Események detektálása, osztályozása és szemantikus szerepeik címkézése

DOKTORI ÉRTEKEZÉS

Subecz Zoltán

2019

Témavezető: Prof. Dr. Csirik János

Szegedi Tudományegyetem
Informatika Doktori Iskola

Tartalomjegyzék

Táblázatok listája

Ábrák listája

Rövidítések listája

ACE	Automatic Content Extraction
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ATIS	Air Traveller Information System
CCG	Combinatory Categorial Grammar
CRF	Conditional Random Fields
DARPA	Defence Advanced Research Projects Agency
EE	Event Extraction
FVG	Funktionsverbgefüge
IE	Information Extraction
LVC	Light Verb Construction
MSD	Morphological Coding System
MUC	Message Understanding Conference
NE	Named Entity
NP	Noun Phrase
NLP	Natural Language Processing
RRM	Robust Risk Minimization
SRL	Semantic Role Labelling
SVM	Support Vector Machine
TAG	Tree Adjoining Grammar
VP	Verb Phrase
WSD	Word Sense Disambiguation

Előszó

Az *eseménykinyerés* (EE, event extraction) az információkinyerés egy fontos részfeladata, célja strukturált információ kinyerése olyan eseményekkel kapcsolatban, amelyek strukturálatlan dokumentumokban találhatóak. Az *eseményi információk* egyre időszerűbbé váltak sok NLP alkalmazás számára, mint például a válaszkeresés, az automatikus összegzés, az információ visszakeresés és az információkinyerés. Az *eseménykinyerést* a mindennapok sok területén felhasználják, például a politika, pénzügy, gazdaság, kereskedelem, piackutatás, döntéstámogatás, egészségügy területeken.

Az *események detektálásának* a feladata az esemény-előfordulások azonosítása a szövegekben. Esemény előfordulásnak tekintünk minden olyan kifejezést, ami olyan eseményt vagy állapotot jelöl, amit egy adott időponthoz, vagy intervallumhoz tudunk kapcsolni. A szövegekben a legtöbb esemény *igékhez* kapcsolódik, és az igék általában eseményeket jelölnek. Az igéken kívül lehetnek események más szófajú szavak is pl. *főnevek, igenevek* stb.

Az események detektálása mellett egy fontos feladat a megtalált események szerepeinek meghatározása, a *szemantikus szerepek címkézése* (SRL, Semantic Role Labeling). Ez a természetesnyelv-feldolgozás azon feladata, ami detektálja egy mondat predikátumának szemantikus argumentumait és osztályozza ezeket speciális szerepek szerint. A szemantikus szerepek az események és a résztvevői közötti logikai kapcsolatok. Egy SRL rendszer eredményét felhasználva javíthatjuk számos magasabb rendű feladat hatékonyságát.

Ez a disszertáció a természetes nyelven kifejezett *események számítógépes feldolgozásával* foglalkozik. Ezen belül az *események detektálásával, osztályozásával*, valamint *az események szereplőinek azonosításával*.

Subecz Zoltán
Szeged, 2019. június 11.

Köszönetnyilvánítás

Elsősorban szeretnék köszönetet mondani témavezetőmnek Csirik Jánosnak a támogatásáért és hasznos tanácsaiért a kutatási időszak folyamán.
Szintén köszönöm munkahelyemnek, a Szolnoki Főiskolának, majd Neumann János Egyetemnek, hogy támogatta ezen kutatásomat anyagi téren és munkaidő-kedvezménnyel is.
Köszönöm Vincze Veronikának és Nagyné Csák Évának tanácsaikat nyelvészeti témákban.
Végül, de nem utolsó sorban köszönöm családomnak támogatásukat és türelmüket ezen hosszú és sok munkát igénylő időszak alatt.

1 Bevezetés

1.1 Motiváció

Gyakran hivatkozunk a mai időkre úgy, mint az *információs kor*. Az információ mennyiségének exponenciális növekedése az egyik legösszetettebb probléma, ami az utóbbi időben jelentkezett. Jelenleg az új információ megjelenési aránya magasabb, mint amit az emberiség kezelni tud. Az Internettel elérhetővé vált információ mennyisége exponenciális mértékben növekszik, például hírek, tudósítások, üzenetek formájában. Ebből kifolyólag egy új kihívást jelent az információ hozzáférése, keresése, a feldolgozás javítása és automatizálása, abból a célból, hogy minél több előnyhöz jussunk ebből az értékes tudásforrásból.
A *természetes nyelv* egy fontos kommunikációs eszköz, amit széleskörűen használunk az ismeretek és adatok megosztására. Az elmúlt évezredekben ez volt a leghatékonyabb forma az egyének közötti információ közvetítésére. Az információ számos fajtája létezik *szöveges formában*. Ezen szöveges információk egy része strukturált szövegekben található (például táblázatokban, adatbázisokban), viszont nagy részük *strukturálatlan szövegekben* érhető el, főleg természetes nyelvű szövegekben (például könyvek, hírek, cikkek, riportok). A strukturált formában lévő információ sokkal könnyebben visszakereshető és feldolgozható.
A számítógépek megjelenésével a természetes szövegek nagy mennyiségét tároljuk digitális formában. Megoldandó feladatként jelentkezik a számítógépek felkészítése a természetes nyelvű információ feldolgozására, hogy azok a szövegeket strukturált információvá alakítsák. Ez a feladat a mesterséges intelligencia (AI, artificial intelligence), speciálisabban a *természetesnyelv-feldolgozás* (NLP, natural language processing) területéhez tartozik. *A természetesnyelv-feldolgozás* az emberi nyelv feldolgozása számítógép segítségével (Jurafsky & Martin, 2009). Ez egy széles terület, ami sok részterületet átfog a beszédfeldolgozástól kezdve a szemantikáig. Olyan hatékony számítógépes algoritmusok megalkotásával foglalkozik, amelyek képesek analizálni, megérteni és generálni természetes nyelveket, beszélt és írott formában (Moreno, Palomar, Molina, & Ferrandez, 1999), (Allen J. F., 1995), (Jurafsky & Martin, 2009). A számítógépes nyelvészet segít megteremteni a kapcsolatot a számítógépek hatékony feldolgozási képességei és a természetes nyelvi kommunikáció között.
Ennek a feladatnak a sokrétűsége miatt sok részproblémát definiáltak a természetesnyelv-feldolgozáson belül. Ilyenek például az információ-visszakeresés, információkinyerés, összegzés, osztályozás, kivonatolás, csoportosítás, válaszkeresés, gépi fordítás, véleménydetektálás, szövegbányászat, szemantikus annotálás.
Az *információkinyerés* (IE, Information extraction) a természetesnyelv-feldolgozás egy fontos területe. Strukturálatlan, vagy félig strukturált dokumentumokból gyűjt ki információkat, amiket strukturált formában állít elő, majd tárol el. Az IE eredményeit sok más alkalmazás felhasználja, mint például az információ visszakeresés, válaszkeresés, gépi fordítás, így az utóbbi időben az információkinyerés egyre népszerűbbé vált, mint sok alkalmazás eszköze (Cowie & Lehnert, 1996). Az információkinyerés is több részfeladatra bontható, mint például a névelem felismerés, *eseménykinyerés*, kereszthivatkozások azonosítása, szereplők azonosítása, szereplők közötti reláció azonosítása, eseménykeretek illesztése.

Az IE technikák a szabályalapú rendszerektől fejlődtek a statisztikai és gépi tanulásos megközelítésekéig. A *szabályalapú* rendszerek „kézzel kódolt" mintákat használnak az információ kinyerésére. Bár ezeket könnyebb implementálni és nyomon követni, de nagymértékben függnek a fejlesztő gondolatmenetétől és sok „kézi munkát" igényelnek (Chiticariu, Li, & Reiss, 2013), valamint új doméneken nehéz azokat alkalmazni. A szabályalapú rendszerekkel általában nagy pontosságot, de emellett kis fedést érnek el. A *gépitanulás-alapú* megközelítések ezzel szemben adaptálhatóak és kiterjeszthetőek. Emberi erő-ráfordítással annotált korpuszok kifejlesztésével a gépi tanuláson alapuló módszerek jelentős fejlődést értek el.
Az emberi nyelvek nem csak entitásokkal (személyek, tárgyak, helyek) foglalkoznak, mint például a névelemek, hanem *szituációkkal, eseményekkel* is. Ezért a szituációk számos aspektusát érdemes analizálni a nyelvészeti jelentés modellezéséhez. Az információnak az *eseményi* dimenziója alapvető fontosságú a világban lévő változások megmagyarázásához. Az emberek gyakran jellemzik a történéseket, tapasztalataikat eseményi, időbeliségi és okozati struktúrákkal. Például egy idősorral, ahol a szövegben események találhatóak a hozzájuk tartozó időpontokkal. Az idő fogalmát használják az események sorba rendezéséhez, az események időpontjának és intervallumának meghatározásához.

Példa a természetes szövegekben lévő események sorozatára.
*De végül is **odaértünk**, mert **jött** az egyik osztálytársam apukája kocsival és **elvitt** minket.*

Ezekből a mondatokból az olvasó újjáalkothatja a következő valóságot: Van egy esemény (jött), ami megtörtént egy adott időpontban. És egy másik esemény (elvitt), amelyik az első esemény után történt meg. A mondatban először említett esemény (*odaértünk*) az előző két esemény után történt meg. Ezek az események állapotváltozásokat jelentenek a történet szereplői számára. Nagyon értékes lenne, ha az eseményeket hatékonyan és automatikusan tudnánk detektálni és kinyerni.
A dolgozat magyar nyelvű mintapéldáit, ahol lehetett, a Szeged Korpuszból vettem. (Csendes, Csirik, & Gyimothy, 2004)

Az *eseménykinyerés* (event extraction, EE) az információkinyerés egy fontos részfeladata. Az eseménykinyerés célja strukturált információ kinyerése olyan eseményekkel kapcsolatban, amelyek strukturálatlan dokumentumokban találhatóak. Népszerűségre az elmúlt évtizedben tett szert a Big Data, valamint a szövegbányászat és természetesnyelv-feldolgozás kapcsolódó területeinek megjelenésével és fejlődésével. A valós világbeli események növekvő számú online megjelenésével az eseménydetektálás egy fontos kutatási területté vált.
Az *eseményi információk* egyre időszerűbbé váltak számos NLP alkalmazás számára, mint például a válaszkeresés (Moldovan, Clark, & Harabagiu, 2005), (Sauri R. , Knippen, Verhagen, & Pustejovsky, 2005) az automatikus összegzés (Mani & Shiffman, 2003), az információ visszakeresés (Alonso, Gertz, & Baeza-Yates, 2007) és az információkinyerés (Surdeanu M. , Harabagiu, Williams, & Aarseth, 2003). A válaszkeresési kutatások (Sauri R. , Knippen, Verhagen, & Pustejovsky, 2005) szerint a legtöbb webes kereső kérdés eseményekkel kapcsolatos. A válaszkereső rendszereknek szükségük van az események feldolgozására olyan kérdések megválaszolásához, amelyek eseményekkel kapcsolatosak. Az automatikus

összegzés szintén igényli az eseményinformációkat, felhasználva az események egymáshoz viszonyított sorrendjét.

Az *eseménykinyerést* a mindennapok számos területén felhasználják, mint például a politika, pénzügy, gazdaság, kereskedelem, piackutatás, döntéstámogatás, egészségügy. A parlamenti választások, bejelentések, igazgatóváltások, felvásárlások is eseményeket jelölnek. Események generálhatnak kereskedési jelzéseket részvénypiacokon, hiszen a pénzügyi piacok nagyon érzékenyek a fontos hírekre (Mitchell & Mulherin, 1994). Az algoritmikus-kereskedésnél számítógépes programokat használnak kereskedések nyitására algoritmusok alapján, amelyek szintén felhasználhatnak eseményi információkat. Döntéstámogatási rendszerekben eseményvezérelt adatintegrációra is van lehetőség. A gazdasági események fontos szerepet játszanak napi döntések meghozatalában, akár emberek akár gépek részéről. A hírek gyorsabb feldolgozása elősegítette, hogy a gépek nagyobb mennyiségű hírrel tudnak foglalkozni, több információhoz férnek hozzá, mint mi emberek, így jobban informált döntéseket tudnak hozni.

Manapság az online közösségi szolgáltatásoknak, mint például a Twitter, Facebook, Google+, LinkedIn fontos szerepük van a valós idejű információ terjesztésében. Mivel az *események* kiemelt szerepet játszanak a mindennapi életben, a közösségi média oldalakra feltöltött információ nagy része is eseményekkel kapcsolatos. Vizsgálatok bizonyítják, hogy sok esemény és hír először a közösségi média csatornáin jelenik meg és csak később a hagyományos online híroldalakon, a blogokon, vagy a televíziók rádiók híreiben (Sakaki, Okazaki, & Matsuo, 2010), (Kwak, Lee, Park, & Sue, 2010), (Sakaki et al., 2010b; Kwak et al., 2010). A szociális média és a közösségi rendszerekben megjelenő események megfigyelésére külön csoportok alakultak a bűnüldözés területén is. Ezeken kívül orvosi és biológiai eseményeket is azonosítanak NLP technikákkal (Yakushiji, Tateisi, & Miyao, 2001).

A szövegekben a legtöbb esemény igékhez kapcsolódik, és az igék általában eseményeket jelölnek. Például: *Este, ahogy **megbeszéltük**, **lementünk** fürödni.* Az igék és főnévi igenevek közül azonban nem mindegyik tekinthető eseményjelölőnek (például: van, volt, lesz, marad, segédigék), így ezek kiszűrésére külön figyelmet kell fordítani. Például *Ebben az évben is oda **akartunk** menni, de a nyaralót eladták, így nem jött össze.* Az igéken kívül lehetnek események más szófajú szavak is pl. főnevek, igenevek stb. Például: *A **futás** után következett a **torna** negyed órán át, amit még közösen csináltunk*. A *főnévi eseményeknek* két nagy csoportja van: igéből képzettek (deverbális) és nem igéből képzettek (nem deverbális). Példa igéből képzett főnevekre: *futás, írás*.

Az igéből képzett főnevek két fő csoportja az események és az eredmények, közöttük gyakori a kétértelműség. Vannak olyan szavak (pl. írás), amelyek egyes mondatokban események, másokban pedig eredmények. Például az *írás* főnév a következő mondatban esemény: *Nagyot lélegzett, és folytatta az **írást***. Viszont a következő mondatban nem esemény, hanem eredmény: *Az RTL Klub az említett **írást** jogsértőnek, az abban szereplő állításokat valótlannak tartja*. A többértelműség miatt nem elég a szóalak vizsgálata, a szövegkörnyezetet is elemezni kell.

Szemantikus szerepek címkézése

Az események detektálása mellett egy másik fontos feladat a megtalált események szerepeinek meghatározása, a *szemantikus szerepek címkézése* (SRL, Semantic Role Labeling). Ez a

természetesnyelv-feldolgozás azon feladata, ami detektálja egy mondat predikátumának szemantikus argumentumait, és osztályozza ezeket speciális szerepek szerint. A *szemantikus szerepek* az események és a résztvevői közötti logikai kapcsolatok.
Az NLP magában foglalja a szövegek struktúrájának feltérképezését morfológiai, szintaktikai és szemantikai szinteken (Jurafsky & Martin, 2009). A szintaktikai elemzés mellett fontos a szemantikai összefüggések feltárása is (Carreras X. , 2005). A *szemantikai* információ a lexikai és szintaktikai alkotórészek és a predikátum között lévő kapcsolatokat írja le. Ezen kapcsolatok azonosítása fontos olyan kérdések megválaszolása szempontjából, mint „Ki?", „Mit?", „Hol?", „Mikor?", „Kivel". A szemantikus entitások pontos azonosítása és a közöttük lévő kapcsolatok egyértelműsítése fontos lépés olyan NLP alkalmazások sikeréhez, mint az összegzés, válaszkeresés, információkinyerés és gépi fordítás. A *szemantikus szerepek címkézése* egy mondatnak a szemantikus feldolgozása, ahol egy adott predikátumhoz azonosítjuk az argumentumokat és osztályozzuk azokat (Gildea & Jurafsky, 2002).

Nézzük a következő példamondatot: *Észre vettem egy bácsit aki éppen újságot olvasott és kiflit* ***evett*** *az autójában*. Az *evett* igének itt három szerepe van. **Az evés cselekvője** = bácsi, **Amit eszik** = kifli és **Ahol eszik** = az autójában. A *szemantikus szerepek meghatározása* félúton helyezkedik el a szintaktika és a szemantika között. Inkább szemantikus feladat, mint a szófajok megállapítása vagy a szintaktikai elemzés, de kevésbé szemantikai, mint az információkinyerés, vagy a válaszkeresés. Előző munkák (Shen & Lapata, 2007), (Christensen, Mausam, & Etzioni, 2010) megmutatták, hogy egy SRL rendszer eredményét felhasználva javíthatjuk számos ilyen magasabb rendű feladat hatékonyságát.

Ez a disszertáció a természetes nyelven kifejezett *események számítógépes feldolgozásával* foglalkozik. Ezen belül az *események detektálásával*, *osztályozásával*, valamint az *események szereplőinek azonosításával*.

1.2 A disszertáció bemutatása

1.2.1 Célok és feladatok

A disszertációban ismertetett kutatás megmutatja, hogy egy szöveg eseményeinek fontos részei automatikusan kinyerhetőek gépi tanulásos módszerekkel, valamint példákat ad arra vonatkozólag, hogy a számítógépek hogyan taníthatóak az eseménykinyerés feladatára, felügyelt gépi tanulásos módszerekkel. Így általános irányelvet nyújt egy adott eseménykinyerő módszer választásához, kiépítéséhez és teszteléséhez. A jelenlegi megközelítések nagy része az eseménydetektálás területén morfoszintaktikai ismeretekre fókuszál. Azonban az események gyakran többértelműek a nyelvi elemzés szintjén, ezért egy jobb kinyerési eredményhez az eseményi információ feldolgozásánál a szemantikát is figyelembe kell venni. Egy olyan modellt mutatok be az eseményi információ feldolgozáshoz, ami a szemantikus szempontokat is figyelembe veszi.
Ez egy automatikus rendszer, ami a morfoszintaktikai jellemzők mellett olyan jellemzőket is használ, amelyek lexikai szemantikán, szemantikai szabályokon és eseményi szemantikán alapulnak.

Mindhárom fő kutatási résznél kiemelt feladatomnak tekintettem olyan jellemzőcsoportok részletes kidolgozását, amelyek figyelembe veszik a magyar nyelv sajátosságait. Ezek a *morfológiai* és a *függőségifa-alapú jellemzőcsoportok* voltak. Mivel a magyar morfológiailag gazdag nyelv, így a *morfológiai jellemzőcsoportra* kiemelt figyelmet fordítottam. És mivel a magyar nyelv szabad szórendű és a függőségi fákkal dolgozó reprezentáció különösen jól használható szabad szórendű nyelvek elemzésére, ezért a *függőségifa-alapú jellemzőcsoportot* is kiemelten kezeltem. Ezek a jellemzőcsoportok jelentősen hozzájárultak az angol nyelvre már alkalmazott jellemzők eredményeinek javításához a magyar nyelvű szövegeken.

Eseményi információkinyerő keretrendszerem magyar szövegeken eseményeket tud detektálni, osztályozni és bejelöli az események szemantikus szerepeit. Az ajánlott módszerek hatékonyságának igazolásához számos tapasztalati eredményt és kiértékelést nyújtok. Mindegyik részfeladathoz új módszereket és jellemző ábrázolási módokat ismertetek, amelyeket a magyar szövegek tulajdonságai alapján dolgoztam ki.

Úgy gondolom, hogy a disszertáció elősegíti az események megértését a nyelvfeldolgozásban, valamint támogatja a számítógépek szöveg-feldolgozási képességének javítását.

1.2.2 Publikációk

A kutatásom eredményeit a következő tudományos publikációkban tettem közzé:

1. Igei és főnévi igenévi események detektálása és osztályozása

Detection and Classification of events in Hungarian natural language texts
17th International Conference on Text, Speech and Dialogue, TSD 2014,
Brno, Czech Republic, September 8–12 2014 (Subecz, 2014)

Event Detection in Hungarian Texts with Dependency and Constituency Parsing and WordNet. Informatics 2017, IEEE 14th International Scientific Conference on Informatics, Poprad Slovakia, November 14–16, 2017. (Subecz Z. , 2017a)

Subecz Zoltán, Nagyné Csák Éva: Igei események detektálása és osztályozása magyar nyelvű szövegekben X. MAGYAR SZÁMÍTÓGÉPES NYELVÉSZETI KONFERENCIA - Szegedi Tudományegyetem - 2014. január 17. (Subecz & Csák, 2014)

2. Főnévi események automatikus detektálása függőségifa- és konstituensfa-alapú szintaktikai reprezentációval és WordNettel

Zoltán Subecz: Automatic Detection of Nominal Events in Hungarian Texts with Dependency Parsing and WordNet
Information and Software Technologies, 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016 (Subecz Z. , 2016)

Zoltán Subecz: Event Detection in Hungarian Texts with Dependency and Constituency Parsing and WordNet. Informatics 2017, IEEE 14th International Scientific Conference on Informatics, Poprad Slovakia, November 14–16, 2017. (Subecz Z. , 2017a)

Subecz Zoltán: Főnévi események automatikus detektálása függőségi elemző és WordNet alkalmazásával magyar nyelvű szövegeken XIII. MAGYAR SZÁMÍTÓGÉPES NYELVÉSZETI KONFERENCIA - Szeged, 2017. január 26-27. (Subecz Z. , Főnévi események automatikus detektálása függőségi elemző és WordNet alkalmazásával magyar nyelvű szövegeken, 2017b)

3. Események szemantikus szerepeinek automatikus címkézése

Zoltán Subecz: Automatic Labeling of Semantic Roles with a Dependency Parser in Hungarian Economic Texts
18th International Conference on Text, Speech and Dialogue, TSD 2015, Brno, Czech Republic, September 14–17 2015 (Subecz Z. , 2015a)

Subecz Zoltán: Szemantikus szerepek automatikus címkézése függőségi elemző alkalmazásával magyar nyelvű gazdasági szövegeken XI. MAGYAR SZÁMÍTÓGÉPES NYELVÉSZETI KONFERENCIA – Szegedi Tudományegyetem - 2015. január 15-16. (Subecz Z. , 2015b)

1.2.3 A disszertáció felépítése

A disszertáció tíz fejezetet tartalmaz. Az első fejezet röviden áttekinti a disszertáció fő témáit.
A második fejezetben bemutatom az események és a szemantikus szerepek jellegzetességeit, ezen belül külön foglalkozok az eseményekkel és a szemantikus szerepekkel.
A harmadik fejezetben áttekintést nyújtok az eseményekhez és a szemantikus szerepek címkézéséhez kapcsolódó korpuszokról. Először bemutatom az elérhető angol nyelvű korpuszokat, majd ismertetem a kutatásomban is felhasznált magyar nyelvű korpuszokat.
A negyedik fejezetben ismertetem a leggyakrabban használt gépi tanulásos technikákat, amelyek nagy részét a kutatásomban alkalmaztam.
A gépi tanulásos modelleknél alkalmaztam a mondatok szavai közötti kapcsolatok elemzéséhez a függőségifa- és a konstituensfa-alapú reprezentációt, valamint a szavak szemantikai jellemzéséhez a magyar WordNetet. Ezeket az eszközöket az ötödik fejezetben mutatom be.
A hatodik fejezetben írom le a kutatásomhoz kapcsolódó előzetes munkákat, amelyeket mások publikáltak az események detektálásának és a szemantikus szerepek címkézésének területén.

A 7-9 fejezetekben ismertetem kutatásom fő eredményeit:
A hetedik fejezetben az *igei és főnévi igenévi események detektálására és osztályozására* fókuszáltam. Ezen belül először az igei és főnévi igenévi események azonosításával foglalkoztam, majd az azonosított eseményeket osztályoztam több szempont szerint.
Az igéken kívül lehetnek események más szófajú szavak is, például főnevek, igenevek, stb. A nyolcadik fejezetben a szövegekben megtalálható *főnévi események detektálásával* foglalkoz-

tam (általános főnevek és igéből képzett főnevek). Az igei, a főnévi igenévi és a főnévi események már lefedik az események jelentős részét.
A kilencedik fejezetben ismertetem eredményeimet az *események szemantikus szerepeinek automatikus címkézése* területén. Ennek keretében igei és főnévi igenévi eseményekhez kerestem szerepeket a vállalati vásárlásokkal, tulajdonváltozásokkal és a tőzsdei hírekkel foglalkozó szövegekben.

A tízedik fejezetben összefoglaltam a jelen értekezésben elért főbb eredményeket és ismertettem jövőbeli terveimet ezen a kutatási területen.

A disszertáció fő témái és a hozzájuk tartozó publikációk közötti kapcsolat (1.1. táblázat**)**

7. fejezet	MSZNY	2014	(Subecz Z. et al., 2014a)
	TSD	**2014**	**(Subecz Z., 2014b)**
	Informatics	**2017**	**(Subecz Z., 2017b)**
8. fejezet	MSZNY	2017	(Subecz Z., 2017)
	ICIST	**2016**	**(Subecz Z., 2016)**
	Informatics	**2017**	**(Subecz Z., 2017b)**
9. fejezet	MSZNY	2015	(Subecz Z., 2015a)
	TSD	**2015**	**(Subecz Z., 2015b)**

1.1. táblázat: A disszertáció témái és a hozzájuk tartozó publikációk közötti kapcsolat

Az értekezés szerzőjének hozzájárulásai az eredményekhez.
Az MSZNY 2014-es publikáción kívül minden publikáció az értekezés szerzőjének önálló munkája. Az MSZNY 2014-es publikációnál a társszerző a kutatás nyelvészeti hátteréért felelt.

2 Az események és a szemantikus szerepek jellegzetességei

Az események és azok szemantikus szerepeinek számítógépes feldolgozásának ismertetése előtt feltétlenül szót kell ejteni az események fogalmáról és a nyelvészeti megközelítési módokról. Ebben a fejezetben bemutatom az események és a szemantikus szerepek megközelítését nyelvészeti szempontból.

2.1 Események

A kezdeti kutatási lépéseket nyelvész kutatók tették meg ezen a tudományterületen is, mint ahogy a számítógépes nyelvészet sok más részénél. A nyelvész szakemberek feltárták, hogy az emberi nyelv hogyan kódolja az *eseményeket*. Miután a természetesnyelv-feldolgozás egy tudományterületté nőtte ki magát, ezen elméletek nagy részét megvalósították, továbbfejlesztették és felhasználták számítógépes módszerekkel is.
A szövegekben az idővel kapcsolatos kijelentéseket három alapfogalommal jellemezhetünk: időpontok, *események* és időbeli kapcsolatok (Moens & Steedman, 1988). Nincs értelme egy eseménydetektáló rendszerről beszélni anélkül, hogy előtte nem határozzuk meg pontosan, hogy mi egy *esemény* az adott munka szempontjából. Bár nincs teljesen általános definíciója az eseménynek, de néhány irányadó definíció adható.
Az eseményekkel kapcsolatban a következő fő fogalmakat használják: *eseményszerűség*[1] (eventuality) (Bach, 1986), vagy *szituáció* (situation) (Bernard, 1976), (Smith, 1991). Az *eseményszerűség* egy gyűjtőfogalom az *állapot*okra (state) és az *esemény*ekre (event) (Bach, 1986). Az *állapot* egy olyan eseményszerűség, amiben nincs lényeges változás amíg az tart (például ismerni valakit, boldognak lenni). Az *esemény* egy olyan eseményszerűség, ami magában foglal állapotváltozást (például tanulni egy nyelvet, építeni egy házat).
A szakirodalomban az *esemény*ekre és *állapot*okra vonatkozó terminológia nincs következetesen használva, amit Tenny és Pustejovsky is elismernek (Tenny & Pustejovsky, 2000), akik ezt az eseményekre vonatkozó terminológiát „nem-stabilnak" nevezik.
Különböző szerzők az *eseményszerűség*eknek számos osztályát különböztették meg, de a tipikus felosztás a következő: *nem-állandó* (események) és *állandó* (állapotok) eseményszerűségek (Vendler, 1967), (Dowty D. , 1979), (Bach, 1981), (Jackendoff R. , 1990), (Verkuyl H. , 1993), (Pustejovsky J. , The generative lexicon, 1991), (Hovav & Levin, 1998). Ez a felosztás a változás szempontját vizsgálja: az esemény magában foglal egy változást a kezdőállapota és a végállapota között (például épít), addig az állapotok nem változnak fennállási időtartamuk alatt (például szeret) (Dowty D. , 1979), (Parsons, 1990), (Pustejovsky J. , 1991).
A Cambridge English Dictionary szerint egy *esemény* „bármi ami megtörténik, különösen ami fontos vagy szokatlan".
Az események al-eseményekre tagolódhatnak. Például az *Arab tavasz* hónapokig tartott és több forradalomból állt, mindegyik egy saját történelemmel, párhuzamos történeti szálakkal, tetőpontokkal és következményekkel. Az eseményeket megfogalmazhatjuk változó hosszúsá-

[1] http://www.nytud.hu/oszt/korpusz/resources/kuti_et_al_WN_2006.pdf

gú kifejezésekkel, dokumentum gyűjteményektől (Ritter, Etzioni, & Clark, 2012) kezdve egyszerű szavakig.

Az eseményekkel kapcsolatban gyakran történik említés a következőkre: *tagadott esemény, feltételes esemény, modális esemény,* amelyekről nem mondhatjuk bizonyosan, hogy „megtörténik vagy előfordul" (Pustejovsky J. , 1991). Az eseményeknek nem kell valósaknak és megfigyelhetőnek lenniük ahhoz, hogy annotálhassuk azokat egy adott szövegben. A nem valós eseményeket is szerepeltetni kell egy dokumentum annotálásánál, mint például az elképzelt vagy modális szövegkörnyezetben lévő események. Jövőbeli események vagy a feltételes történések szintén események, és ezeket is fel kell dolgozni egy vizsgálatnál.

Az *eseményekkel* kapcsolatos nyelvészeti munkák nagy része a *szituáció* (situation) típusokkal volt kapcsolatos. Ezekben az irodalmakban a *szituációkat* belső struktúrájuk alapján több szempontból csoportosították (Dowty D. , 1979), (Saeed, 1977), (Vendler, 1957), (Verkuyl H. J., 1972). Számos csoportosítás született, amelyek között a legalapvetőbbek a *statikus és dinamikus, folyamatos és időponthoz kötött* (durative, punctual), *telikus és atelikus* (telic, atelic). Ezen terület egyik legmeghatározóbb munkájában Vendler (Vendler, 1957) a predikátumokat (predicate) négy szituációs típusba sorolta: *állapotok, cselekmények, teljesítmények, eredmények* (states, activities, accomplishments, achievements). Számos nyelvészeti munka a Vendler osztályozására épült, és a kutatók, mint Dowty (Dowty D. , 1979) és Pustejovsky (Pustejovsky J. , 1991) ezt a kutatást kiterjesztve, további alcsoportokra bontotta Vendler négy osztályát. Például Pustejovsky a predikátumokat hierarchikus esemény struktúrák szerint csoportosította, azonban Vendler négy alap szituációs típusa hosszabb távon hatott, különösen a természetesnyelv-feldolgozás statisztikai kutatásaiban.

Az eseménysémák, mint pl. a cselekvés, mozgás, állapotváltozás, elhelyezkedés megnyilvánulásait a nyelvhasználó tapasztalati úton érzékeli és hasonlósági viszonyok alapján alkot meg fogalmi kategóriákat, így az eseményeket a kognitív lingvisztika, azon belül a kognitív szemantika kutatási elveibe ágyazottan kell kezelni.

Az események leírásához nem elegendő a szintaktikai szerkezetek vizsgálata. A magyar nyelvű szövegek esetében is szükség van a szemantikai, sőt részben a világ-ismereti leírásra.

A jelentés kognitív nyelvészeti megközelítésben – a kognitív szemantika

A Chomsky-féle generatív grammatika strukturális leíró módszere elégtelennek bizonyult a nyelvhasználat változatainak bemutatására. A generatív nyelvészet a nyelvet a grammatikai szabályokat követő, ideális nyelvhasználatra alapozva vizsgálja (Bańczerowski, 1994). Figyelmen kívül hagyja a nyelvi megnyilatkozások nyelven kívüli aspektusait, a megnyilatkozás kontextusának összetevőit. A kognitív nyelvészet fogalmának megalkotása George Lakofftól származik (Lakoff & Thompson, 1975). Elmélete szerint az ember egy fogalmi rendszert hoz létre, amely összefügg a valósággal és ennek képe tükröződik a nyelvben.

A lexikai egységek kialakulására vonatkozólag is más szempontból tekint Langacker. Több helyütt megfogalmazza, hogy az emberi megismerés folyamatainak modellálása során képeződnek le a fogalmi tartalom egységei. Meglátása szerint különböző módon lehet kifejezni egy adott valóságtartalom-egységet, és minden egyes megfogalmazás különálló jelentést képvisel. Egy adott kifejezés bizonyos képeket hoz létre az általa idézett tartalomra, tehát a jelentés nem más, mit fogalmi tartalom-meghatározás. A kognitív grammatika azt állítja, hogy a nyelvi kifejezések jelentésének teljes tartalma a megismerés teljes körét fedi le, vö.:

„Cognitive Grammar claims that […], a full account of the meaning of linguistic expressions would mean a full account of cognition." (Langacker 1987: 154)

A magyar nyelvészek szintén egyetértenek abban, hogy a jelentés fogalma nagyon nehezen határozható meg. Kiefer leszögezi, hogy a nyelvi jelentés szorosan összefügg a környezetünk észlelési, megismerési folyamataival (Kiefer, 2007). Meglátása szerint a kognitív szemantika a nyelvi kifejezések jelentését kognitív oldalról közelíti meg, a jelentés eszerint összefügg azzal, hogy a világ dolgait hogyan érzékeljük (percepció) és hogyan dolgozzuk fel (kogníció). A nyelvhasználó tapasztalatait mentálisan rendszerezi és kategóriákat hoz létre. Az egyes tartalmi egységek jelentése tehát az ember kategorizálási képességén alapul, de az egyes jelentések közé nem húzható éles határ. A szemantika feladata eszerint nem a jelentéskomponensek meghatározásából áll, hanem egy prototípus kijelöléséből. A prototípuselméletből indul ki Tolcsvai is vizsgálódásainál. Feltételezi, hogy egy bizonyos prototípus egy jelenség bizonyos kategóriához való kapcsolódása és a közös tapasztalat révén megállapított tulajdonságok alapján alakul ki (Tolcsvai, 2009). Központi állításának tekinthető, hogy egy prototípust bizonyos tipikalitási feltételek határozzák meg. Minél kevesebb tipikalitási feltétel teljesül, annál bizonytalanabbá válik egy szó jelentése. A prototipikus főnév *madár* kategóriában az a tipikalitási feltétel, hogy szárnya van, fontosabb, mint az, hogy repül, hiszen nem minden szárnnyal rendelkező madár tud repülni. A prototipikus ige egy eseményt időben körülhatárolt módon fejez ki, térbeli helye a résztvevők viszonylatában határozható meg. A folyamat tetten érhető az időben egymás után következő állapotok sora nyomon követésével (Langacker, 1987).

