Félig kompozicionális szerkezetek automatikus azonosítása magyar és angol nyelven

Vincze Veronika[1], Nagy T. István[2], Zsibrita János[2]

[1]Magyar Tudományos Akadémia, Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103., e-mail:vinczev@inf.u-szeged.hu
[2]Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport
Szeged, Árpád tér 2., e-mail:{nistvan,zsibrita}@inf.u-szeged.hu

Kivonat Jelen munkában bemutatjuk szabályalapú és gépi tanult módszereken alapuló megközelítéseinket, melyek mind angol, mind magyar nyelven képesek a félig kompozicionális szerkezetek folyó szövegben történő automatikus azonosítására. Eredményeink azt igazolják, hogy a sekély morfológiai elemzésre épülő módszereink mellett a szintaktikai információ is nagyban képes segíteni a félig kompozicionális szerkezetek automatikus azonosítását. Cikkünkben kitérünk a feladat angol és magyar nyelvű sajátságaira is.

Kulcsszavak: többszavas kifejezések, lexikális szemantika, többnyelvűség, FXtagger

1. Bevezetés

A természetes nyelvi feldolgozásban, különösen a gépi fordítás és fordítástámogatás területén az egyik legnehezebb problémát a többszavas kifejezések megfelelő kezelése jelenti. A többszavas kifejezések sikeres kezelésének első lépése, hogy felismerjük őket a folyó szövegben. Ebben a munkában a többszavas kifejezések egy altípusának, a félig kompozicionális szerkezeteknek automatikus felismerésére koncentrálunk.

A félig kompozicionális szerkezetek (FX-ek) olyan, főnévből és igéből álló többszavas kifejezések, ahol a szemantikai fej a főnév, míg az ige pusztán csak a szerkezet igeiségéért felel. Mivel jelentésük nem teljesen kompozicionális, a szerkezet elemeinek egyenkénti lefordítása nem (vagy csak nagyon ritkán) eredményezi a szerkezet idegen nyelvű megfelelőjét. Emellett a félig kompozicionális szerkezetek (*választ kap*) szintaktikailag hasonló felépítéssel bírnak, mint más, produktív (kompozicionális) szerkezetek (*pulóvert kap*), illetve idiómák (*vérszemet kap*) [1], így azonosításuk nem valósulhat meg pusztán szintaktikai mintákat figyelembe véve. Végül, mivel a szerkezet szintaktikai és szemantikai feje nem azonos, a szerkezet nyelvi elemzésekor célszerű a főnevet és az igét egy komplex egységként kezelni - az angol vonzatos igékhez (phrasal verbs) hasonlóan.

A fenti okokból kifolyólag a félig kompozicionális szerkezetek kezelése különleges figyelmet érdemel a természetes nyelvi alkalmazásokban. Ennek első lépéseként azonosítani kell őket, mely célhoz különféle algoritmusok fejlesztése segíthet hozzá. Ennek megfelelően először szabályalapú megközelítéseket definiálunk, majd ezek eredményeire alapozva gépi tanuló módszerek segítségével is azonosítjuk a félig kompozicionális szerkezeteket.

2. Kapcsolódó munkák

A félig kompozicionális szerkezetek automatikus azonosítására, illetve a főnév + ige szerkezetek osztályokba sorolására már több szerző is kísérletet tett.

Van de Cruys és Moirón [2] szemantikai alapokon nyugvó rendszere igeprepozíció-főnév kombinációkat azonosít holland szövegekben. Módszerük az ige és a főnév szelekciós megkötéseire épül, illetve az igével együtt előforduló főnevek szemantikai osztályát is figyelembe veszik.

Cook és munkatársai [3] angol ige + főnév szerkezetek szó szerinti és idiomatikus használatát különítik el egymástól. Feltevésük szerint idiomatikus használatban főként a szerkezet szótári alakja fordul elő, míg szó szerinti használatban a szerkezet nagyobb szintaktikai változatosságot mutat. A szerkezet szintaktikai rögzítettségét kihasználó felügyelet nélküli osztályozó módszerük 72%-os eredményt ér el.

Bannard [4] szintén angol nyelvű ige + főnév szerkezeteket osztályoz szintaktikai rögzítettségük alapján. Az általa használt jellemzők közé tartozik a főnév névelőzhetősége, módosíthatósága, a szerkezet szenvedő szerkezetben való előfordulása stb.

Samardžić és Merlo [5] angol-német párhuzamos korpuszokban előforduló félig kompozicionális szerkezeteket vizsgálnak. Eredményeik szerint a szerkezetek párhuzamosításánál különösen nagy szerepet játszanak a gyakorisági adatok mellett a szerkezetek nyelvi jellemzői is, például a kompozicionalitás foka.

