ha ez az információvesztés kicsi, a szétbontás-

sal keletkezett alállapotok nem hordoztak elég hasznos információt, ezért összevonhatjuk őket. A feladat jelenleg ugyanez, azaz meg kell találni a megfelelő összevonásokat a morfológiai jellemzők értékeire. Ennek következtében a preterminális szinten – a Berkeley Parser által létrehozott alállapotok helyett – a morfológiai jellemzők értékeire meghívjuk az előbb említett kiértékelő eljárást. Így a Berkeley Parser bináris elágazású véletlenül szétbontott hierarchiája helyett, a jelenlegi feladat alállapot-keresési tere egy háromszintes hierarchia lesz, ahol az első szinten a fő szófaji kódok, a másodikon a morfológiai jellemzők és a harmadikon az egyes jellemzők értékei találhatók. Mivel ez a hierarchia nem bináris elágazású, ezért módosítottam a Berkeley Parser idevonatkozó implementációját.

A morfológiai értékek összevonását végző eljárást az 1. algoritmus írja le.

algoritmus 1 Algoritmus a preterminális halmaz összevonására.

1. Berkeley Parser tanítása a fő szófaji kódok felhasználásával
 2. a preterminális szinten szereplő alállapotok összevonása
 3. minden szófaj - morfológiai jellemző párra
 - (a) a szófajhoz tartozó preterminális szétbontása a morfológiai jellemző értékei szerint
 - (b) a preterminális jobboldalú szabályok újraszámítása a preterminális alállapotai felett egyenletes eloszlással
 - (c) lexikai szabályok valószínűségének újraszámítása a morfológiai értékek treebanken leszámolt relatív gyakorisága alapján
 - (d) Expectation-Maximization (10 iteráció) futtatása a teljes treebanken az előző lépésben inicializált értékek megtartása mellett
 - (e) teljes gráf készítése, melyben az egyes csúcsok a megadott morfológia jellemző értékei
 - (f) a gráf minden élére kiszámítjuk a csúcsai összevonásából következő információvesztést (a Berkeley Parser összevonó eljárásával megegyező módon)
 4. az összes gráfra nézve a nagy veszteséggel járó élek eltávolítása (egy *th* paraméterrel szabályozva)
 5. az egy összefüggő komponensben szereplő morfológiai értékek összevonása
-

4.3.2. Baseline preterminális halmazok létrehozása

A javasolt megoldást négy módszerrel állítottam szembe. A két legegyszerűbb irány preterminális halmaz készítésére a fő szófaji kódok és a teljes morfológiai leírás használata. Ezen felül magyar nyelvre rendelkezésre áll egy köztes méretű kódhalmaz is, melyet a magyarlanc fejlesztésekor nyelvészeti szempontokat figyelembe véve, kézzel hoztak létre [27]. Ez a manuálisan létrehozott kódhalmaz eltérő szófaji kódok esetén eltérő morfológiai jellemzőket tartalmaz, és az összevonások benne a morfológiai értékek szintjén történtek, ami alapján nem lehet meglepő, hogy az előző szakaszban bemutatott automatikus összevonó eljáráshoz ezen korábbi kézi megvalósítás is erős inspirációként szolgált.

Az utolsó baseline a Dehdari [21] által javasolt kísérlet magyar nyelvre való megismétlése. Itt a teljes morfológiai jellemző-halmazból kiindulva mindig töröltem egy-egy jellemzőt, és az így kapott új halmazokkal újrataníttam a konstituenselemzőt. Azt tapasztaltam, hogy a leghatározottabb visszaesést a PARSEVAL statisztikában a ‘Num’ jellemző törlése okozta, míg a legenyhébbet a ‘Form’ törlése mellett kaptuk. A ‘Num’ az igéknek, illetve névszóknek a számát jelöli. A ‘Form’ kötőszavak esetén megkülönbözteti, hogy az adott szó egyszerű vagy összetett (ami egy lexikai jellemző), számok esetében pedig azt jelöli, hogy az adott szám betűkkel, arab vagy római számmal van-e leírva. Érdekes volt látni, hogy azok a törlések, amik rontottak a PARSEVAL eredményeken, javítani tudtak a hibátlanul elemzett mondatok arányán, ez valószínűleg annak a ténynek köszönhető, hogy az eltérő helyesírási változatai egy számnak (például 6 és VI) csak összezavarják az elemzőt. Emellett ott vannak az összetett kötőszavak, amelyek semmilyen módon nem kapcsolódnak egymáshoz a fában és pont úgy viselkednek, mint az egyszerű kötőszavak, így ez a megkülönböztetés szintén problémás lehet az elemző számára.