A továbbiakban az igék és a főnevek jelentéseinek leírási módjait emelem ki.

Az ige és a főnév szemantikája –– az eseményszerkezet

A jelentésleírás módozatai kiindulópontjaként jegyezzük meg, hogy minden konkrét főnév egy tárgy (Bańczerowski, 1999) vagy egy dolog (Tolcsvai, 2009) sémának, minden konkrét ige pedig egy folyamat- vagy eseménysémának egy bizonyos megnyilvánulása. Az ige cselekvés-, történés-, állapot- vagy létfogalmat fejez ki. Az állapotok és a létfogalmak az eseményszerűségek (eventualities) (Bach, 1986) között külön kategóriát képviselnek és szemben állnak a valódi eseményekkel. Az állapotok sajátosságai közül csaknem mindig a homogenitás a legjellemzőbb sajátosság, nincs eseményszerkezetük, nem különíthetők el egymástól különböző szakaszok. Elemzésemben olyan cselekvésekre és történésekre fókuszálok, amelyek eseményszerkezettel rendelkeznek.

Az igék

Az igei jelentések, azon belül az eseményszerűségek elkülönítéséhez figyelembe kell venni az aspektuális viszonyokat, azaz a vizsgált események beszédidőtől független belső, mondatban kifejezett időszerkezetét. A magyar nyelvben a legszembetűnőbb aspektuális viszony a folyamatos és a befejezett aspektus közötti különbség, mivel bizonyos igék gyakran morfológiailag is jelölik azt, pl. *Könyvet olvasott*: folyamatos aspektus, *elolvasta a könyvet*: befejezett aspektus. A folyamatos és befejezett aspektus elkülönítésére Kiefer vállalkozik:

Egy esemény akkor és csakis akkor folyamatos, ha az esemény által kifejezett cselekvés, történés vagy folyamat az adott időtartomány legtöbb résztartományára, osztatára érvényes. Egy esemény akkor és csakis akkor befejezett, ha az esemény által kifejezett cselekvés, történés

vagy folyamat csak az egész szóban forgó idő tartományra vonatkozik, más szóval, az adott időtartománynak nincs egyetlen olyan résztartománya, osztata sem, amelyre külön érvényes lenne." (Kiefer, 2007) (265.o)

A folyamatos aspektus egy külön altípusának tekinthető a progresszív altípus, ha a cselekvés folyamat jellegének megléte egy másik, általában perfektív, befejezett szemléletű esemény viszonylatában nyer hangsúlyt. A jelenséget bizonyos szintaktikai jegyek is kísérik, pl. az *amikor* kötőszó: *Amikor ebédeltünk, megszólalt a csengő.* Az *ebédeltünk* folyamatjellege nem a beszédidőben, hanem az egyszeri, pontszerű esemény – megszólalt a csengő – viszonylatában válik hangsúlyossá.

Az ige járulékos tulajdonsága továbbá az akcióminőség, pl. következő járulékos jelentések, jelentéstöbbletek: az iteratív akcióminőség (gyakorítás), pl. *küldözget*, frekventatív akcióminőség (ismétlődés), pl. *be-beugrik*, inchoatív akcióminőség (kezdet), pl. *felcsattan*.

Nem minden nyelvben különíthető el akcióminőség kategória. A franciában egyáltalán nem beszélhetünk akcióminőségről, az angolban csak korlátozott mértékben, a németben azonban már tetten érhető legalább öt akcióminőség, a magyarban tizenhárom, a szláv nyelvben számuk tizenöt fölött mozog. Az aspektualitás és akcióminőség tárgyalása túlmutatna e dolgozat keretein. Összességében azonban elmondható, hogy az aspektuális viszonyok az eseményszerkezet feltérképezésénél jelentős szerepet játszhatnak. Minthogy az aspektussal foglalkozó munkák tekintélyes része a Zeno Vendler-féle osztályozást tekinti mérvadónak (Vendler, Linguistics in Philosophy, 1967), az igei események, eseményszerűségek és nem események osztályozásánál is ezt a rendszert tekintem célravezetőnek:

- állapotok (states), pl. *birtokol, tartalmaz, érez, gyűlöl;*
- cselekvések (activities), pl. *ír, olvas, fut, úszik, főz, takarít;*
- teljesítmények (accomplishments), pl. *megír, elolvas, befut, átúszik;*
- eredmények (achievements), pl. *megérkezik, megold, rájön, elér,megtalál.*

Az angolban az állapotok és a cselekvések, illetőleg a teljesítmények és az eredmények a progresszív aspektuson alapuló vizsgálat segítségével különböztethetők meg. Az állapotok és eredmények általában nem teszik lehetővé a progresszív aspektust. Igei jelentések meghatározásához a magyarban is szükséges figyelembe venni az aspektuális tulajdonságokat, habár a magyarban a progresszív aspektus némileg másképpen viselkedik (Kuti, és mtsai., 2006).

Mindegyik eseménytípus különböző eseményszerkezettel rendelkezik. Az eseményszerkezet minimálisan az igével jelölt esemény részeseményeiből és azok egymáshoz való időbeli viszonyából áll. Az időbeli viszony többféle lehet, leglényegesebb a megelőzés és az időbeli egybeesés vagy átfedés viszonya. A *megáll* ige eseményszerkezete az időmódosítókkal való kompatibilitás alapján a következőképpen néz ki: a *megáll* esemény feltételez egy a megállást megelőző mozgáseseményt, tartalmazza a megállás pillanatnyi eseményét és a megállás utáni nyugalmi állapotot A *megír* ige eseményszerkezete két részeseményből áll, az egyik az írás folyamata, a másik az írás befejezése utáni állapot. Az *elér* ige eseményszerkezete három részeseményből áll: egy folyamatból, egy pontszerű eseményből és egy utóállapotból. A *feljajdul* ige egyetlen pillanatnyi eseményből áll, az igének nincs belső eseményszerkezete. Az eseményszerkezet ezek alapján az igével jelölt esemény részeseményeiből és azok egymáshoz való időbeli viszonyából tevődik össze. Ebből is látszik, hogy szoros összefüggés érzékelhető az eseményszerkezet és az aspektus között.

A módbeli és időbeli segédigék esetében nem beszélhetünk eseményről: pl. *tud, kell, szabad, ismer, van, volt, lesz, fog, marad,* stb. Viszont az igekötővel jelzett perfektív jelentéstartalmú ige, pl. *megismer* már eseménynek tekinthető.

A főnevek

A főnév bármilyen élőlényt, élettelen tárgyat, gondolati és elvont dolgot jelentő szó, eseményszerkezete igei szófajra történő leképezés által határozható meg. A főnév általában a következő esetekben örökli meg az alapige esemény-szerkezetét vagy eseményszerkezet nélküliségét:

- *-ás/-és* képző deverbális főnevek esetében: pl. *festés – A festés két napig tartott, két napig festettek.*
  Bizonyos helyzetekben azonban az-*ás/és* képző eredményt jelöl, pl. *írás*, mint cikk, levél vagy kész dokumentum, *mosás*, mint mosnivaló ruha.
- *-ó/ő* képzős deverbális főnevek, főként főnévvé vált melléknevek.
  Az -ó/ő képzős alakok egy része alanyi, másik része eszközhatározós szerepű, pl. aki sportol, aki ír, aki könyvel, az sportoló, író, könyvelő, amivel ásunk, amivel evezünk, az az ásó, illetve evező.
- Deverbális szóösszetételek, pl. alanyi összetételek: hóesés, kutyaharapás, tárgyas összetételek: levélírás, favágás, hóesés, rendszerváltás.
- Nem deverbális, eseményszerkezet nélküli egyszerű események pl. a *zaj, villám.*

Ige-főnév kapcsolatok – funkcióigés szerkezetek – félig kompozícionális főnév + ige szerkezetek

Az ige+főnév kapcsolatok számos nyelvészeti elemzés tárgyát képezik és jelentős szerephez jutnak a számítógépes nyelvészet különböző területein, de kezelésük nem problémamentes. A szerkezetek elnevezésében eltérések figyelhetők meg. A német szakirodalomban a *Funktionsverbgefüge* ill. *Funktionsverbformel* honosult meg Engelen és Heringer nyomán (Engelen, 1968), (Heringer, 1968). Funkcióigés szerkezeten olyan kéttagú konstrukciókat értenek, amelyekben a főnév szemantikai, az ige pedig funkcionális szintaktikai funkciót tölt be, de egy gondolategységet alkot.

Kearns hasonló eredményre jut elemzései során (Kearns, 1998). A *make the claim* konstrukció mentén mutatja be az adott kollokáció-típus szintaktikai tulajdonságait és a *Funktionsverbgefüge* (FVG)-hez hasonló *light verb construction* (LVC) terminust használ.

Az orosz nyelvészeti körökben leginkább elterjedt meghatározás az állandósult ige-főneves szerkezetek: *устойчивые глагольно-именные словосочетания* (Deribas, 1968). Ezeket a „terpeszkedő szerkezeteket” (Grétsy & Kemény, 1996) félig kompozícionális főnév + ige szerkezeteknek tekintjük, mivel Vincze is (Vincze V. , 2009) Langer nyomán (Langer, 2004) megállapítja, hogy ezek a kollokációk egyik altípusaként kezelendők, amelyekben a kifejezés szemantikai tartalmát nagyrészt a főnév hordozza, ugyanakkor az ige vállal főszerepet a szerkezet szintaxisának kialakításában.

E szerkezetek, mint egész jelentése nem határozható meg az alkotóelemek jelentéskomponenseinek összeadásával, mivel egyjelentéses lexikai egységek. Noha két különálló szövegszó kapcsolódik egymáshoz, a szótárakban nem egyszerű kollokációként, hanem egy lexikai egységként kell feltüntetni. Ezekkel az állandósult szókapcsolatokkal leképezett jelentésről úgy dönthető el, hogy esemény-e, vagy sem, hogy megvizsgáljuk a funkcióige eseményszerkeze-

tét. A *döntést hoz, szerződést köt, részt vesz* esetében az igei komponens eseményszerkezettel rendelkezik, így a teljes szerkezet eseménynek tekinthető. Külön figyelmet érdemes szentelni a félig kompozícionális főnév + ige szerkezetekből képzett főnevekre is. Az igei komponensből különböző képzőkkel képzett főnevek, mint pl. *döntéshozatal, szerződéskötés, részvétel* szintén rendelkeznek eseményszerkezettel.

A határozói igenevek (gerundiumok)
A határozói igenév az ige egyik személytelen alakja, amely szintaktikailag általában valamilyen határozói szerepet tölt be. A magyarban a végződése *-va, -ve* vagy *-ván, -vén*, más nyelvekben több különböző alakban és funkcióban is előfordulhat.
Amennyiben a határozói igenév a létige bármely alakjával együttesen jelenik meg a szövegkörnyezetben, pl. *el van intézve, az arcára volt írva, nincs megoldva*, stb., az igenév nem tekinthető eseménynek. Ha a határozói igenév igével történő átfogalmazásával a jelentés nem változik, pl. *Az elnököt idézve – az elnököt idézzük*, vagy *indulástól számítva – az indulás időpontjától számítjuk*, akkor az igenév eseménynek tekinthető.

A melléknévi igenevek (participiumok)
A melléknévi igenév az ige olyan személytelen alakja, amely szintaktikailag általában jelzői szerepet tölt be. Különböző nyelvekben többféle melléknévi igenév is létezhet, a magyarban megkülönböztethetünk folyamatos (egyidejű), befejezett (előidejű) és beálló (utóidejű) melléknévi igeneveket. A folyamatos melléknévi igenév képzője az *-ó/ő*, a befejezetté a múlt idő jelével többnyire megegyező alakú *-(o)t(t)*, a beállóé az *-andó/endő*. A határozói igenevekhez hasonlóan dönthető el a melléknevekről, hogy eseményt jelölnek-e. Szemantikájukat tekintve nem változnak a következő szerkezetek igésítve: *láttuk a lángoló házat – a házat, ami lángolt, a tegnap elmondott beszédben –a beszédben, amit tegnap elmondtak, az elolvasandó könyv – a könyv, amit el kell olvasni*.

Az *események* időbeli helyének és időbeli tulajdonságainak nyelvtani kifejezését a nyelvekben az ***igeidővel, az aspektussal és a modalitással*** (tense, aspect, modality) fejezik ki (Lyons, 1981). Ezért e tulajdonságok elemzése szükséges egy eseményekkel kapcsolatos számítógépes információs rendszer fejlesztéséhez. Vizsgáljuk meg ezen tulajdonságokat részletesebben is.

Igeidő (tense)
Az *igeidő* egy speciális szerkezet az események időbeli elhelyezkedésének kifejezésére (Comrie, 1985), (Crystal, A Dictionary of Linguistics and Phonetics, 2011). Reichenbach volt az első és egyben az egyik legnagyobb hatású közreműködő az igeidők szemantikus analízisében (Reichenbach, 1966). Reichenbach három időpont felhasználásával elkészített egy időbeli modellt az igeidők ábrázolására. Ezen időpontok a következők: Az *esemény ideje* (Event Time, E): amikor az esemény történt. *A beszéd ideje* (Speech Time, S): az esemény elmondásának az ideje. *Referencia időpont* (Reference Time, R): egy időpont, amihez viszonyítunk. Az igeidők reprezentálására Reichenbach definiált számos kapcsolatot ezen időpontok között. Például E–R–S, Befejezett múlt (Past Perfect), *Letettem a vizsgát a nyár végéig. (I had passed the exam by the end of the summer.)* Reichenbach munkája hatással volt a későbbi formális és

számítógépes nyelvészetre, különösen azokra amelyek az elbeszélések időbeli struktúráját elemezték (Webber, 1988). Mivel Reichenbach néhány relációja többértelmű (Comrie, 1985), ezért Song és társa (Song & Cohen, 1991) egyértelmű kapcsolatok újabb halmazát dolgozta ki.

Aspektus (aspect)
Az *aspektus* a mondat belső időszerkezete (Kiefer, 2006). A *lexikális aspektus* az eseményeknek a következő időbeli tulajdonságaival foglalkozik: *dinamikusság, telikusság és folyamatosság* (Vendler, Linguistics in Philosophy, 1967), (Dowty D. , 1979). Moens és társa (Moens & Steedman, 1988) készítettek egy Vendler osztályozásán alapuló esemény ontológiát.

Időbeli következés (Temporal reasoning)
Az *időbeli következés* magában foglalja az *események* és azok *időbeli kapcsolatainak* időbeli ábrázolását. Az idővel kapcsolatos hatékony számítógépes modellek fejlesztése a mesterséges intelligencia (AI) fontos területe volt az 1960-as évek óta. Allen *Interval Algebra* című munkája (Allen J. , 1983) az egyik legnagyobb hatású ebben a témában, modelljében az *eseményeket időintervallumokkal* ábrázolta. Allen két intervallum között tizenhárom lehetséges kapcsolatot definiált: egyenlő (=), előtt (<), után (>), kezdődik (s), folyamán (d), befejeződik (f) átfedi (o), átfedve a másik által (oi), stb.

A beszélgetés időbeli struktúrája (Temporal Structure of Discourse)
A nyelv helyes megértéséhez szükség van a beszélgetés szövegkörnyezetének ismeretére is. Ebbe tartozik az *események időbeli sorrendje* is. A *beszélgetés időbeli struktúrájával* kapcsolatban a következő témákban jelentek meg jelentős munkák:

- időbeli párbeszéd értelmezési alapelvek (Dowty D. R., 1986),
- tanulmány a párbeszéd anaforáról (Webber, 1988),
- igeidő fák (Hwang & Schubert, 1992),
- referencia pont (Kamp & Reyle, 1983),
- dinamikus aspektus fák (Meulen, 1995),
- időbeli konceptuális gráfok (Moulin, 1997).

Lascarides és társa (Lascarides & Asher, 1993) rámutatott, hogy a valós világról alkotott ismeret szükséges az időbeli kapcsolatok helyes értelmezéséhez. A statisztikai kutatás egyik legkorábbi példája Siegel és társától származik (Siegel & McKeown, 2000). Munkájukban angol szövegek igéit vizsgálták és manuálisan jelölték mindegyikről, hogy statikus vagy dinamikus. Egy másik munkájukban az igék mellett jelölték, hogy telikusak vagy nem. Ezt az annotált korpuszt használták később statisztikai gépi tanulásos vizsgálatokhoz.
A nyelvészeti kutatásoknak és elméleteknek ma is nagy hatása van a természetesnyelv-feldolgozás közösségére. 2002-ben kezdték el a már bemutatott TimeML nyelvészetileg motivált kutatási projektet (Pustejovsky, és mtsai., 2003), aminek a célja *időbeli és esemény kifejezésekhez* egy jelölő nyelv kidolgozása volt. A TimeML egy gazdag és összetett specifikációs nyelv természetes nyelvű szövegekben található események, időbeli kifejezések és a közöttük lévő kapcsolatok annotálásához. Jelenleg a TimeML-t használja a kutatók többsége időbeli információk annotálására.

A TimeML annotációs irányelveket felhasználva Pustejovsky és kollégái annotálták a 183 dokumentumot tartalmazó TimeBank korpuszt 7935 eseménnyel és 6418 időbeli kapcsolattal (Pustejovsky, és mtsai., 2006). Később számos kutató használta a korpuszt események számítógépes detektálásának kiértékelésére.
A TimeML sémát eredetileg angol nyelvre fejlesztették ki. Később számos nyelvre elkészítették azt, mint például a francia (Bittar, 2009), olasz (Caselli, 2009), koreai (Im, You, Jang, Nam, & Shin, 2009), román (Forascu, 2008), portugál, török (Seker & Diri, 2010), spanyol (Sauri R. , Tempeval 2. spanish data release., 2010) és kínai (Xue & Zhou, 2010) nyelvekre.

Magyar szerzők eredményei
Az *események* nyelvészeti elemzésének lehetőségét megalapozta a szövegek jelentésének, szemantikájának részletes vizsgálata. Ezen a területen Kiefer Ferenc, az általános nyelvészet és a nyelvi szemantika kutatója ért el jelentős eredményeket (Kiefer, 2006). Leszögezte, hogy a nyelvi jelentés szorosan összefügg a környezetünk észlelési, megismerési folyamataival. Definíciót adott a folyamatos és befejezett aspektus elkülönítésére (Kiefer, 2007), megállapítja, hogy az aspektus a mondat belső időszerkezete (Kiefer, 2006).
Kuti elemzésében megállapítja, hogy az igei jelentések meghatározásához a magyarban is szükséges figyelembe venni az aspektuális tulajdonságokat, habár a magyarban a progresszív aspektus némileg másképpen viselkedik (Kuti, és mtsai., 2006).
Tolcsvai vizsgálódásainál a prototípuselméletből indul ki. A prototípusok kialakulásáról feltételezi, hogy azok egy jelenség kategóriához való kapcsolódása és a közös tapasztalat révén keletkező tulajdonságok alapján alakulnak ki (Tolcsvai, 2009). Központi állításának tekinthető, hogy bizonyos tipikalitási feltételek határoznak meg egy prototípust.

2.2 Szemantikus szerepek

A *szemantikus szerepek* eredete a szanszkrit Panini (i.e. 4. század) mélyeset (karaka) elméletéből származik. Panini hat mélyesetet (szemantikus kapcsolatok az ige és a főnevek között) különböztet meg: Agent, Goal, Recipient, Instrument, Locative, Source (Kiparsky, 2002).
Az azóta eltelt időben sokat tanulmányozták a szemantikus szerepeket. A modern vizsgálatok közül az elsők az 1960-as években történtek (Gruber J. S., 1965), (Fillmore C. , 1968). Ettől kezdve sokfajta szerepcsoportot definiáltak. Léteznek ajánlások néhány absztrakt, általános szerepre (például agent, patient) és sok speciális szerepre is (például utazó, utazási eszköz, utazási idő, távolság). A különböző szerepeket definiálni tudjuk a részletesektől kezdve (például utazó), a csak két elemet tartalmazó proto szerepig: proto-agent és proto-patient (Dowty D. , 1991).
Manapság az NLP-ben a legáltalánosabban használt szemantikus szerepcsoportokat a FrameNet projektben (Johnson, Fillmore, Petruck, & Baker, 2002), (Ruppenhofer, Ellsworth, Petruck, & Johnson, 2005) és a Proposition Bank projektben (Palmer, Gildea, & Kingsbury, 2005) definiálták.

Szemantikus kapcsolatok és szemantikus szerepek

A nyelvészek sokszor megkísérelték rekonstruálni a természetes nyelvek felépítését. Egy tipikus felosztás szerint a két fő összetevő a nyelv struktúrája (szintaxis) és a nyelv jelentése (szemantika). A szemantikus kapcsolatok az egyik legrégebbi osztályai a *nyelvészeti elmélet*nek (linguistic theory), ami első kísérletként próbált kapcsolatot teremteni a morfológia és a szemantika között. A *nyelvészeti elmélet* a több ezer éves *Panini karaka* elméletére tekint vissza (Misra & Niwas, 1966), (Rocher, 1964), (Dahiya, 1995).

A nyelvészek kidolgozták a *szemantikus szerepek* elméletét, ami egy kapcsolódási pont a szintaxis és a szemantika között (Belletti & Rizzi, 1988), (Sciullo & E., 1987), (Levin & Rappaport, 1986), (Marantz, 1984), (Rappaport & Levin, 1988), (Williams, 1981), (Zubizarreta, 1987). A *szemantikus szerep* egy predikátum és az argumentuma közötti kapcsolatot írja le.

A szemantikus elemzés egy fontos közbülső lépés a természetes nyelvek megértéséhez. Az utóbbi években a gépi tanulás gyors fejlődésével és a szintaktikai elemzés javulásával növekvő igény támadt mélyebb és szélesebb körű szemantikus adatábrázolásra és annotációra. A szemantikus reprezentáció kialakításánál szemantikus egyedeket kell azonosítani, egyedek és predikátumok között kell kapcsolatokat megnevezni. A nyelvészek között sincs teljes egyetértés a szemantikus szerepek azonosításával kapcsolatban (Baker M. , 1988), (Dowty D. R., 1989), (Fillmore C. , 1968), (Gruber J. , 1967), (Jackendoff R. , 1987), (Marantz, 1984), (Nishigauchi., 1984), (Rappaport & Levin, 1988), (Riemsdijk & Williams, 1986), (Rozwadowska, 1988), (Talmy, 1985). A *szemantikus szerepek címkézése* (SRL, Semantic Role Labeling) a természetesnyelv-feldolgozás azon feladata, ami detektálja egy mondat predikátumának a szemantikus argumentumait és osztályozza ezeket speciális szerepek szerint.

A szintaxis és a szemantikus szerepek közötti kapcsolat.

A *linking* argumentum-realizációs elmélet (linking theory) is foglalkozik a szemantikus szerepekkel. A *linking* elmélet egy nyelvtani elmélet, ami leírja a kapcsolatot szintaxis és a szemantika között. A szemantikus szerepek által adunk jelentést a szintaktikai összetevőknek. A *linking* elmélet központi kérdése, hogy ezek a szerepek hogyan következnek a szintaxisból. A mai modern elemző programok képesek pontosan kinyerni a szintaxist a szövegekből, ez az utóbbi időkben hozzájárulhatott a szemantikus szerepek címkézésével kapcsolatos megújult érdeklődéshez.

A természetes nyelvek megértésének két kiemelt feladata a *szintaktikai függőségi elemzés* és a *szemantikus szerepek címkézése*. Ezek közeli kapcsolatban vannak egymással és úgy tekinthetjük őket, mint a mondat felső szintű elemzése. Ezeket a részfeladatokat általában egymás után következő osztályozókkal oldják meg. A szintaktikai elemző fut első lépésben, majd az adott predikátumhoz a szemantikus szerepeket azonosítják és osztályozzák (Gildea and Jurafsky, 2002). Léteznek olyan elemzők is amelyek a nyelvi szinteket egyszerre elemzik (joint learning), illetve elemezetlen szövegből közvetlenül, egy lépésben frame-szemantikai reprezentációt előállító elemzők is egyben.

Ezen feladatokban való előrehaladásból profitálhat a természetesnyelv-feldolgozás számos területe. Az utóbbi években több rendszer született, ami a függőségi reprezentációt és az SRL eredményét használta fel. Ilyenek például az automatikus összegzés (Melli, Shi, Wang, Liu, Sarkar, & Popowich, 2006), válaszkeresés (Narayanan & Harabagiu, 2004), információkinye-

rés (Surdeanu M. , Harabagiu, Williams, & Aarseth, 2003), kereszthivatkozások azonosítása (Kong, Li, Zhou, Zhu, & Qian, 2008), (Marquez, Recasens, & Sapena, 2013) és gépi fordítás (Boas, 2002).
Tehát a predikátum-argumentum kapcsolatok szoros összefüggésben vannak a mondat szintaktikai struktúrájával. A szemantikus szerepek az absztrakció egy fontos szintjét képviselik. Míg az argumentumok szintaktikai funkciói változnak az esemény formája szerint (például cselekvő vagy szenvedő szerkezeti formák), az argumentumok szemantikus szerepei változatlanok maradnak.

Nézzük például a következő mondatokat:
(a) Róbert kitöltötte a papírokat
(b) A papírok ki lettek töltve Róbert által.
(c) Ezek azok a papírok, amelyeket Róbert kitöltött.
(d) Ezek azok a papírok, amelyek ki lettek töltve Róbert által.
(e) Ezek azok a papírok, amelyeket Róbert elfelejtett a kitöltés után.

Ezekben a mondatokban a *kitölt* predikátum és a *papírok* argumentum ugyanaz marad, annak ellenére, hogy a nyelvtani kapcsolat változik. Az (a) mondatban a szerep tárgy pozícióban van. A (b) szenvedő mondatban ez alany pozícióban van. A szemantikus kapcsolat nem változik akkor sem, ha az argumentum a mellékmondat feje (c,d), vagy a még összetettebb (c) esetben, ahol az argumentum távol van az predikátumtól.

Szemantikus szerep típusok
Történetileg két fő típusa alakult ki a *szemantikus szerepeknek*.

A, Domén független szemantikus szerepek
Kezdetben domén független általános szerepeket definiáltak. A szemantikus reprezentáció legegyszerűbb formájánál csak két szerepet adtak meg: Proto-Agent és Proto-Patient (Dowty D. , 1991), de a legtöbb elmélet ennél több szerepet használ. Példák ilyen domén független szerepekre: ágens (agent), elszenvedő (patient), téma (theme), cél, forrás, tapasztaló (experiencer), eszköz, idő, hely. Ezek a szerepek általános szemantikai fogalmat rendelnek a predikátumokhoz. Például a cselekvő predikátum alanya mindig az ágens (agent). Sok fajta hasonló szemantikus szerepet definiáltak. Például Liu és társa (Liu & Soo, 1993) 13 szerepet, míg Rosa és társa (Rosa & Francozo, 1999) 23 szerepet definiált.

B, Domén-specifikus szemantikus szerepek – Keretek
Sok információ kinyerő rendszer ezzel ellentétben *domén-specifikus* szerepeken alapul. Egy repülőgép helyfoglalásával foglalkozó rendszer például a következő szerepeket használja: *indulási állomás, érkezési állomás, indulási idő*, stb (Stallard, 2000). Vállalat felvásárlásokkal foglalkozó rendszerek például a következő szerepeket használják: *kapcsolat, mennyiség, vásárló, eladó*.
Nem gazdaságos ha szerepeinket csak egy doménen tudjuk használni, hiszen vannak olyan szerepek, amelyek hasonló doménekhez is kapcsolhatóak. Ehhez alkották meg a szemantikus keretek (frames) fogalmát, amelyek nem kívánnak új szerephalmazt minden új domén esetén,

a hasonló doméneket csoportosítva kezelik. Ezeket a szerepeket a szemantikus keretek szintjén határozzuk meg (Fillmore C. J., 1976). Egy keret helyzetek szemantikus ábrázolása számos résztvevővel és szerepekkel (Fillmore C. J., 1976). Az ÁTRUHÁZÁS kereten belül például a *küld* és a *kap* igékhez ugyanazok a szemantikus szerepek tartoznak. (küldő, fogadó, termékek, ár, stb.). A szemantikus szerepek definiálása ezen a közbülső keretszinten segít elkerülni azokat a nehézségeket, amelyek a domén-független túl általános szerepekkel járnak. Valamint segítenek egy adott kereten belüli általánosításra.

Példamondat a szemantikus szerepekre.
[A svéd Ericsson] $_{Eladó}$ *bejelentette, hogy [a német Infineonnak]* $_{Vevő}$ ***adja el*** *[chipgyártó részlegét]* $_{Áru}$ *, [400 millió euróért]* $_{Ár}$ *.*

Szemantikus szerepek címkézése

A *szemantikus szerepek címkéit* a sekélyes szemantikus elemzéssel kapjuk meg. A sekélyes szemantikus elemzés feladata megtalálni egy mondat egyszerű szemantikai struktúráját. Például „Ki tette?, Mit tett?, Kinek?, Mikor?, Miért?, Hogyan?" (Pradhan S. , Ward, Hacioglu, Martin, & Jurafsky, 2004). Ez foglalja össze a mondat alapvető jelentését. A sekélyes szemantikus elemzés a jelentésnek egy magasabb szintű megértését nyújtja, mint a hagyományos szintaktikai elemzés.

A *szemantikus szerepek címkézésének* a célja megtalálni egy mondatban a predikátumhoz minden argumentumot a helyes szóhatárokkal (argumentum azonosítás), majd osztályozni ezeket a szemantikus szerepek szerint (argumentum osztályozás). Egy predikátum argumentuma helyes akkor és csak akkor, ha a szóhatárok és a címkék is helyesek. Ezeket az osztályozásokat Fillmore (Fillmore C. , 1968) és Jackendoff (Jackendoff R. , 1975) ajánlotta a szemantikai argumentumok szintaktikai megvalósulásának bemutatásához.

A *szemantikus szerepek címkézése* egy konverzió a szintaktikai és a szemantikai struktúrák között. Két lépést tartalmaz: a szintaktikai struktúrát olyan szemantikai struktúrává alakítani, ami még nem címkézett, majd kitölteni ezt a struktúrát szemantikus szerepekkel. A szemantikus szerepek címkézése a jelentésnek egy magasabb szintű megértését adja és kevésbé nyelvfüggő, mint a szintaktikai elemzés.

Egy SRL rendszer bemenete egy mondat és a mondat predikátuma, kimenete a szemantikus szerepekkel címkézett mondat.

A szemantikus szerepeket többféle módon ábrázolhatjuk. Ezekre néhány példa:

Példa egyszerű szerepnevekkel

A Bell Atlantic Corp. jövőre felvásárolja a Control Data Corp. számítógép-javító üzletágát New Yorkban.

predikátum: *felvásárolja*
Ki: *Bell Atlantic Corp.*
Mikor: *jövőre*
Hol: *New York*
Mit: *a Control Data Corp. számítógép-javító üzletága*

Példa bővebb szerepnevekkel

Melyik Lufthansa repülő érkezik Budapestre 10 óra után?

Cél: repülő
Célállomás neve = *Budapest*
Légiközlekedési Vállalat = *Lufthansa*
Érkezési idő = *10 óra*
Érkezési idő - relatív = *után*

Példa táblázatos ábrázolással (2.1. táblázat**)**

Tüntetők egy csoportja köveket dobott a kivezényelt rendőrökre a belváros közelében.

Típus	Támadás	
Trigger	dobtak	
Argumentumok	Argumentum szerep	Előfordulás
	Támadó	tüntetők egy csoportja
	Támadott	a kivezényelt rendőrök
	Eszköz	köveket
	Hely	a belváros közelében

2.1. táblázat: Szemantikus szerepek táblázatos ábrázolása

Magyar szövegeken elért eredmények

Kiefer Ferenc foglalkozott tematikus szerepekkel, definiálta is ezeket: „*A vonzatok viselkedését azonos módon behatároló fogalmi szerepeket ugyanazon általánosabb szereptípus képviselőinek tekinthetjük, s ezeket a szereptípusokat tematikus szerepeknek nevezzük*" (Kiefer, 1992). Megjegyzi, hogy a *tematikus szerepekből következtetni lehet a szintaktikai funkciókra*. Valamint, hogy *a szintaktikai funkciójukat tekintve az NP-k (a vonzatok) nem egységesek, vannak közöttük alanyok, tárgyak, határozók, és ezeket a funkciókat a régens rendeli hozzájuk.*

Makrai megállapítja, hogy *az argumentumok szemantikai szerepe (pl. ágens) és szintaktikai tulajdonságai között rendszeres, és több esetben különböző nyelvekben is felbukkanó megfelelések vannak* (Makrai, 2015).

2.3 Összegzés

Az események és azok szemantikus szerepeinek számítógépes feldolgozásának ismertetése előtt ebben a fejezetben az események fogalmáról és a nyelvészeti megközelítési módokról írtam. Bemutattam az események és a szemantikus szerepek megközelítését nyelvészeti szempontból.

3 Esemény és SRL korpuszok

A szintaktikai elemzés felhasználása lehetővé tette olyan alkalmazások készítését, amelyek annotált korpuszokon alapuló gépi tanulásos módszereket valósítottak meg (Charniak, 2000), (Collins, 2003). Jelenleg számos ilyen korpuszt lehet felhasználni. Például: TimeBank (Pustejovsky, és mtsai., 2006), PropositionBank (Palmer, Gildea, & Kingsbury, 2005), FrameNet (Baker, Fillmore, Lowe, & B., 1998), VerbNet (Schuler, 2005) és az Interlingua Annotation of Parallel Corpora (Dorr, Farwell, Green, Habash, & Helmreich, 2004).
Ebben a fejezetben először áttekintem legfőbb angol nyelvű nyelvészeti erőforrásokat, majd ismertetem a kutatásomban is felhasznált magyar nyelvű nyelvészeti adatbázisokat.