Gurrutxaga és Alegria [6] baszk nyelvű szövegekből nyernek ki idiomatikus és félig kompozicionális főnév + ige szerkezeteket statisztikai módszerek segítségével. Mivel a baszk szabad szórendű nyelv, azzal az előzetes feltételezéssel éltek, hogy az ige tágabb környezetét nézve javulni fognak az eredmények, azonban kísérleteik ezt nem támasztották alá.

Tu és Roth [7] ige + főnév párokat osztályoznak aszerint, hogy félig kompozicionális szerkezetek-e vagy sem. Mind környezeti, mind statisztikai jellemzőkkel dolgoznak, és megállapításuk szerint a többértelmű példákon a lokális környezeti jellemzők használata vezet a legjobb eredményhez.

Sass Bálint [8] beszámol egy igei szerkezetek párhuzamos korpuszból való kinyerésére szolgáló eljárásról, mely egy korábbi, igéket és azok bővítményeit kinyerő algoritmusra épül. A módszer lényege, hogy a tagmondatok igéit egymás mellé rendelve egy komplex ige jön létre, melyhez a bővítményeket halmazként rendeljük hozzá, felcímkézve őket aszerint, hogy melyik nyelvű részkorpuszból származnak. Az így kapott reprezentációból az eredeti algoritmus segítségével lehet kigyűjteni az egyes nyelvekre jellemző igei szerkezeteket.

A félig kompozicionális szerkezetek automatikus azonosítását célzó módszerek nagy része kiindulási alapnak tekinti a szintaxist, azaz általában ige-tárgy párokat osztályoznak [3,4,9,7]. Ezzel szemben mi nem a szintaktikai mintázatok alapján megszűrt FX-jelölteket szeretnénk osztályozni, hanem folyó szövegben szeretnénk azonosítani őket, nem feltétlenül szintaktikai információk segítségével. Kísérleteink közben azonban kiemelt figyelmet szentelünk a szintaktikai információk hozzáadott értékének.

3. A félig kompozicionális szerkezetek automatikus felismerése

A félig kompozicionális szerkezetek automatikus azonosítására szabályalapú és gépi tanulási módszereket is definiáltunk. Angol és magyar nyelvre alapjában véve ugyanazokat az eljárásokat alkalmaztuk, természetesen figyelembe véve az adott nyelv sajátosságait.

Módszereink kiértékeléséhez három korpuszt használtunk. A SzegedParalellFX párhuzamos korpusz [10] angol és magyar nyelven ugyanazokat a szövegeket tartalmazza, melyekben összesen 1100 angol nyelvű és 1112 magyar nyelvű FX található. A Szeged Korpuszban szintén be vannak jelölve a félig kompozicionális szerkezetek [11]. Kísérleteinkhez a sajtónyelvi részkorpuszokat használtuk. Az angol nyelvű Wiki50 korpuszban [1] többszavas kifejezések és névelemek vannak annotálva, így a félig kompozicionális szerkezetek is be vannak jelölve. Noha a korpuszokban a félig kompozicionális szerkezetek melléknévi igenévi és főnévi alakjai is be vannak jelölve, jelen munkánkban csak az igei alakok felismerésére koncentrálunk. A felhasznált korpuszok adatait az 1. táblázat mutatja.

1. táblázat. A felhasznált korpuszok adatai

Korpusz	**Mondat**	**Token**	**Igei FX**
Wiki50 (angol)	4.350	114.570	368
SzegedParallelFX (angol)	14.262	298.948	745
SzegedParallelFX (magyar)	14.262	240.399	753
Szeged Treebank (újságcikkek - magyar)	10.210	182.172	458

3.1. Szabályalapú módszerek

Számos szabályt fogalmaztunk meg a félig kompozicionális szerkezetek automatikus azonosítására. Az angol nyelvű szövegeket a Stanford elemzési lánc segítségével tokenizáltuk, majd elemeztük szófajilag [12] és szintaktikailag [13]. A SzegedParalellFX magyar nyelvű szövegeit a `magyarlanc` [14] csomaggal tokenizáltuk és elemeztük szófajilag. A Szeged Korpuszból származó szövegek esetén az etalon szófaji és dependenciaelemzésekre hagyatkoztunk, illetve az összevethetőség kedvéért a `magyarlanc` által nyújtott szófaji elemzésekkel is végeztünk kísérleteket.

A **POS-szabályok** („POS”) módszer esetében különféle szófaji mintákat adtunk meg, például VB.? (NN|NNS) angolra vagy N V a magyarra. Amennyiben ezek illeszkedtek a szöveg egy szegmensére, azt megjelöltük mint félig kompozicionális szerkezetet. Mivel további módszereink morfológiai információkra épülnek, pontosabban az ige vagy a főnév természetére tesznek megszorításokat, a POS-szabályokra való illeszkedés előfeltétele a többi módszer alkalmazhatóságának.

A **végződés** („vég”) módszer alapja, hogy az FX-ek főnévi komponense legtöbbször igéből képzett főnév. Ebben az esetben azokat az FX-jelölteket fogadtuk el, amelyekre illeszkedett egy szófaji minta, és a főnév az előre definiált n-gramok (képzők) egyikében végződött.