4.3.3. Eredmények különböző preterminális halmazokkal

A 4.2. táblázat összesítve tartalmazza a baseline módszerekkel és az automatikus összevonó megoldás által megkapott preterminális halmazokkal mért eredményeket. Az összevonó algoritmussal két különböző címkehalmazt is megadtam, melyek eltérő küszöbérték (*th*) mellett lettek összevonva.

4.2. táblázat. PARSEVAL eredmények és a hibátlanul elemzett mondatok aránya (EX) eltérő preterminális halmazok mellett.

	#pt	PARSEVAL	EX
fő szófaji kód	16	82.36	5.52
manuális	72	85.38	9.23
teljes	680	88.29	15.22
teljes - Num	479	87.43	14.49
teljes - Form	635	88.24	15.73
összevont ($th = 0.5$)	378	88.36	15.92
összevont ($th = 0.1$)	642	88.52	15.44

A fő szófaji kódok és a teljes morfológiai leírás közötti különbség meglepően magas, ebből következik, hogy a preterminálisok által hordozott morfológiai információk rendkívül hasznosak a konstituenselemző számára, és hogy a Berkeley Parser képes sok száz elemű preterminális halmazok kezelésére. Magyarra azt találtam, hogy az egyes jellemzők teljes eltávolításától az eredmények nem javulnak. Ez a felfedezés szögesen ellentmond Dehdari [21] arab nyelvre tett megfigyeléseivel, ahol a ‘Case’ eltávolításától a PARSEVAL eredmény 1%-kal lett jobb. Megfigyelhető, hogy a baseline eredmények is teljesen eltérnek a két nyelv között, míg magyarra a teljes morfológiai leírás sokkal eredményesebbnek bizonyult a fő szófaji kódoknál, addig ugyanez a két érték arabra Dehdari eredményei alapján közel azonos volt.

A táblázat szintén tartalmazza az általam tervezett eljárás két eredményét. A $th = 0.1$ esetben csak néhány morfológiai jellemző érték került összevonásra, és ez enyhe javulást eredményezett a teljes kódhalmazt tartalmazó konfigurációval szemben. A másik esetben, ahol a th értéke 0.5, közel azonos eredményt kaptam a teljes morfológiai leírással, miközben feleannyi preterminálist használtam (ráadásul a hibátlanul elemzett mondatok aránya releváns javulást mutatott). Így, bár statisztikailag nem lett jobb az eredmény, mint a legjobb baseline esetében, de az elemzés futási ideje majdnem a felére csökkent.

Összességében az összevonó megoldásom a teljes morfológiai leírásnál jobb preterminális halmazokat talált meg, melyek az új címkék számától függően javítottak az eredményeken vagy gyorsították az elemzést.

4.4. Morfológián alapuló jellemzők használata az eredmények újrarangsorolására

Újrarangsoroló alatt általában egy olyan rendszert értünk, ami gazdag jellemzőkészlet segítségével az első elemző által megadott n legjobb kimenethez új sorrendet állít elő. Az első elemző gyors, a lehetséges megoldások teljes teréből dolgozik, kiválasztja a legjobb megoldások egy halmazát. A újrarangsoroló több millió jellemzőt használ, ennek következtében lassú, de a részletesebb információknak köszönhetően az első elemző által adott n legjobb megoldást át tudja rangsorolni.

Újrarangsoroló rendszereket [29, 4] a múltban már többször használtak konstituenselemzők hatékonyságának javítására.

A továbbiakban olyan jellemzősablonokat adok meg, melyek morfológiai információkon alapulnak, illetve megvizsgálom ezek hatását az eredetileg angolra fejlesztett jellemzőkészletekkel szemben. Baseline-ként a Charniak [4] és Yannick [30] által megadott jellemzőkészleteket (kivéve a külső korpuszra építő jellemzőket) használtam.