3.1 Angol nyelvű nyelvészeti erőforrások

3.1.1 Nyelvészeti adatbázisok az események detektálásához - TimeML, TimeBank

A *TimeML* specifikációs nyelvet azért készítették, hogy segítségével olyan gépi tanulásos alkalmazásokat lehessen készíteni, amelyek időkkel és eseményekkel kapcsolatos kifejezéseket képesek azonosítani (Pustejovsky, és mtsai., 2003). A TimeML az *eseményeket* úgy tekinti, mint „egy gyűjtő fogalom szituációkra, amelyek történnek, vagy bekövetkeznek”.
A *TimeBank* (Pustejovsky, és mtsai., 2006) egy olyan annotált nyelvészeti erőforrás, ami a TimeML specifikációs nyelv alapján készült. Időkifejezéseket, eseményeket és időbeli kapcsolatokat jelöl 183 hír cikkben. Tanító és kiértékelő adatbázist tartalmaz a gépi tanulásos kutatás számára.
A *TimeML események* kifejezhetőek igékkel, főnevekkel, melléknévi igenevekkel, állítmányi mellékmondatokkal és elöljárós kifejezésekkel. A TimeML annotációs szabvány a következő alap XML címkéket definiálja az annotáláshoz: **események:<EVENT>**, idők: <TIMEX3>, időpontok közötti kapcsolatok:<TIMEX3>, <TLINK>, <SLINK> és <ALINK>. Ezen XML címkék gyakorisága a TimeBank-ban (3.1. táblázat):

XML címkék	Gyakoriság
EVENT	7 935
TLINK	6 418
TIMEX3	1 414
SLINK	2932
SIGNAL	688
ALINK	265
Összesen	19652

3.1. táblázat: Az XML címkék gyakorisága a TimeBank-ban

A TimeML ezen címkék mellett osztályokat is jelöl. A TimeML az *eseményekre* (EVENT) a következőosztályokat tartalmazza:
OCCURRENCE: valami, ami történik, vagy bekövetkezik. *die, crash, build, merge, sell*
STATE: olyan eset, amiben valami igaznak érvényes. *kidnapped, love, on board*

REPORTING: Egy személy vagy szervezet bejelentése, kinyilvánítása, informálása egy eseményről. *say, report, announce*
PERCEPTION: olyan esemény, amelyek egy másik esemény fizikai észlelést jelenti. *See, hear, watch, feel*
ASPECTUAL: olyan predikátum, amelyik más esemény aspektuális szemléletére utal. *begin, finish, stop, continue.*
I_ACTION (intensional action): olyan cselekvés, amiből következtetni tudunk valamire egy másik eseménnyel kapcsolatban. *attempt, try, promise, offer*
I_STATE: olyan állapotokat tartalmaz, ami más eseményekre utalnak. *believe, intend, want*

Ezen eseményosztályok megoszlása a TimeBank-ban (3.2. táblázat):

Esemény osztály	**Gyakoriság**	**Arány**
Occurence	4215	53,1%
State	1117	14,1%
Reporting	1028	13,0%
I_Action	681	8,58%
I_State	584	7,36%
Aspectual	262	3,3%
Perception	48	0,6%
Összesen	**7935**	

3.2. táblázat: Az eseményosztályok megoszlása a TimeBank-ban

Az események szófajainak megoszlása a TimeBank-ban (3.3. táblázat):

Esemény szófaj	**Gyakoriság**	**Arány**
Ige	5122	64,5%
Főnév	2225	28,0%
Egyéb	299	3,7%
Melléknévi igenév	266	3,3%
Elöljárós kifejezés	28	0,3%
Összesen	**7940**	

3.3. táblázat: Az események szófajainak megoszlása a TimeBank-ban

TimeBank annotációs példák:
All 75 passengers<EVENT eid="e1" class="OCCURRENCE" tense="past" aspect="NONE">*died*</EVENT>.
<TIMEX3 tid="t1" type="DURATION" value="P2M" temporalFunction="false">*Two months* </TIMEX3>*before the attack …*
<TIMEX3 tid="t2" type="DATE" value="2002-07-08" anchorTime="t0" temporalFunction="true" temporalFunctionID="tf1">*2 days before yesterday*</TIMEX3>

A TimeBank hivatalos **annotátorok közötti egyetértést** mutató számai (3.4. táblázat) (Boguraev, Pustejovsky, Ando, & Verhagen, 2007) felfedik, hogy az események annotálása nem egyszerű és egyértelmű feladat még embereknek sem.

Címke neve	Egyetértés
EVENT	*0,78*
TIMEX3	0,83
SIGNAL	0,77
ALINK	0,81
SLINK	0,85
TLINK	0,55

3.4. táblázat: Annotátorok közötti egyetértés TimeBank

3.1.2 Korpuszok a szemantikus szerepek címkézéséhez

Ha rendelkezésre állnak nagyméretű annotált korpuszok predikátumok argumentumaihoz, akkor tudunk készíteni gépi tanulásos módszereket felhasználó, szemantikus kapcsolatokat azonosító alkalmazásokat.

A szemantikus szerepekhez több számítógép-orientált korpusz is készült. Ezek közül a jelentősebbeket mutatom be: A *Berkeley FrameNet Project* (Baker, Fillmore, Lowe, & B., 1998) a keret(frame) szemantikán alapul, ami arra épít, hogy egy predikátum és argumentumainak jelentése döntően függ annak a szövegkörnyezetnek az ismeretétől, amelyben azokat használjuk. Ez más szóval azt jelenti, hogy nem tudjuk megérteni a *vesz, elad, bérel* szavak argumentumainak jelentését anélkül, hogy tudnánk, hogy hogyan működik a kereskedés. (Próbáljuk meg elmagyarázni ezeket a szavakat egy olyan személynek, aki egész életében egy pénzmentes közösségben élt). Ennek a megközelítésnek a hátránya, hogy csak az elkészített doméneken belül használható. A *Propbank project* (Palmer, Gildea, & Kingsbury, 2005) nem foglalkozik a predikátumok közötti kapcsolatokkal. Ehelyett minden predikátum minden jelentéséhez külön keres szerepeket. A *VerbNet* (Schuler, 2005) az angol igéket rendszerezi szintaktikai és szemantikai jellemzők alapján. Ehhez hasonló a *Nombank* (Meyers, Reeves, Macleod, Szekely, Zielinska, & Young, 2004), ami főnevekhez annotál szemantikus argumentumokat a Wall Street Journal korpuszon.

A következőkben ezeket a korpuszokat jellemzem részletesen.

FrameNet

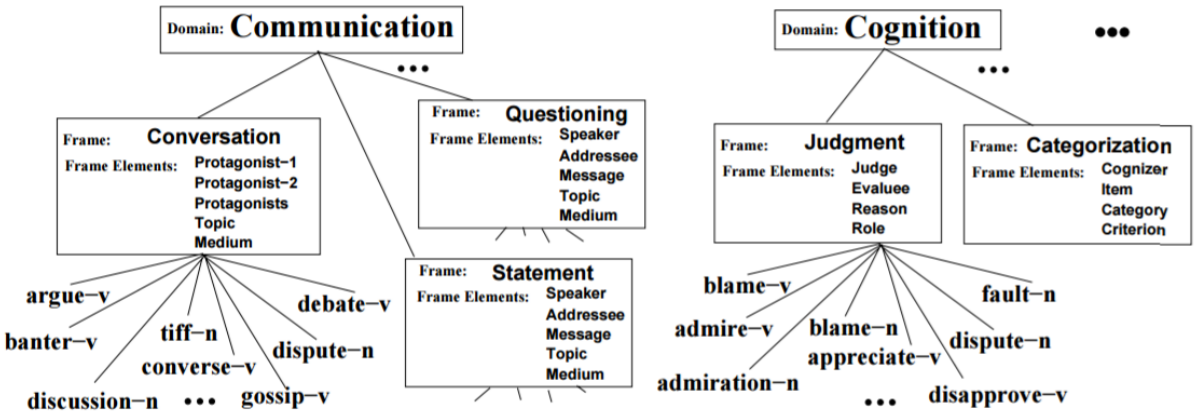
A FrameNet project (Baker, Fillmore, Lowe, & B., 1998) olyan szemantikus szerepeket használ, amelyek nem olyan általánosak, mint az absztrakt szerepek, és nem olyan specifikusak, mint az ige-specifikus szerepek ezrei. A FrameNet szerepek a szemantikus keretekhez lettek definiálva. Például az *utazás* keret olyan szemantikus szerepeket tartalmaz, mint *utazó, utazási eszköz, útvonal, időtartam, ár* (keret elemek) és olyan predikátumokkal állnak kapcsolatban, mint *utazik, kirándul, vándorol*. Ezeknek az igéknek hasonló argumentumaik vannak, ezért van értelme együtt kezelni ezeket.

A FrameNet a következő fő részeket tartalmazza:

- *Lexikai egység adatbázis*: szó-keret párok. Egy lexikon, ami minden szó jelentéséhez megad egy keretet.
- *Keret adatbázis*: keretek halmazát tartalmazza, szerepekkel (keretelemekkel) és a keretek közötti kapcsolatokkal kiegészítve. Olyan keretek közötti kapcsolatokkal, mint az öröklődési-, rész-, ok-kapcsolat.
- *Mintamondat adatbázis*: Hasznos tanító halmaz címkézéshez a British National Corpusból (könyvek, novellák, hirdetések, stb.).

A FrameNet közel 800 szemantikus keretet tartalmaz, több mint 10000 célszóval, 135000 annotált példamondattal kiegészítve. A fő domének alatt aldoméneket találunk. Például a TRANSPORTATION domén közvetlen „gyerekei" a DRIVING és a RIDING. A gyerek domén örökli a szülő domén szerepeit és lehetnek ezeken kívül más szerepei is. A Framenet Project annotál igei, főnévi predikátumokat is.

Példa doménekre és keretekre a FrameNet lexikonból (3.1. ábra).



3.1. ábra: Minta domének és keretek a FrameNet lexikonból

Ábra forrása: (Gildea & Jurafsky, 2002)

A minta alapján a Párbeszéd (Conversation) keret a következő igékhez kapcsolódik: *argue*, *banter*, *debate*, *converse*, és *gossip*, és a következő főnevekhez: *dispute*, *discussion*, és *tiff*. Olyan szituációkhoz tartozik, amikor emberek beszélnek egymással.
A szerepek ehhez a kerethez: Szereplő1, Szereplő2, vagy egyszerűen Szereplők (Protagonist1, Protagonist2, Protagonists), Téma (Topic), Közvetítő eszköz (Medium).

A keret egy példamondata:
[Mary$_{\text{Prot-1}}$] and [Peter$_{\text{Prot-2}}$] **argued** [angrily$_{\text{Manner}}$] [over who was the real "Prince of Sleaze"$_{\text{Topic}}$].

PropBank (University of Colorado)
A PropBank (Palmer, Gildea, & Kingsbury, 2005) vagy PropositionBank a szemantikus szerepeket igénként adja meg: minden predikátum minden jelentéséhez külön keres szerepeket.

Az igék argumentum struktúrájára fókuszál, ezért *ige-orientált* forrásnak tekinthető (Palmer, Gildea, & Kingsbury, 2005), főnévi eseményekkel nem foglalkozik. A Penn Treebank II korpusznak a Wall Street Journal részét annotálja (Marcus, Santorini, & Marcinkiewicz, 1994), 3 256 különböző igére 112 000 szerepet annotál.
A Propbank a szerepeket a következő módon csoportosítja: Az argumentumok első típusa (mag argumentumok, sorszámozott argumentumok) *ige-specifikus* szerepeket tartalmaz ARG0-tól ARG5-ig jelölve. Általában az Arg0 az ágens (agent) az Arg1 pedig a téma (theme). Az ARG2-ARG5 argumentumok használata igénként különböző. Például az „eszik" igénél az ARG0 az evő, ARG1 a dolog, amit megeszik. A „vesz" igénél az ARG0 a vevő, ARG1 a dolog, amit vesz, ARG2 az eladó, ARG3 a fizetett ár.
Az argumentumok másik típusai (segéd argumentumok) nem jellegzetesek egy adott igéhez, sokfajta igéhez tartozhatnak. Ezek jelentése minden igére megegyezik (ige módosítók). Példák:ARGM-TMP: *idő,* ARGM-LOC: *hely,* ARGM-CAU: *ok,* ARGM-DIR: *irány*, ARGM-NEG: *tagadás*,

Példa igék leírására a Propbankban:
Az accept ige 01-es jelentése: "szívesen vállalja"
Frameset accept.01 "take willingly"
Arg0: Acceptor
Arg1: Thing accepted
Arg2: Accepted-from
Arg3: Attribute
Example: [He]$_{Arg0}$ [would]$_{ArgM-MOD}$ [n't]$_{ArgM-NEG}$ **accept** [anything of value]$_{Arg1}$ [from those he was writing about]$_{Arg2}$.

VerbNet
A VerbNet (Kipper, Dang, & Palmer, 2000) egy szemantikus-szerep lexikon, ami Levin elméleti munkáján alapul (Levin B. , 1995). Csak igékkel foglalkozik és az igék hierarchikus osztályokba vannak sorolva. 5800 igét és 270 ige-osztályt tartalmaz. Néhány VerbNet szerep: Ágens (Agent), Téma(Theme), Ok(Cause), Forrás(Source), Cél(Destination), Hely(Location), Eszköz(Instrument), Idő(Time), Kedvezményezett(Beneficiary).

NomBank (New York University)
A NomBank (Meyers, Reeves, Macleod, Szekely, Zielinska, & Young, 2004) egy annotációs projekt, amely kapcsolatban áll az előzőleg bemutatott PropBank-al. A NomBank célja a *főnévi események* argumentumainak címkézése. A két korpusz szerkesztői összehangolták, hogy a szerepek definíciói azonosak legyenek mindkét korpusznál. Például a Propbank a *decide* igére hasonló argumentumokat használ, mint a NomBank a *decision* főnévre (Gerber, Chai, & Meyers, 2009). A NomBank és a PropBank közötti kompatibilitás azért fontos, mert az igei és főnévi eseményekhez általában ugyanazok a szerepek tartoznak.

3.2 Korpuszok a magyar szövegekhez

A következő részben bemutatom azokat a korpuszokat, amelyeket a disszertációban ismertetett alkalmazásokhoz felhasználtam.

3.2.1 Ige és főnévi igenévi események detektálása és osztályozása

A következő szövegeket az *ige és főnévi igenévi események detektálására és osztályozására* használtam fel. Ehhez a Szeged Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, Hungarian Dependency Treebank, 2010) egy részét használtam, amely 5000 mondatot tartalmaz a következő doménekről: *üzleti rövidhírek, szépirodalom-fogalmazás, számítógépes szövegek, újsághírek, jogi szövegek.* Mind az öt doménről 1000 mondatot választottam ki. A mondatokat két annotátor annotálta, az annotátorok közötti egyetértés a detektálásra 87%-os, az osztályozásra 81%-os volt (ilyen százalékban jelölték azonosan a jelölteket). Detektálásnál a jelöltek az igék és a főnévi igenevek voltak. Osztályozásnál a már detektált eseményeket osztályoztam több szempont szerint.
Az 5000 mondat 10 628 igét és főnévi igenevet tartalmazott, ezek voltak az esemény jelöltek. Az annotátorok ezek közül 6479-et jelöltek eseménynek.

A, Ige és főnévi igenévi események detektálása
Példák:
Végre rákerült a sor, ***bement*** *és 15 perc múlva* ***kijött*** *mosolyogva.* (események)

Azonban nem minden ige és főnévi igenév esemény (például segédigék), így speciális kezelés szükséges ezek kiválogatásához.
Ezt ***akartam*** *neked mondani.* (nem esemény)

További példák a többértelműségre: ***dob*** (ige, főnév), ***vár*** (ige, főnév).

Annotációs példák:
<E>: esemény, *<NE>*: nem esemény

Ha <E>elvégzem</E> ezt az iskolát, akkor még tovább <NE>szeretnék</NE> <E>tanulni</E> , de ez még <E>elválik</E>.

<NE>Megpróbálom</NE> úgy <E>nevelni</E>, hogy <NE>szeressen</NE> <E>tanulni</E>, <NE>szeresse</NE> az embereket és ha <NE>lehet</NE> minél kevesebb előítélettel <NE>lássa</NE> a világot .

B, Ige és főnévi igenévi események osztályozása
Az igei események detektálása után **osztályoztam** azokat. Az osztályozást több szempont szerint is elvégeztem. Az *első csoportnál* az igék alapkategóriáit vizsgáltam meg: cselekvés, történés, létezés, állapot. Ezek közül az eseményeknél a **cselekvésnek és a történésnek** van fő szerepe, így ezt a két kategóriát emeltem ki. Az 5000 mondaton belül a 6479 esemény között 4158 cselekvés és 1752 történés típusú esemény volt.

A **cselekvést jelentő ige** olyan tevékenységet nevez meg, amely az alany akaratától függ.
A **történést kifejező ige** olyan változás, folyamat megnevezésére szolgál, amely független az alany akaratától.

Példák
Cselekvés: *Négy lovas harcos **közeledett** felém.*
Történés: *Hát igen, de nekem megbicsaklódott a lábam és a második lépcsőfokról **leestem**.*

Annotációs példák:
<*CS*>: cselekvés, <*T*>: történés

Arról, hogy postán <CS>küldjön</CS> neki levelet, egyszerűen szó sem lehetett .
Ha egy közönséges ember egy kapitalistával <CS>beszélt</CS>, alázatosan meg kellett<CS>hajolnia</CS> előtte , le kellett <CS>vennie</CS> a kalapját , és uramnak kellett<CS>szólítania</CS> .

A következő pillanatban iszonyú csattanás <T>hallatszott</T>.
A helyiség túlsó végében folyó célbadobós játék is <T>félbeszakadt</T> egy fél percre .

A cselekvés és a történés kategóriák együtt lefedik az események legnagyobb részét. Modellemet, az előző osztályozástól függetlenül, olyan kategóriákon is teszteltem, amelyek ezeknél jelentősen kevesebb elemet tartalmaznak, de még gyakoriak. Így a következő vizsgálathoz kiválasztottam két kisebb, de még gyakori kategóriát: a *mozgást* és a *kommunikációt*. A vizsgált korpuszon 586 mozgás és 1120 kommunikáció típusú esemény volt.
A *mozgás* egy test, tárgy, személy helyének megváltozása az idő és egy viszonyítási pont viszonylatában. A *kommunikáció*: információcsere, közlés, tájékoztatás.

Példák
Mozgás: *A gyerek **elment** az iskolába.*
Kommunikáció: *Tegnap telefonon **beszélgettünk**.*

Annotációs példák:
<*MOZ*>: mozgás, <*KOM*>: kommunikáció

Mire végzek, este lesz, nem tudok <MOZ>menni</MOZ> sehova .
Egy kék overallos alak <MOZ>közeledett</MOZ> a járdán, vagy tízméternyire tőle.

A Pick Szeged Rt. 70.000 darab saját részvényt vásárolt 2.850 forintos árfolyamon – <KOM>közölte</KOM> a társaság a Magyar Tőkepiac szerdai számában.
Az Adobe Systems csütörtökön <KOM>bejelentette</KOM>, hogy mintegy 25 százalékkal csökkenti a harmadik negyedévre elvárt profitját , és 10 százalékkal az éves elvárást , mivel a cég részvényeinek árfolyama meredeken zuhant az első félév folyamán .

3.2.2 Főnévi események automatikus detektálása

A következő szövegeket *főnévi események* automatikus detektálására használtam fel. A szövegekben általában a legtöbb esemény igékhez kapcsolódik és az igék általában események is. De más szófajú szavak is lehetnek események, például főnevek, melléknévi igenevek. Ebben a részben a *főnévi események detektálásával* foglalkoztam. Néhány példa a főnévi eseményekre: *futás, építés, írás, háború, ünnep*.
A mondatokban az esemény jelöltek a főnevek voltak.
A főnévi eseményeknek két nagy csoportja van: *igéből képzettek* (deverbal) és *nem igéből képzettek* (non-deverbal). *Igéből képzett* főnévi események: írás, futás; *nem igéből képzett* főnévi események: háború, ünnep. Az igéből képzett főnévi eseményeknek két nagy csoportja van: események és eredmények. Ezek a főnevek általában többértelműek, néhány mondatban események, másokban pedig eredmények (nem események).

Példa:
*Nagyot lélegzett, és folytatta az **írás**t. (itt esemény)*
*Az RTL Klub az említett **írást** jogsértőnek, az abban szereplő állításokat valótlannak tartja. (itt nem esemény, hanem eredmény)*

Alkalmazásomhoz a Szeged Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, Hungarian Dependency Treebank, 2010) egy részét használtam, amelyik 10000 mondatot tartalmaz a következő doménekről: *üzleti rövidhírek, szépirodalom-fogalmazás, számítógépes szövegek, újsághírek, jogi szövegek.* A szövegeket két nyelvész annotálta, az annotátorok közötti egyetértés Kappa = 0,7 volt.

Statisztikai adatok a feldolgozott főnevekről (3.5. táblázat**)**

Mondatok száma	10 000
A jelöltek száma (főnevek)	48 388
A pozitív jelöltek száma (esemény főnevek)	7626
Igéből képzett főnevek	5325
Igéből képzett főnévi események	4169
Nem igéből képzett főnevek	43 063
Nem igéből képzett főnévi események	3457

3.5. táblázat - Statisztikai adatok a feldolgozott főnevekről

Annotációs példák:
<E>: esemény, *<NE>*: nem esemény

A hazai <NE>helyzet</NE><E>áttekintése</E> után lépjünk globális <NE>szintre</NE>.
Tizennyolc <NE>hónappal</NE> ezelőtt még online <E>beszerzés</E> sem létezett.

3.2.3 Szemantikus szerepek automatikus címkézése

A következő szövegeket szemantikus szerepek automatikus címkézésére használtam fel. Igei és főnévi igenévi célszavakhoz kerestem szerepeket. A vizsgálat első részében a *vállalat fel-*

vásárlások keretével foglalkoztam. A következő célszavakat vizsgáltam az adott kereten belül: *ad, árul, bekebelez, beruház, elad, értékesít, gyarapít, kap, kereskedik, szerez, vásárol, vesz.* A következő szerepeket kerestem a célszavakhoz: *Vevő, Eladó, Árucikk, Ár, Dátum.*
A vizsgálat második részében modellemet a *részvénypiaci hírek* doménen teszteltem. A következő célszavakat vizsgáltam: *befejez, csökken, csúszik, emelkedik, esik, gyengül, kezd, növel, nő, nyer, nyit, szerez, ugrik, változik, veszít, zár*. Ezekhez a célszavakhoz a következő szerepeket kerestem: *Instrumentum, Ár, Elmozdulás-irány, Elmozdulás-érték, Piac, Dátum, Idő, Mennyiség*.
A két vizsgált keretnél ezek voltak a leggyakoribb célszavak és szerepek, azért választottam ezeket.

Példák a célszavakra és a hozzájuk tartozó szerepekre a két doménen. A példákban a ***célszavakat*** vastagon emeltem ki, a *[szerepek]* pedig szögletes zárójelben láthatóak. Az szerepek típusa $_{\text{alsóindexben}}$ van feltüntetve.

1. [A svéd Ericsson]$_{Eladó}$ bejelentette, hogy [a német Infineonnak]$_{Vevő}$ ***adja el*** *[chipgyártó részlegét]$_{Áru}$, [400 millió euróért]$_{Ár}$.*
2. [A japán Hitachi Ltd.]$_{Vevő}$ [2,05 milliárd dollárért]$_{Ár}$ ***megveszi*** *[az IBM merevlemez-meghajtókat gyártó üzletágát]$_{Áru}$ - jelentette a Bloomberg.*
3. [A Budapesti Értéktőzsde]$_{Piac}$ [hivatalos részvény indexe, a BUX]$_{Instrumentum}$ [36,54 pontos]$_{Elmozdulas-érték}$ [csökkenéssel]$_{Elmozdulás-irány}$, [7.376,30 ponton]$_{Ár}$ ***nyitott***$_{Idő}$ *[csütörtökön]$_{Dátum}$.*
4. [Hétfőn]$_{Dátum}$ [5871,3 ponton]$_{Ár}$ ***zárt***$_{Idő}$ *[az FTSE-100]$_{Instrumentum}$, [8,5 ponttal, azaz 0,1 százalékkal]$_{Elmozdulás-érték}$ [alacsonyabban]$_{Elmozdulas-irány}$ a pénteki zárónál.*

A példáknál láthatjuk, hogy a szerepek gyakran több szóból állnak és a mondatok általában nem tartalmaznak minden szerepet. Alkalmazásom teszteléséhez a Szeged Dependency Treebank-nek (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) az üzleti rövidhírek doménjét használtam fel, aminek egyik verziójában annotálva vannak a szemantikus szerepek. 1000 – 1000 mondatot használtam fel mindkét doménről.

Statisztikai adatok

A vállalat-felvásárlások doménen:
Mondatok száma: 1000
Mondatok száma, amelyek tartalmazzák az adott szerepet (3.6. táblázat):

Szerep	Darab
Vásárló	783
Eladó	579
Áru	1025
Ár	299
Dátum	312

3.6. táblázat: Mondatok, amelyek tartalmazzák az adott szerepet a vállalat-felvásárlások doménen

A tőzsdei hírek doménen:
Mondatok száma: 1000
Mondatok száma, amelyek tartalmazzák az adott szerepet (3.7. táblázat):

Szerep	**Darab**
Instrumentum	787
Ár	530
Elmozdulás-irány	431
Elmozdulás-érték	683
Piac	485
Dátum	436
Idő	109
Mennyiség	302

3.7. táblázat: Mondatok, amelyek tartalmazzák az adott szerepet a tőzsdei hírek doménen

4 Gépi tanulási technikák

Ebben a fejezetben rövid áttekintést adok a gépi tanulásról, bemutatva a gépi tanulás néhány alapvető módszerét és azok kiértékelési elveit.

4.1 A gépi tanulás alapelvei

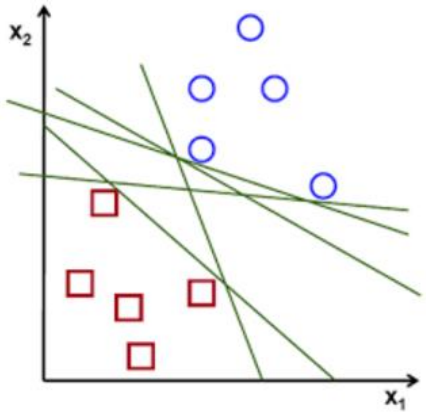
A gépi tanulás a számítógép tudomány egyik területe, ami képességet ad számítógépeknek a tanulásra. Arthur Samuel a számítógépes játékok és a mesterséges intelligencia egyik amerikai úttörője használta először a „Gépi tanulás" fogalmát 1959-ben[1]. A mintafelismerésből és a mesterséges intelligenciából kialakulva, a gépi tanulás olyan számítógépes algoritmusokat használ fel, amelyek képesek tanulni, következtetéseket levonni és előrejelzéseket tenni az adatok alapján. A gépi tanulást a számítógépes feladatok széles körében felhasználják, olyan területeken, ahol a jó képességű explicit algoritmusok kifejlesztése nehéz, vagy nem lenne megoldható.
A gépi tanulás közel áll a számítógépes statisztikához, ami szintén készít előrejelzéseket számítógépek segítségével. Szoros kapcsolatban van a matematikai optimalizálással is, ami algoritmusokat, módszereket fejlesztett ezen a területen és gyakran összeolvad az adatbányászattal.
A gépi tanulás összetett modelleket és algoritmusokat készít, amelyek képesek összefüggések feltárására és előrejelzésekre. Ezeket a modelleket kutatók, mérnökök felhasználják döntések meghozatalára, amelyekben rejtett összefüggéseket tárnak fel az ismert adatokon való tanulás által. A keresett minták megtalálását gyakran megnehezíti az elegendő tanító adatok hiánya.
A gépi tanulás két fő kategóriára osztható: *felügyelt tanulás* (supervised learning) és *felügyelet nélküli tanulás* (unsupervised learning). A *felügyelt tanulásnál* a számítógépet ellátjuk megfelelő tanító adattal. A cél, hogy a számítógép felfedezze a bemeneti és a kimeneti adatok közötti összefüggéseket. *Felügyelet nélküli* tanulásnál a számítógépnek címkézett tanító adatok hiányában kell a minták közötti összefüggéseket felismernie. Speciális esetekben a tanító adat csak részben áll rendelkezésre (*félig felügyelt tanulás*, semi-supervised learning), ahol a számítógépnek címkézett és címkézetlen adatokkal is kell dolgoznia.
A gépi tanulásos algoritmusokat sok fajta problémára alkalmazzák, például a számítógépes látás, beszédfelismerés, gépi fordítás, orvosi diagnózis, adatbányászat, szövegbányászat.
A következőkben a leggyakrabban használt gépi tanuló algoritmusokat mutatom be.

4.2 Szupportvektorgépek

A szupportvektorgépek (Support Vector Machine, SVM) felügyelt tanítási modellek, hozzákapcsolódó tanító algoritmusokkal. Ez az egyik leggyakrabban alkalmazott gépi tanulásos osztályozó módszer (Cortes & Vapnik, 1995). A tanító adatok halmazán minden vizsgálandó adatnál be van jelölve, hogy egy adott kategóriába, vagy a másikba tartozik-e. Ezek alapján az

[1] http://robotics.stanford.edu/~ronnyk/glossary.html

SVM tanító algoritmusa egy modellt készít, ami az új egyedeket besorolja vagy az egyik, vagy a másik kategóriába.
AZ SVM modellnél a vizsgált egyedeket a tér pontjaiként ábrázolhatjuk. Így a pont egy **p** dimenziós jellemző vektor. A jellemző vektorokat valamelyik kategóriához rendeljük.
A feladat egy p-1 dimenziós hipersík, vagy határterület megtalálása, ami elválasztja a két kategória pontjait egymástól. Általában sok fajta ilyen hipersík lehet, ezért az elválasztott kategóriák közötti hipersíkot (vagy határterületet) úgy kell kialakítani, hogy az minél szélesebb legyen. Azt a hipersíkot választjuk, amelyikhez a legközelebbi pont távolsága a maximális (4.1. ábra).



4.1. ábra: Hipersíkok felvétele

Ábra forrása: https://www.quora.com/Why-is-a-support-vector-machine-called-a-machine

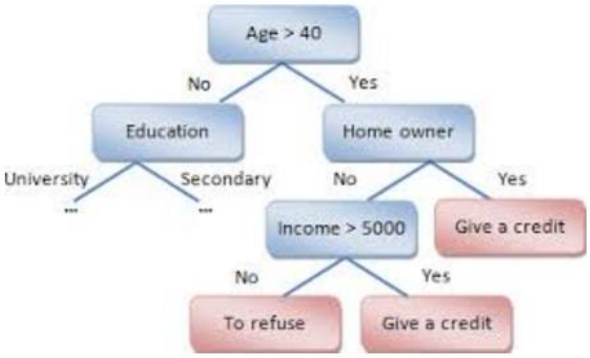
A tanítási fázis után az új vizsgálandó egyedeket szintén leképezzük ennek a térnek a pontjaiként és eldöntjük, hogy az elválasztó határterület melyik oldalára esik. Ezzel hozzuk meg osztályozási döntésünket.

A pontok felvétele és a hipersík kiválasztása részletesebben:
Adott egy n pontból álló tanító adathalmaz. ($\underline{\mathbf{x}}_1$, y_1),…,($\underline{\mathbf{x}}_n$, y_n). Ahol y_i értéke vagy 1 vagy -1, jelöli azt az osztályt, amihez az $\underline{x}_i$ tartozik. Mindegyik $\underline{\mathbf{x}}_i$ egy p dimenziós vektor. Keressük azt a legszélesebb határterületű hipersíkot, amelyik elválasztja egymástól az $y_i = 1$ pontokat az $y_i = -1$ pontoktól. A keresett hipersík és a legközelebbi pont távolsága legyen maximális.
Egy hipersík azon $\underline{x}$ pontok halmaza, amelyekre igaz: $\underline{w} \cdot \underline{x} - b = 0$, ahol a $\underline{w}$ a hipersík normálvektora. Ha a tanító adat lineárisan elválasztható, akkor kiválaszthatunk két egymással párhuzamos hipersíkot, melyek elválasztják az adatok két csoportját, úgy, hogy a két hipersík között a távolság a legnagyobb legyen. A keresett legnagyobb határterületű hipersík ezen két párhozamos hipersík között helyezkedik el a kettő közötti távolság felénél. A két hipersíkot megadhatjuk a következő két egyenlettel:
$\underline{w} \cdot \underline{x} - b = 1$
$\underline{w} \cdot \underline{x} - b = -1$
A közöttük lévő távolság: $\frac{2}{||\underline{w}||}$. Így a két hipersík közötti távolság maximalizálásához a $||\underline{w}||$ értékét kell minimalizálni. A legnagyobb határterületű hipersíkot azon $\underline{x}_i$ pontok határozzák

meg, amelyek legközelebb esnek hozzá. Ezeket az $\underline{x}_i$ pontokat nevezzük szupport vektoroknak. (Tikk, 2007)
Disszertációmban a Weka **SMO** implementációját felhasználtam a 8. fejezetben (Főnévi események automatikus detektálása természetes nyelvű szövegekben).

4.3 Döntési fák

A döntési fa egy vizuális döntést támogató eszköz, ami egy fa-struktúrájú gráfot használ döntések meghozatalához és egyik módja az algoritmusok ábrázolásának. A döntési fa a gyökérelemből indul ki és a levélelemeken végződik. A fa csomópontjainál döntési pontok vannak, ahol szabályokat alkalmazunk. Először nézzünk meg a döntési fára egy általános példát (4.2. ábra).