A **leggyakoribb ige** („ige”) módszer azon megfigyelésen alapszik, hogy általában a leggyakoribb igék szerepelnek funkcióigeként (az angolban a *do*, *make*, *take* stb., míg a magyarban *ad*, *vesz*, *hoz* stb.). így azokat az FX-jelölteket fogadtuk el, amelyek illeszkedtek a szófaji mintákra, és az igei komponens lemmája megegyezett az előre megadott leggyakoribb igék egyikével.

A **szótő** („tő”) módszer a főnév szótövét vizsgálja. Mint fentebb említettük, a főnévi komponens igen gyakran igéből származik, így az angolban azt néztük meg a Porter stemmert használva [15], hogy a főnév szótöve egybeesik-e egy igei szótővel (*to make a decision - to decide*) vagy maga a főnév egybeesik-e egy igével (*to have a walk - to walk*). A magyarban pedig a hunmorph elemző [16] segítségével állapítottuk meg a főnév szótövét, és vizsgáltuk meg, hogy annak van-e igei elemzése.

A félig kompozicionális szerkezetek azonosításában a szintaktikai információk is hasznosak lehetnek. Az angolban a szerkezet két tagja között általában `dobj` vagy `prep` viszony szerepel (tárgyi vagy prepozíciós vonzat esetében), míg a magyarban `obj` vagy `obl` (tárgy vagy egyéb argumentum). A **szintaxis** módszert alkalmazva azokat az FX-jelölteket fogadtuk el, amelyek tagjai a fenti relációk egyikében álltak egymással.

A fenti módszereket kombináltuk is egymással: vagyis vettük a különféle módszerek unióját $\cup$ (egy potenciális FX jelölt abban az esetben került elfogadásra, amennyiben legalább az egyik módszer elfogadta azt), és a metszetüket $\cap$ (csak akkor jelöltünk szóösszetételt FX-nek, amennyiben minden szabály elfogadta azt). Eredményeinket a 2. táblázat szemlélteti.

3.2. A szabályalapú módszerek eredményei

A 3. táblázat mutatja a szabályalapú módszereink eredményét a négy felhasznált korpuszon. Jól látszik, hogy három korpusz esetében a leggyakoribb ige módszer bizonyul a legsikeresebbnek, jóval magasabb F-mértéket ér el, mint a többi módszer vagy azok kombinációi. Az egyetlen kivételt a SzegedParalellFX angol állománya jelenti, ahol is az ige és tő módszerek metszete a legeredményesebb. Ez valószínűleg annak köszönhető, hogy a korpuszban nagy arányban fordulnak elő tipikus főnév + tipikus ige kombinációk. A végződés jellemző a SzegedParalellFX-en bizonyul hasznos információnak, a másik két korpuszon önmagában még ront

2. táblázat. Szabályalapú megközelítések eredményei, fedés/pontosság/F-mérték.

Megközelítés	Wiki50			ParalellFX angol			ParalellFX magyar			Szeged Treebank		
POS	77,14	6,32	11,68	79,40	5,07	9,52	65,55	7,67	13,74	74,56	5,75	10,69
Vég	17,14	9,47	12,20	15,24	10,5	12,43	21,45	12,79	16,02	19,30	6,53	9,76
Ige	55,24	34,32	42,34	54,56	28,81	37,73	43,83	30,19	35,76	58,77	24,28	34,36
Tő	54,29	7,72	14,64	61,55	7,66	13,62	21,05	16,14	18,27	16,67	7,85	10,67
Vég ∩ Ige	9,52	43,48	15,64	10,24	48,31	16,90	15,15	40,36	22,03	18,42	32,81	23,60
Vég ∪ Ige	62,86	19,64	29,93	59,64	19,02	28,84	50,13	18,21	26,71	59,65	12,39	20,51
Vég ∩ Tő	14,29	10,79	12,30	11,07	11,14	11,10	19,30	16,31	17,68	15,79	8,37	10,94
Vég ∪ Tő	57,14	7,60	13,42	65,71	7,74	13,84	23,19	12,90	16,58	20,18	6,32	9,62
Ige ∩ Tő	40,95	42,57	41,75	43,45	38,87	41,03	15,01	46,09	22,65	16,67	35,19	22,62
Ige ∪ Tő	68,57	8,93	15,81	72,74	8,25	14,82	49,87	20,52	29,07	58,77	14,44	23,18
Vég ∩ Ige ∩ Tő	8,57	52,94	14,75	7,62	47,41	13,13	13,67	46,36	21,12	15,79	39,13	22,50
Vég ∪ Ige ∪ Tő	70,48	8,70	15,48	74,29	8,05	14,53	50,54	17,77	26,30	59,65	11,97	19,94

is az eredményeken, viszont kiegészítve a leggyakoribb ige jellemzővel már mindenütt javít a rendszer teljesítményén. A szótő jellemző pedig a Szeged Korpusz kivételével mindenhol javulást eredményezett: feltehetőleg arányaiban kevesebb a tipikus (igéből képzett) főnévi komponenst tartalmazó félig kompozicionális szerkezet ebben a korpuszban, mint a többiben.