A kísérletekhez minden esetben a Berkeley Parser által adott legjobb 50 elemzést használtam, és az újrarangsorolóhoz szükséges tanító adatok előállításához 5-szörös keresztvalidációt alkalmaztam.

A kiindulásként használt jellemzősablon a preterminálisokat atomi jellemzőkként tekintti. Első lépésként megvizsgáltam, hogy melyik preterminális halmaz a legideálisabb a baseline jellemzőkészlet számára. A kísérletekhez már az összevont morfológiai kódokat alkalmazó Berkeley Parsertől kértem le a legjobb elemzéseket, így kétféle morfológiai kódhalmazt tudtam a jellemzők közé felvenni. Egyik lehetőségként az automatikusan összevont kódok halmazát használtam (összevont) a másik opció a csak fő szófaji kódokat (fő szófaj) tartalmazó jellemzőhalmaz volt. A 4.3. táblázatból látszik, hogy a fő szófaji kódok alkalmazása felülmúlta a teljes összevont kódokkal épített rendszert. Ez valószínűleg a kevés adat problémájának a következménye.

Ezen megfigyelésekre alapozva a fő szófaji kódokat használtam a baseline jellemzőkészlet elemeiként és a morfológiai elemzésből megszerzett információkra alapozva új jellemzősablonokat készítettem az alábbiak szerint.

Az egyes preterminálisoknál az összes morfológiai jellemzőhöz tartozó értékére fel-

vettem egy olyan jellemzőt, ami a preterminálishoz tartozó szóból és a morfológiai jellemző értékéből állt. Például, ha a preterminális $N\#Cas=n|Num=s\#$ és a hozzá tartozó szó a *macska*, akkor két új jellemző lesz: $macska-N-Cas=n$ és $macska-N-Num=s$.

Hasonlóan a frázisok esetén is felvettem új jellemzőket, úgy, hogy a frázis mellett annak a fejéhez tartozó preterminális morfológiai jellemzőinek az értékeit adtam meg. Például az $(NP (A\#SubPOS=f|Deg=p\# fekete) (N\#Cas=n|Num=s\# macska))$ frázisban, ahol az NP feje a 'macska', az alábbi jellemzők kerültek felvételre: $NP-N-Cas=n$, $NP-N-Num=s$. (Itt a fejek meghatározására egyszerű, általános szabályokat alkalmaztam. Nyelvspecifikus szabályok bevezetésével még lehetne javítani a fejek meghatározásának pontosságán.)

Minden egyes frázis esetén, ha a fejet tartalmazó ágon kívül van más gyereke is a frázisnak, akkor felvettem a frázis fejében, illetve a gyerek fejében található morfológiai jellemzőkhöz tartozó értékek kombinációit is. Azaz az előbbi $(NP (A\#SubPOS=f|Deg=p\# fekete) (N\#Cas=n|Num=s\# macska))$ példa esetén négy új jellemző lesz: $NP-Cas=n-A-SubPOS=f$, $NP-Cas=n-A-Deg=p$, $NP-Num=s-A-SubPOS=f$, $NP-Num=s-A-Deg=p$.

Ezzel analóg módon a fejek és azok testvéreik közötti kapcsolat alapján is adtam hozzá új jellemzőket a modellhez.

Az első két jellemzősablon lehetővé teszi az újrarangsorolónak, hogy információkat gyűjtsön a nyelv gazdag morfológiájáról lexikai és konstituens szinten egyaránt. Míg az utolsó két sablon motivációja a morfológiailag gazdag nyelvek szabad szórendjében rejlő összefüggések megtalálása volt.

A 4.3. táblázatban látható a jellemzősablonok (kiterjesztett) által elért növekedés a fő szófaji kódokhoz képest. A hibátlanul elemzett mondatok arányában ismét határozott különbség tapasztalható.

Bár a morfológiai adatok szintaktikai elemzésben történő alkalmazásának ötlete korábbi tanulmányokra épül, például a Bohnet parser [31] – ami jelenleg az egyik leghatékonyabb eszköz függőségi fák elemzésére – az általam megvalósítottéhoz hasonló jellemzősablonokat alkalmaz, de információim szerint ez az első olyan kísérlet, amely a morfológiai adatokat mint atomi jellemzőket használt fel konstituensfák elemzése közben.