4.2. ábra: Döntési fa

Ábra forrása:
http://help.prognoz.com/en/mergedProjects/Lib/06_datamining/lib_decisiontree.htm

Azokat a fa modelleket, amelyeknél a célváltozó értékek diszkrét halmazából vehet fel értéket, *osztályozó fáknak* nevezzük. Ezeknél a fa struktúráknál a levelek az osztály-címkéket reprezentálják, az ágak pedig jellemzőik olyan kapcsolatait, amelyeken keresztül eljuthatunk az osztálycímkékhez. Azokat a döntési fákat, amelyeknél a célváltozók folyamatos értékek közül vehetnek fel értéket (tipikusan valós számok), *regressziós fáknak* nevezzük.
A döntésifa-alapú tanulás egy módszer a diszkrét értékű célfüggvények közelítésére, ahol a tanult műveletet egy döntési fával ábrázoljuk. A cél egy olyan modell elkészítése, amelyik megjósolja egy több bemeneti változón alapuló célváltozó értékét.
A bemeneti adatok a következő formában érkeznek: $(\underline{\mathbf{x}}, Y) = (x_1, x_2, \ldots x_k, Y)$
A függő változó az Y, amit osztályozni szeretnénk. Az $\underline{x}$ vektort az $x_1, x_2, \ldots, x_k$ jellemzőkből képezzük.
A döntési fa tanításakor a tanító adathalmazt kisebb részhalmazokra bontjuk egy attribútum értékének vizsgálata alapján. A vizsgálatot a döntési fának egy nem-levél csomópontján végezzük el. A vizsgálati lépés során azt az attribútumot keressük, amelyik a legjobban darabolja a vizsgált részhalmazt. Az attribútumok kiválasztására szolgáló eljárás arra irányul, hogy minimalizáljuk az eredményül kapott fa mélységét. Azt az attribútumot választjuk, amellyel a

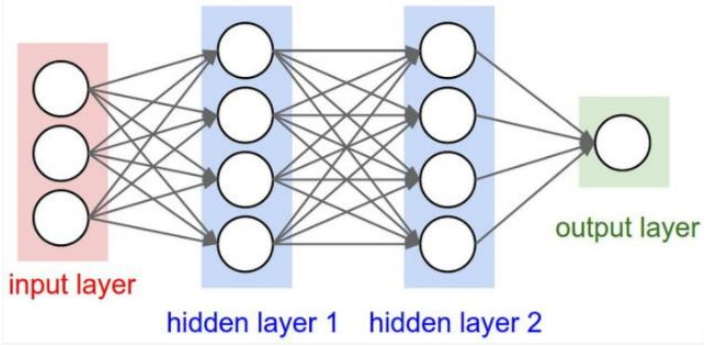
lehető legmesszebbre jutunk a példák pontos osztályozásában. Egy tökéletes attribútum a példákat egy csupa pozitív és egy csupa negatív példát tartalmazó halmazra osztja. Tökéletes attribútummal csak ritkán találkozunk, de vannak olyan attribútumok, amelyek jobbak, vagy kevésbé jók. Az attribútum kiválasztásánál egy olyan mértéket keresünk, ami akkor éri el maximumát, amikor az attribútum tökéletes, és akkor minimális, amikor az attribútumnak egyáltalán nincs semmi haszna. A vizsgálat eredményét az ágak képviselik. Ezt a folyamatot rekurzív módon ismételjük top-down módszerrel, mohó algoritmussal (Tikk, 2007).
A részhalmazon végrehajtott rekurziót megállítjuk és a csomópont egy levélben végződik:

- ha a részhalmaz minden eleme ugyanahhoz az osztályhoz tartozik. A levelet az osztállyal címkézünk.
- ha már nincs több kiválasztható attribútum. A levelet a leggyakoribb osztállyal címkézzük.

Disszertációmban a Weka csomag (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009) J48 osztályozóját használtam fel, ami a C4.5 –ös (Quinlan, 1993) döntési fa algoritmust implementálja. Ezt alkalmaztam a 7, 8, 9 fejezetekben.

4.4 Neurális hálózatok

A neurális hálózatok (ANN, Artificial Neural Networks) olyan számítógépes rendszerek, amelyek kifejlesztését az élőlények agyában lévő biológiai neurális hálózatok inspirálták. A neurális hálózatok összekapcsolt egységek (neuronok) hálózatán alapul, az agyban lévő neuronok analógiája alapján. A neuronok közötti kapcsolatok jeleket tudnak továbbítani más neuronok felé. A fogadó neuron feldolgozza a jeleket és továbbít egy jelet más neuronoknak. A neuronok közötti kapcsolatoknak súlyuk van, amit a tanítási folyamatban változtatunk a bemeneti és kimeneti adatok alapján.
A neuronok rétegekbe (layer) szervezhetőek. A különböző rétegek eltérő átalakítást végezhetnek a bemenő jeleken. A jelek az első rétegtől (input) haladnak az utolsó rétegig (output), miközben több rejtett rétegen mehetnek át (4.3. ábra).



4.3. ábra: Neurális hálózat

Ábra forrása: https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network

A tanulási algoritmus módosítja a neurális hálózat paramétereit (súlyok és küszöbértékek) úgy, hogy a hálózat adott bemeneti jelekre adott kimeneti jeleket produkáljon. A hálózat taní-

tása hibavezérelt visszacsatolásos módszerrel történik. Ha egy egyedet nem jól kategorizál a hálózat, akkor módosítja a súlyok értékét a hiba csökkentének érdekében. A bemeneti (és a tárolt értékekből) az aktuális kimeneti értéket egy tipikusan nemlineáris függvény alkalmazásával hozza létre, melyet aktiváló vagy aktivációs függvénynek nevezünk (activation function).
A legegyszerűbb bináris osztályozást végző hálózat a *perceptron*, aminek algoritmusa leképezi a bemeneti vektort a kimeneti vektorra. Függvénye kiszámítja a j. neuronra érkező $p_j(t)$ jelet az előző neuronokról érkező $o_i(t)$ kimeneti jelekből a súlyok segítségével, tipikusan a következő formában:

$$p_j(t) = \sum_i o_i(t) w_{ij}$$

Az osztályozás kimenete a függvény előjelétől függ: ha az pozitív, akkor c_1 kategóriába, egyébként c_2 kategóriába sorolja a vizsgált egyedet. A neuronokhoz be lehet állítani egy küszöbértéket, hogy a bemenetére érkező jelekből képzett összegzett jelet csak egy adott küszöbérték felett dolgozza fel. (Tikk, 2007)
Az utóbbi néhány év ígéretei kiemelten ilyen eszközökre alapozódnak és egyre több eredményt mutatnak fel (Nguyen & Grishman, 2015) (Feng, Qin, & Liu, 2018).

4.5 Kiértékelési elvek

A különböző gépi tanulási módszerek hatékonyságának méréséhez és összehasonlításához több szempontot is figyelembe kell venni. Az eseménykinyerő rendszerek értékelése általában a klasszikus pontosság (precision), fedés (recall) és F1-mérték (Rijsbergen, 1979) értékek szerint történik. A *fedés* (felidézés, recall) megadja, hogy az összes releváns osztályozandó eset közül mennyi szerepel a találatok között. A találati halmaznak általában nem minden eleme releváns. Minél több a releváns találat, annál nagyobb a *pontosság* (precision). E két mérték meghatározásához a következő adatokat kell ismernünk:
Igaz pozitív (true positive, TP): a találati listában szereplő releváns találatok száma
Igaz negatív (true negative, TN): a találati listában nem szereplő nem releváns találatok száma.
Hamis pozitív (false positive, FP): a találati listában szereplő nem releváns találatok száma
Hamis negatív (false negative, FN): a találati listában nem szereplő releváns találatok száma

Például az események detektálása feladatnál ez a következőket jelenti:
Igaz pozitív: azon elemek száma, amelyeket az események közül helyesen eseménynek jelöltünk.
Igaz negatív: azon elemek száma, amelyeket a nem-események közül helyesen nem-eseménynek jelöltünk.
Hamis pozitív: azon elemek száma, amelyeket a nem-események közül helytelenül eseménynek jelöltünk.
Hamis negatív: azon elemek száma, amelyeket a események közül helytelenül nem-eseménynek jelöltünk.

Ezek alapján a pontosság (precision): P= $\frac{TP}{TP+FP}$, a fedés (recall): R = $\frac{TP}{TP+FN}$.
Egy adott C osztályra az 1.0 *pontosság* érték azt jelenti, hogy minden esetet, amit a C osztályhoz tartozónak jelöltünk valóban a C osztályhoz tartozik. Ez azonban nem mond semmit arról, hogy a C osztály elemei közül hányat jelöltünk be helyesen és hányat hamisan. Az 1.0 *fedés* érték azt jelenti, hogy a C osztály minden elemére helyesen bejelöltük, hogy a C osztályhoz tartozik. Ez azonban nem mond semmit arról, hogy hány más elemet jelöltünk be hamisan a C osztályhoz tartozónak, ami valójában nem tartozik a C osztályhoz.
Jó osztályozás esetén a pontosság és a fedés is magas értékű. Gyakran az egyiket csak a másik kárára tudjuk növelni. A *fedés* értékét például 1.0 értékre be tudjuk úgy állítani, hogy minden elemet a C halmazhoz tartozónak jelölünk. Ekkor a hamis negatív (FN) érték 0, hiszen mindegyiket pozitívnak jelöltük. Ekkor azonban a pontosság értéke nagyon alacsony, hiszen a hamis pozitív (FP) esetek száma nagy. A pontosság értékét magas értékűre be tudjuk állítani, ha csak azt a néhány esetet jelölünk pozitívnak, amelyekről nagy valószínűséggel tudjuk, hogy a C halmazhoz tartoznak. Ilyenkor a hamis pozitív (FP) esetek száma alacsony, vagy nulla. De viszont a fedés alacsony értékű, a hamis negatív (FN) esetek nagy száma miatt.
Ezért sem a pontosságot, sem a fedést nem érdemes csak önállóan vizsgálni, a kettőt egy mérőszámmal kombináljuk. A kettő harmonikus közepével, az osztályozó rendszer hatékonysága egyetlen mérőszámmal jellemezhető, ezt F1-mértéknek nevezzük: F1= $\frac{2 \cdot P \cdot R}{P+R}$
Ha a pontosság és fedés értékeit nem azonos súllyal akarjuk figyelembe venni, akkor súlyozhatjuk azokat.
Ebben az esetben $F_{beta} = \frac{(\beta^2+1) \cdot P \cdot R}{(\beta^2 \cdot P+R)}$

4.6 Összegzés

Ebben a fejezetben bemutattam az alapvető gépi tanulási elveket, amelyeket az események detektálása és osztályozása, valamint az események szemantikus címkézése feladatokban alkalmaztam. Bemutattam a szupportvektorgépeket, a döntési fákat és a neurális hálózatokat. Ezek közül a szupportvektorgépeket és a döntési fákat alkalmaztam a feladataimban. Bemutattam azokat a kiértékelési elveket, amelyet a feladataimban felhasználtam rendszereim hatékonyságának mérésére. A méréseket és eredményeket majd a 7, 8, 9 fejezetekben ismertetem.

5 A függőségifa- és konstituensfa-alapú elemzés és a WordNet

A kutatási terület minden részénél felhasználtam a függőségi reprezentációt és az ismert lexikai adatbázist, a magyar WordNetet. A konstituensfa-alapú reprezentációt egy témánál használtam fel. Ebben a fejezetben bemutatom mind a két reprezentációt és a WordNetet.

5.1 Függőségi reprezentáció

A függőségi nyelvtan (Dependency grammar) a modern szintaktikai elméletek egy olyan csoportja, amelyik a függőségi kapcsolatokon alapul (ellentétben a konstituensfa-alapú kapcsolatokkal). A nyelvtani egységek közötti függőség elve már több évszázaddal ezelőtt felmerült. A modern függőségi nyelvtanok Lucien Tesnière (1893–1954) munkájával kezdődtek. A függőségen keresztül lehet a nyelvi elemeket (például szavakat) egymáshoz kapcsolni közvetlen kapcsolatokkal. A mondat fő igéje képezi a struktúra kiinduló elemét. Minden más szintaktikai egység (szavak) vagy közvetlenül, vagy közvetve kapcsolódnak a főigéhez. A függőségi nyelvtan különbözik a frázis strukturált nyelvtanoktól (konstituens nyelvtanok), mivel frázis kapcsolatokat nem jelöl. A szerkezetet a szó és a függőségei alapján lehet kialakítani.
A függőségi reprezentáció a szavakat a közöttük lévő kapcsolatok alapján kapcsolja össze és egy fastruktúrában ábrázolja. A fa minden csomópontja egy szót reprezentál, a gyerekcsomópontok azon szavak, amelyek függnek a szülőcsomóponttól és az ágat a kapcsolattal címkézzük. A fő ige a gyökér elem. Ha egy kapcsolati-egység több szót tartalmaz, akkor ezek a szavak egy részfát alkotnak a fő fán belül. A részfa a fejszaván (headword) keresztül kapcsolódik a fő fához. A függőségi elemezés során a mondatokban azonosítjuk a szavak közötti nyelvtani (mondattani, szintaktikai) viszonyokat: pl. jelzők, módosítók, állítmány, alany, tárgy stb.
A függőségi elemzők olyan szintaktikai elemző programok, amelyek a mondatok elemzését olyan fastruktúrával reprezentálják, ahol a gyökér a mondat főigéje és a szavak között a nyelvtani fejektől mutatnak irányított élek a bővítmények felé. A függőségi reprezentáció segítségével jobban megérthetők a mondatot alkotó szavak közötti viszonyok, pontosabb tartalmi elemzést lehetővé téve.
A magyar nyelv gazdag morfológiával és szabad szórenddel rendelkezik. A függőségi fákat felhasználó reprezentáció különösen jól használható szabad szórendű nyelvek elemzésére, így a magyarra is, ezek ugyanis könnyebben teszik lehetővé az egymással nem szomszédos, de összetartozó szavak kapcsolatának megtalálását. Ezért a magyar nyelvű szövegekre általában függőségi reprezentációt használtam.

Példa a Szeged Dependency Treebank-ból:
A magyar Országgyűlés az év első 35 hetében összesen 8 héten át ülésezett.

A mondat függőségi reprezentációja:

1	A	a	T	T	SubPOS=f	3	DET
2	magyar	magyar	A	A	SubPOS=f|Deg=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none	3	ATT

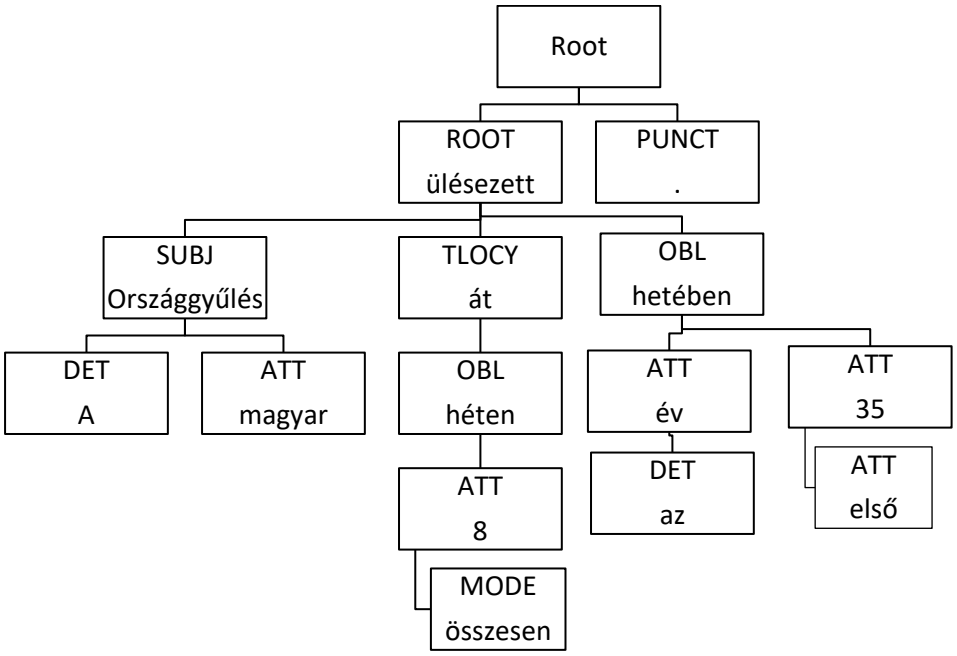
```
3   Országgyűlés Országgyűlés N   N
    SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none 13  SUBJ
4   az  az  T   T   SubPOS=f   5   DET
5   év  év  N   N
    SubPOS=c|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none 8   ATT
6   első    első    M   M
    SubPOS=o|Num=s|Cas=n|Form=l|NumP=none|PerP=none|NumPd=none 7   ATT
7   35  35  M   M
    SubPOS=c|Num=s|Cas=n|Form=d|NumP=none|PerP=none|NumPd=none 8   ATT
8   hetében hét N   N
    SubPOS=c|Num=s|Cas=2|NumP=s|PerP=3|NumPd=none   13  OBL
9   összesen    összesen R   R   SubPOS=x|Deg=none|Num=none|Per=none 10
    MODE
10  8   8   M   M
    SubPOS=c|Num=s|Cas=n|Form=d|NumP=none|PerP=none|NumPd=none 11  ATT
11  héten   hét N   N
    SubPOS=c|Num=s|Cas=p|NumP=none|PerP=none|NumPd=none 12  OBL
12  át  át  S   S   SubPOS=t   13  TLOCY
13  ülésezett   ülésezik    V   V
    SubPOS=m|Mood=i|Tense=s|Per=3|Num=s|Def=n 0   ROOT
14  .   .   .   _   _   0   PUNCT
```

Az elemzésben egy sor egy tokennek felel meg. Az első oszlopban a token mondatbeli sorszáma, a másodikban a szóalak, a harmadikban a lemma, a negyedikben az MSD-kód, az ötödikben a szófaj, a hatodikban a morfológiai jegyek, a hetedikben a szülőcsomópont sorszáma, a nyolcadikban pedig a függőségi él-címke látható.

Az előző reprezentációt egyszerűsítve, a szavak közötti kapcsolatok kiemeléséhez csak a sorszámot, a szóalakot a szülőcsomópont sorszámát és a függőségi él-címkét megtartva:

```
1   A   3   DET
2   magyar  3   ATT
3   Országgyűlés    13  SUBJ
4   az  5   DET
5   év  8   ATT
6   első    7   ATT
7   35  8   ATT
8   hetében 13  OBL
9   összesen    10  MODE
10  8   11  ATT
11  héten   12  OBL
12  át  13  TLOCY
13  ülésezett   0   ROOT
14  .   0   PUNCT
```

A példamondat függőségi reprezentációja fastruktúrában ábrázolva (5.1. ábra**):**



5.1. ábra: A példamondat függőségi reprezentációja fastruktúrában ábrázolva

5.2 Konstituensfa-alapú reprezentáció

A szintaktikai elemzés célja a mondat szavai közt fennálló kapcsolatok azonosítása. Ezen kapcsolatok reprezentálásának egyik elterjedt módja a konstituens (constituent) fák alkalmazása. Ebben egy konstituens egy szó, vagy szavak csoportja, ami egy egységként funkcionál egy hierarchikus struktúrán belül. Szavak csoportjai (eredeti sorrendben) egységeket alkotnak. A frázisok (phrase) több mint csak szavak csoportja. Olyan szavak csoportja, amelyek együtt töltenek be egy speciális szerepet a mondaton belül. Ezen szócsoportok együtt mozgathatóak, vagy helyettesíthetőek, miközben a mondat jól olvasható és nyelvtanilag helyes marad. A konstituens reprezentáció célja a mondatok ilyen rész-frázisokra való bontása. A természetes nyelvekben a frázisok egymásba ágyazva helyezkednek el, így a mondatokat fastruktúrában tudjuk ábrázolni.

A nyelvészetben a frázis-strukturált nyelvtanok azok a nyelvtanok, amelyek a konstituens kapcsolatokon alapulnak. A frázis-strukturált nyelvtant először Noam Chomsky mutatta be (Chomsky, 1957). Az egységek osztályokba sorolhatóak, amelyek külső és belső szempontokból jól definiáltan viselkednek. Ilyenek például a főnévi csoportok (Noun Phrase, NP) és az igei csoportok (Verb Phrase, VP).

A kontituensfában a csomópontok (nem-terminálisok) a kifejezések típusai, a levelek (terminálisok) a mondat szavai, az ágak nincsenek címkézve. Ez a reprezentáció a konstituens kapcsolatokon alapul. Az elemzőfa az S szimbólummal kezdődik (gyökér csomópont) és a mondat szavaival, a levelekkel végződik. A közbülső csomópontokhoz egy vagy több gyerek-csomópont kapcsolódhat. Egy mondatban csak egy gyökércsomópont lehet. Az ige-argumentum viszonyokat címkék kódolják.

A konstituens-nyelvtanok általában kötött szórendű nyelvekre alkalmazhatóak jól. A magyar nyelv gazdag morfológiával és szabad szórenddel rendelkezik, ezért a magyar nyelvű szövegekre minden feladatnál a függőségi reprezentációt használtam fel. Egy feladatnál (a főnévi események detektálásánál) emellett alkalmaztam a konstituens-alapú reprezentációt is.

A függőségi reprezentáció bemutatásánál ezt a mondatot vizsgáltuk meg:
A magyar Országgyűlés az év első 35 hetében összesen 8 héten át ülésezett.

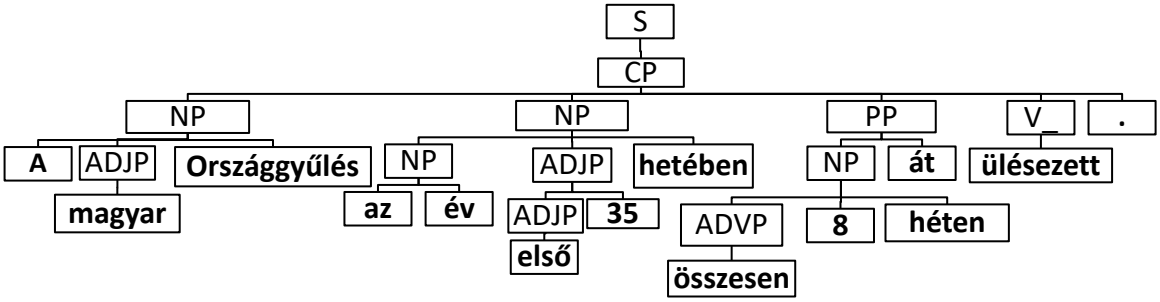
A mondat konstituensfa-alapú reprezentációja a Szeged Treebank-ban:

```
<s id="HVG.1.2.1">A magyar Országgyűlés az év első 35 hetében összesen 8 héten át ülésezett.
    <CP id="HVG.1.2.1.1">
        <NP id="HVG.1.2.1.2">
            <w>A</w>
            <ADJP>
                <w>magyar</w>
            </ADJP>
            <w>Országgyűlés</w>
        </NP>
        <NP id="HVG.1.2.1.3">
            <NP>
                <w>az</w>
                <w>év</w>
            </NP>
            <ADJP>
                <ADJP>
                    <w>első</w>
                </ADJP>
                <w>35</w>
            </ADJP>
            <w>hetében</w>
        </NP>
        <PP id="HVG.1.2.1.4">
            <NP>
                <ADVP>
                    <w>összesen</w>
                </ADVP>
                <w>8</w>
                <w>héten</w>
            </NP>
            <w>át</w>
        </PP>
        <V_ id="HVG.1.2.1.5">
            <V0>
                <w>ülésezett</w>
            </V0>
            <CHILDREN>
                <NODE idref="HVG.1.2.1.2" type="NP" role="NOM"></NODE>
                <NODE idref="HVG.1.2.1.3" type="NP" role="INE"></NODE>
                <NODE idref="HVG.1.2.1.4" type="PP" role="TLOCY"></NODE>
            </CHILDREN>
        </V_>
```

```
                <c>.</c>
        </CP>
</s>
```

A reprezentációban a konstituens-részek és azok kapcsolatai mellett jelölve vannak az ige (ülésezett) szintaktikai kapcsolatai is: <CHILDREN>…</CHILDREN>

A példamondat konstituensfa-alapú reprezentációja (5.2. ábra**):**



5.2. ábra: A példamondat konstituensfa-alapú reprezentációja

5.3 WordNet

A WordNet (Miller George, 1995) fogalmak adatbázisa, egy gráf struktúrában tárolt szótár, amelyben a fogalmak össze vannak kapcsolva szemantikai kapcsolatokon keresztül. Ezek a fogalmak lehetnek általánosak („egyed”, „állapot”), kevésbé általánosak („állat”, „mozgás”) és nagyon specifikusak is („Lánchíd”). A szinonim fogalmak jelentése hasonló egymáshoz. Egy szinonim halmaz (synonym set, synset) szinonim fogalmak csoportja. A WordNet olyan lexikai adatbázis, amiben a szinonim fogalmakat un. synsetek halmazaiba csoportosítják. A synsetek lehetnek egyszerű fogalmak, vagy többszavas kifejezések. Egy többértelmű szó különböző jelentéseit különböző synsetekhez sorolják.

Az első WordNetet a Princeton University-n készítették angol nyelvre 1985-ben. Az angol mellett később sok más nyelvre is elkészítették a WordNetet, többek között a magyarra is. A jelentésük bemutatásához a synsetek tartalmaznak egy rövid definíciót és példamondatokat is. A synsetek kapcsolatban állnak egymással szemantikai és lexikai kapcsolatokon keresztül, mint például a hipernim, hiponim, meronim, holonim kapcsolatok.

A synsetek között vannak olyan kapcsolatok, amelyek fogalmak hierarchiáját alkotják. Ilyenek a hiponimia-hipernimia relációk, amelyek alá-fölérendeltséget fejeznek ki. Például vizsla-kutya, fa-növény. Hipernimok az általánosabb synsetek, hiponimok a specifikusabb synsetek. A meronim, holonim relációk rész-egész viszonyokat írnak le. Például ablak-épület, kerék-autó.

Ezek a kapcsolatok az azonos szófajú szavak között lettek meghatározva. Így a WordNet valójában négy alhálózatot tartalmaz: főnevek, igék, melléknevek, határozók. Ezeknek a csoportoknak a tagjai hierarchikus rendbe rendezhetőek ezen kapcsolatok alapján.

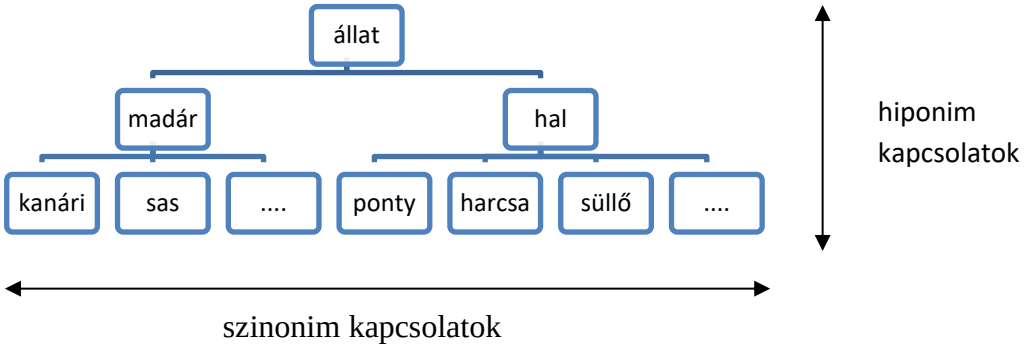
Bár a WordNet egy olvasható formájú adatbázis, mégis leginkább számítógépes programok használják. A WordNet struktúrája jól használhatóvá teszi a számítógépes nyelvészet és a természetesnyelv-feldolgozás területein.
Léteznek jelentés-egyértelműsítő eljárások (WSD, word sense disambiguation), amelyek a WordNetre épülnek. Az egyértelműsítés feladata egy többértelmű szó jelentésének meghatározása az adott szövegkörnyezetben. Erre az egyik módszer a Lesk algoritmus (Jurafsky & Martin, 2009). Ennek során a vizsgált szó környezetét összevetik a szóhoz tartozó több synsetnél megadott rövid definícióval és példamondatokkal. Azt a synsetet választják, amelyikben a legtöbb a közös szó a vizsgált szó környezete és a definíció és példamondatok között.
A WordNetet a szóegyértelműsítés mellett széleskörűen használják többek között információvisszakeresésre, szövegek osztályozására, szövegek összegzésére, gépi fordításra, információkinyerésre. Van olyan felhasználás is, mely a szavak hasonlóságát vagy mondatok jelentésbeli hasonlóságát a WordNet segítségével próbálja kifejezni. A synseteket egy gráfban tároljuk, így a synsetek hasonlóságát mérni tudjuk a legközelebbi útvonallal a gráfban (synsetek közötti élek száma). Ezt útvonalhasonlóságnak nevezzük (path similarity), és azon az előfeltételezésen alapul, hogy minél közelebb van két szó a gráfban, annál közelebbi a jelentésük is.

Statisztikai adatok a magyar WordNetről (5.1. táblázat):

szófajok	**synsetek száma**	**arány**
főnév	33530	79,3 %
melléknév	4112	9,7 %
ige	3607	8,5 %
határozószó	1039	2,5 %
Összesen	**42288**	

5.1. táblázat: Statisztikai adatok a magyar WordNetről

Szavak kapcsolatai a WordNetben (5.3. ábra**):**



5.3. ábra: Szavak kapcsolatai a WordNetben

5.4 Összegzés

Kutatásom minden részénél felhasználtam a függőségi reprezentációt és a magyar WordNetet. A konstituensfa-alapú reprezentációt egy témánál használtam fel. Ebben a fejezetben bemutattam mind a két reprezentációt és a WordNetet. A két reprezentációt részletesen jellemeztem, összehasonlítottam azokat egymással. Egy példamondaton keresztül mindkettőnél megadtam a mondat szöveges és grafikus reprezentációját a Szeged Treebank és a Szeged Dependency Treebank elemzései alapján. Bemutattam, hogy a két reprezentáció közül melyiket milyen környezetben és mely nyelveknél érdemes használni. Jellemeztem a WordNet felépítését, megadva a fogalmak ábrázolási módszerét, a közöttük lévő kapcsolatok típusait és az összekapcsolási módokat.

6 Kapcsolódó munkák - általános áttekintés

Mielőtt bemutatom új kutatási eredményeimet az *események* kinyerése területén, fontos, hogy szóljak a témában a múltban tett kutatási erőfeszítésekről is. Ebben a fejezetben *általános áttekintést* nyújtok a disszertációmhoz kapcsolódó megelőző munkákról. A következő fejezetekben majd részletezem az adott területekhez tartozó munkákat.

6.1 Események detektálása és osztályozása

Az *eseménykinyerés*, az információkinyerés egy speciális formája, az 1980-as évekre vezethető vissza. Népszerűségét a múlt évtizedekben nyerte el a „big data" valamint a szövegbányászat és természetesnyelv-feldolgozás kapcsolódó területeinek megjelenésével. Módszereit sok területen alkalmazzák, mint például a híroldalak, bioinformatika, biztonság, orvostudomány és blogok esetében. (Grishman, Huttunen, & & Yangarber, 2002), (Grishman, Huttunen, & & Yangarber, 2002), (Ji & Grishman, 2008), (King & Lowe), (Naughton, Kushmerick, & Carthy, 2006), (Yangarber, 2005). Tanev és társai (Tanev, és mtsai., 2009) mutattak példát az eseménykinyerő rendszerek többnyelvű alkalmazására.
A 2000-es évek eleje óta az általános *információkinyerés* területéről – ami elsősorban a névelemek detektálással foglalkozott, mint a személyek, helyek, szervezetek – áthelyeződött a hangsúly az adatbányászat összetettebb területeire, mint például az *eseménykinyerés*. Ehhez többek között már az egyedek közötti összetettebb kapcsolatok azonosítására van szükség (Björne, Ginter, Pyysalo, Tsujii, & Salakoski, 2010). Ez a fejlődés a szövegbányászat és a természetes nyelv feldolgozás módszereinek folyamatos javulásával, valamint a „big data" és az annotált korpuszok megjelenésével vált lehetségessé.
Az *események kinyerése* az 1990-es évek óta vált népszerű kutatási területté, amikor az első Üzenet-értési konferenciát (Message Understanding Conference, MUC-1) szervezte a DARPA (Defense Advanced Research Projects Agency), az Egyesült Államok Védelmi Minisztériumának kutatásokért felelős részlege. A konferencia célja katonai szöveges üzenetek automatikus analízisével kapcsolatos kutatások támogatása volt. Később annotált korpuszokat készítettek a további MUC konferenciákhoz is (MUC 1,…, MUC 7).
Yang és társai is 1998-ban az *események detektálásával* foglalkoztak (Yang, Pierce, & Carbonell, 1998). Fő feladatként *új eseményeket detektáltak* hírek szövegeiből. A rendszer kiértékelésével 82%-os F1-mértéket értek el offline detektálásra és 42%-ot online detektálásra. A rendszer jó teljesítménye mutatta, hogy az eddig már más területeken alkalmazott technikák hatékonyan felhasználhatóak *események detektálására* is.
Két évvel később Allan és társai az *eseménykinyerés* következő területeit vizsgálták: (a) detektálás, (b) első-esemény detektálás, (c) események közötti kapcsolat detektálás (Allan, Lavrenko, & Jin, 2000). Az eredmények elfogadhatóak voltak a három kiértékelt feladatra, de nem olyan mértékben, mint ahogy a készítői elvárták, mivel az első-esemény detektálás gyenge eredményt ért el.
2004-től kezdődően az ACE program (Automatic Content Extraction) jelentős haladást ért el az *eseményekkel* kapcsolatos területeken. Ez a már említett MUC program folytatása volt. Több nyelvvel is foglalkozott (angol, kínai és arab) és számos területet fedett le (katonai, jogi,

üzleti), biztosítva tanító és kiértékelő adatot általános eseménydetektálási kutatási feladatokhoz. A 2005-ös ACE programnak 8 eseménytípusa volt 33 altípussal. A program feladatai voltak: (a) egyed detektálás, (b) kapcsolat detektálás, (c) esemény detektálás. Az ACE feladat magába foglalta az eseményekkel kapcsolatban a modalitás, az igeidő és a polaritás detektálását is.

TimeML-motivált kutatás

A már bemutatott TimeML és a TimeBank kifejlesztése elősegítette az *események annotációjának és azonosításának* munkáját. Számos kutató használta ezt a korpuszt események számítógépes azonosításának tanítására és kiértékelésére.

A TARSQI Tool Kit (Verhagen, Mani, Sauri, Knippen, Jang, & Littman, 2005) egy teljes rendszer, ami a TimeML annotáción alapul. Öt *szabályalapú* modult implementál: GUTime, *EVITA*, GUTenLINK, Slinket és SputLink. A GUTime modul TIMEX3 *idő elemeket* ismer fel és a TempEx címkézőre épül (Mani & Wilson, 2000). Az EVITA modul az *események felismerését* végzi el. Az alkalmazott *szabályok* felhasználták a szófaji, lemmatizálási információkat, a környezeti elemzést és a WordNet adatokat. Ezeket az adatokat egy szófaji és egy szintaktikai elemző szolgáltatta. Az *EVITA* volt az első rendszer ami képes volt TimeML elvek szerint annotált *események detektálására*. Igei, főnévi és melléknévi eseményeket is azonosítottak. A *főnévi események* detektálását lexikai kereséssel valósította meg a WordNetben. Ezeket az eseményeket a WordNet eseményekkel kapcsolatos 25 rész-fájában keresték. A *főnévi események* detektálását jelentés-egyértelműsítés követett olyan esetekben, amikor a főnévhez tartozott eseményi és nem eseményi jelentés is. Az EVITA biztató eredményeket ért el az *események azonosításában*: 74% pontosság, 87% fedés, 80 F-mérték, de mivel a kiértékelésre és a tanításra is ugyanazt a korpuszt használták, túlbecsülték az Evita teljesítményét nem látott szövegek esetére. Ezeket az eredményeket az igékre, főnevekre és igenevekre együtt érték el. A modul további eredményei: 90% a szófajra, 92% az igeidőre, 98% a polaritásra, 97% a modalitásra. Az annotálást egyetemi hallgatók végezték, az annotátorok közötti egyetértés igékre 80%, főnevekre 64% volt. A GUTenLINK modul teremtett kapcsolatot a szintaktikai és a lexikai információk között. A Slinket modul *eseménypárok közötti kapcsolatokat* azonosított.