Míg a leggyakoribb ige az igei komponensre, a szótő és végződés pedig a főnévi komponensre tesz megszorításokat. Így a módszerek uniója a fedésre van jó hatással, hiszen a nem tipikus főnév + tipikus ige és a tipikus főnév + nem tipikus ige párokat egyaránt meg lehet találni. A módszerek metszete pedig a pontosságot javítja, hiszen így csak a tipikus főnév + tipikus ige párokat találjuk meg.

3. táblázat. Szabályalapú megközelítések eredményei a Szeged Treebanken, fedés/pontosság/F-mérték.

Megközelítés	pred. POS			etalon POS			pred. POS + szint.			etalon POS + szint.		
POS	74,56	5,75	10,69	84,21	6,70	12,41	76,32	6,92	12,69	85,09	7,77	14,23
Vég	19,30	6,53	9,76	21,93	7,35	11,01	19,30	7,64	10,95	21,93	8,56	12,32
Ige	58,77	24,28	34,36	69,30	28,11	40,00	60,53	26,44	36,80	70,18	29,20	41,24
Tő	16,67	7,85	10,67	20,18	9,35	12,78	16,67	9,00	11,69	20,18	10,80	14,07
Vég ∩ Ige	18,42	32,81	23,60	20,18	35,38	25,70	18,42	35,00	24,14	20,18	35,94	25,84
Vég ∪ Ige	59,65	12,39	20,51	71,05	14,57	24,18	61,40	14,31	23,22	71,93	16,33	26,62
Vég ∩ Tő	15,79	8,37	10,94	18,42	9,55	12,57	15,79	9,68	12,00	18,42	11,11	13,86
Vég ∪ Tő	20,18	6,32	9,62	23,68	7,38	11,25	20,18	7,35	10,77	23,68	8,54	12,56
Ige ∩ Tő	16,67	35,19	22,62	19,30	38,60	25,73	16,67	38,00	23,17	19,30	40,00	26,04
Ige ∪ Tő	58,77	14,44	23,18	70,18	17,02	27,40	60,53	16,35	25,75	71,05	18,75	29,67
Vég ∩ Ige ∩ Tő	15,79	39,13	22,50	17,54	41,67	24,69	15,79	41,86	22,93	17,54	42,55	24,84
Vég ∪ Ige ∪ Tő	59,65	11,97	19,94	71,05	14,14	23,58	61,40	13,81	22,54	71,93	15,83	25,95

A Szeged Korpusz etalon szófaji annotációja lehetővé tette azt is, hogy összevethessük a `magyarlanc` által elemzett és az etalon szófaji kódokat tartalmazó szövegeken a szabályalapú módszerek teljesítményét. Az eredményeket a 3. táblázat első két oszlopa mutatja. Egyértelműen kiderül, hogy jobb eredményeket lehet elérni, ha az etalon kézi címkéket használjuk, hiszen így a szófaji egyértelműsítés hibái kiküszöbölődnek. Különösen látványos javulás érhető el a leg-

gyakoribb ige jellemző esetében, ami valószínűleg arra vezethető vissza, hogy a `magyarlanc` gyakran minősíti hibásan melléknévnek a múlt idejű igéket (amelyek homonímek az ige befejezett melléknévi igenévi alakjával), például *adott*. Az etalon címkék használata átlagosan 2,75% javulást eredményezett az F-mértékben.

4. táblázat. Szabályalapú megközelítések eredményei szintaktikai információval (fedés/pontosság/F-mérték).

Megközelítés	**Wiki50**			**ParalellFX**			**angol Szeged Treebank**		
POS	73,33	8,85	15,79	72,98	6,89	12,59	76,32	6,92	12,69
Vég	15,24	11,03	12,80	14,52	12,82	13,62	19,30	7,64	10,95
Ige	53,33	42,11	47,06	51,19	34,82	41,45	60,53	26,44	36,80
Tő	51,43	10,87	17,94	56,19	10,16	17,21	16,67	9,00	11,69
Vég ∩ Ige	7,62	38,10	12,70	9,76	55,03	16,58	18,42	35,00	24,14
Vég ∪ Ige	60,95	24,90	35,36	55,95	23,06	32,66	61,40	14,31	23,22
Vég ∩ Tő	13,33	12,73	13,02	10,60	14,02	12,07	15,79	9,68	12,00
Vég ∪ Tő	53,33	10,53	17,58	60,12	10,18	17,40	20,18	7,35	10,77
Ige ∩ Tő	40,00	50,00	44,44	40,48	44,04	42,18	16,67	38,00	23,17
Ige ∪ Tő	64,76	12,45	20,89	66,90	10,99	18,88	60,53	16,35	25,75
Vég ∩ Ige ∩ Tő	7,62	50,00	13,22	7,26	53,98	12,80	15,79	41,86	22,93
Vég ∪ Ige ∪ Tő	66,67	12,15	20,56	68,33	10,64	18,42	61,40	13,81	22,54