	PARSEVAL	EX
Újrarangsoroló (összevont)	89.05	18.45
Újrarangsoroló (fő szófaji)	89.33	18.64
Újrarangsoroló (kiterjesztett)	89.47	20.35

4.3. táblázat. Az újrarangsorolással – beleértve a morfológiai alapokon nyugvó jellemző bővítést – elért eredmények a magyar validációs halmazon.

4.5. A teljes rendszer teljesítménye

4.5.1. Korpuszok

A teljes rendszer vizsgálatához a 2013-as SPMRL (Statistical Parsing of Morphologically Rich Languages) elnevezésű verseny korpuszait használtam. Az eredeti tanító és teszt felosztás meghagyása mellett a baszk [32], a francia [33], a héber [34], a magyar [8] és a német [35] korpuszokon vizsgáltam az eredményeket. A 4.4. táblázat a korpuszok alapvető statisztikáit mutatja. A magyar korpusz megegyezik az eddigiekben használttal, de a validációs halmaz helyett a teszhalmazon végeztem el a méréseket.

	Baszk	Francia	Német	Héber	Magyar
mondatok száma a tanító halmazban	7577	14759	40472	5000	8146
mondatok száma a teszhalmazban	946	2541	5000	716	1009
tokenek átlagos száma mondatonként	12.92	30.13	17.51	25.33	21.76
nemterminális címkék száma	3000	770	994	1196	890
fő szófaji kódok száma	16	33	54	46	16

4.4. táblázat. A korpuszok alapvető statisztikái.

4.5.2. Eredmények

Az egyes módszerek fejlesztése után arra koncentráltam, hogy a teljes rendszer hatékonyságát lépésről lépésre lemérjem több morfológiailag gazdag nyelv esetén is. A 4.5. táblázat összesítve tartalmazza a végső eredményeket. Kiindulópontként a módosíthatlan Berkeley Parsert vettem a teljes morfológiai leírást tartalmazó preterminális réteg használata mellett (Berkeley Parser). Ezek után a lexikai modellt gazdagítottam a teszhalmaz automatikus szófaji egyértelműsítéséből származó gyakoriságokkal (+ Külső lexikon). A harmadik lépésben lecseréltem a teljes morfológiai leírást az automatikus preterminális

	Baszk	Francia	Német	Héber	Magyar
Berkeley Parser	79.21/19.03	79.53/18.46	74.77/26.56	87.87/14.53	88.22/26.96
+ Külső lexikon	82.02/25.69	78.91/17.87	75.64/28.36	88.53/13.69	89.09/26.76
+ Morf. összevonó	83.19/24.74	79.53/18.58	77.12/ 30.02	88.07/13.83	89.15/28.05
+ Újrarangsoroló	83.81/25.66	80.31/18.91	77.78 /29.80	88.38/15.12	89.57/30.23
+ Morf. jellemzők	84.03/26.28	80.41/20.07	77.74/29.23	88.55/15.24	89.91/30.55

4.5. táblázat. PARSEVAL érték és mondat szintű teljes egyezés a teszhalmazokon vizsgálva.

összevonó eljárás eredményére (+ Morf. összevonó). Az egyes nyelvek esetén az összevonáshoz használt küszöbértéket a validációs halmazon mérték szerint állítottam be. Az utolsó két sor pedig az 50 legjobb elemzést felhasználó újrarangsoroló felhasználásával kapott eredményeket tartalmazza. Az előbbi sorban az eredeti jellemzőkészlettel (+ Újrarangsoroló), míg az utolsó sorban a morfológiai információkkal kiterjesztett jellemzőkkel (+ Morf. jellemzők) mért eredmények szerepelnek.

A kiterjesztett lexikai modell használatával a legnagyobb különbség a baszk nyelv esetén volt tapasztalható (körülbelül 2,8 százalékpont), ezen felül releváns javulás volt tapasztalható a magyar, a német és a héber nyelv esetén is, miközben a francia korpuszon enyhén romlottak az eredmények.