Ugyanabban az időben Boguraev és társa hasonló rendszert fejlesztett (Boguraev & Ando, 2005). Ők RRM (Robust Risk Minimization) osztályozót alkalmaztak TimeML *események azonosításához*. Névelemek azonosításához használt jellemzőket (például szófaj, nagybetűs, szomszédos szavak) és szavak közötti szintaktikai kapcsolatok azonosítására használt jellemzőket (például alany, tárgy) használtak. A szöveget tokenek sorozataként értelmezték és szócsoportokat (chunk) is vizsgáltak. Egy szónak a helyzetét egy szócsoporthoz viszonyítva háromféleképpen jellemezték: E: a szó a szócsoport végén van, I: a szó a szócsoporton belül van, O: a szó kívül van a szócsoporton. 80,3%-os F-mértékkel tudták az *eseményeket azonosítani* és 64%-al *osztályokba sorolni* azokat.

Mani és társai *időbeli kapcsolatok kategorizálásával* foglalkoztak. A következő gépi tanulásos módszereket kombinálták: SVM (support vector machine), Maximum entropy és Naive-Bayes osztályozók, morfo-szintaktikai tulajdonságokat felhasználva (Mani, Verhagen, Wellner, Lee, & & Pustejovsky, 2006).

Hasonló módszereket alkalmaztak a következő kutatásokban is: (Bethard & Martin, 2006) , (Bethard S. , 2007), (March & Baldwin, 2008), ahol az *események azonosítását és osztályozását* gépi tanulásos módszerrel valósították meg. A STEP rendszerben, amit *események felismerésére és osztályozására* készítették, SVM modellt építettek morfológiai, szintaktikai függőségi és WordNet hipernim jellemzők segítségével (Bethard & Martin, 2006). Minden szóhoz BIO címkével jelezték, hogy a szó belül vagy kívül van egy adott eseményen. Kiértékelésnél két baseline mérést alkalmaztak. Egyiknél azt a címkét fogadták el, ami a leggyakoribb volt az adott szóra a TimeBank korpuszban, a másiknál az EVITA rendszert szimulálták, de már nem ugyanazt az adatokat használva tanításra és kiértékelésre. A két baseline módszerrel kiértékelésnél az esemény és osztály azonosításra 50,2% és 50,9%-os F-mértéket kaptak.
March és társa (March & Baldwin, 2008) *események felismerésére* Biased SVM osztályozót alkalmazott, a következő fő jellemzők felhasználásával: egy adott környezeti ablakban a szavak és szófajok, stop-szó eltüntetés, szócsoportok kezelése. A rendszer 76,4%-os F-mértéket ért el.
A bemutatott rendszerek eredményeit nehéz összehasonlítani, mert mindegyik saját kiértékelési keretrendszert használt. Ennek megoldására nemzetközi kiértékelő fórumokat rendeztek, ahol az idő-információ feldolgozással kapcsolatos rendszereket össze tudták hasonlítani. Két ilyen nagy kiértékelési feladat volt: a TempEval-1 (2007), és a TempEval-2 (2010). A TempEval-1 (Verhagen, Gaizauskas, Hepple, Schilder, Katz, & Pustejovsky, 2007) a TimeML *időbeli kapcsolatainak* automatikus kategorizálására fókuszált. Ebben még eseményekkel nem foglalkoztak és csak egynyelvű (angol) szövegeket használtak.
A TempEval-2 (Verhagen, Sauri, Caselli, & Pustejovsky, 2010) már foglalkozott az idők mellet az *eseményekkel* és az *idő-esemény kapcsolatokkal* is. Az angol mellett spanyol, olasz, francia, koreai és kínai nyelvű szövegeket is feldolgozott.
A TimeML annotálással kapcsolatban más nyelvekre is születtek még rendszerek (Caselli, dell'Orletta, & Prodanof, 2009), (Robaldo, Caselli, Russo, & Grella, 2011), (Bittar, 2009).
A legtöbb rendszer morfo-szintaktikai jellemzőket használt, a szemantikai jellemzők használata korlátozott volt.

Eseménykinyerés a közösségi médiában
Reuter és társai a közösségi alkalmazásokban (Flickr, Youtube, Panoramino) megjelenő, ugyanarról az *eseményről* szóló tartalmakat csoportosítottak (Reuter, Cimiano, Drumond, Buza, & Schmidt-Thieme, 2011), (Reuter & Cimiano, 2012).
Petrovic és társai *új eseményeket detektáltak* Twitter üzenetekből (Petrovic, Osborne, & Lavrenko, 2010). A rendszerrel 160 millió üzenetet dolgoztak fel, és jobb eredményt értek el, mint az addig legmodernebbnek számító első történet (first story) detektáló rendszerrel (Allan, Lavrenko, Malin, & Swan, 2000). *Algoritmusuk* nagyobb adathalmazok feldolgozását tűzte ki célul, adatáramlási modellt felhasználva.
McClosky és társa rendszere egy többosztályos osztályozót használt *események célszavainak detektálására*, függőségi elemzőfákat és esemény modell jellemzőket alkalmazva (McClosky & Surdeanu, 2011).
He és társai jeleket elemzett a frekvenciatérben (He, Chang, & Lim, 2007), diszkrét Fourier Transzformációt (DFT) alkalmazva, a jeleket az idősíkról a frekvenciasíkra konvertálta. Egy kiugrás a frekvenciatérben egy *eseményt jelzett* az időtérben.

Wengés társa Twitter *eseménydetektálási feladattal* foglalkoztak (Weng & Lee, 2011). Új *eseményeket detektáltak és csoportosították* azokat és szóspecifikus jeleket vizsgáltak az időtérben.
Li és társai csoportosított Twitter üzeneteket rangsoroltak új *események detektálásához* (Li, Sun, & Datta, 2012).

Orvos-biológiai alkalmazások
Az *eseménykinyerést* széleskörűen alkalmazták az orvos-biológiai doménen is (Björne, Ginter, Pyysalo, Tsujii, & Salakoski, 2010), (Chun, Hwang, & Rim, 2004), (Cohen, Verspoor, Johnson, Roeder, Ogren, & Baumgartner, 2009), (Miwa, Saetre, Kim, & Tsjujii, 2010), (Riedel, Chun, Takagi, & Tsujii, 2009), (Landeghem, Björne, Wei, Hakala, Pyysalo, & Ananiadou, 2013), (Yakushiji, Tateisi, & Miyao, 2001). Például a *molekuláris, gén események azonosítására*, amit az orvos-biológiai kutatásokban használnak fel. Yakushijiés társai *szabályalapú* rendszerében nyelvtani elemző módszert alkalmazott a mondatok fő argumentumainak megtalálásához (Yakushiji, Tateisi, & Miyao, 2001). Az *eseményinformációkat* doménspecifikus *szabályokkal* nyerték ki keret (frame) ábrázolásokkal.
A BioNLP az *eseménykinyerés* egy kiértékelési feladata (Shared Task) volt a biológiai doménen (Kim, Ohta, Pyysalo, Kano, & Tsujii, 2002). A következő területeket vizsgálták: a) magesemény és elsődleges argumentumok felismerése, b) másodlagos argumentumok kinyerése, c) tagadás felismerése.

Egyéb területek
Az *eseménykinyerés* sok más domén irányába is elmozdult, mint például a *politika* területe (Jungermann & Morik, 2008). Itt a vizsgált *események*, mint például a parlamenti választások, bejelentések, vezetőváltások, melynek szereplői az egyedek (személyek, kormányok, országok, vállalatok) és a köztük lévő kapcsolatok (miniszterek, pártvezetők) (IJntema, Sangers, Hogenboom, & Frasincar, 2012).
A *pénzügyi* felhasználás is az *eseménykinyerés* egy másik népszerű alkalmazási területe (Kakkonen. & Arendarenko, 2012), (IJntema, Sangers, Hogenboom, & Frasincar, 2012), (Li, Sheng, & Zhang, 2002). Erre egy példa a Hermes News Portal, ahol pénzügyi eseményeket gyűjtenek ki napi kereskedési döntések meghozatalához (IJntema, Sangers, Hogenboom, & Frasincar, 2012). Az eseményeket előre definiált lexikai-szemantikai minták segítségével nyerik ki.
Az 1980-as évek óta nagy igény mutatkozott eseményalapú megoldásokra olyan *biztonsággal* kapcsolatos felhasználási területekre, mint a terrorizmus, fegyveres konfliktusok, járványok. Ezek a problémák még ma is generálnak új kutatási vizsgálatokat (Atkinson, Piskorski, Tanev, Goot, Yangarber, & Zavarella, 2009), (Atkinson, Du, Piskorski, Tanev, Yangarber, & Zavarella, 2013), (Piskorski, Tanev, & Wennerberg, 287–300), (Tanev, Piskorski, & Atkinson, 2008).
Több *eseménykinyerő* alkalmazás született *újsághírek* szövegeinek elemzésére. Általában az *eseménykinyerés* eredményeinek összegzésre használták fel (Lee, Chen, & Z., 2003), ahol nagy hírüzenetekből generáltak kisebb üzeneteket az azonosított események alapján. Ez hasznosnak bizonyult hírekben lévő személyek azonosítására, ahol adott személyekhez keresnek releváns híreket *események* alapján. *Híresemény detektáló* alkalmazásokat találunk több he-

lyen algoritmikus kereskedésnél (Nuij, Milea, Hogenboom, Frasincar, & Kaymak, 2014), kockázat analízisnél (Capet, Delavallade, Nakamura, Sandor, Tarsitano, & Voyatzi, 2008) és döntéstámogató rendszereknél is (Wei & Lee, 2004).
A legtöbb *hírorientált eseménykinyerő alkalmazás* általános hírfeldolgozási célokat szolgál (Aone & Ramos-Santacruz, 2000), (Lee, Chen, & Z., 2003), (Lei, L., Zhang, & Liu, 2005), (Liu, Liu, Xiang, Chen, & Yang, 2008), (Naughton, Kushmerick, & Carthy, 2006), (Tran, Nguyen, Nguyen, Nguyen, & Phan, 2012), de alkalmaztak eseménykinyerést például tudományos (Vargas-Vera & Celjuska, 2004) és kitüntetésekkel kapcsolatos hírekre egyaránt (Xu, Uszkoreit, & Li, 2006).
Eseménykinyeréssel kapcsolatban találunk még alkalmazásokat jogi dokumentumoknál (Lagos, Segond, Castellani, & O'Neill, 2010), történelmi archívumoknál (Cybulska & Vossen, 2011) és blogoknál (Y. Nishihara, 2009) is. Újabban az eseménykinyerés nem korlátozódik dokumentumok írott szövegeire, hanem képeket, televízió-közvetítéseket és videókat is feldolgoz (Chen, Zhang, & Chen, 2007), (Kamijo, Matsushita, Ikeuchi, & Sakauchi, 2000).

Magyar szövegeken elért eredmények
Gábor és társa *események detektálásánál* csoportosító algoritmust használtak magyar igékhez (Gábor & Héja, 2007). Más szófajú események detektálásával nem foglalkoztak.
Szőts és társai a MASZEKER projekt keretében egy olyan új elveken alapuló integrált kereső rendszert fejlesztettek ki, amely adaptált (statisztikai és szimbolikus alapú) technológiák és újszerű megoldások kombinálásán keresztül teszi lehetővé a természetes nyelvű dokumentumtárakban (szövegekben) történő tartalmi keresést. Ennek keretében *események detektálásával* is foglalkoztak (Szőts, Csirik, Gergely, & Karvalics, 2010).

6.2 Szemantikus szerepek címkézése

Több munka foglakozott az elmúlt években a szemantikus szerepek címkézésével, amelyeket ebben a fejezetben mutatok be.
A szemantikus szerepek elméleti eredményei alapján kutatók számítógépes módszereket is kifejlesztettek szemantikus szerepek automatikus címkézéséhez, többek között Gildea és társa (Gildea & Jurafsky, 2002), valamint Punyakanok és társai (Punyakanok V. , Roth, Yih, Zimak, & Tu, 2004). Az itt elért eredményeket számos más NLP alkalmazás is felhasználta, mint például az összegzés (Melli, Shi, Wang, Liu, Sarkar, & Popowich, 2006), információ visszakeresés (Moreda, Navarro, & & Palomar, 2007) és a válaszkeresés (Moreda, Llorens, Saquete, & Palomar, 2011).
Az események és szerepeinek kinyerését is általában osztályozók sorozatának problémájaként definiálják (Ahn, 2006), (Chen & Ji, 2009), (Chen & NG, 2012). Először az eseményjelző (trigger) azonosítás történik, majd ezekhez azonosítják az argumentumokat. Ennek első lépéseként eldöntik, hogy az adott szó vagy kifejezés argumentum-e, és ha igen, akkor osztályozzák azokat, meghatározzák az argumentum szerepet (Grishman, Westbrook, & Meyers, 2005). SVM és Maximum Entropy osztályozók a legnépszerűbb algoritmusok ezen a területen.
A szemantikus szerepek címkézésének korai kísérletei csak korlátozott doménekkel foglalkoztak, mint például az Air Traveler Information System (ATIS) (Hemphill, Godfrey, &

Doddington, 1990). Ezekben speciális igékhez kerestek szemantikus szerepeket, előre definiált sablonok kitöltéséhez. (I want a fight to [*to city*] from [*from city*] on [*flight date*]). Az ATIS rendszerben Miller és társai kiszámították, hogy például a légi utazások szemantikus keretében egy komponens (például Atlanta) milyen valószínűséggel célállomás szerep (Miller, Stallard, Bobrow, & Schwartz, 1996).
Riloff épített egy mintaszótárat szerepek megtalálásához a terrorista-támadások speciális doménre (Riloff, 1993), (Riloff & Schmelzenbach, 1998).
Később Blaheta és társa egy doménfüggetlen rendszert tanított a Penn Treebank korpuszon olyan szerepekre, mint mód (MANNER) és idő (TEMPORAL) (Blaheta & Charniak, 2000).
Mint sok más NLP feladatban, a szemantikus szerepek címkézésénél is a kutatás a *szabályalapú* rendszerektől fejlődött a *statisztikai rendszerekig*, amelyek már felügyelt vagy felügyelet nélküli gépi tanulásra épülnek.

Szabályalapú rendszerek
A nyelvi szemantika kezdeti modelljei emberi erőfeszítéssel összeállított *szabályalapú módszereken* és az adott domén részletes ismeretén nyugszanak. Például Hirst munkája nagyrészt szabályalapú szintaktikai elemzésen és keretismereten alapult (Hirst, 1987), hasonlón a Fillmore által kifejlesztett rendszerhez (Fillmore C. J., 1976). Hirst összekapcsolta a szintaktikai összetevőket a hozzájuk tartozó keretszerepekkel. Emberi erőfeszítéssel összeállított lexikonok és nyelvtanok használatát lehet látni Pustejovsky (Pustejovsky J. , 1991) valamint Copestake és társa (Copestake & Flickinger, 2000) munkáiban is. Dahl és társai (Dahl, Palmer, & Passonneau, 1987), Hull és társa (Hull & Gomez, 1996) és Meyers és társai (Meyers, Macleod, Yangarber, Grishman, Barrett, & Reeves, 1998) szabályok halmazát alkalmazta szintaktikai alkotóelemek társításához főnévi predikátumokhoz.
Statisztikai SRL
Sok kutató 2000 óta *gépi tanítási módszerrel* tanulmányozta az SRL problémáját. (Jurafsky & Martin, 2009), (Gildea & Palmer, 2002), (Gildea & Jurafsky, 2002), (Chen & Rambow, 2003), (Gildea & Hockenmaier, 2003), (Hacioglu, Pradhan, Ward, Martin, & Jurafsky, 2004), (Moschitti, 2004), (Pradhan S. , Ward, Hacioglu, Martin, & Jurafsky, 2004), (Punyakanok V. , Roth, Yih, Zimak, & Tu, 2004), (Yi & Palmer, 2004), (Pradhan S. , Ward, Hacioglu, Martin, & Jurafsky, 2005), (Punyakanok, Roth, & Yih, 2005), (Toutanova, Haghighi, & D., 2005). Két évig a CoNLL kiírt feladatának is ez volt a témája (Carreras & Marquez., 2004), (Carreras X. , 2005). Számos gépi tanulási algoritmust kipróbáltak és sok lexikai jellemzőt mutattak be ezen a területen.
A FrameNet korpusz elkészítése a szabályalapú szemantikus feldolgozás irányából a statisztikai tanulás alapú megközelítések felé mozdította el a kutatási irányt.

Gildea és Jurafsky munkája
Gildea és társa az SRL problémáját felügyelt tanítási feladatként határozta meg és a FrameNet korpuszt használta tanító adatként (Gildea & Jurafsky, 2002). Ez a munka volt az első komolyabb próbálkozás a szemantikus szerepek címkézésére, ami statisztikai rendszert használt a FrameNet korpusszal.
Valószínűségi statisztikát alkalmaztak a szövegeken belüli keret (frame) elemek határainak azonosításához és szemantikai szerep címkék hozzárendeléséhez ezen elemekhez (Gildea &

Jurafsky, 2002). Ehhez számos lexikai és szintaktikai jellemzőt felhasználtak. A tanulmány eredményei bíztatóak voltak: 63%-os F-mértéket értek el az elemazonosítás és címkézés közös feladatán. Ehhez jellemzőket a szintaktikai elemzőfából generálták.
Főbb módszereik, eredményeik, amelyek a későbbi munkákra is hatással voltak:

- A szintaktikai információ alapvető a jó minőségű SRL-hez. Sok nyelvészeti elmélet állítja, hogy erős kapcsolat van a szintaktika és a szemantika között.
- A predikátumok szintaktikai és szemantikai korlátozásokat határoznak meg a szemantikus szerepekre, amelyekkel kapcsolatba kerülhetnek.
- Több lépeses módszer szükséges a feladat megoldásához. Különálló osztályozókat használtak a keretelemek azonosításához majd a címkéknek azokhoz való rendeléséhez.

A jelöltekhez a következő jellemzőket rendelték:

- Kifejezés típus (Phrase Type). A jelölt és célszó közötti szavak csoportjának szintaktikai kategóriája. Például NP, PP, S.
- Uralkodó kategória (Governing Category). Az elemzőfában a jelölt és a célszó közötti útvonalon a legfelső csomópont szintaktikai kategóriája.
- Fa útvonal (Tree Path). A jelölt és a célszó közötti útvonalon a kategóriák listája, nyilakkal tagolva, jelölve a fel- vagy lefelé mozgást az elemzőfában. Ez egy hatékony jellemző a predikátum és az argumentumai közötti nyelvtani kapcsolat bemutatásához.
- Helyzet (Position). Bináris indikátor, ami jelzi, hogy a jelölt a predikátum előtt vagy után van.
- Szemlélet (Voice). Bináris jellemző, ami megmutatja, hogy a megfigyelt predikátum szenvedő vagy cselekvő szemléletben van-e.
- Fejszó (Headword). A lexikai kifejezés fejszava.

Egy másik gyakran használt jellemző még az élek száma (length) az útvonalon (Pradhan S. , Ward, Hacioglu, Martin, & Jurafsky, 2005).

További SRL rendszerek

Gildea és társa később bemutatta, hogy ugyanaz a rendszerfelépítés jól teljesít nem csak a FrameNet szerepek, hanem a Propbank szerepekre is (Gildea & Palmer, 2002). Nem sokkal a megjelenése után a PropBank a statisztikai SRL kutatóinak népszerű forrásává vált, támogatva sok tanulmányt és motiválva számos kiértékelési versenyt: CoNLL Shared Tasks (Carreras & Marquez., 2004), (Carreras & Marquez, 2005), (Surdeanu, Johansson, Meyers, Marquez, & Nivre, 2008).
Bár a különböző SRL rendszerek különböző gépi tanítási megközelítést alkalmaztak és különböző lexikai jellemzőket használtak, általánosságban ugyanazokat a lépéseket alkalmazták: előfeldolgozás, argumentum azonosítás, argumentum osztályozás, utófeldolgozás. Ezen a 4-lépéses osztályozási megközelítésen kívül voltak akik BIO címkézést alkalmaztak (Marquez, Comas, Gimenez, & Catala, 2005), vagy CRF-et a fastruktúrán (Cohn & Blunsom, 2005), (Moreau & Tellier, 2009).
Angol nyelvű szövegekre az SRL területén sokáig konstituens reprezentációt alkalmaztak, mint például a Charniak elemző (Charniak parser). De az utóbbi években számos munkához

függőségi reprezentációt használtak az SRL feladathoz. Johansson és Nugues (Johansson, 2008) egy függőségi reprezentáción alapuló SRL rendszert készített. Mindkét reprezentáció jól alkalmazható SRL feladatokra, közel azonos eredményeket lehet velük elérni.
Gildea és társa CCG nyelvtant (Combinatory Categorial Grammar) (Steedman, 2000) használt az alapjellemzők teszteléséhez (Gildea & Hockenmaier, 2003). A CCG nyelvtan közvetlen kapcsolatot próbál feltérképezni a szintaktikai struktúra és a szemantikai jelentés között. Ez a rendszer a fő (számozott) argumentumokon jobban teljesített, mint Gildea és társa rendszere (Gildea & Palmer, 2002).
Chen és társa épített egy rendszert a TAG nyelvtanon (Tree Adjoining Grammar) (Chen & Rambow, 2003), aminek segítségével mélyebb szintaktikai szintekről gyűjtötték a jellemzőket.
Surdeanu és társai döntésifa algoritmust használtak az SRL feladathoz (Surdeanu M. , Harabagiu, Williams, & Aarseth, 2003).
Pradhan és társai SVM (Support Vector Machines) algoritmust használtak az SRL-hez, (Pradhan, Hacioglu, Krugler, Ward, Martin, & Jurafsky, 2003). Jellemzőik: igei csoportok, névelemek, részleges útvonal és a fejszó szófaja.
Thompson és társai egy generatív modellt készítettek az SRL feladatra a FrameNet-el (Thompson, Levy, & Manning, 2003). Ez a típus különbözik a Gildea és társa által alkalmazott diszkriminatív modelltől (Gildea & Jurafsky, 2002).
Xue és társa a hagyományosan alkalmazottak mellett más szintaktikai jellemzőket is használtak az SRL feladathoz, ezzel javítva az eredményeken (Xue & Palmer, 2004).

Kiértékelési feladatok (Shared Task) a Szemantikus szerepek címkézése témakörben
CoNLL
A 2004-es és a 2005-ös CoNLL kiértékelési feladatának témája a szemantikus szerepek címkézés volt (Carreras & Marquez., 2004), (Carreras & Marquez, 2005). Az detektálandó argumentumok között szerepeltek az általános összetevők (például agent, goal, patient) mellett a kiegészítő összetevők is (például manner, temporal, locative). A CoNLL szervezői biztosították a résztvevők számára a számítógépes rendszerek kifejlesztéséhez az adatokat a PropBankból. Punyakanok és társai rendszere végzett az első helyen a CoNLL-2005 kiértékelésen (Punyakanok, Roth, & Yih, 2005). Egy alap SRL rendszert fejlesztettek tovább úgy, hogy osztályozójukat a Charniak elemzőnek (Collins, 2003) nem csak a legjobb eredményt adó elemzésével futtatták le, hanem a további négy legjobb elemzésével is. A végső osztályozási döntéseket ezen elemzések alapján hozták meg az osztályozók előrejelzéseinek kombinálásával.

Senseval
A Senseval-33 (Litkowski, 2004) témája szintén egy SRL feladat volt angol nyelvre a FrameNet felhasználásával. A feladat Gildea és társa munkájára alapult (Gildea & Jurafsky, 2002). Mint a CoNLL feladatnál, a szervezők itt is biztosították a mondatokat és a cél igéket, a résztvevőknek itt is az argumentumokat kellett felismerni.
Toutanova és társai egy egyesített modellt mutattak be SRL-re (Toutanova, Haghighi, & D., 2005), ami egy algoritmust használt a kiválasztott szemantikus szerepek rendezéséhez.

Fillmore és társai is készítettek egy módszert az eseménykinyeréshez a FrameNet korpuszon (Fillmore, Narayanan, & Baker, 2006). Kerethordozó szavakat (igéket, főneveket és melléknévi igeneveket), valamint ezen szavakra épülő nyelvtani szerkezeteket (szerepek) detektáltak. Surdeanu és társai három különböző SRL rendszert építettek (Surdeanu, Marquez, Carreras, & Comas, 2007). Az első két modell (M1 & M2) olyan szekvencia címkézési technikákat alkalmazott, amelyek már ismertek a névelemek detektálásánál. Ez a BIO (beginning-inside-outside) rendszer első lépésben egy szintaktikai feldolgozót használ szintaktikai jellemzők kinyeréséhez, majd egy lineáris dekódolással választja ki a legjobb szerepeket. Az M3 rendszer osztályozta a Charniak elemző által generált összetevőkkel. A végső szerep előjelzését ezen három rendszer kombinálásával kapták.
Toutanova és társai egy összekapcsolt modellt készített SRL-hez (Toutanova, Haghighi, & Manning, 2008). A minőség javításához az ige argumentumai közötti egymásra hatást tanulmányozták és figyeltek meg szabályszerűségeket, mint például, hogy a legtöbb argumentum csak egyszer szerepel egy adott igéhez (például általában csak egy ágens [Agent] tartozik egyes igékhez). Más bonyolultabb minták is előfordulnak. Például, ha van egy címzett egy adással kapcsolatos mondatban, akkor valószínű, hogy van adomány is.
Vickrey és társa is készített egy módszert az SRL feladathoz (Vickrey & Koller, 2008). Először kézzel írt szabályokat készítettek a mondatok egyszerűsítéséhez, majd ezen mondatokkal log-linear (Maximum Entropy) osztályozót tanítottak az SRL címkézéshez.
Punyakanok és társai SRL rendszere csak az elemzőfában az ige alatti közvetlen elemeket tekintette lehetséges argumentumoknak (Punyakanok, Roth, & Yih, 2008), ezzel kiszűrve sok valószínűtlen jelöltet. Valamint a szemantikus szerepeket nem csak elszigetelten kezelték, hanem egymásra hatásukat is vizsgálták.
Llorens és társai felhasználták a szemantikus szerepek címkézését a TimeML eseményfelismeréshez és -osztályozáshoz (Llorens, Saquete, & Navarro-Colorado, 2010). A CRF (Conditional Random Fields) (Lafferty, McCallum, & Pereira, 2001) tanítási algoritmust alkalmazták és a következő jellemzőcsoportokat használták: morfo-szintaktikai, WordNet-alapú, szemantikai szerep jellemzők. Rendszerük a felismeréshez 81.40%-os és az osztályozáshoz 64.20%-os F-mértéket ért el.
Naradowsky és társai egy Markov modellen alapuló módszert alkalmaztak SRL-re (Naradowsky, Riedel, & David, 2012). Tackstrom és társai egy dinamikus programozási módszert alkalmaztak SRL-re (Tackstrom, Ganchev, & Das, 2015).

Magyar szövegeken elért eredmények

Szemantikus szerepek címkézésére magyar nyelvű szövegekre is készültek már munkák. Farkas és társai (Farkas, Konczer, & Szarvas, 2004) a szemantikuskeret-illesztésre *szabályalapú* módszert használtak. A szabályalapú módszerrel ellentétben a gépi tanulásos módszer nem igényel annyi erőforrást és előfeldolgozást és automatikusan alkalmazható más doménekre is.
Ehmann és társai (Ehmann, Lendvai, Miháltz, Vincze, & László, 2013) pszichológiai témájú szövegeken szemantikus szerepek címkézésénél csak két általános szerepet keresnek: az ágens és az elszenvedő szerepeket.
Szőts és társai a MASZEKER projekt keretében egy új elveken alapuló integrált kereső rendszert fejlesztett ki, amely adaptált (statisztikai és szimbolikus alapú) technológiák és újszerű megoldások kombinálásán keresztül teszi lehetővé a természetes nyelvű dokumentumtárakban

(szövegekben) történő tartalmi keresést. Ennek keretében tematikus szerepeket is feldolgoztak, témakörönként és kontextusokként definiáltak szereprelációkat (Szőts, Csirik, Gergely, & Karvalics, 2010).
Prószéky és társai eseménykinyerésre és a résztvevő azonosítására fejlesztették a NewsPro projektet (Prószéky, NewsPro: automatikus információszerzés gazdasági rövidhírekből, 2003) (Prószéky & Kis, 2003). A projekt célja a szöveges üzleti hírekből való információkinyerés. A NewsPro rendszer a feldolgozandó szövegek – a jelenlegi prototípusrendszerben gazdasági rövidhírek – tartalmi előkészítését végzi. Strukturált, adatbázisba illeszthető formába alakítja az elemzett szöveget. A bevitt rövidhírből XML-formában ábrázolt adatstruktúra keletkezik, amelynek elemei például adatbányász-alkalmazásokban használhatók fel. Azonosítja a mondatban az eseménysémákat, amelyek meghatározzák az esemény fajtáját és a résztvevőket.
Siklósi és társai munkája a DigiCons rendszer (Siklósi & Novák, 2015), (Siklósi, Novák, & Prószéky, 2014), (Siklósi & Novák, 2014), amelynek keretében kórházi zárójelentésekből történt az eseménykivonatolás. A kórházi dokumentumok elkészítésének és tárolásának módja gyakran akadályozza a tartalmuk hozzáférhetőségét. A DigiCons rendszer ezen dokumentumok sorait egyezteti a dokumentumok címeivel és alcímeivel. Emellett a klinikai szövegekben lévő rövidítéseket egyértelműsíti és oldja fel, valamint többszavas kifejezéseket azonosít és közöttük lévő hasonlóságot állapít meg.
Miháltz és társai bemutatták igei vonzatkereteket azonosító kereslet-kínálat alapú nyelvi elemzőjük működését, valamint megvizsgálták a MetaMorpho és a VerbNet összekapcsolásának és a tematikus szerepek gépi úton történő átvitelének lehetőségét (Miháltz, Indig, & Prószéky, 2015). Az utóbbi keretében *szabályokat* készítettek események tematikus szerepeinek azonosítására, amely szabályokat a későbbiekben fel lehet használni SRL alkalmazásokban.

6.3 Összegzés

Ebben a fejezetben bemutattam a disszertációmhoz kapcsolódó megelőző munkákat. Először az események detektálása és osztályozása, majd a szemantikus szerepek címkézése területén elért eredményeket ismertettem. A kezdeti kutatási lépéseket nyelvész kutatók tették meg ezeken a tudományterületeken is, mint ahogy a számítógépes nyelvészet sok más részénél, ezért mindkét esetben először a nyelvészeti eredményeket ismertettem, majd a számítógépes és gépi tanulásos eredményeket. Külön kiemeltem a magyar szövegeken elért eredményeket.

A következő fejezetekben saját eredményeimet ismertetem:

7. fejezet: Igei és főnévi igenévi események detektálása és osztályozása természetes nyelvű szövegekben
8. fejezet: Főnévi események detektálása magyar nyelvű szövegekben függőségifa- és konstituensfa-alapú reprezentációval és WordNettel
9. fejezet: Események szemantikus szerepeinek automatikus címkézése

7 Igei és főnévi igenévi események detektálása és osztályozása természetes nyelvű szövegekben

Ebben a fejezetben bemutatom eredményeimet, amiket az igei és főnévi igenévi események detektálása és osztályozása területén értem el.

7.1 Bevezetés

A természetes szövegekből történő információkinyerés egyik fontos részterülete a névelemek azonosítása mellett az események detektálása (Jurafsky & Martin, 2009). Szövegekben lévő események detektálása és analizálása fontos szerepet tölt be számos számítógépes nyelvészeti alkalmazásban, mint például a kivonatolás és a válaszkeresés. Az események felismerése, analizálása, és hogy hogyan viszonyulnak egymáshoz időben, fontos a szöveg tartalmának megismerésében.
A szövegekben a legtöbb esemény igékhez kapcsolódik, és az igék általában eseményeket jelölnek, ezért jelen fejezetben külön foglalkoztam az igei és főnévi igenévi eseményekkel. Például: *S természetesen ismét* ***megrendezik*** *a Forma-1-es futamot a Hungaroringen.*
Az igék és főnévi igenevek közül azonban nem mindegyik tekinthető eseményjelölőnek (például: van, volt, lesz, marad, segédigék), így ezek kiszűrésére külön figyelmet kell fordítani. Például: *De ha a politikai vezetés meg* ***akarja*** *őrizni a légiót, nem szabad a reformmal kikezdenie értékrendjét és összetartását.* A szavak elemzése nem elegendő, a szó szövegkörnyezetét is vizsgálni kell. További példák a többértelműségre: *dob* (ige, főnév), *vár* (ige, főnév).
Vannak olyan események, amelyeket két szóval fejezünk ki (pl. *döntést hoz*), ezek szintén külön kezelést igényelnek. Több munka is foglalkozott már részletesen a többszavas igei kifejezésekkel (Vincze V. , 2009), (Vincze, Zsibrita, & Nagy, 2013), (Subecz & Csák, 2014), ezek eredményeit felhasználtam.
Jelen fejezetben bemutatom gépi tanuló módszeremet, amely automatikusan képes magyar nyelvű szövegekben igei és főnévi igenévi események detektálására és osztályozására függőségi reprezentáció és WordNet alkalmazásával. Modellemet a gépi tanulás mellett kiegészítettem szabályalapú módszerekkel is.
A rendszer bemenete egy tokenszinten címkézett tanító korpusz, modellem jelöltjei a mondatokban lévő igék és főnévi igenevek voltak. A feladatot három részre osztottam. A szövegekben először az egy- és többszavas főnév + igei és főnévi igenévi kifejezéseket válogattam ki, majd a kiválogatottak közül detektáltam az eseményeket. A megtalált eseményeket ezután osztályoztam. A feladat megoldásához statisztikai és szabályalapú módszereket is alkalmaztam.
A gépi tanulásos modellhez gazdag jellemzőkészleten alapuló osztályozót használtam. Módszeremet a Szeged Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) öt különböző doménjén vizsgáltam meg.
Magyar nyelvű szövegeimen *függőségi reprezentációt* használtam fel, mivel ez jól használható szabad szórendű nyelvek elemzésére, így a magyarra is.

Az általam megvalósított megközelítés *gépi tanuló módszer* alapján detektálja és osztályozza az eseményeket, amit *szabályalapú módszerrel* is kiegészítettem. Megoldásomban a vizsgált szavak szemantikai jellemzéséhez felhasználtam a magyar **WordNetet** (Miháltz, és mtsai., 2008). Mivel egy szóalakhoz több jelentés is tartozhat a WordNetben, ezért az egyes jelentések között egyértelműsítést (Word Sense Disambiguation, WSD) végeztem a **Lesk algoritmussal** (Jurafsky & Martin, 2009).