Mivel számos korábbi munka szintaktikai információból kiindulva kísérelte meg a félig kompozicionális szerkezetek automatikus felismerését, mi is fokozott figyelmet fordítottunk a szintaxis szerepére. Legjobb tudomásunk szerint magyar nyelvű dependenciaelemző még nem áll rendelkezésre, így magyar nyelvi méréseinkhez a Szeged Korpusz etalon dependenciaannotációját használtuk fel.

Amennyiben pusztán szintaktikai információt használunk fel a félig kompozicionális szerkezetek azonosítására, azaz a korpuszban előforduló ige-tárgy párokat minősítünk annak, csupán 17,69-es F-mértéket érünk el a Wiki50 korpuszon (fedés: 59,51 és pontosság: 10,39). Mivel módszereink arra épülnek, hogy a baseline módszer által meghatározott lehetséges FX-ek köréből további megszorítások segítségével válasszuk ki a tényleges FX-eket, így olyan baseline-t érdemes választani, amely nagy fedéshez vezet. E célnak pedig a POS-szabályok sokkal inkább megfelelnek (76,63-as fedés a Wiki50 korpuszon), így a továbbiakban a szintaktikai információk hozzáadott értéket vizsgáljuk meg az egyes korpuszokon.

A 3. és 4. táblázat összevetéséből látszik, hogy a szintaktikai információ javít a rendszer teljesítményén, különösen a leggyakoribb ige (és kombinációi) esetében. Az átlagos javulás F-mértékben 2,3% a Wiki50, 2,26% a SzegedParalellFX és 1,52% a Szeged Korpusz esetében. A 4. táblázat utolsó oszlopa azt is mutatja, hogy a Szeged Korpuszon akkor érjük el a legjobb eredményeket, ha etalon szófaji kódokat és szintaktikai információt használunk az FX-ek azonosításában, átlagosan 4%-kal javítva az F-mértéket a predikált szófaji kódokra épülő rendszerhez képest.

3.3. Gépi tanulási módszerek

Szótárillesztéses megközelítéseket használtunk baseline megoldásnak a gépi tanulási módszerek esetében. Mivel mindkét nyelven rendelkezésünkre állt két annotált korpusz, ezért az ezeken előforduló FX-ekből lemmatizált listákat hoztunk létre. Az azonos nyelvű korpuszokra a másikról gyűjtött listát jelöltük rá. Így például a Wiki50 esetében az angol SzegedParallelFX-ről gyűjtött lista került illesztésre. A különböző korpuszokon így elért eredmények a 5. táblázatban láthatók.

5. táblázat. A szótáralapú megközelítés eredményei.

Korpusz	**Fedés**	**Pontosság**	**F-mérték**	**Szótárméret**
Wiki50	8,57	81,81	15,51	587
SzegedParalellFX angol	9,01	73,07	16,04	287
SzegedParalellFX magyar	29,5	40,14	34,01	1215
Szeged Treebank	30,7	39,77	34,65	578

Az eddig ismertetett megközelítéseken túl implementáltuk az FXtagger nevű, gépi tanuló alapú megközelítésünket is. Vizsgálatainkban a Conditional Random Fields (CRF) [17] szekvenciális tanuló MALLET [18] implementációját használtuk, az alábbi alapjellemzőkkel ([19] alapján a feladat sajátságaira szabva):

- **Felszíni jellemzők:** kis/nagybetűs kezdet, szóhossz, a szó belsejében előforduló különleges karakterek (számok, nagybetűk stb.), karakter bi- és trigramok, toldalékok;
- **Szótárak:** személynevek, cégnevek, helynevek, a leggyakoribb funkcióigék, főnevek szótövei;
- **Gyakorisági jellemzők:** a token gyakorisága, a kis- és nagybetűs alakok előfordulásának aránya, a nagybetűs és mondatkezdő alakok előfordulásának aránya;
- **Nyelvi jellemzők:** szófaj, függőségi viszonyok;
- **Környezeti jellemzők:** mondatbeli pozíció, a szó környezetében előforduló leggyakoribb szavak, idézőjelek a szó körül stb.

Ezt az általános jellemzőteret egészítettük a szabályalapú megközelítések jellemzőkre transzformált verzióival. így a leggyakoribb ige és a szótő módszereket szótáralapú jellemzőként, a POS-szabályokat és a mondat szavai közti szintaktikai kapcsolatokat nyelvi jellemzőként, míg a végződés megközelítést felszíni jellemzőként alkalmaztuk a CRF tanítása során. Mivel a magyar nyelv részletesebb morfológiai elemzést tesz lehetővé, ezért magyar nyelvű gépi tanulás során a jellemzőket még kiegészítettük ezekkel a részletesebb jellemzőkkel. Továbbá minden esetben szótáralapú jellemzőként használtuk a szótárillesztés baseline megközelítésnél használt listákat.