Bár az újrarangsoroló eljárás minden nyelv esetén javított az eredményeken (átlagosan 0,5 százalékpontot), ez a javulás elmarad az angol esetében tapasztaltaktól (1,5 százalékpont). Ez valószínűleg annak köszönhető, hogy a morfológiailag gazdag nyelvekben nagyszámú eltérő szóalak található, miközben az egyes jellemzősablonok általában ezekre a szóalakokra épülnek, aminek következtében ismét előjön a kevés adat problémája. A morfológiai információkon alapuló új jellemzők enyhe növekedés hoztak 4 nyelv esetén, de úgy gondolom, hogy ezen a területen még maradtak bőven kiaknázható lehetőségek.

5. fejezet

Összegzés

Dolgozatomban bemutatam a konstituenselemzésre legáltalánosabban használt eljárásokat, azok működését, és az azokhoz szükséges erőforrásokat. Ezenfelül három olyan módszert [23] mutattam be, amelyek az elemzés javítására szolgálnak morfológiailag gazdag nyelveken.

Fő eredményem a preterminális összevonó eljárás, ami az előző munkáknál egy általánosabb és gyorsabb megoldást ad köszönhetően annak, hogy nincs szükség a konstituenselemző lehetséges preterminális halmazonkénti újratanítására. Az összevonások a legalacsonyabb szinten, a morfológiai jellemzők értékének a szintjén történnek, ami segítségével az eddigi megoldásoknál mélyebb kapcsolatok feltárására alkalmas. Az így kapott eljárásnak köszönhetően javítani tudtam az elemzés pontosságán és sebességén is.

Kísérleteztem külső korpuszok felhasználásával is a lexikai modellben. Megmutattam, hogy felügyelt szófaji egyértelműsítés használatával jelentős javulást lehet elérni a rendszer pontosságában, amit külső annotálatlan nyelvi erőforrás segítségével még tovább lehet növelni.

Ezen felül az elemzések további javítására felhasználtam egy újrarangsoroló rendszert, aminek a hatékonyságát tovább javítottam a jellemzők közé felvett morfológiai információk segítségével.

A különböző eljárások összeillesztésével egy olyan rendszert kaptam, amely relevánsan javít az általam vizsgált morfológiailag gazdag nyelveken történő elemzés hatékonyságán.

Köszönetnyilvánítás

Jelen kutatási eredmények megjelenését „Telemedicina fókuszú kutatások Orvosi, Matematikai és Informatikai tudományterületeken” című, TÁMOP-4.2.2.A-11/1/KONV-2012-0073 számú projekt támogatja. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Irodalomjegyzék

- [1] Fraser, A., Schmid, H., Farkas, R., Wang, R., Schütze, H.: Knowledge Sources for Constituent Parsing of German, a Morphologically Rich and Less-Configurational Language. *Computational Linguistics* **39**(1) (2013) 57–85
- [2] Tsarfaty, R., Seddah, D., Kübler, S., Nivre, J.: Parsing Morphologically Rich Languages: Introduction to the Special Issue. *Computational Linguistics* **39**(1) (2013) 15–22
- [3] Charniak, E.: A maximum-entropy-inspired parser. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. (2000) 132–139
- [4] Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. *ACL '05* (2005) 173–180
- [5] Petrov, S., Barrett, L., Thibaux, R., Klein, D.: Learning Accurate, Compact, and Interpretable Tree Annotation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. (2006) 433–440
- [6] Huang, L.: Forest Reranking: Discriminative Parsing with Non-Local Features. In: Proceedings of ACL-08: HLT. (2008) 586–594
- [7] Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of English: the penn treebank. *Computational Linguistics* **19**(2) (1993) 313–330