Sok kutatás foglalkozik az események detektálásával. A legtöbb munkában csak adott eseményekkel foglalkoznak (például üzleti), vagy még azon belül is csak kiemelt eseményekkel (például cégfelvásárlás). Jelen munkámban **minden** igei és főnévi igenévi esemény detektálásával és osztályozásával foglalkoztam.

A fejezetben ezen a területen alkalmazott új módszereket mutatok be. Algoritmusaimat tesztadatbázisokon kiértékelve versenyképes eredményeket érnek el az eddig bemutatott angol nyelvű eredményekkel összehasonlítva.

A **detektálásnál** 95,5-ös, a négy **osztályozásnál** 87,63; 74,04; 69,20 és 82,34-es F-mértéket értem el. Tudomásom szerint magyar szövegekben található *minden igei és főnévi igenévi esemény* detektálására és osztályozására ez az első angol nyelvű kutatási eredmény.

7.2 Kapcsolódó munkák

Több kutatás foglalkozott már angol nyelvű szövegekre igei események detektálásával és osztályozásával.

Bethard statisztikai jellemzők alapján detektált eseményeket (Bethard S. , 2002), figyelembe véve többszavas kifejezéseket is. A következő jellemzőcsoportokat használta fel a modelljéhez: az adott szó, trigramok a szó elején, végén, morfológiai jellemzők, szófaj, szintaktikai jellemzők, időbeliség kifejezése, tagadási jellemző, WordNet hipernim jellemző. Nem csak a vizsgált szóra, hanem a környező néhány szóra is kigyűjtötte ezeket a jellemzőket. Detektálásra a modell 88,3-os F-mértéket ért el, osztályozásra 70,7-ot.

Llorens és társai CRF modellt alkalmazott események detektálásához és osztályozásához szemantikai szabályok felismerésével (Llorens, Saquete, & Navarro-Colorado, 2010). Morfológiai, szintaktikai, szemantikai jellemzőket használtak fel, egyes jellemzőket nem csak az adott szóhoz, hanem néhány szavas környezetükhöz is kigyűjtöttek. Detektálásra a modell 91,33-os F-mértéket ért el, osztályozásra 73,51-ot.

Marsic csak igei események detektálásával és osztályozásával foglalkozott (Marsic, 2011), statisztikai módszereket használva a feladathoz, morfológiai és szintaktikai jellemzők segítségével. Detektálásra a modell 86,49-os F-mértéket ért el.

Jacobs és társai (Jacobs, Lefever, & Hoste, 2018) felügyelt gépi tanító algoritmust alkalmazott eseménydetektálásra angol gazdasági újságcikkeken. Tíz fajta gazdasági eseménytípust detektáltak, amihez kétfajta tanító algoritmust implementáltak: szupportvektorgépeket (SVM) és egy szó-vektor alapú neurális hálózatot (RNN-LSTM). Ehhez lexikai és szintaktikai jellemzőket alkalmaztak.

Az előző publikációk angol nyelvre készültek. Bittar francia nyelvű szövegekhez végzett eseménydetektálást (Bittar, 2009), detektálásra a modell 88,8-os F-mértéket ért el. Gábor és társa csoportosító algoritmust használtak magyar igékhez (Gábor & Héja, 2007).

7.3 Az igei és főnévi igenévi események

A szövegekben a leggyakoribb események az igei és főnévi igenévi események. Példák igei és főnévi igenévi eseményekre: *olvas, olvasni, alszik, aludni*. Az eseményjelleggel kapcsolatban beszélhetünk többértelműségről is és nem minden igét és főnévi igenevet tekinthetünk esemény-indikátornak. Például nem események: *tud, akar* (segédigék) stb., ezért különös figyelem szükséges ezek kiszűrésére. A szavak elemzése nem elegendő, a szó **szövegkörnyezetét is vizsgálni kell**. További példák a többértelműségre: *dob* (ige, főnév), *vár* (ige, főnév).

7.4 A Korpusz és az alkalmazott programcsomagok

Alkalmazásomban a **Szeged Dependency Treebank** (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) egy részét használtam fel, ami *5000 mondatot* tartalmaz a következő területekről: *üzleti rövidhírek, szépirodalom, jogi szövegek, újsághírek, fogalmazás*. Mind az öt területre az első 1000 mondatot választottam ki, tanításhoz és kiértékeléshez tízszeres keresztvalidációt alkalmaztam.
A mondatokat nyelvész segítségével két személy *annotálta* a detektáláshoz és az osztályozáshoz is. A detektálásnál az eseményjelöltekhez jelölték be, hogy események vagy sem, az osztályozásnál pedig a megtalált eseményeket sorolták adott kategóriákhoz. Az annotátorok közötti egyetértés a detektálásnál 87%-os volt, az osztályozásnál 81% (ilyen százalékban jelölték azonosan a jelölteket).
A feladatokat bináris osztályozásra vezettem vissza. Az osztályozáshoz a *Weka programcsomagnak* (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009) a C4.5 döntési fa algoritmust implementáló J48 tanuló algoritmusát alkalmaztam. A szavak morfológiai elemzésére, majd szófaji egyértelműsítésére és a mondatok függőségi nyelvtan szerinti szintaktikai elemzésére a Szeged Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) annotált elemzését használtam fel.

7.5 Többszavas kifejezések detektálása

A feladat első részeként detektáltam a szövegekben a *többszavas kifejezéseket*. A többszavas kifejezések detektálásának eredményét nem használtam fel az események felismeréséhez, ez egy kiegészítő szolgáltatása az alkalmazásnak az események detektálása mellett. Az előző részben bemutatott annotált korpusz tartalmazta a többszavas kifejezések jelölését is, ezt használtam fel a tanításra és kiértékelésre.

Pontosság	Fedés	***F-mérték***
90,48	41,30	***56,72***

7.1. táblázat: Többszavas kifejezések detektálása - alapjellemzőkkel

Az 5000 mondatot tartalmazó korpuszom 100291 tokent, és ezen belül 542 többszavas kifejezést tartalmazott. A feladathoz felhasználtam a Nagy és társai által bemutatott alkalmazás jellemzőit (Nagy, Vincze, & Zsibrita, 2013). Ők a következő *alapjellemzőket* használták fel:

felszíni jellemzők, lexikai jellemzők, morfológiai jellemzők, szintaktikai jellemzők. Ezeket alkalmazva a következő eredményeket értem el (7.1. táblázat).
Ezután modellemet kiegészítettem még egy jellemzővel: a jelöltekhez *frekvenciainformációkat* vettem fel. Minden főnév + ige párhoz meghatároztam, hogy milyen arányban volt a tanító korpuszon többszavas kifejezés. A tanításnál és a kiértékelésnél felhasználtam ezt az arányt is, mint jellemzőt. Ezzel a kiegészítéssel a következő eredményt értem el (7.2. táblázat):

Pontosság	Fedés	***F-mérték***
96,43	58,70	***72,97***

7.2. táblázat: Többszavas kifejezések detektálása - alapjellemzőkkel és frekvenciainformációkkal

Ez a jellemző jelentősen javította az eredményt.

7.6 Igei és főnévi igenévi események detektálása

Ebben a modulban az igei és főnévi igenévi eseményeket **detektáltam**. A feladatot bináris *osztályozásra* vezettem vissza, amit *szabályalapú módszerrel* is kiegészítettem. Ehhez a modulhoz külön osztályozót készítettem, ahol az eseményjelöltek az igék és a főnévi igenevek voltak. Az 5000 mondatom 10628 igét és főnévi igenevet tartalmazott, ezek voltak az eseményjelöltek. Az annotátorok ebből 6479-et jelöltek eseménynek.
Kiemelt feladatomnak tekintettem olyan jellemzőcsoportok részletes kidolgozását, amelyek figyelembe veszik a magyar nyelv sajátosságait. Ezek a *morfológiai* és a *függőségifa-alapú jellemzőcsoportok* voltak.

7.6.1 Jellemzőkészlet

A jelöltekhez a következő *jellemzőcsoportokat* definiáltam:

- Felszíni jellemzők
- Lexikai jellemzők
- Morfológiai jellemzők
- Szintaktikai jellemzők (Függőségi reprezentáció)
- Szemantikai jellemzők (WordNet)

- **Felszíni jellemzők:** bigramok, trigramok: A vizsgált szavak elején és végén lévő 2-es, 3-as betűcsoportok. Ezeken kívül: szóhossz, lemmahossz, valamint a szó sorszáma a mondaton belül.
- **Lexikai jellemzők** (bináris jellemzők): Az adott szó létige, vagy segédige-e? Egy-egy listába kigyűjtöttem a létigéket és a segédigéket. Ez a jellemző jelezte, hogy az adott szó szerepel-e valamelyik listában. Mivel egy szónak az eseményjellegét meghatározhatja az is, hogy előtte, vagy utána áll-e létige vagy segédige, ezért ezt a négy bináris jellemzőt is felhasználtam.
- **Morfológiai jellemzők**: Mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért számos morfológiaalapú jellemzőt definiáltam. Felvettem a jellemzők közé a jelölt *lem-*

máját. Alkalmaztam az adott szó tövét és toldalékait (prefixum, szufixum) *szózsák modellel*. A szózsákba kigyűjtöttem a jelöltekhez a szótövet és a toldalékokat az elemző igei é főnévi igenévi elemzéseiből. A szótő és igekötő meghatározásához felhasználtam a magyarlanc nyelvészeti program RFSA morfológiai elemzőjét (Zsibrita, Vincze, & Farkas, 2013). Ha az elemző egy jelölthöz több ilyen elemzést is megad, akkor mindegyiket a szózsákba tettem. A szózsák modellhez a következő módszert használtam: Először a tanító halmaz alapján kigyűjtöttem minden toldalékhoz, hogy milyen valószínűséggel tartozik eseményjelölthöz. Majd ezen értékek alapján minden jelölthöz két jellemzőt határoztam meg: MorfológiaiSzózsákÁtlag és MorfológiaiSzózsákLegnagyobb. A *MorfológiaiSzózsákÁtlag* jellemző esetén a jelölthöz meghatároztam a szótőre és a toldalékaira kiszámolt valószínűségek átlagát. Nagy átlag arra utal, hogy a jelölt szótöve és toldalékai között fontos elemek vannak az eseményjelleg szempontjából. A *MorfológiaiSzózsákLegnagyobb* jellemző esetén hasonlóan az előzőhöz, de itt minden jelöltnél a szótő és a toldalékok közül a legnagyobb valószínűséget választottam ki. Nagy maximális érték arra utal, hogy a jelölt szótöve és toldalékai közül legalább az egyik fontos az eseményjelleg szempontjából.

Az eseményjelöltek MSD-kódját (morfológiai kódrendszer, morphological coding system) felhasználva a következő morfológiai jegyeket definiáltam: típus(SubPos), mód(Mood), eset(Cas), idő(Tense), személy(PerP), szám(Num), határozottság(Def). Ezeken kívül a következő jellemzőket is definiáltam: az adott szó valamint az előtte és az utána álló szó szófaja.

Szintaktikai jellemzők (Függőségi reprezentáció): Definiáltam a vizsgált eseményjelölt gyerekeinek szintaktikai jellemzőit a függőségi elemzőfa alapján *szózsák modellel*. A szózsákba tettem a kapcsolatok címkéit (például alany, tárgy, …) és a kapcsolatban lévő szavak lemmáját. A Morfológiai jellemzőcsoportnál bemutatott szózsák módszerrel készítettem el a *SzintaktikaiSzózsákÁtlag* és a *SzintaktikaiSzózsákLegnagyobb* jellemzőket.

- **Szemantikai jellemzők (WordNet)**: Ehhez a Magyar WordNetet (Miháltz, et al., 2008) használtam fel, ami összesen 42288 synsetet tartalmaz, amiből 3611 az igei synset. Ezen jellemzőnél a WordNet hipernim hierarchiájában található szemantikai kapcsolatokat használtam fel. A korábbi munkákhoz képest a következő *új módszert* alkalmaztam. Először egy *külön modellel* a tanító halmaz alapján kiválogattam azokat a synseteket, amelyek alá jellemzően események tartoznak. Ennél a modellnél minden eseményjelölthöz jellemzőként kigyűjtöttem a hipernimáit. A modell a tanítóhalmazon a jellemzők alapján kiválogatta *döntési fába* azokat a synseteket, amelyek alá jellemzően események tartoznak. A vizsgált 3611 synsetből 95-öt gyűjtött ki a döntési fába. Ezek a synsetek fontosak a jelöltek eseményjellegének eldöntéséhez, mert alájuk jellemzően események tartoznak. A *fő modellnél* ezt a 95 synsetet a tanításnál a jellemzőhalmazhoz adtam bináris jellemzőként. Ezzel minden eseményjelölthöz jellemzőként definiáltam, hogy szerepel-e valamelyik kiválogatott synset hiponimái között. Módszerem *egyik előnye* a megfelelő synsetek *automatikus* kigyűjtése. A WordNet szerteágazó hipernim kapcsolatrendszerében az események nem csak néhány synset alá tartoznak, ezért az összes megkereséséhez összetett vizsgálat szükséges. *Másik előnye*, hogy általánosan, változtatás nélkül alkalmazható más olyan feladatokban is, ahol közös hipernim csomópontokat, kapcsolatokat kell megkeresni a WordNet hierarchiában adott szavak csoportjához. Ezt a módszert alkalmaztam később az események osztályozásánál is. Mivel egy szóalakhoz több jelentés is tartozhat a WordNetben, ezért az egyes jelentések között *jelentésegyértelműsítést* végeztem (WSD, word sense disambiguation) a Lesk algoritmussal (Jurafsky & Martin, 2009)

a következő módszerrel: A WordNetben a synsetekhez definíció és példamondatok tartoznak. Többjelentésű eseményjelölt esetén megszámoltam, hogy az eseményjelölt szintaktikai környezetében lévő szavak közül hány található meg az egyes WordNet jelentések definíciói és példamondatai között (stopszó szűrés után). Azt a jelentést választottam, amelyik a legtöbb közös szót tartalmazta.

A *vektortér méretét csökkentettem* a következő módszerrel: csak azokat a *jellemző-előfordulásokat* vettem fel az osztályozáshoz, amelyek a tanító halmazon *legalább háromszor* szerepeltek. Ezzel csak az osztályozás szempontjából jelentéktelen jellemző-előfordulásokat hagytam ki.

Jellemzők száma az egyes csoportokban (7.3. táblázat):

Felszíni	7
Lexikai	6
Morfológiai	12
Szintaktikai (Függőségi reprezentáció)	4
Szemantikai	1-10

7.3. táblázat: jellemzők száma az egyes csoportokban - detektálás

A gépi tanuló módszeremet kiegészítettem **szabályalapú módszerrel** is. A jogi korpuszon sok olyan kifejezés volt, amelyekben a vizsgált ige más szövegekben általában eseményt jelöl, de ebben a szövegkörnyezetben nem. Például: *A törvény **kimondja**, hogy… Az okirat **meghatározza**, hogy…* Ezekhez az esetekhez definiáltam a következőhöz hasonló szabályokat: Ha Alany=”törvény” És Jelölt=”kimondja” Akkor Jelölt ≠ Esemény.

A kiértékelés során a pontosság(P), fedés(R) és F-mérték(F) metrikákat használtam.
Először porlasztásos méréssel vizsgáltam meg az egyes *jellemző csoportok* jelentőségét az adott feladathoz. Majd megvizsgáltam az alkalmazás működését külön-külön az öt *részkorpuszon* is.
Modellem teljesítményének kiértékeléséhez *két baseline* megoldást vizsgáltam. Az egyikben minden igét és főnévi igenevet eseménynek tekintettem (Baseline 1), a másikban csak azokat az igéket és főnévi igeneveket tekintettem eseménynek, amelyek nem létigék és nem segédigék (Baseline 2). A mérésekhez tízszeres keresztvalidációt használtam.

7.7 Eredmények – Eseménydetektálás

Baseline méréseim eredményeit a következő táblázat tartalmazza (7.4. táblázat).

	Pontosság	Fedés	**F-mérték**
Baseline 1	67,15%	100%	**79,45 %**
Baseline 2	75,23%	97,16%	**84,37 %**

7.4. táblázat: Baseline eredmények F-mérték - detektálás

A további eredményeken látni fogjuk, hogy *gépi tanulási módszerem jóval felülteljesítette a Baseline mérés eredményét*.

A modellem eredményei

Teljes jellemzőkészlettel, a következő eredményeket értem el (7.5. táblázat):

Pontosság	Fedés	**F-mérték**
94,89	96,47	**95,67**

7.5. táblázat: Eredmények teljes jellemzőkészlettel

Következő lépésként megvizsgáltam, hogy az egyes **jellemzőcsoportok** hogyan befolyásolják a gépi tanulórendszer eredményeit. Ehhez *porlasztásos mérést* végeztem, ahol a teljes jellemzőkészletből elhagytam az egyes jellemzőcsoportokat, majd a maradék jellemzőkre támaszkodva tanítottam (7.6. táblázat). Az eredmények alapján a leghasznosabbnak a *morfológiai, a szintaktikai* és a *szemantikai* jellemzők bizonyultak.

Elhagyott jellemzők	**Pontosság**	**Fedés**	**F-mérték**	**Eltérés**
Felszíni	94,52	96,50	*95,50*	+0,02
Lexikai	94,67	96,16	*95,41*	-0,07
Morfológiai	93,74	95,17	*94,13*	**-1,45**
Szintaktikai (Függőségi)	93,35	95,12	*94,27*	**-1,21**
Szemantikai	94,63	96,06	*95,34*	**-0,84**

7.6. táblázat: A porlasztásos mérés eredményei

A *szemantikai* jellemzőknél *ha nem alkalmaztam a Lesk algoritmust*, akkor kissé gyengébb eredményt kaptam: a „teljes jellemzőkészlettel" mérésnél a 95,67 helyett 95,23 F-mértéket. A *morfológiai* jellemzőknél ha nem alkalmazom a szózsák módszert, akkor az F=95,67 érték helyett csak F=94,92 értéket kaptam volna. Ha a *szintaktikai* jellemzőknél nem alkalmaztam volna a szózsák módszert, akkor az F=95,67 érték helyett csak F=94,61 értéket kaptam volna. Ezen eredményeken látszik, hogy *a szózsák módszer* alkalmazása szócsoportok jellemzéséhez hasznos ezen a területen is.

Majd modellemet csak az igéken teszteltem, a főnévi igenevek nélkül. A következő eredményeket értem el szabályalapú módszerrel és anélkül (7.7. táblázat):

szabályalapú módszerrel	95,84 %
szabályalapú módszer nélkül	95,20 %

7.7. táblázat: Eredmények csak az igékre F-mérték

Az eredményeken látjuk, hogy a modell az igéknél jobb eredményt ért el, mint a főnévi igeneveknél (95,84, 95,67 F-mértékek).
A továbbiakban alkalmaztam a szabályalapú módszert is és az igék mellett a főnévi igeneveket is vizsgáltam.

Korpuszonként is megvizsgáltam az alkalmazás működését (7.8. táblázat). Legjobban az *Üzleti rövidhírek* doménen teljesített a modell, leggyengébben pedig a *Jogi* doménen.

Korpusz	Pontosság	Fedés	F-mérték
Fogalmazás	96,08	98,00	*97,03*
Jogi	89,74	86,42	***88,05***
Szépirodalom	95,45	97,35	*96,39*
Üzleti rövidhírek	97,86	98,56	***98,21***
Újsághírek	96,71	97,35	*97,03*

7.8. táblázat: Eredmények az egyes részkorpuszokon

7.7.1 Kiegészítő mérések az esemény detektáláshoz

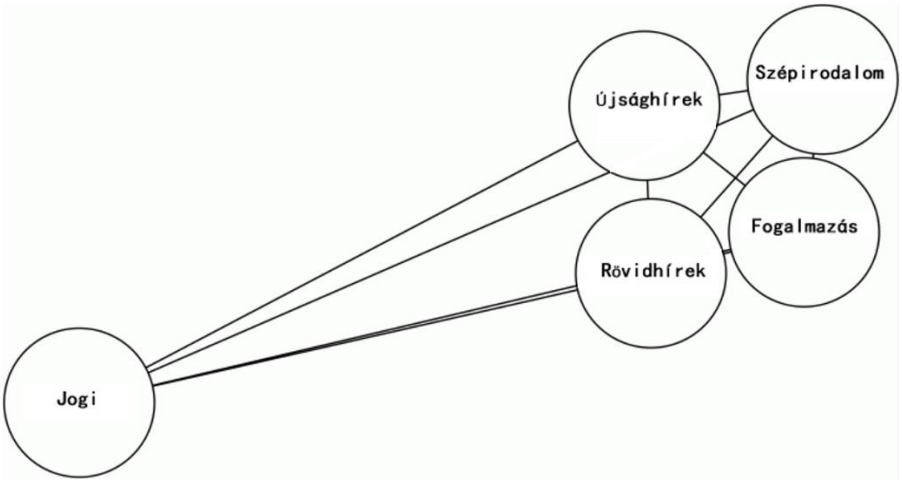
Korpusz	Pontosság	Fedés	F-mérték
Fogalmazás	96,08	98,00	***97,03***
Jogi	68,56	99,09	*81,04*
Szépirodalom	92,04	97,58	*94,73*
Üzleti rövidhírek	92,73	98,04	*95,32*
Újsághírek	91,26	98,62	*94,80*
Jogi	89,74	86,42	***88,05***
Fogalmazás	81,70	72,67	*76,92*
Szépirodalom	88,19	72,34	*79,48*
Üzleti rövidhírek	94,72	76,47	*84,62*
Újsághírek	90,74	71,90	*80,23*
Szépirodalom	95,45	97,35	***96,39***
Fogalmazás	94,68	95,64	*95,16*
Jogi	67,05	97,79	*79,56*
Üzleti rövidhírek	92,91	96,16	*94,51*
Újsághírek	91,38	96,22	*93,74*
Üzleti rövidhírek	97,86	98,56	***98,21***
Fogalmazás	92,63	95,86	*94,21*
Jogi	69,83	96,74	*81,11*
Szépirodalom	91,26	96,48	*93,79*
Újsághírek	91,29	95,86	*93,52*
Újsághírek	96,71	97,35	***97,03***
Fogalmazás	93,33	97,60	*95,42*
Jogi	72,13	98,05	*83,11*
Szépirodalom	90,83	98,09	*94,32*
Üzleti rövidhírek	93,48	98,04	*95,71*

7.9. táblázat: Keresztmérések eredményei az egyes részkorpuszokon

Első esetben arra kerestem választ, hogy az egyik korpuszon tanított modellem hogy teljesít egy másik korpuszon. Valamint, hogy ezek alapján a vizsgált korpuszok közül melyek hasonlóak egymáshoz az eseménydetektálás szempontjából és van-e olyan, amelyik jobban eltér a

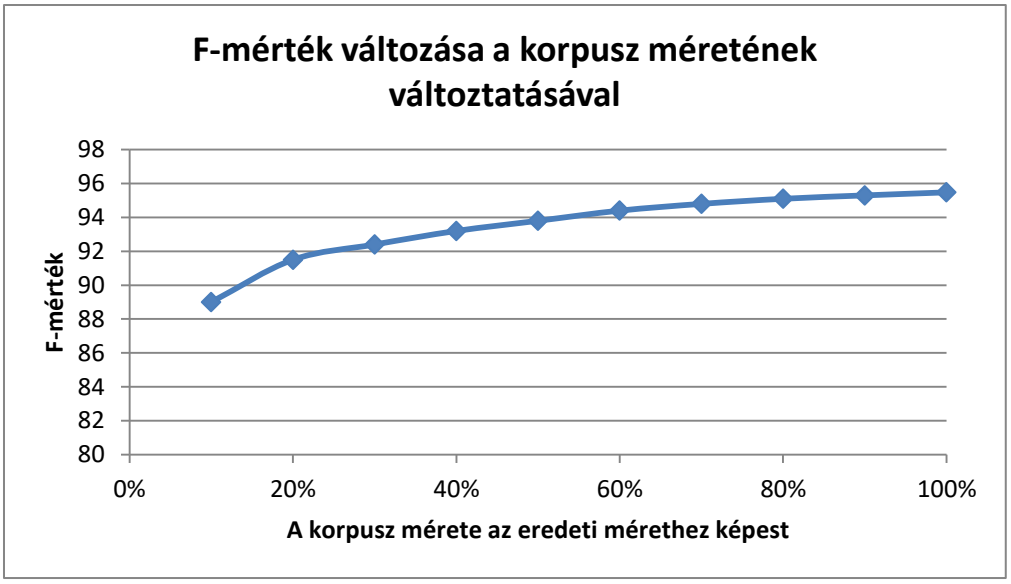
többitől. Így a **domének közötti keresztméréseknél** a forráskorpuszon tanított modellt értékeltem ki a célkorpuszon (7.9. táblázat).
Magyarázat a táblázathoz: Az öt korpusz közül az első a Fogalmazás korpusz. Ehhez tartozik a táblázat 2-6 sora. Mind az öt esetben a tanítás a Fogalmazás korpuszon és a kiértékelés az adott sor első cellájában megadott korpuszon volt. Tehát például a 3. sor esetén a tanítás a Fogalmazás, a kiértékelés a Jogi korpuszon volt.
A fogalmazás korpuszon az újsághírek doménen tanított modell teljesített a legjobban 95,42-es F-mértéket elérve. A jogi korpuszon szintén az újsághírek doménen tanított modell teljesített a legjobban 83,11-es F-mértéket elérve. A szépirodalom korpuszon a fogalmazás doménen tanított modell teljesített a legjobban 94,73-es F-mértéket elérve. Az üzleti rövidhírek korpuszon az újsághírek doménen tanított modell teljesített a legjobban 95,71-es F-mértéket elérve. Az újsághírek korpuszon a fogalmazás doménen tanított modell teljesített a legjobban 94,80-es F-mértéket elérve.

A keresztmérések eredményei alapján az egyes *domének közti hasonlóságokat* megjelenítettem egy *irányítatlan súlyozott gráf* segítségével. (7.1. ábra) Az ábrán látható, hogy a jogi korpusz a legkevésbé hasonló a többihez e szempontok alapján, a többi hasonlít egymásra.



7.1. ábra: Doménhasonlósági gráf a keresztmérések eredményei alapján

A *következő mérésben* megvizsgáltam, hogy a korpusz méretének változtatása hogyan befolyásolja az eredményeket teljes jellemzőkészlettel. A mondatok számát csökkentve romlott az eredmény. (7.2. ábra). A vízszintes tengelyen a korpusz mérete látható az eredeti (100%) mérethez képest, a függőleges tengelyen az F-mérték van ábrázolva. Látható, hogy a korpusz méretének növelése javítja az eredményeket, de az ezzel hozzáadott érték folyamatosan csökken.



7.2. ábra. Az F-mérték változása a korpusz méretének változtatásával

7.8 Igei és főnévi igenévi események osztályozása

Az igei események detektálása után **osztályoztam** azokat. Az osztályozást több szempont szerint is elvégeztem. Az *első csoportnál* az igék alapkategóriáit vizsgáltam meg: cselekvés, történés, létezés, állapot. Ezek közül az eseményeknél a **cselekvésnek és a történésnek** van fő szerepe, így ezt a két kategóriát emeltem ki. Az 5000 mondaton belül a 6479 esemény között 4158 cselekvés és 1752 történés típusú esemény volt.

Példák
Cselekvés: *Az idén is **megrendezték** az építészeti diákkonferenciát, amelynek én is meghívottja voltam.*
Történés: *Annyira izgultam, hogy ez ne történjen meg, hogy véletlenül **megcsúsztam** és **leestem** a völgybe.*

A cselekvés és a történés kategóriák együtt lefedik az események nagy részét. Modellemet, az előző osztályozástól függetlenül, olyan kategóriákon is szerettem volna tesztelni, amelyek ezeknél jelentősen kevesebb elemet tartalmaznak, de még gyakoriak. Így a *következő vizsgálathoz* kiválasztottam két kisebb, de még gyakori kategóriát, a mozgást és a kommunikációt. A korpuszon 586 mozgás és 1120 kommunikáció típusú esemény volt.
Példák
Mozgás: *Múlt év szeptemberében az osztállyal **elmentünk** kirándulni a Balatonra.*
Kommunikáció: *Este, ahogy **megbeszéltük**, lementünk fürödni.*

A többkategóriás osztályozás egyik célja az volt, hogy a detektálásra kialakított modellemet teszteljem a detektálás mellett más feladatokra is. Az osztályozásokhoz ugyanazt a **jellemzőkészletet** használtam fel, mint a detektálásnál. A feladatra négy külön bináris osztályozót építettem. A szemantikai jellemzőknél itt is a WordNet alapján először külön modellekkel kiválogattam az adott osztályra jellemző synseteket. Olyan synseteket kerestem, amelyek

hiponimái között jellemzően az adott osztály szavai szerepelnek. Ezeket a synseteket egy listában felvéve a fő modellnél jellemzőként, definiáltam, hogy az adott szó szerepel-e valamelyik ilyen synset hiponimái között.

A modell által kiválasztott synsetek száma az egyes osztályozási eseteknél (7.10. táblázat):

cselekvés	112 db
történés	88 db
mozgás	31 db
kommunikáció	53 db

7.10. táblázat: A modell által kiválasztott synsetek száma az osztályozásnál

A gépi tanuló módszeremet a mozgás vizsgálatánál kiegészítettem **szabályalapú módszerrel** is. Itt sok olyan kifejezéssel találkoztam, amelyeknél az esemény más szövegkörnyezetben mozgást jelöl, de itt nem. Például: *A részvényárak sokat **mozogtak** a nap folyamán.* Ilyen esetek detektálásához definiáltam szabályokat, mint például: Ha Alany=részvényár És jelölt=mozog Akkor jelölt-csoportja ≠ Mozgás
Az osztályozásokhoz is készítettem *baseline* megoldásokat. A cselekvés-történés osztályozásnál a *baseline* modellem minden eseményt cselekvésnek tekintett (Baseline-1). A mozgás és kommunikáció osztályozásnál a *baseline* modellemhez kiválasztottam gyakori mozgást (11 szó) (Baseline-2), illetve gyakori kommunikációt (16 szó) jelentő szavakat (Baseline-3). A modell csak ezeket a szavakat vette az adott kategóriához tartozónak, a többit nem.
A méréseket tízszeres keresztvalidációval végeztem el.

7.9 Eredmények – Esemény osztályozás

A *baseline mérésekre* a következő eredményeket kaptam (7.11. táblázat):

Baseline-1	78,38 %
Baseline-2	49,15 %
Baseline-3	45,07 %

7.11. táblázat: Baseline mérés eredményei F-mérték - osztályozás

A további eredményeken látni fogjuk, hogy *gépi tanulási módszerem jóval felülteljesítette a Baseline mérés eredményét*.

A modellem eredményei
Ha csak a *WordNet jellemzőt* alkalmaztam önállóan, a következő F-értéket értem el (7.12. táblázat):

cselekvés	86,63
történés	66,00
mozgás	65,64
kommunikáció	81,24

7.12. táblázat: Események osztályozása - csak a WordNet jellemzővel

Teljes jellemzőkészlettel a következő eredményt értem el az F-mértékre (7.13. táblázat):

cselekvés	87,18
történés	73,55
mozgás	68,64
kommunikáció	81,68

7.13. táblázat. Események osztályozása - teljes jellemzőkészlettel

Továbbiakban megvizsgáltam, hogy az egyes *jellemzőcsoportok* hogyan befolyásolják a gépi tanulórendszerem eredményeit. Ehhez *porlasztásos mérést* végeztem, aminek keretében a teljes jellemzőkészletből elhagytam az egyes jellemzőcsoportokat, majd a maradék jellemzőkre támaszkodva tanítottam (7.14. táblázat). Az eredmények alapján a leghasznosabbnak a *morfológiai*, *a szintaktikai* és a *szemantikai jellemzők* bizonyultak.

Elhagyott jellemzők	Cselekvés	Történés	Mozgás	Kommunikáció	Eltérés	Eltérés átlag
Felszíni	87,02	73,58	68,40	81,13	-0,04/+0,15/-0,11/-0,44	**-0,11**
Lexikai	86,90	73,09	68,37	80,32	-0,16/-0,34/-0,14/-1,25	**-0,47**
Morfológiai	**84,65**	**70,58**	**59,54**	**78,91**	**-1,50/-2,35/-7,83/-1,72**	**-3,35**
Szintaktikai (függőségi)	85,58	73,54	68,54	80,74	**-1,48/-1,37/-3,24/-1,43**	**-1,88**
Szemantikai	**86,21**	**72,52**	**66,02**	**80,22**	**-0,85/-0,91/-2,49/-1,35**	**-1,40**

7.14. táblázat: A porlasztásos mérés eredményei - F-mérték – Esemény osztályozás

A Szemantikai jellemzőknél *ha nem alkalmaztam a Lesk algoritmust*, akkor kissé gyengébb eredményt értem el: az eltérések átlagánál a -1,40-es érték helyett csak -1,32-es értéket értem el. A Morfológiai jellemzőknél ha nem alkalmaztam a szózsák módszert, akkor az eltérések átlagánál a -3,35-ös érték helyett csak -3,19-es értéket kaptam. Ha a Szintaktikai jellemzőknél nem alkalmaztam a szózsák módszert, akkor az eltérések átlagánál a -1,88-as érték helyett csak -0,54-es értéket kaptam. Ezen eredményeken látszik, hogy *a szózsák módszer* alkalmazása szócsoportok jellemzéséhez hasznos ezen a területen is.

Korpuszonként is megvizsgáltam az alkalmazás működését (7.15. táblázat). A *mérések átlagát tekintve* legjobban az *Üzleti rövidhírek* doménen teljesített a modell, leggyengébben pedig az Újsághírek doménen.