Kísérleteinkhez a korpuszokat 70%:30% arányban osztottuk fel tanító és kiértékelő adatbázisra. Mivel a korpuszok több témában is tartalmaznak szövegeket (újságcikkek, szépirodalom, tankönyvi mondatok stb.), minden egyes dokumentumot a fenti arányoknak megfelelően osztottunk fel a tanító és a kiértékelő adatbázis között. Eredményeink a 6. táblázatban láthatók.

6. táblázat. A gépi tanult megközelítés eredményei a különböző korpuszokon.

Korpusz	**Fedés**	**Pontosság**	**F-mérték**
Wiki50	42,86	56,96	48,91
SzegedParalellFX angol	37,91	55,55	45,07
SzegedParalellFX magyar	61,0	67,78	64,21
Szeged Treebank etalon	44,73	62,96	52,03
Szeged Treebank predikált	43,86	56,82	49,51

3.4. A gépi tanulási módszerek eredményei

A szótáralapú megközelítések eredményeiben igen nagy kontraszt mutatkozott a két vizsgált nyelvben. Ez a módszer magyar nyelvű korpuszokon kétszer jobb F-mértéket ért el, mint az angol nyelvűeken. Ugyanakkor az angol nyelvű korpuszokon a megközelítés pontossága jóval magasabb volt, mint a magyarokén. A fedésben mutatkozó különbségeket az magyarázhatja, hogy a magyar nyelvű korpuszok jóval homogénebbek voltak az angolokénál. Az enciklopédia domén (Wiki50), mely több különböző témát ölel fel, egészen más jellegű, mint a homogénebb SzegedParallelFX, nagyrészt újságcikből és regényekből álló domén, mely hatással lehet az FX-ek eloszlására is. Mivel a két magyar nyelvű korpusz mindegyikében találhatók újságcikkek, ezért a belőlük kinyert FX-listák kevésbé voltak eltérőek. A SzegedParalellFX korpuszon mért eredmények közti különbségeket magyarázhatja az alkalmazott listák mérete. Mivel a Szeged Treebank jóval nagyobb, mint a Wiki50, ezért az ezekből a korpuszokból összeállított listák mérete is nagyon eltérő. Ugyanakkor ezen baseline megközelítés pontossági értékei szerint a félig kompozicionális szerkezetek kevésbé többértelműek angolban, mint a magyar nyelvben, azaz a listákban előforduló FX-jelölt nagyobb valószínűséggel lesz a valóságban is FX.

Az 5. táblázat pontossági értékei is igazolják, hogy a félig kompozicionális szerkezetek automatikus azonosítása során hasznos információ lehet a kontextus is. Így például a *titokban tartja a kapcsolatot Imrével* szövegrészletben a *titokban tarja* és a *tartja a kapcsolatot* is lehetséges FX. Ebben az esetben a szövegkontextus segíthet eldönteni, hogy melyik szekvencia az adott szövegben az FX. A folyó szövegekben előforduló félig kompozicionális szerkezetek automatikus azonosítása így nagyban segítheti az olyan alkalmazásokat, mint a gépi fordítás vagy az információkinyerés. Ugyanakkor előfordulhat olyan eset is, amikor a felhasználót alapvetően a szövegből kigyűjthető FX-ek listája érdekli alapvetően. Ebben az esetben elegendő minden potenciális FX azonosítása a szövegben, nem

szükséges annak eldöntése, hogy az adott szekvencia FX-ként viselkedett-e az adott kontexusban.

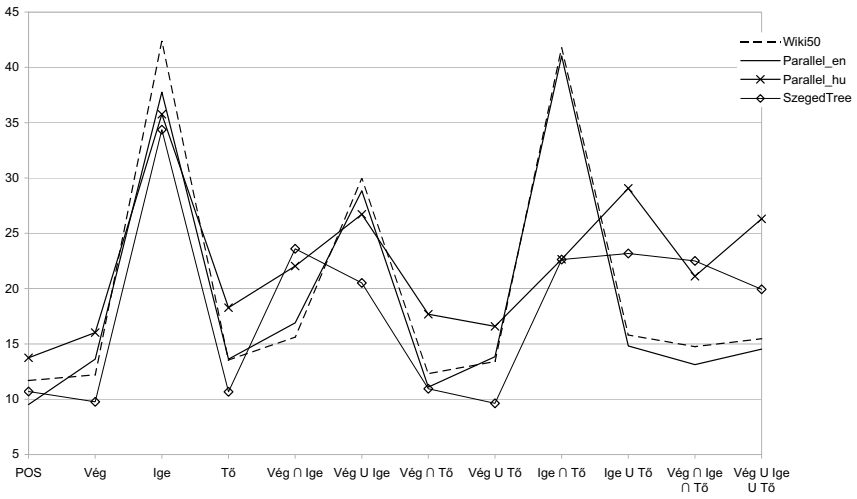
Az FXtaggerrel elért eredmények az 6. táblázatban láthatóak. A gépi tanuló megközelítéssel elért eredmények minden korpuszon meghaladták mind a szótáralapú baseline módszer, mind a szabályalapú rendszerek eredményeit. Vagyis a félig kompozicionális szerkezetek automatikus azonosítására hatékony reprezentációt voltunk képesek adni a CRF lineáris tanuló számára kibővített jellemzőtér segítségével. Mint ahogy megfigyelhettük, a korpuszokról gyűjtött szótárak kedvező hatással voltak a pontosságra, míg a POS-szabályok a fedést javították. A gépi tanuló módszerek ezen jellemzők kedvező kombinálásával érhették el a legjobb eredményeket a különböző korpuszokon.