- [8] Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: TSD. (2005) 123–131
- [9] Iván, S., Ormándi, R., Kocsor, A.: Magyar mondatok SVM alapú szintaxis elemzése. In: V. Magyar Számítógépes Nyelvészeti Konferencia. (2007) 281–283
- [10] Goldberg, Y., Elhadad, M.: Word Segmentation, Unknown-word Resolution, and Morphological Agreement in a Hebrew Parsing System. *Computational Linguistics* **39**(1) (2013) 121–160
- [11] Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J.D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galleitebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Marton, Y., Nivre, J., Przepiórkowski, A., Roth, R., Secker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A.: Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In: *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*. (2013) 146–182
- [12] Csendes, D., Hatvani, C., Alexin, Z., Csirik, J., Gyimóthy, T., Prószy, G., Kocsor, A.: Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. In: *Magyar Számítógépes Nyelvészeti Konferencia*. (2003) 238–247
- [13] Johnson, M.: PCFG models of linguistic tree representations. *Computational Linguistics* **24**(4) (1998) 613–632
- [14] Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. (2003) 423–430
- [15] Matsuzaki, T., Miyao, Y., Tsujii, J.: Probabilistic CFG with latent annotations. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. (2005) 75–82
- [16] Baker, J.: Trainable grammars for speech recognition. In: *97th meeting of the Acoustical Society of America*. (1979) 547–550
- [17] Petrov, S.: Coarse-to-Fine Natural Language Processing. PhD thesis, University of California at Berkeley, Berkeley, CA, USA (2009)

- [18] Choi, K.S., Han, Y.S., Han, Y.G., Kwon, O.W.: KAIST tree bank project for Korean: Present and future development. In: Proceedings of the International Workshop on Sharable Natural Language Resources, Citeseer (1994) 7–14
- [19] Le Roux, J., Sagot, B., Seddah, D.: Statistical Parsing of Spanish and Data Driven Lemmatization. In: Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages. (2012) 55–61
- [20] Marton, Y., Habash, N., Rambow, O.: Improving Arabic Dependency Parsing with Lexical and Inflectional Morphological Features. In: Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages. (2010) 13–21
- [21] Dehdari, J., Tounsi, L., van Genabith, J.: Morphological Features for Parsing Morphologically-rich Languages: A Case of Arabic. In: Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages, Dublin, Ireland, Association for Computational Linguistics (2011) 12–21
- [22] Chen, X., Kit, C.: Higher-order Constituent Parsing and Parser Combination. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). (2012) 1–5
- [23] Szántó, Z., Farkas, R.: Special Techniques for Constituent Parsing of Morphologically Rich Languages. In: 14th Conference of the European Chapter of the Association for Computational Linguistics. (2014) 135–144
- [24] Abney, S., Flickenger, S., Gdaniec, C., Grishman, C., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., Strzalkowski, T.: Procedure for quantitatively comparing the syntactic coverage of English grammars. In Black, E., ed.: Proceedings of the workshop on Speech and Natural Language. (1991) 306–311

- [25] Goldberg, Y., Adler, M., Elhadad, M.: EM Can Find Pretty Good HMM POS-Taggers (When Given a Good Start). In: Proceedings of ACL-08: HLT. (2008) 746–754
- [26] Mueller, T., Schmid, H., Schütze, H.: Efficient Higher-Order CRFs for Morphological Tagging. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. (2013) 322–332
- [27] Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP. (2013)
- [28] Váradi, T.: The Hungarian National Corpus. In: In Proceedings of the Second International Conference on Language Resources and Evaluation. (2002) 385–389
- [29] Collins, M.: Discriminative Reranking for Natural Language Parsing. In: Proceedings of the Seventeenth International Conference on Machine Learning. ICML '00 (2000) 175–182
- [30] Versley, Y., Rehbein, I.: Scalable Discriminative Parsing for German. In: Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09). (2009) 134–137
- [31] Bohnet, B.: Top Accuracy and Fast Dependency Parsing is not a Contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). (2010) 89–97
- [32] Aduriz, I., Aranzabe, M.J., Arriola, J.M., Atutxa, A., Díaz de Ilarraza, A., Garmendia, A., Oronoz, M.: Construction of a Basque Dependency Treebank. In: TLT-03. (2003) 201–204
- [33] Abeillé, A., Clément, L., Toussanel, F.: Building a Treebank for French. In Abeillé, A., ed.: Treebanks. Kluwer, Dordrecht (2003)
- [34] Sima'an, K., Itai, A., Winter, Y., Altman, A., Nativ, N.: Building a Tree-Bank for Modern Hebrew Text. In: Traitement Automatique des Langues. (2001)

- [35] Brants, S., Dipper, S., Hansen, S., Lezius, W., Smith, G.: The TIGER Treebank. In Hinrichs, E., Simov, K., eds.: Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002). (2002) 24–41