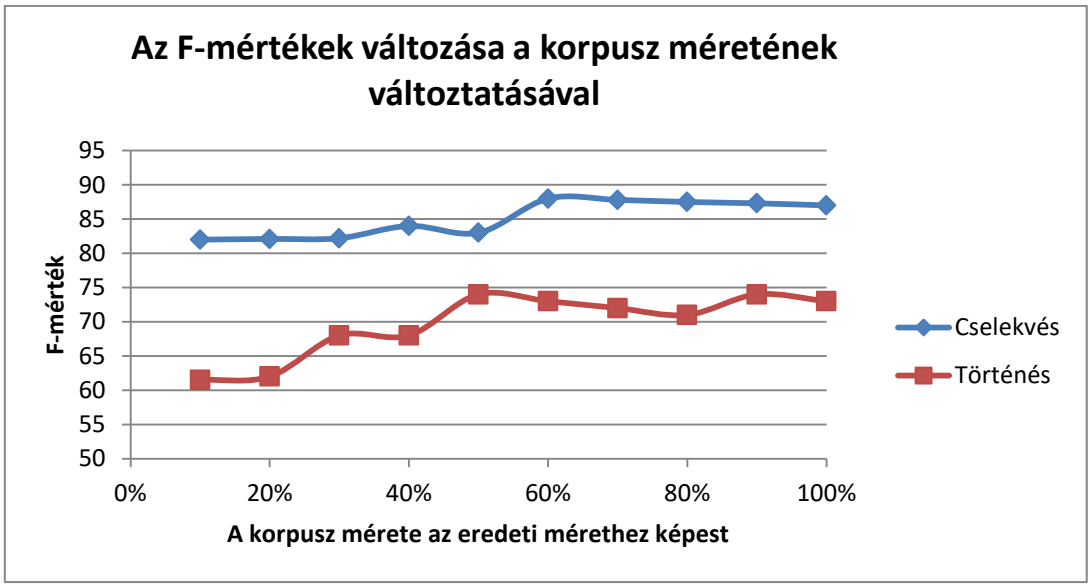
Korpusz	Cselekvés	Történés	Mozgás	Kommunikáció
Fogalmazás	85,32	56,67	86,96	75,68
Jogi	84,40	71,43	66,67	84,85
Szépirodalom	85,71	60,32	70,27	72,34
Üzleti rövidhírek	**88,89**	**92,86**	**62,37**	**85,71**
Újsághírek	83,09	47,76	58,22	70,18

7.15. táblázat: Eredmények az egyes részkorpuszokon - F-mérték

7.9.1 Kiegészítő mérések az események osztályozásához

A detektáláshoz hasonlóan az események osztályozásánál is végeztem **domének közötti keresztméréseket**, ahol a forráskorpuszon tanított modellt értékeltem ki a célkorpuszon. Legjobb eredményt a cselekvések osztályozásánál értem el, a szépirodalom doménen tanított modellel a fogalmazás korpuszon 85,5-os F-mértékkel. A leggyengébb eredményt pedig a történések osztályozásánál a fogalmazás doménen tanított modellel a szépirodalom korpuszon 53,91-os F-mértékkel.

A következő mérésben megvizsgáltam az események osztályozásánál is, hogy a korpusz méretének változtatása hogyan befolyásolja az eredményeket. Csökkentve a korpusz méretét mindkét osztályozásnál romlottak az eredmények (7.3. ábra). A vízszintes tengelyen a korpusz mérete látható az eredeti (100%) mérethez képest, a függőleges tengelyen az F-mérték van ábrázolva. Látható, hogy a korpusz méretének növelése javítja az eredményeket, de az ezzel hozzáadott érték folyamatosan csökken.



7.3. ábra: Az F-mértékek változása a korpusz méretének változtatásával

7.10 Összegzés

A fejezetben bemutattam gazdag jellemzőtéren alapuló gépi tanuló megközelítésemet, amely automatikusan képes magyar nyelvű szövegekben **igei eseményeket azonosítani és azokat osztályozni**. A problémát három lépésben oldottam meg. Először detektáltam a többszavas főnévi+igei és főnévi+főnévi-igenévi kifejezéseket, majd detektáltam az igei és főnévi-igenévi eseményeket, és végül osztályoztam azokat. Módszeremet a Szeged Dependency Treebank öt doménjén teszteltem.

Mindegyik részfeladathoz gazdag jellemzőtéren alapuló bináris osztályozót alkalmaztam, modellemet kiegészítettem szabályalapú módszerekkel is. Kiemelt feladatomnak tekintettem olyan jellemzőcsoportok részletes kidolgozását, amelyek figyelembe veszik a magyar nyelv sajátosságait. Ezek a *morfológiai* és a *függőségifa-alapú jellemzőcsoportok* voltak. Ezen a területen *új módszereket* is mutattam be. Az igei és főnévi igenévi események detektálására és

osztályozására magyar nyelvű szövegekre ismereteim szerint ez az első kutatási eredmény. Az események osztályozását több szempont szerint végeztem el: cselekvés és történés; mozgás és kommunikáció.
A modellem jellemzőkészletét teszteltem porlasztásos módszerrel, majd az öt doménen egyesével is. Algoritmusaimat tesztadatbázisokon kiértékelve *versenyképes eredményeket* érnek el az eddig bemutatott angol nyelvű eredményekkel összehasonlítva. A detektálásra 95,67-es F-mértéket, a négy szempont szerinti osztályba sorolásra pedig 87,18; 73,55; 68,64 és 81,68-as F-mértéket értem el.

7.11 A fejezet eredményei

A fejezet fő eredményeinek összefoglalása:

- Két fő területtel foglalkoztam: az **igei és főnévi-igenévi események detektálásával** és azok több szempont szerinti **osztályozásával**. Ezek mellett foglalkoztam a **többszavas kifejezések detektálásával**.
- Alkalmaztam továbbá **statisztikai és szabályalapú módszereket**.
- Morfológiai elemzéshez magyar nyelvre kialakított **morfológiai elemzőt**, szintaktikai jellemzéshez **függőségi reprezentációt**, szemantikai jellemzéshez a magyar **WordNetet** használtam fel. A WordNetben az egyes jelentések között **egyértelműsítést** végeztem a **Lesk algoritmussal**.
- Modellemben gazdag jellemzőtéren alapuló osztályozót használtam a következő **jellemzőcsoportokkal**: felszíni, lexikai, morfológiai, szintaktikai (függőségifa-alapú reprezentáció) és szemantikai (WordNet) jellemzők.
- A WordNet jellemzőnél egy **külön modellt** is készítettem, ami kiválogatja azokat a synseteket, amelyek alá jellemzően események tartoznak, majd a kiválogatott elemeket felhasználtam a fő osztályozónál. Ugyancsak a WordNet jellemzőnél kipróbáltam a **Lesk algoritmus** alkalmazásával és anélkül is a modellemet.
- Morfológiai elemzéshez felhasználtam még a magyarlanc nyelvészeti programcsomag **RFSA morfológiai elemzőjét** (Zsibrita, Vincze, & Farkas, 2013).
- A morfológiai és a szintaktikai (függőségifa-alapú) jellemzőknél alkalmaztam a **szózsák modellt** szócsoportok jellemzésére a következő szócsoportokra: a szó töve és toldalékai; a kapcsolatok címkéi és a kapcsolatban lévő szavak lemmája a függőségi reprezentációnál.
- A detektálásnál megvizsgáltam ***külön az igékre és külön a főnévi igenevekre*** a modell teljesítményét.
- **Domének közötti keresztmérést** is végeztem, ennek során a forráskorpuszon tanított modellt értékeltem ki a célkorpuszon. A domének közötti hasonlóságot gráfban ábrázoltam.
- Mérésekkel megvizsgáltam, hogy a **jelöltek számának változása** hogyan befolyásolja a detektálás és az osztályozás eredményeit.
- Az igei események detektálása után **osztályoztam** azokat. Az osztályozást több szempont szerint is elvégeztem. Az első csoportnál az **igék alapkategóriáit** vizsgáltam meg: cselekvés, történés, létezés, állapot. Ezek közül az eseményeknél a cselekvésnek

és a történésnek van fő szerepe, így ezt a két kategóriát emeltem ki. Modellemet **két kisebb, de még gyakori kategórián** is megvizsgáltam: a mozgás és a kommunikáció kategóriákon.

Eredmények a tézispontokon:
Igazoltam a következőket az igei és főnévi igenévi események detektálásánál és osztályozásánál (1. tézispont):

- *Bizonyítottam, hogy ezen a területen a legjobban teljesítő jellemzőcsoportok a morfológiai, a függőségifa-alapú szintaktikai és a szemantikai csoportok.*
- *Igazoltam, hogy a szabályalapú módszer alkalmazása a jogi korpuszon javítja a gépi tanulási rendszer eredményeit.*
- *Megmutattam, hogy a WordNet jellemzőcsoportnál a Lesk algoritmus alkalmazása javítja az eredményeket.*
- *Megmutattam, hogy a morfológiai és a szintaktikai (függőségifa-alapú) jellemzőknél a szózsák modellt hatékonyan lehet alkalmazni a következő szócsoportokra: a szó töve és toldalékai; függőségi reprezentációnál a kapcsolatok címkéi és a kapcsolatban lévő szavak lemmája.*
- *Igazoltam, hogy a detektálásnál az igékre jobb eredményt ad a modell, mint a főnévi igenevekre.*
- *Megmutattam, hogy a detektálás és az osztályozás szempontjából a Fogalmazás, Szépirodalom, Üzleti rövidhírek és az Újsághírek domének hasonlítottak legjobban egymásra, ezektől jelentősen eltért a Jogi domén.*
- *Bizonyítottam, hogy a Detektálásnál és osztályozásnál is a korpusz méretének növelése javítja az eredményeket, de a hozzáadott érték folyamatosan csökken.*

8 Főnévi események detektálása magyar nyelvű szövegekben függőségifa- és konstituensfa-alapú szintaktikai reprezentációval és WordNettel

Ebben a fejezetben bemutatom eredményeimet, amit a főnévi események automatikus detektálása területén elértem.

8.1 Bevezetés

Jelen fejezetben bemutatom gazdag jellemzőtéren alapuló gépi tanuló megközelítésemet, amely automatikusan képes magyar nyelvű szövegekben főnévi események detektálására függőségifa- és konstituensfa-alapú reprezentáció és WordNet alkalmazásával. A feladathoz gazdag jellemzőkészleten alapuló osztályozót használtam, modellem jelöltjei a mondatok főnevei voltak. A jellemzők mellé kiegészítő módszereket is alkalmaztam, amelyek javították az eredményeket. Módszeremet a Szeged Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) öt különböző doménjén vizsgáltam meg.
Vizsgálatomban *függőségifa-* és *konstituensfa-alapú* szintaktikai reprezentációt is használtam és azok eredményeit összehasonlítottam. Hipotézisem az volt, hogy jobb eredményt lehet elérni a függőségifa-alapú reprezentációval, mint a konstituensfa-alapú reprezentációval, hiszen a függőségifa-alapú reprezentáció jól használható szabad szórendű nyelvek elemzésére, így a magyarra is.
Megoldásomban a vizsgált szavak szemantikai jellemzéséhez felhasználtam a magyar *WordNet*et (Miháltz, et al., 2008). Mivel egy szóalakhoz több jelentés is tartozhat a WordNetben, ezért az egyes jelentések között egyértelműsítést (word sense disambiguation, WSD) végeztem a *Lesk algoritmussal* (Jurafsky & Martin, 2009).
Számos kutatás témája volt már az eseménydetektálás, azonban a legtöbb munka csak adott típusú eseményekkel foglalkozott (például üzleti események). Munkámban én minden fajta főnévi esemény detektálásával foglalkoztam. Ismereteim szerint főnévi események detektálására, függőségifa- és konstituensfa-alapú reprezentáció és WordNet alkalmazásával, magyar nyelvű szövegekre ez az első angol nyelven publikált kutatási eredmény. Algoritmusaimat tesztadatbázisokon kiértékelve versenyképes eredményeket érnek el az eddig bemutatott angol és más nyelvű eredményekkel összehasonlítva.

8.2 Kapcsolódó munkák

Az EVITA (Sauri R. , Knippen, Verhagen, & Pustejovsky, 2005) volt az első eseményfelismerő eszközök egyike, ami nyelvészeti és statisztikai technikák kombinálásával ismeri fel az eseményeket. Nyelvészeti ismereteken alapuló *szabályokat* használ fő jellemzőként és WordNet osztályokat is alkalmaz a főnévi események felismeréshez. A főnevek szemantikai egyértelműsítésére Bayes osztályozót használ.

Boguraev és társa (Boguraev & Ando, 2007) egy gépi tanuláson alapuló módszert mutatott be automatikus események annotáláshoz. A feladatot osztályozási problémaként kezelték és egy RRM (robust risk minimization) osztályozót használtak a megoldáshoz. Lexikai, morfológiai és szintaktikai attribútumokat használtak két- és háromszavas ablakokkal.
Bethard és társa (Bethard & Martin, 2006) esemény-felismerésre fejlesztették a STEP rendszert. Szintaktikai és szemantikai jellemzőket alkalmaztak és az esemény-felismerési feladatot osztályozásként oldották meg. Gazdag jellemzőkészletet használtak: lexikai, morfológiai, szintaktikai függőségi és WordNet osztályokat, valamint SVM (Support Vector Machine) modellt implementáltak e jellemzőkre alapozva.
Llorens és társa (Llorens, Saquete, & Navarro-Colorado, 2010) egy eseményfelismerő alkalmazást mutatott be. A jellemzőkhöz szemantikus szerepeket és szemantikai szabályokat is felvettek és az események detektálására egy CRF (Conditional Random Field) modellt építettek.
Jeong és társa (Jeong & Myaeng, 2012) függőségi reprezentációt használt, de csak a közvetlen kapcsolatokat vizsgálta a jelölt főnév és a hozzá kapcsolódó ige között. Összetett jellemzőket használtak: az ige + a kapcsolat típusa párokat. Használták a WordNetet is, de jelentés-egyértelműsítés nélkül. A MaxEnt osztályozó algoritmust alkalmazták a következő jellemző csoportokkal: lexikai, szemantikai és függőségifa-alapú jellemzők, a jellemzőket súlyozva a Kullback-Leibler divergencia módszerrel.
Sprugnoli és társa (Sprugnoli & Tonelli, 2019) igei események mellett foglalkoztak főnévi események detektálásával is történelmi szövegeken. Ehhez két módszert alkalmaztak: egy hagyományos Conditional Random Fields (CRFs) és egy neurálishálózat-alapú (NN) módszert. Csak a következő jellemzőket használták fel a +/- 4 szavas szövegkörnyezeti ablakban: lemma, szófaj, dokumentum típusa.
Spanyol szövegekre Peris és társa (Peris, Taule, Boleda, & Rodriguez, 2010) csak igéből képzett főnévi eseményekkel foglalkozott. Osztályozásra a Weka döntési fa osztályozóját alkalmazták és külső főnévi lexikont használtak fel. Függőségi reprezentációt használtak, de csak a jelölt főnév és a közvetlenül ahhoz kapcsolódó ige közötti kapcsolatot vizsgálták. Felhasználták a jelölt argumentum struktúráját is.
Német nyelvű szövegekre Gorzitze és társa (Gorzitze & Pado, 2012) bootstrapping módszert használt esemény-felismerésre. Idővel kapcsolatos kifejezéseket és aspektuális igéket kerestek a jelölt főnév közelében. Vizsgálták a jelölt és a közvetlen ige kapcsolatát és *szabályalapú* függőségi reprezentációt használtak.

8.3 Főnévi események

Példák főnévi eseményekre: *futás, építés, írás, háború, ünnepség*.
A főnévi eseményeknek két nagy csoportja van: igéből képzettek (deverbális) és nem igéből képzettek (nem deverbális). Példa igéből képzett főnevekre: *futás, írás*, nem igéből képzett főnevekre: *háború, ünnepség*. Az igéből képzett főnevek két fő fajtája az események és az eredmények. Ezeknél a főneveknél gyakori a kétértelműség is. Vannak olyan szavak (például *írás*), amelyek egyes mondatokban események, másokban pedig eredmények. Például az *írás* főnév a következő mondatban esemény: *Azonban az idő hamar elszaladt, a várakozás és a felvételi* ***írása*** *közben egyaránt.* Viszont a következő mondatban nem esemény, hanem ered-

mény: *Ezután megnéztük a vár alatt lévő múzeumot, ahol különféle fegyvereket, harci eszközöket,* ***írás****okat lehetett látni.* A többértelműség miatt nem elég a szóalak vizsgálata, a szövegkörnyezetet is elemezni kell.

8.4 Környezet

Alkalmazásomban a Szeged Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) egy részét használtam fel a következő területekről: *üzleti rövidhírek, szépirodalom-fogalmazás, számítógépes szövegek, újsághírek, jogi szövegek*, tanításhoz és kiértékeléshez tízszeres keresztvalidációt alkalmaztam. A mondatokat két nyelvész annotálta, az annotátorok közötti egyetértés Kappa = 0,7 volt.

A feladatokat *bináris osztályozásra* vezettem vissza, az osztályozáshoz a Weka programcsomagnak (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009) a J48-as döntési fa elemzőjét használtam fel. A szavak morfológiai elemzésére, a szófaji egyértelműsítésére és a mondatok függőségi nyelvtan szerinti szintaktikai elemzésére a Szeged Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) annotált elemzését alkalmaztam. A mondatok konstituensfa-alapú reprezentációjára a Szeged Treebank (Csendes D. , Csirik, Gyimóthy, & Kocsor, 2005) annotált elemzését használtam fel.

A Szeged Dependency Treebank is tartalmaz a szavakhoz morfológiai elemzést, de a HunMorph elemzőcsomag (Tron, Kornai, Gyepesi, Németh, Halácsy, & Varga, 2005) sok esetben részletesebb elemzést készít, ezért ezt is felhasználtam, így a feladathoz két morfológiai elemzést is alkalmaztam. Ezenkívül felhasználtam a Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoportjának *Névelem-felismerő alkalmazását* (Szarvas, Farkas, & Kocsor, 2006).

Ahogy láttuk a kapcsolódó munkáknál, mások is használtak függőségi reprezentációt, de az elemzőfában mindenki csak a jelölt és a vele közvetlen szülő és gyerek kapcsolatban lévő szavakat vizsgálta. Én vizsgáltam a jelölt és a fában tőle távolabbi igék kapcsolatát is.

Rendszeremben alkalmaztam a magyar *WordNet*et (Miháltz, és mtsai., 2008) a vizsgált szavak szemantikai jellemzéséhez. Ehhez a WordNet hipernim hierarchiájában található szemantikai kapcsolatokat használtam fel.

Statisztikai adatok

A tanító és kiértékelő halmaz statisztikai adatait a következő táblázat tartalmazza (8.1. táblázat):

mondatok száma	10000 db
jelöltek száma	48388 db
események	7626 db

8.1. táblázat: Statisztikai adatok

A jelölteket a hasonló tulajdonságok alapján két fő részre osztottam fel, az igéből képzett főnevek és a nem igéből képzett főnevek csoportjára (8.2. táblázat).

Igéből képzett főnevek	jelöltek	5325
	pozitív jelöltek	4169
Nem igéből képzett főnevek	jelöltek	43063
	pozitív jelöltek	3457

8.2. táblázat: Főnévi jelöltek adatai

8.5 Az osztályozás bemutatása

Az osztályozáshoz bináris osztályozót használtam, amihez a Weka adatbányászati programcsomag J48-as *Döntési fa* és az SVM *Support Vector Machine* algoritmusait alkalmaztam. A mondatok főnevei voltak a jelöltek, amelyek a függőségi reprezentációban egy-egy csomópontot jelentenek. Módszeremet gazdag jellemzőtérrel valósítottam meg. Az eseménydetektálásos feladatokban gyakran használt jellemzőket én is alkalmaztam, ezeken kívül *új jellemzőkkel* is kibővítettem a jellemzőkészletemet. Az új jellemzőket a *magyar szövegek tulajdonságai alapján* választottam ki, az [ÚJ] jelölést használtam az új jellemzőkhöz és módszerekhez a főnévi események detektálása területen.

A konstituensfa- és a függőségifa-alapú reprezentáció összehasonlítása
Az 5. fejezetben részletesen bemutattam a két reprezentációt, itt csak a lényeget emelem ki:
A *konstituensfa-alapú reprezentáció* a szöveget részkifejezésekre, frázisokra bontja. A fában a csomópontok a kifejezések típusai, a levelek a mondat szavai, az ágak nincsenek címkézve. Ez a reprezentáció a konstituens kapcsolaton alapul, az elemzőfa az S szimbólummal kezdődik és a mondat szavaival, a levelekkel végződik.
A *függőségifa-alapú reprezentáció* a szavakat a közöttük lévő kapcsolatok alapján kapcsolja össze. A fa minden csomópontja egy szót reprezentál, a gyerek csomópontok azon szavak, amelyek függnek a szülő csomóponttól és az ágat a kapcsolattal címkézzük. A főige a gyökér elem. Ha a jelölt több szót tartalmaz, akkor ezek a szavak egy részfát alkotnak a fő-fán belül. A részfa a fejszaván (headword) keresztül kapcsolódik a fő-fához.

8.5.1 A jellemzőkészlet

A jelöltekhez a következő *jellemzőcsoportokat* definiáltam:

- Felszíni jellemzők
- Morfológiai jellemzők
- Szintaktikai jellemzők (Függőségi reprezentáció)
- Szemantikai jellemzők (WordNet)
- Konstituensfa jellemzők
- Szózsák jellemzők
- Lista jellemzők
- Kombinált jellemzők

Felszíni jellemzők: *Bigramok, trigramok*: A vizsgált szavak végén lévő 2-es, 3-as betűcsoportok. *PositionInSentence*: a jelölt hányadik szó a mondatban. *NagyBetuNemMondatElejen*: Nem a mondat elején lévő nagybetűs szavak legtöbbször névelemek, így ez utal a jelölt nem-

esemény jellegére. *Névelem [ÚJ]*: A vizsgált szó névelem, vagy nem (A névelem detektáló alkalmazással eldöntve). Ez utal a jelölt nem-esemény jellegére.

Morfológiai jellemzők-1: Mivel a magyar nyelv morfológiailag gazdag nyelv, ezért számos morfológiaalapú jellemzőt definiáltam, ezekhez a Szeged Dependency Treebank annotált morfológiai elemzését használtam fel. Jellemzőként definiáltam az eseményjelöltek MSD morfológiai kódját, felhasználva a következő morfológiai jegyeket: *típus*(SubPos), *mód*(Mood), *eset*(Cas), *idő*(Tense), *személy*(PerP), *szám*(Num), *határozottság(*Def). Ezeken kívül a következő jellemzőket definiáltam ebben a csoportban. *Lemma*: a jelölt lemmája. *hasVerbRoot*: igéből képzett-e a jelölt. *SzofajElotte* és *SzofajUtana*: a jelölt előtti és utáni szó szófaja. *LegkozelebbiIgeMondatbanLemma*: a jelölthöz a mondatban legközelebb álló ige lemmája. *Igeto*: igéből képzett főnév esetén az alapige.

Morfológiai jellemzők-2: [ÚJ: együtt két morfológiai elemzés] Ebben a csoportban a HunMorph morfológiai elemzőjét használtam fel. *IgetoVan*: Igéből származik-e a főnév. *IgebolFonevKepzo*: Igéből képzett főneveknél a képző. *IgeToHunMorph*: igéből képzett főnév esetén az alapige.

Morfológiai jellemzők-3: A HunMorph elemző a morfológiailag többértelmű szavak esetén minden esethez megad külön morfológiai elemzést. Ebben a csoportban minden elemzés esetén megadtam a hozzá tartozó ragokat, képzőket, jeleket.

Függőségifa jellemzők-1 [ÚJ]: Ezeket a jellemzőket a függőségi elemzőfa alapján készítettem el. *JeloltEdgeType*: A jelölt és az elemzőfában a felette levő szó közötti kapcsolat típusa. (például SUBJ, OBJ, COORD) *JeloltEdgeTypeNE*: A jelölt NE (névelem) típussal kapcsolódik-e a felette levő szóhoz, ami utal a jelölt nem esemény jellegére. A kapcsolat típusa NE a több szóból álló névelemek esetén, ez esetben a jelölt általában nem-esemény. *JeloltFelettLemmaFaban*: A jelölt feletti szó lemmája az elemzőfában. *JeloltFelettIgeLemmaFaban*: Az elemzőfában közvetlenül a jelölt felett lévő ige (ha van) lemmája. *KozvetlenSzintaktikaiKapcsolat*: Ha a jelölt fölött van közvetlenül ige az elemzőfában, akkor a kettő közötti szintaktikai kapcsolat típusa. (Például: SUBJ, OBJ) *LegkozelebbiIgeFeletteFabanLemma, LegkozelebbiIgeFelette-TavolsagFaban*: Az elemzőfában a jelölt feletti legközelebbi ige lemmája és annak távolsága a fában a jelölttől. *JeloltReszfaTokenekSzama*: Az elemzőfában a jelölt alá tartozó részfa elemeinek száma. *FeletteSzoEdgeType*: Az elemzőfában a jelölt feletti szó és az a feletti szó közötti kapcsolat típusa (Például TLOCY). Az elemzőfában az időhatározók, időbeliséget kifejező szavak az események felett helyezkednek el, ezek függőségi címkéje jelzi, hogy ez időbeliséget kifejező szó. Az időbeliséget kifejező kifejezés jelenléte utalhat a jelölt eseményjellegére.

Függőségifa jellemzők-2 [ÚJ]: (útvonalak az elemzőfában az ágak mentén) Ha a jelölt nem közvetlenül kapcsolódik a felette levő igéhez az elemzőfában, akkor részletesen jellemeztem a jelölt és az ige közötti útvonalat. *SzófajÚtvonal*: Egymás után írtam a jelölt és az ige közötti csomópontok szófaját, jelölve a haladás irányát a fában, például: C↑S↑V↑C↑V↑V. *Lemmaútvonal*: Itt a jelölt és az ige közötti lemmákat írtam egymás után, például: napoztatás↑és↑törölgetés↑hajszárító↑megszárít. *SzintaktikaiKapcsolat-Útvonal*: A jelölt és az ige közötti útvonalon a szintaktikai kapcsolatok típusai egymás után, például: OBL↑COORD↑SUBJ↑COORD↑CONJ↑.

Konstituensfa jellemzők [ÚJ] : Ezeket a jellemzőket a konstituensfa alapján gyűjtöttem ki. A konstituensfából kevesebb jellemzőt tudtam kiválasztani, mint a függőségi fából.

ConstJelöltFelettiCsomópont: A jelölt feletti csomópont típusa. *ConstSzavakSzáma*: A jelölt szavainak a száma. A többszavas névelemek az elemzőfában együtt vannak ábrázolva egy levélként, ami utal a jelölt nem-esemény jellegére. *ConstCsomópontÚtvonal*: A jelölt és a legközelebbi ige közötti csomópontok egymás után írva, jelölve a fában a haladás irányát is (például: NP↑NP↑V). *ConstUralkodóKategória*: A jelölt és a legközelebbi ige közötti útvonalon a legmagasabb szinten lévő csomópont típusa.

Szózsák jellemzők-1 [ÚJ]: Ezekhez a jellemzőkhöz az adatok a függőségi-fából lettek kigyűjtve a szózsákba, ahol a szózsák modellt használtam fel szócsoportok jellemzésére. *ReszfaLemmak-SzozsakAtlag:*A jelöltekhez gyakran tartozik egy részfa az elemzőfában. Ez a részfa a függőségi fának az a része, amelyiknek csúcsa a jelölt. A vizsgált jellemző nem csak a részfa fejszavát (headword) jellemzi, hanem a részfa többi szavát is. Ennél a jellemzőnél a jelölthöz tartozó részfa szavainak lemmáit reprezentáltam szózsák modellel. Először a tanító halmazon minden lemmához kiszámítottam, hogy milyen valószínűséggel tartozott pozitív jelölt részfájához. Majd minden jelölthöz kiszámítottam a részfáját alkotó lemmákhoz tartozó valószínűségek átlagát. Nagy átlag arra utal, hogy a jelölt részfájában fontos szavak vannak az eseményjelleg szempontjából. *ReszfaLemmak-SzozsakLegnagyobb*: Hasonló az előzőhöz, de itt a második lépésnél minden jelöltnél a részfájához tartozó lemmák közül azt választottam ki, amelyikhez legnagyobb valószínűség tartozott. Nagy maximális valószínűség utal arra, hogy a jelölt részfájában van legalább egy olyan lemma, ami erősen fontos az eseményjelleg szempontjából. Ez a jellemző segít a részfa egy-egy fontos szavának felismerésében. *KozvetlenAlattaLemmak-SzozsakAtlag* és *KozvetlenAlattaLemmak-SzozsakLegnagyobb*: Az előzőkhöz hasonlóan a szózsák modellt alkalmaztam, de itt nem a jelölt részfájához tartozó minden szót vizsgáltam, hanem csak a részfa azon szavait, amelyek szintaktikailag kapcsolódnak a jelölthöz az elemzőfában. *KozvetlenAlattaEdgeType-SzozsakAtlag* és *KozvetlenAlattaEdgeType-SzozsakLegnagyobb*: Az előzőhöz hasonlóan, itt a jelölt és a hozzá szintaktikailag kapcsolódó szavak közötti kapcsolat típusát vizsgáltam szózsák modellel. *LemmaParseTreePathIgeigLemmak-SzozsakAtlag* és *LemmaParseTreePathIgeig-Lemmak-SzozsakLegnagyobb*: Ezeknél a szózsákba a jelölt és az elemzőfában felette lévő ige közötti útvonalon található lemmák kerültek.

Szózsák jellemzők-2 [ÚJ]: Ezekhez a jellemzőkhöz az adatok a kostituensfából lettek kigyűjtve a szózsákba. *ConstFelett1SzintRészfaLemmá-kSzozsakAtlag és ConstFelett1SzintRészfaLemmák-SzozsakMax:* A jelölt feletti első szintű csomópont részfájának lemmái jellemezve szózsák modellel. *ConstFelett2SzintRészfaLemmák-SzozsakAtlag és ConstFelett2SzintRészfaLemmák-SzozsakMax:* A jelölt feletti második szintű csomópont részfájának lemmái jellemezve szózsák modellel.

Szózsák jellemzők-3 [ÚJ]: Ezekhez a jellemzőkhöz a lemmák a mondatból és nem az elemzőfákból lettek kigyűjtve a szózsákba. *MondatbanKornyezet-N-LemmakSzozsakAtlag* és *MondatbanKornyezet-N-Lemmak-SzozsakLegnagyobb*: A mondatban a jelölt N távolságú környezetét jellemeztem szózsák modellel, N=3 és N=5 esetekben. *MondatbanLeközelebbiIgéig-Lemmak-SzozsakAtlag és MondatbanLeközelebbiIgéig-Lemmak-SzozsakLegnagyobb:* A jelölt és a legközelebbi ige közötti lemmák jellemezve szózsák modellel.

WordNet jellemző csoportok: Ezekhez a jellemzőkhöz felhasználtam a magyar WordNet (Miháltz, et al., 2008) hipernim hierarchiájában található szemantikai kapcsolatokat. Mivel

egy szóalakhoz több jelentés is tartozhat a WordNetben, ezért az egyes *jelentések között egyértelműsítést (word sense disambiguation,* WSD*) végeztem a Lesk algoritmussal* (Jurafsky & Martin, 2009).

WordNet jellemzők-1 [ÚJ]: Ebben a csoportban a szózsák modellt alkalmaztam a WordNet synset-jeire. *WordNet-SzozsakAtlag* és *WordNet-SzozsakLegnagyobb*: A szózsák jellemzőkhöz hasonlóan itt a szózsákba a WordNetben a jelölt hipernim hierarchiájába tartozó szavakat vettem fel. Ezek azok a szavak, amelyek a WordNetben az adott jelentés felett helyezkednek el a hipernim hierarchiában. *WordNetSzozsakLegnagyobbSynset*: Megadtam a jelölt hipernim hierarchiájában lévő synset-ek közül azt, amelyik a legnagyobb valószínűséggel tartozik események hipernim hierarchiájába.

WordNet jellemzők-2 [ÚJ]: *WordNetHipernimSynsetekTanulobol* (bináris): Készítettem egy halmazt, amibe kigyűjtöttem a tanító halmazból az esemény jelöltek hipernim hierarchiájának synset-jeit, majd minden jelölthöz megadtam, hogy a hipernim hierarchiájának synset-jei közül tartozik-e legalább egy ebbe a halmazba.

WordNet jellemzők-3 [ÚJ]: *WordNetLegjobbLemmakAlatt*: Először kigyűjtöttem azokat a lemmákat, amelyek alatt a WordNet hipernim hierarchiájában tanító halmazon legalább 80%-ban voltak események és legalább háromszor fordultak elő. Majd ezek alapján a jelölteknél jelöltem, hogy lemmája alatta van-e valamelyik ilyen kiemelt lemma hiponím hierarchiájának.

Szózsák jellemzők-4 [ÚJ]: Először a Szózsák jellemzők 1-3 csoportoknál bemutatott minden esethez itt kiválasztottam a legjobb elemeket a szózsákokból 1-1 halmazba, azokat, amelyek legnagyobb valószínűséggel tartoztak eseményekhez (LegjobbWordNetSynsetek, LegjobbRészfaLemmák, LegjobbÚtvonalLemmák, LegjobbMondatbanKornyezet-N-Lemmak halmazok). Legjobbnak azokat választottam, amelyek legalább 80%-ban tartoztak pozitív jelöltekhez és legalább háromszor szerepeltek a tanító halmazon. Majd a következő jellemzőkkel jelöltem, hogy hozzá tartozó szózsák tartalmaz-e az adott halmaz elemei közül legalább egy elemet. *LegjobbWordNetSynsetek*: A jelölt hipernim hierarchiájába tartozó synsetek között van-e ami szerepel a LegjobbWordNetSynsetek halmazban. *LegjobbReszfaLemmak*: A jelölt részfáinak lemmái között van-e olyan lemma, ami szerepel a LegjobbRészfaLemmák halmazban. *LegjobbLemmakUtvonalIgeig*: A jelölt és az elemzőfában a legközelebbi ige közötti lemmák között van-e olyan lemma, ami szerepel a LegjobbÚtvonalLemmák halmazban. *LegjobbMondatbanKornyezet-N-Lemmak*: A mondatban a jelölt N távolságú környezetében van-e olyan lemma, ami szerepel a LegjobbMondatbanKornyezet-N-Lemmak halmazban. Ezt megnéztem N=3 és N=5 esetekre is.

Lista-jellemzők: *FeletteLemmaIdohatarozoListabol*: Először listába kigyűjtöttem időhatározókat (például előtt, folyamán), amik alatt az elemzőfában gyakran események vannak. Majd jellemzőben megadtam, hogy a jelölt felett van-e ilyen idővel kapcsolatos kifejezés. *FeletteIgeAspektualisListabol*: Listába kigyűjtöttem gyakori aspektuális igéket (például elkezd, folytatódik). Ezen igék alá tartozó főnevek gyakran események. Jellemzőben jelöltem, hogy a jelölt felett az elemzőfában van-e ilyen ige.

Kombinált jellemzők-2 eleműek: Ebben a csoportban az előző jellemzők közül kombináltam össze kettőt, egymás után másolva. *JeloltFelettLemmaFaban+JeloltEdge-Type*: Egy szó eseményjellegét gyakran pontosabban jelzi, ha a felette levő lemmát és a kettőjük közötti kapcsolatot együtt vizsgáljuk, mintha csak külön-külön vizsgálnánk azokat.