Szembetűnő, hogy az angol nyelvű korpuszokon elért eredmények szerényebbek a magyar nyelven elérteknél. Ezt magyarázhatja, hogy megközelítéseink alapvetően a morfológiai jellemzőkre támaszkodnak, így hatékonyabbnak bizonyultak a morfológiailag jóval gazdagabb magyar nyelv esetében. Az etalon POS-címkék pozitív hatását jól mutatja a Szeged Treebanken mért két eredményünk. A SzegedParalellFX korpusz magyar nyelvű változatán elért legmagasabb F-mértéket többek közt az ebben az esetben alkalmazott nagyobb FX-lista magyarázhatja.

4. Eredmények

Az általunk definiált szabályalapú megközelítések eredményei azt igazolják, hogy már sekély morfológiai elemzések segítségével is versenyképes eredményeket lehet elérni félig kompozicionális szerkezetek automatikus azonosítása során. Hatékony jellemzőnek bizonyult a lemmatizálás, szótövesítés, szófaji egyértelműsítésen kívül egy funkcióige-lista is. Ugyanakkor a szintaktikai információk integrálása tovább javítja a rendszer teljesítményét. A félig kompozicionális szerkezetek felismerése ennélfogva leghatékonyabban a szintaktikai elemzést követően, egy utófeldolgozó lépésben valósulhat meg, annak végeredményét pedig jól tudják hasznosítani a magasabb rendű alkalmazások, például az információkinyerés és a gépi fordítás.

A különböző szabályalapú módszerek jellemzőkre való transzformálásával megvizsgáltuk a gépi tanuló algoritmusok hatékonyságát is. Általánosan elmondható, hogy a gépi tanuló módszerekkel magasabb F-mértéket tudtunk elérni, mint a szabályalapú megközelítésekkel. Ugyanakkor az eredményekből kitűnik, hogy a szabályalapú módszerek jobb fedést tudnak elérni, míg a gépi tanuló megközelítés jórészt jó pontosságának köszönheti sikerét. Ahogy a 6. táblázatban is látszik, a gépi tanuló megközelítés mind a négy vizsgált korpuszon 50% fölötti pontosságot volt képes elérni, míg a szabályalapú megközelítések vagy egyáltalán nem képesek ilyen magas pontosságra, vagy csak igen alacsony fedés mellett.



1. ábra. Szabályalapú eredmények a korpuszokon.

5. Az angol és magyar eredmények összevetése

Az angol és magyar korpuszokon elért eredményeket az 1. ábra szemlélteti. Bizonyos módszerek esetében alapvető különbségeket figyelhetünk meg a nyelvek között. érdekes módon a leggyakoribb ige és a szótő metszete sokkal jobb eredményt ért el az angol korpuszokon, mint a magyarokon, ugyanakkor e két módszer uniója a magyar korpuszokon teljesít sokkal jobban. Ennek az lehet az oka, hogy feltehetőleg az angol korpuszokban több olyan FX fordul elő, amelyek tipikus ige és tipikus főnév kombinációja, míg a magyarokban a tipikus ige + nem tipikus főnév párok vannak túlsúlyban.

További számottevő eltérést figyelhetünk meg mindhárom módszer metszete kapcsán: sokkal jobb eredményhez vezet a magyarban, mint az angolban. Ez talán azzal magyarázható, hogy a metszet megköveteli, hogy egy igei tövű főnév adott képzőben végződjön. A magyarban ez definíció szerint megvalósul (igéből képzők segítségével tudunk főnevet képezni: *dönt - döntés*), ugyanakkor az angolban a konverzió művelete is létrehozhat igéből főnevet (például *walk - walk*). Utóbbi megfelel a szótő definíciójának, de a végződésének már nem, így az ilyen típusú főneveket tartalmazó FX-eket nem lehetséges azonosítani a módszerek metszetével.

A nyelvek közti eltérések egy újabb vetületét jelenti a leggyakoribb igék száma. Míg az angolban a 12 leggyakoribb igével lehetett 40% körüli eredményeket elérni, addig a magyarban nagyobb (17 elemű) igelistával is szerényebb eredményekhez jutottunk. E jelenség magyarázatát keresve összevetettük a SzegedParalellFX két részében található FX-igék számát. Míg angolban összesen 100 ige fordult elő, melyek eloszlása megfelel a Zipf-törvénynek, addig a magyarban 179 ige fordult elő, kiegyenlítettebb eloszlásban. Tehát az angolban kevesebb

ige is nagyobb hányadát fedi le az FX-eknek, mint a magyarban. Mindez azt is mutatja, hogy az FX-igelisták bővítésével várhatóan jobb eredményeket lehet elérni mindkét nyelven.

6. Összegzés

Ebben a cikkben bemutattuk szabályalapú és gépi tanult módszereken alapuló megközelítéseinket, melyek mind angol, mind magyar nyelven képesek a félig kompozicionális szerkezetek automatikus azonosítására sekély morfológiai jellemzők segítségével. Eredményeink összevethetők más, szintaxison alapuló megközelítésekkel. Módszereinket két különböző nyelven és három korpuszon teszteltük, melyeken hasonló eredményeket értünk el. Eredményeink azt mutatják, hogy mind angol, mind magyar vonatkozásban egy adott nyelvre és doménre szabott funkcióige-lista és a főnév szótöve bizonyul a leghasznosabb jellemzőnek, illetve az angol anyagban a szintaktikai jellemzők beépítése is számottevően javít a rendszer teljesítményén. Gépi tanult megközelítésnek lineáris CRF tanuló algoritmust alkalmaztunk, melynek alap jellemzőterét kiegészítettük a szabályalapú módszerek jellemzőkre transzformált verzióival. FXtagger nevű, gépi tanuló megközelítésünk érte el a legmagasabb F-mértékeket az összes vizsgált korpuszon.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER és BELAMI kódnevű projektek keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Vincze, V., Nagy T., I., Berend, G.: Multiword expressions and named entities in the Wiki50 corpus. In: Proceedings of RANLP 2011, Hissar, Bulgaria (2011)
2. Van de Cruys, T., Moirón, B.n.V.: Semantics-based multiword expression extraction. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. MWE '07, Morristown, NJ, USA, Association for Computational Linguistics (2007) 25–32
3. Cook, P., Fazly, A., Stevenson, S.: Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. MWE '07, Morristown, NJ, USA, Association for Computational Linguistics (2007) 41–48
4. Bannard, C.: A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. MWE '07, Morristown, NJ, USA, Association for Computational Linguistics (2007) 1–8

5. Samardžić, T., Merlo, P.: Cross-lingual variation of light verb constructions: Using parallel corpora and automatic alignment for linguistic research. In: Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, Uppsala, Sweden, Association for Computational Linguistics (2010) 52–60
6. Gurrutxaga, A., Alegria, I.n.: Automatic Extraction of NV Expressions in Basque: Basic Issues on Cooccurrence Techniques. In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, Portland, Oregon, USA, Association for Computational Linguistics (2011) 2–7
7. Tu, Y., Roth, D.: Learning English Light Verb Constructions: Contextual or Statistical. In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, Portland, Oregon, USA, Association for Computational Linguistics (2011) 31–39
8. Sass, B.: Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In Tanács, A., Vincze, V., eds.: VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2010) 102–110
9. Tan, Y.F., Kan, M.Y., Cui, H.: Extending corpus-based identification of light verb constructions using a supervised learning framework. In: Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts, Trento, Italy, Association for Computational Linguistics (2006) 49–56
10. Vincze, V., Felvégi, Z., R. Tóth, K.: Félig kompozicionális szerkezetek a Szeged-Paralell angol–magyar párhuzamos korpuszban. In Tanács, A., Vincze, V., eds.: MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary, University of Szeged (2010) 91–101
11. Vincze, V.: Félig kompozicionális szerkezetek a Szeged Korpuszban. In Tanács, A., Szauter, D., Vincze, V., eds.: VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2009) 390–393
12. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of EMNLP 2000, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 63–70
13. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Annual Meeting of the ACL. Volume 41. (2003) 423–430
14. Zsibrita, J., Vincze, V., Farkas, R.: Ismeretlen kifejezések és a szófaji egyértelműsítés. In Tanács, A., Vincze, V., eds.: MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary, University of Szeged (2010) 275–283
15. Porter, M.F.: An algorithm for suffix stripping. In Sparck Jones, K., Willett, P., eds.: Readings in information retrieval. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997) 313–316
16. Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L., Varga, D.: hunmorph: Open Source Word Analysis. In: Proceedings of the ACL Workshop on Software, Ann Arbor, Michigan, Association for Computational Linguistics (2005) 77–85
17. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289
18. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (2002)
19. Szarvas, G., Farkas, R., Kocsor, A.: A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: Discovery Science. (2006) 267–278