Hasonlóan együtt vizsgáltam a következőket:

JeloltFelettIgeLemmaFaban+JeloltEdgeTypeOBJ,
JeloltFelettIgeLemmaFaban+JeloltEdgeTypeSUBJ,
JeloltFelettLemmaFaban+LegjobbWordNetSynsetek,
JeloltFelettIgeLemmaFaban+LegjobbWordNetSynsetek
Kombinált jellemzők - 3 eleműek: Az előző kételemű jellemzőkhöz hasonlóan itt három jellemzőt másoltam egymás után:
JeloltFelettLemmaFaban+EdgeType+WordNetLegjobbSynset,
JeloltFelettIgeLemmaFaban+JeloltEdgeType+WordNetLegjobbSynset,

A *vektortér méretét csökkentettem* a következő módszerrel: csak azokat a *jellemző-előfordulásokat* vettem fel az osztályozáshoz, amelyek a tanító halmazon *legalább háromszor* szerepeltek. Ezzel csak az osztályozás szempontjából jelentéktelen jellemző-előfordulásokat hagytam ki.

8.5.2 **Kiegészítő módszerek**

[ÚJ] Mindegyik hasznos volt az eredménye alapján, így más NLP feladatoknál is hasznosak lehetnek.

A ,
A jelöltek csoportosítása. Az osztályozó hasonló tulajdonságú adathalmazon könnyebben találja meg a szabályokat, mint olyan halmazon, ami sokféle adatot tartalmaz, ezért érdemes a jelölteket kisebb, hasonló tulajdonságú csoportokra bontani. Már az is a csoportosítás egyik lépése, hogy az igei, főnévi igenévi és főnévi főcsoportokat egy-egy fejezetben külön vizsgáltam. A főcsoportokon belüli további csoportosítással is megkönnyíthetjük az osztályozó döntését, majd a csoportok eredményeit a TP, TN, FP, FN eredmények alapján összegezzük. Ennek megfelelően a főnévi jelöltjeimet két fő szempont szerint csoportosítottam. *Első lépésként a jelölteket két csoportra bontottam*: igéből képzett (deverbális) és nem igéből képzett (nem deverbális) főnevek, hiszen e két csoport tagjai eltérően viselkednek. Az igéből képzett főnevek között sokkal nagyobb arányban vannak események. *Másik csoportosítás* a jelöltek lemmái alapján történt, itt 3 alcsoportot képeztem. Első csoportba azok a lemmák kerültek, amelyek legalább 80%-ban voltak események és legalább háromszor fordultak elő a tanító halmazon. A másik csoportba a többi jelölt lemmája a tanító halmazról, a harmadik csoportba a kiértékelő halmazon azon jelöltek lemmái, amelyek nem szerepeltek a tanító halmazon. *Így összesen 2*3=6 csoportot alakítottam ki, és mindegyikre külön-külön végeztem el a tanítást és kiértékelést*. Minden csoportra 10-szeres keresztvalidációt alkalmaztam, majd összegeztem a csoportok TP, TN, FP, FN eredményeit. Az összegzett TP, TN, FP, FN eredmények alapján számítottam ki a Pontosság, Fedés és F-mérték értékeket.

B,
Jellemzők súlyozása. A jellemzők közül kiemeltem a jelölt lemmáját, mert azt feltételeztem, hogy ha egy lemma legalább 80% valószínűséggel esemény a tanító halmazon és ott legalább háromszor előfordult, akkor az nagy valószínűséggel esemény lesz a kiértékelő halmazon is. Így azon jelölteknél, amelyek lemmája legalább 80% valószínűséggel esemény a tanító hal-

mazon és ott legalább háromszor előfordult és ennek ellenére a kiértékelésnél nem-eseménynek jelölte az osztályozó, a végső kiértékelésnél a jelöltet eseménynek jelöltem.
Hasonlóan: azon jelölteknél, amelyek lemmája legalább 80% valószínűséggel nem-esemény a tanító halmazon és ott legalább háromszor előfordult és ennek ellenére a kiértékelésnél eseménynek jelölte az osztályozó, a végső kiértékelésnél a jelöltet nem-eseménynek jelöltem. A módszer alkalmazása javított az eredményen.
Majd az eredményeknél látni fogjuk, hogy ezek a kiegészítő módszerek javították az eredményeimet.

8.6 Eredmények

A kiértékelés során a pontosság, fedés és F-mérték metrikákat használtam.

A, Baseline módszerek
Modellem hatékonyságának vizsgálatához Baseline mérést végeztem. Ennek keretében a jelöltek közül az igei alapúakat vettem pozitív esetnek a többit pedig negatívnak. Az eredményeket a következő táblázatban találjuk (8.3. táblázat):

Pontosság	**Fedés**	**F-mérték**
66,67%	47,57%	55,52%

8.3. táblázat: Baseline mérés eredményei - F mérték

A további eredményeken látni fogjuk, hogy *gépi tanulási módszerem jóval felülteljesítette a Baseline mérés eredményét*.

B, Módszerem eredménye
Gépi tanulásos módszerem a következő eredményt érte el a teljes korpuszon az adott jellemzőhalmazzal és a kiegészítő módszerekkel (8.4. táblázat).

	Pontosság	**Fedés**	**F-mérték**
Döntési fa	81,31	68,16	**72,83**
Support Vector Machine	78,37	65,25	**69,35**

8.4. táblázat: Eredmények a teljes korpuszon (%)

A döntésifa-alapú osztályozóval jobb eredményt értem el, ezért a következő részletes elemzésekhez ezt használtam.
Kiegészítő módszerek alkalmazása nélkül a következő eredményt kaptam (8.5. táblázat):

Pontosság	**Fedés**	**F-mérték**
71,4	61,49	**65,95**

8.5. táblázat: Eredmények főnevekre a kiegészítő módszerek nélkül

Az eredményen látható, hogy a *kiegészítő módszerekkel jelentős javulást értem el*. A javulás 80%-a jelöltek csoportosításából eredt („A” kiegészítő módszer), a javulás maradék 20%-a pedig a jellemzők súlyozása eredményezte („B” kiegészítő módszer).

A jelöltek csoportosításánál, ha csak az első szempont szerint csoportosítottam, akkor az igéből képzett főnevek esetén (F-mérték 84,62) a modell sokkal jobb eredményt ért el, mint a nem igékből képzett főneveknél (F-mérték 39,52).
Modellemet lefuttattam külön-külön az öt részkorpuszon is. Ennek eredménye a következő táblázatban látható (8.6. táblázat). Legjobb eredményt az *Újsághírek doménen* kaptam, a legrosszabbat a *Jogi szövegeken*.

Részkorpusz	**Pontosság**	**Fedés**	**F-mérték**
Szépirodalom-fogalmazás	83,36	72,18	**76,24**
Újsághírek	84,27	73,41	**77,31**
Üzleti rövidhírek	83,27	72,38	**76,12**
Számítógépes szövegek	79,83	68,75	**72,57**
Jogi szövegek	76,62	65,59	**69,74**

8.6. táblázat: Eredmények az alkorpuszokon (F-mérték, %)

C, Porlasztásos mérés eredményei
Megvizsgáltam, hogy az egyes *jellemzőcsoportok* hogyan befolyásolják a gépi tanulórendszer eredményeit a teljes doménen, amihez *porlasztásos mérést* végeztem. Ekkor a teljes jellemzőkészletből elhagytam az egyes jellemzőcsoportokat, majd a maradék jellemzőre támaszkodva tanítottam. A mérés eredményei a következő táblázatban találhatóak (8.8. táblázat). Az adatok azt mutatják, hogy a jellemzőcsoportot elhagyva hogyan változott az eredmény. A csökkenő (negatív) eredmény azt jelzi, hogy a vizsgált jellemzőcsoportnak pozitív hatása van az esemény-felismerésben.

Elhagyott jellemzők	**Változás az F-mértékben**
Felszíni jellemzők	**-0,28**
Morfológiai jellemzők-1	**-2,51**
Morfológiai jellemzők-2	**-0,52**
Morfológiai jellemzők-3	**-2,01**
Elemzőfa jellemzők-1	**-1,92**
Elemzőfa jellemzők-2	**-0,52**
Szózsák jellemzők-1	**-1,34**
Szózsák jellemzők-2	**-2,42**
Szózsák jellemzők-3	**-0,57**
WordNet jellemzők-1	**-6,51**
WordNet jellemzők-2	**-0,53**
WordNet jellemzők-3	**-0,2**
Lista jellemzők	**0.0**
Kombinált jellemzők – 2 eleműek	**-0,79**
Kombinált jellemzők – 3 eleműek	**+0,1**

8.7. táblázat: A porlasztásos mérés eredményei (%)

Ha a hasonló jellemzőcsoportokat összevontam, akkor a következő eredményeket kaptam a csoportokra (8.8. táblázat).

Elhagyott jellemzők	**Változás az F-mértékben**
Felszíni jellemzők	**-0.28**
Morfológiai jellemzők	**-1.63**
Függőségi fa jellemzők	**-1.56**
Konstituens fa jellemzők	**-1.1**
Szemantikai (WordNet) jellemzők	**-7.7**
Szózsák jellemzők	**-4.0**
Lista jellemzők	**0.0**
Kombinált jellemzők	**-0.95**

8.8. táblázat: A porlasztásos mérés eredményei - összevonással(%)

A Szemantikai jellemzőknél *ha nem alkalmaztam a Lesk algoritmust*, akkor kissé gyengébb eredményt kaptam: a „kiegészítő módszerek nélkül" mérésnél a 65,95 helyett 65,32 F-mértéket.
Az eredményeken látszik, hogy majdnem minden jellemzőcsoportnak kedvező hatása van a modell teljesítményére. Legjobb a hatása a *Szemantikai* a *Szózsák* és a *Függőségi elemzőfa* jellemzőknek, mindkét morfológiai elemzés hatása pozitív. Mint alcsoportnak, a WordNet jellemzők-2 részcsoportnak a legjobb a hatása (-6.51%), ebben használtam együtt a WordNetet a szózsák modellel. A Lista jellemzőknek nincs hatása. A konstituens fa jellemzőknek is pozitív volt a hatása, de a *függőségi fa jellemzőknek a hatása ennél jobb volt*. Negatív a hatása a 3 elemű kombinált jellemzőknek, de a 2 elemű kombinált jellemzők hasznosak.

Az eredmények összehasonlítása a kapcsolódó munkákkal
Angol szövegekre Jeong és társa (Jeong & Myaeng, 2012) 71,8%-os, Romeo és társai (Romeo, Lebani, Bel, & Lenci, 2014) 67%-os F-mértéket értek el. Olasz nyelvre Caselli és társai (Caselli, Russo, & Rubino, 2011) 69%-os, spanyol nyelvre Peris és társai (Peris, Taule, Boleda, & Rodriguez, 2010) 59,6%-os F mértéket értek el. Ezekkel összehasonlítva, eredményeim (F-mérték = 72,83%) jónak számítanak.

8.7 Összegzés

Ebben a fejezetben bemutattam gazdag jellemzőtéren alapuló gépi tanuló megközelítésemet, amely automatikusan képes magyar nyelvű szövegekben főnévi eseményeket detektálni. Modellem teljesítményét a Szeged Dependency Treebank öt doménjén teszteltem összesen 10 000 mondattal.
Jellemzőkészletemben felszíni, morfológiai, függőségifa, konstituensfa, szózsák, WordNet, lista és kombinált jellemzőket használtam fel. Ezen jellemzőcsoportok mellett kiegészítő módszereket is alkalmaztam, amelyek javították modellem hatékonyságát.

Döntésifa-alapú és Support Vector Machine (SVM) algoritmusokat használtam. Az algoritmusokat az alapbeállításokkal használtam paraméter-optimalizálás nélkül. Mérésem alapján ezen a területen a döntésifa-alapú osztályozó teljesített jobban.
A legjobb eredményeket az újsághírek doménen, a legrosszabbat a jogi szövegeken értem el. Ennek az oka, hogy az újsághírek szövegei egyszerűbb mondatokat tartalmaznak, mint a jogi szövegek. A WordNet, az Elemzőfa és a Szózsák jellemzőknek volt a legjobb hatása az eredményekre. Ez megerősíti, hogy a WordNet hatékony szemantikai feladatokra és a Szózsák modell jól jellemzi a vizsgált szócsoportokat (a jelölt részfájának szavait; a jelölt hipernim hierarchiájának synsetjeit; a jelölt előtti és utáni lemmákat N-méretű ablakban).
Csak a 3 szavas kombinált jellemzők elhagyásával javultak az eredmények. Talán ezek hatása már beépült más jellemzők hatásába, különösen a 2 szavas kombinált jellemzők csoportjába. Így az ezzel való kiegészítés csak az osztályozó döntési nehézségét növelte.
A főnévi események detektálására, függőségi reprezentáció és konstituensfa-alapú reprezentáció és WordNet alkalmazásával, magyar nyelvű szövegekre ismereteim szerint ez az első kutatási eredmény. Algoritmusaimat tesztadatbázisokon kiértékelve versenyképes eredményeket érnek el az eddig bemutatott angol és más nyelvű eredményekkel összehasonlítva.

8.8 A fejezet eredményei

A fejezet fő eredményeinek összefoglalása:

- A **főnévi események automatikus detektálására** készítettem alkalmazást függőségifa- és konstituensfa-alapú reprezentáció és WordNet alkalmazásával.
- Modellemben gazdag jellemzőtéren alapuló osztályozót használtam a következő **jellemzőcsoportokkal**: felszíni, morfológiai, függőségifa-alapú, konstituensfa-alapú, szemantikai (WordNet), szózsák, lista és kombinált jellemzők.
- A főnévi események detektálásához Névelem-felismerő alkalmazást is implementáltam.
- Szintaktikai jellemzéshez **függőségifa-** és **konstituensfa-alapú reprezentációt** is alkalmaztam és azok hatékonyságát összehasonlítottam.
- Szemantikai jellemzéshez a magyar **WordNet**et használtam fel. A WordNetben az egyes jelentések között egyértelműsítést végeztem a **Lesk algoritmussal**.
- A **szózsák modellt** alkalmaztam szócsoportok jellemzésére a következő területeken: egy részfa tokenei; az elemzőfában két csomópont közötti tokenek; a WordNet hiperním hierarchiájában két synset közötti synsetek; a jelöltek környezetében lévő szavak a mondatban.
- A morfológiai elemzéshez **két morfológiai elemzőt** használtam fel.
- Modellem teljesítményét megvizsgáltam az **igéből képzett és a nem igéből képzett** főnévi eseményekre is.
- **Két osztályozó algoritmust** használtam és hasonlítottam össze az adott feladatra: a **döntési fa** és a **szupportvektorgépek** (SVM) algoritmusokat.
- Megvizsgáltam, hogy a **jelöltek csoportosításával** vagy anélkül lehet jobb eredményeket elérni.

- Az alapjellemzők mellé a következő kiegészítő módszereket is alkalmaztam, amelyek javították az eredményeket: **jelöltek csoportosítása**; **jellemzők súlyozása**.

Eredmények a tézispontokon:
Bizonyítottam a következőket a főnévi események detektálásánál (2. tézispont):

- *Igazoltam, hogy a legjobban teljesítő jellemzőcsoportok a szemantikai és a szózsák csoportok.*
- *Megmutattam, hogy ezen a területen a szózsák modellt hatékonyan lehet alkalmazni a következő szócsoportok esetében: egy részfa tokenei; az elemzőfában két csomópont közötti tokenek; a WordNet hiperním hierarchiájában két synset közötti synsetek; a jelöltek környezetében lévő szavak a mondatban.*
- *Igazoltam, hogy ezen a területen jobb eredményt lehet elérni a függőségifa-alapú szintaktikai reprezentáció használatával, mint a konstituensfa-alapú reprezentáció használatával.*
- *Bizonyítottam, hogy ha a főnévi események detektálását a jelöltek csoportosításával végezzük el, akkor jobb eredményeket lehet elérni, mintha egy csoportban kezelnénk minden jelöltet.*

9 Események szemantikus szerepeinek automatikus címkézése

Ebben a fejezetben bemutatom eredményeimet az események szemantikus szerepeinek automatikus címkézése területén.

9.1 Bevezetés

Az események detektálása mellett fontos azok szemantikus kapcsolatainak, *szemantikus szerepeinek* megtalálása is (*Szemantikus szerepek címkézése*, Semantic Role Labeling, SRL). Az események és azok szemantikus szerepeinek detektálását a természetesnyelv-feldolgozás sok területén lehet hasznosítani, például az összegzéskészítés, gépi fordítás és a válaszkeresés területein.
Ebben a fejezetben a *szemantikus szerepek címkézésével* foglalkoztam. Ez a szemantikus kapcsolatok azonosítását jelenti egy *szemantikus kereten* belül (semantic frame). A *keretek* eseményeket írnak le azok szereplőinek szintaktikai és szemantikai megkötésein keresztül. Munkámban a *vállalati vásárlások, tulajdonváltozások* és a *tőzsdei hírek* kereteivel foglalkoztam.
A *szemantikus szerepek címkézése* napjainkban a természetesnyelv-feldolgozás (NLP) egyik legdinamikusabban fejlődő területe.
Magyar nyelvű szövegeimen *függőségi reprezentációt* használtam fel, mivel ez jól használható szabad szórendű nyelvek elemzésére, így a magyarra is.
A *szerepek* a legegyszerűbb esetekben a célszó *szintaktikai kapcsolatai* voltak, de voltak ettől eltérő esetek is. Sokszor a keresett szerep távolabb helyezkedett el a függőségi fában a célszótól, gyakran a mondat másik felében. És volt olyan eset is, ahol a szintaktikai kapcsolat alapján várt helyen nem a keresett szerep volt. Így a feladat a függőségi fában a célszótól távolabbi szerepek megkeresése és a közelebbi hamis pozitív jelöltek kiszűrése volt.
A feladatra gépi tanulásos módszert alkalmaztam, ami a szabályalapú módszerrel ellentétben nem igényel annyi erőforrást és előfeldolgozást, valamint automatikusan alkalmazható más doménekre is.
Az igei és főnévi igenévi célszavakhoz kerestem szerepeket. Nem csak a két általános szerepet hanem több domén-specifikus szerepet is címkéztem. A vállalatfelvásárlások keretnél öt szerepet, a tőzsdei híreknél nyolc szerepet vizsgáltam. Megoldásomban a vizsgált szavak szemantikai jellemzéséhez felhasználtam a magyar *WordNet*et (Miháltz, et al., 2008). Mivel egy szóalakhoz több jelentés is tartozhat a WordNetben, ezért az egyes *jelentések között egyértelműsítést* (word sense disambiguation, WSD) végeztem a *Lesk algoritmussal* (Jurafsky & Martin, 2009).
A következő *kiegészítő méréseket* végeztem el: célszavak csoportosítása; ritka jellemzőelőfordulások elhagyása.
Ismereteim szerint *szemantikus szerepek automatikus címkézésére domén-specifikus szerepekre, függőségi reprezentáció alkalmazásával, magyar nyelvű szövegekre,* ez az első kutatási eredmény.

9.2 Kapcsolódó munkák

Kezdetben az SRL munkákban csak igékkel foglalkoztak, az igéket önállóan vizsgálták és csak általános szerepeket kerestek (például Agent, Patient, Instrument) hozzájuk. Ehhez a PropBank korpusz (Palmer, Gildea, & Kingsbury, 2005) szövegeit használták fel, amiben igék és a hozzájuk tartozó szemantikus szerepek vannak annotálva. Ezzel a témával foglalkoztak a 2004-es és 2005-ös CoNNL kiértékelési feladatokban is (Carreras & Marquez., 2004), (Carreras & Marquez, 2005).
Később az igéket már nem önállóan vizsgálták, hanem tématerületenként csoportosították azokat (keretek) és az általános szerepek mellett már vizsgáltak domén-specifikus szerepeket is. Ehhez a FrameNet korpusz (Baker, Fillmore, Lowe, & B., 1998) szövegeit használták fel, amiben angol nyelvű szövegek vannak szemantikus szerepek szerint annotálva. Ezek is elsősorban igékkel foglalkoznak, de keresnek nem igei célszavakra is. Egy fontos alaptanulmányt készített D. Gildea és D. Jurafsky (Gildea & Jurafsky, 2002) az SRL témában. A Senseval-3 kiértékelési feladat (Litkowski, 2004) és az ACE program (Ahn, 2006), más NLP feladatok mellett, SRL témával is foglalkozik.
Strubell és társai (Strubell, Verga, & Andor, 2018) neurális hálózatot (NN) használtak SRL feladatra. Többrétegű neurális hálózatukkal több feladatot végeznek el az egyes rétegekkel: szófaj és predikátum detektálás, szintaktikai elemzés, szemantikus szerepek címkézése. Az egyik réteg kimenete lesz a bemenete a következő rétegnek. A feladathoz jellemzőként a lemma mellett szintaktikai információkat és a szintaktikai függőségeket használták fel.
Szemantikus szerepek címkézésére magyar nyelvű szövegekre is készültek már munkák. Farkas és társai (Farkas, Konczer, & Szarvas, 2004) a szemantikuskeret-illesztésre *szabályalapú* módszert használtak. Ehmann és társai (Ehmann, Lendvai, Miháltz, Vincze, & László, 2013) pszichológiai témájú szövegeken szemantikus szerepek címkézésénél csak két általános szerepet keresnek: az ágens és az elszenvedő szerepeket.

9.3 Szemantikus keretek és a szemantikus szerepek

Sok információkinyerő rendszer manapság *tárgykör (domén)* specifikus *keretekkel* dolgozik. Egy-egy tárgykör eseményeit célszerű egy *kereten* belül vizsgálni, hiszen ugyanazok a *szerepek* tartoznak egy adott csoport minden eseményéhez. Ha a célszavakat önállóan dolgozzuk fel, akkor jóval kevesebb tanító adattal tudunk dolgozni. A célszavak *keretekben történő csoportosítása* jelentősen csökkenti ezt a problémát, hiszen a több célszó tanító adatai összeadódnak.
Igei és főnévi igenévi célszavakhoz kerestem ki a szerepeket. Munkám első részében a *vállalati vásárlások, tulajdonváltozások* keretével foglalkoztam. A következő igei célszavakat vizsgáltam meg az adott kereten belül: *vesz, vásárol, szerez, bekebelez, gyarapít, ad, átruház, értékesít, forgalmaz*, valamint e célszavak minden igekötős, módbeli és időbeli változatát is. A főnévi igenévi célszavak a felsorolt igei célszavak főnévi igenévi alakjai voltak. A célszavakhoz a mondatokon belül a következő szerepeket kerestem meg: *vevő, eladó, áru, ár, idő*.
Modellem működését megvizsgáltam a tőzsdei hírek doménen is. Ott a következő célszavakat vizsgáltam: *befejez, csökken, emelkedik, erősödik, esik, gyengül, indul, kezdődik, mérséklődik, nő, nyit, süllyed, ugrik, változik, végez, zár, zuhan*. A célszavakhoz a következő szerepeket

kerestem meg: *instrumentum, ár, elmozdulás-irány, elmozdulás-értek, piac, dátum, idő, forgalom*.
A legtöbb mérést a *vállalati vásárlások* doménen végeztem el, a *tőzsdei hírek* doménen csak az alapbeállításokkal vizsgáltam meg a modell működését.

Példa a célszóra és a szerepekre:
Vastag betűvel van kiemelve a ***célszó*** és szögletes zárójelben a *szerepek* találhatóak. Alsó indexben szerepel az adott szerep típusa.

[A Royal Dutch Shell csoport]$_{Vevő}$ [400 millió dollárért]$_{Ár}$ ***megvenni*** *készül [a legnagyobb kínai offshore-földgáz- és olajmező 20 százalékát]$_{Áru}$.*

A példán is látszik, hogy egy szerep általában több szóból áll és a mondatok általában nem tartalmazzák az összes szerepet.

9.4 A korpusz és a programok

Az alkalmazásom teszteléséhez a Szeged Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) rövidhírek csoportjának egy olyan változatát használtam fel, amelyikben annotálva vannak a szemantikus szerepek. Ezek közül 1000-1000 mondatot használtam fel a két doménen. A tanításhoz és kiértékeléshez 10-szeres keresztvalidációt alkalmaztam.
A feladatokat *bináris osztályozásra* vezettem vissza, az osztályozáshoz a *Mallet* programcsomagnak (McCallum, 2002) a C4.5-ös döntési fa elemzőjét használtam fel. A szavak morfológiai elemzésére, majd szófaji egyértelműsítésére és a mondatok függőségi nyelvtan szerinti szintaktikai reprezentációjára is a Szeged Dependency Treebank annotált elemzését alkalmaztam.

9.5 Az osztályozás

Minden bemeneti mondatnál adott a *célszó* és a feladat az adott szerepek megkeresése volt. Az osztályozóknál a *jelöltek* a függőségi elemzőfa csomópontjai voltak. Egy mondaton belül általában egy csomópont a keresett szerep kiemelt szava (a szerep feje, headword), az osztályozásnál ezek a *pozitív* esetek, a többi csomópont pedig a *negatív* eset. Az osztályozáshoz *bináris osztályozót* használtam, az osztályozó az adott mondatnál bejelöli a keresett szerepet. Minden szerepre külön osztályozót alkalmaztam. Az osztályozónak *nem adtam meg*, hogy az adott mondat tartalmazza-e az adott szerepet, vagy sem.
A kiértékelésnél *szigorú szabályt* alkalmaztam: csak azt a döntést fogadtam el, amelyik pontosan az annotált szerepet jelöli meg, sem az ezt tartalmazó fákat, sem ennek a részfáit nem fogadtam el pozitív döntésnek. Ha ennél enyhébb szabályt alkalmaznék, akkor az eredmények mérőszámai jobbak lennének.
A későbbiekben tervezem megoldani azt a feladatot, hogy ha több jelölt is ugyanazt a címkét kapja egy mondatban, akkor a modell válasszon a jelöltek közül az adott címkéhez.

9.5.1 Jellemzőkészlet

A tanító és a kiértékelő halmazon a jelöltekhez jellemzőket vettem fel. Az SRL feladatokban használt általános jellemzőket (Gildea & Jurafsky, 2002) én is alkalmaztam, ezeken kívül újakkal is kibővítettem a jellemzőkészletemet. Kiemelt feladatomnak tekintettem olyan jellemzőcsoportok részletes kidolgozását, amelyek figyelembe veszik a magyar nyelv sajátosságait. Ezek a *morfológiai* és a *függőségifa-alapú jellemzőcsoportok* voltak.

A jelöltekhez a következő *jellemzőcsoportokat* definiáltam:

- Felszíni jellemzők
- Morfológiai jellemzők
- Szintaktikai jellemzők (Függőségi reprezentáció)
- Szemantikai jellemzők (WordNet)
- Szózsák jellemzők

Felszíni jellemzők: *Bigramok, trigramok*: A vizsgált szavak végén lévő 2-es, 3-as betűcsoportok. *Pozíció*: a jelölt a célszó előtt vagy után áll-e a mondatban. *Távolság-mondatban*: a jelölt és a célszó szótávolsága a mondaton belül.
Morfológiai jellemzők: Mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért számos morfológiaalapú jellemzőt definiáltam. Jellemzőként definiáltam az eseményjelöltek MSD morfológiai kódját felhasználva a következő morfológiai jegyeket: *típus*(SubPos), *mód*(Mood), *eset*(Cas), *idő*(Tense), *személy*(PerP), *szám*(Num), *határozottság(*Def). *Szófaj,*
Ezeken kívül felhasználtam még a jelölt szótövét és toldalékait (prefixum, szuffixum) szózsák modellel. A szótő és a toldalékok meghatározásához felhasználtam a magyarlanc nyelvészeti program RFSA morfológiai elemzőjét (Zsibrita, Vincze, & Farkas, 2013).
A szózsákba kigyűjtöttem a jelöltekhez a szótövet és a toldalékokat az elemző igei és főnévi igenévi elemzéseiből. Ha az elemző egy jelölthöz több ilyen elemzést is megad, akkor mindegyiket a szózsákba tettem. A szózsák modellhez a következő módszert használtam: Először a tanító halmaz alapján kigyűjtöttem minden toldalékhoz, hogy milyen valószínűséggel tartozik eseményjelölthöz. Majd ezen értékek alapján minden jelölthöz két jellemzőt határoztam meg: MorfológiaiSzózsákÁtlag és MorfológiaiSzózsákLegnagyobb. A *MorfológiaiSzózsákÁtlag* jellemző esetén a jelölthöz meghatároztam a szótőre és a toldalékaira kiszámolt valószínűségek átlagát. Nagy átlag arra utal, hogy a jelölt szótöve és toldalékai között fontos elemek vannak az eseményjelleg szempontjából. A *MorfológiaiSzózsákLegnagyobb* jellemző esetén hasonlóan az előzőhöz, de itt minden jelöltnél a szótő és a toldalékok közül a legnagyobb valószínűséget választottam ki. Nagy maximális érték arra utal, hogy a jelölt szótöve és toldalékai közül legalább az egyik fontos az eseményjelleg szempontjából.
Az előző jellemzőkhöz hasonlóan készítettem el a jelölt környezetében lévő szavak lemmáihoz tartozó szózsák-átlag és szózsák-legnagyobb jellemzőket. *Mondatban-környezet-N-lemmák-szozsák-átlag* és *Mondatban-környezet-N-lemmák-szozsák-legnagyobb*: A mondatban a jelölt N távolságú környezetét jellemeztem szózsákkal, N=3 és N=5 esetekben. Az előző jellemzőkhöz hasonlóan készítettem el az ehhez tartozó szózsák-átlag és szózsák-legnagyobb jellemzőket.

Jellemzők a függőségfa alapján-1: Ebbe a csoportba azokat a jellemzőket soroltam, amelyeket az SRL feladatokhoz általában felhasználnak (Gildea & Jurafsky, 2002). A jelölt és a célszó viszonyát vizsgáltam a függőségi elemzőfában. *Szófaj-útvonal*: Egymás után írtam a jelölt és a célszó közötti csomópontok szófaját, feljegyezve azt is, hogy az elemzőfában felfelé, vagy lefelé haladtam az adott kapcsolatnál. Például: C↑S↑V↑C↑V↑V↓V↓N↓N↓A. *Uralkodó-kategória-szófaja*: A jelölt és a célszó közötti útvonalon a legmagasabban fekvő csomóponthoz tartozó szó szófaja.

Jellemzők a függőségfa alapján -2: Ebbe a csoportba az új jellemzőket soroltam fel, amelyeket a magyar mondatokhoz készítettem a függőségi elemzőfa alapján és a magyar nyelv sajátosságai szerint lettek beállítva. *Közvetlen-szintaktikai-kapcsolat*: Ha van a célszó és a jelölt között szintaktikai kapcsolat, akkor annak a típusa. *Jelölt-célszó-távolság-elemzőfában*: A jelölt és a célszó csomópontjai közötti csomópontok száma az elemzőfában. *Lemma-útvonal*: Mint a Szófaj-útvonal jellemzőnél, de itt a jelölt és a célszó közötti útvonalon a csomóponti szavak lemmáját jegyeztem fel. Például: Budapesti↑Értéktőzsde↑honlap↑közöl↓megvásárol. *Szintaktikai-kapcsolat-útvonal*: Az előzőhöz hasonlóan a jelölt és a célszó közötti útvonalon a szintaktikai kapcsolatokat jegyeztem fel. Például: ↑COORD*SUBJ↓ATT↓INF↓OBJ↓ATT. *Jelölt-alatti-részfában-van-e-névelem*: A Szeged Dependency Treebank az elemzésében jelöli, ha van több szóból álló névelem a mondatban (például Deutsche Börse AG). Mivel a vállalati tulajdonváltozások témakörében gyakran találkozunk vállalati névelemekkel, ezért rögzítettem, hogy a jelölt, vagy az alatta levő részfa tartalmaz-e több szóból álló névelemet. *Jelölt-alatti-részfában-névelem-távolság*: a jelölt és a névelem közötti távolság, ha van. *Jelölt-felett-lemma-fában*: Az elemzőfában a jelölt feletti szó lemmája.

Jellemzők a függőségfa alapján - 3 - szózsák modellel: Itt is új jellemzőket soroltam fel, amelyeket a magyar mondatokhoz készítettem a függőségi elemzőfa és a mondatok alapján. *Jelölt-részfa-lemmák-szózsák-átlag*: Minden jelölt 1-1 részfát alkot. Ez a jellemző nem csak a részfa kiemelt fejszavát vizsgálja (headword), hanem a részfa többi szavát is. Ezeknek a szavaknak a lemmáit szózsák modellel vizsgáltam: Először a tanító halmaz minden lemmájához kiszámítottam, hogy az adott lemma milyen valószínűséggel tartozik pozitív jelölt részfájához, majd megadtam a jelölthöz tartozó részfa lemmáihoz tartozó ezen valószínűségek átlagát. Ezzel jól jellemezhetőek a részfa szavai szózsákként. *Jelölt-részfa-lemmák-szózsák-legnagyobb*: Az előző jellemzőhöz hasonló, de itt a jelölt részfájának lemmáihoz tartozó előzőleg kiszámolt valószínűségek legnagyobbikát adtam meg. Ez a jellemző segít a részfa események szempontjából legfontosabb szavának felismerésében, hiszen gyakran vannak olyan kiemelt szavak a pozitív jelöltek részfájában, amelyek más pozitív jelölt részfájában is előfordulnak. *Lemma-útvonal-szózsák-átlag* és *Lemma-útvonal-szózsák-legnagyobb*: Az előző csoportban bemutatott *Lemma-útvonal* jellemzőhöz hasonló. De míg annál a jellemzőnél a célszó és a jelölt közötti csomópontok lemmáit kötött sorrendű listával jellemeztem, így az erősen függött a lemmák sorrendjétől. Az az ábrázolás két lemma-sorozatot teljesen függetlennek tekint, akkor is, ha csak egy elemben térnek el egymástól, ezért csak a teljesen megegyező útvonalak felismerésére jó. E helyett ennél a jellemzőnél az útvonalon lévő lemmákat szózsák modellel jellemeztem, ahol már a sorrend nem számít, így az útvonalak hasonlóságait is kimutatja.

WordNet jellemzők-szózsákkal: Itt is új jellemzőket definiáltam, amelyeket a magyar mondatok tulajdonságai alapján készítettem el a *Magyar WordNet* (Miháltz, et al., 2008) segítségével szózsák modellel. Ezen jellemzőknél a WordNet hipernim hierarchiájában található szemantikai kapcsolatokat használtam fel a következő új módszerrel. *Jelölt-WordNet-szózsák-átlag* és *Jelölt-WordNet-szózsák-legnagyobb*: Szózsák modellel a tanító halmaz alapján kigyűjtöttem a jelölt részfájához tartozó minden lemmához a WordNet hipernim hierarchiában a felette található synseteket. Mivel egy szóalakhoz több jelentés is tartozhat a WordNetben, ezért az egyes jelentések között jelentésegyértelműsítést (*word sense disambiguation* WSD) végeztem a Lesk algoritmussal (Jurafsky & Martin, 2009). Az egyértelműsítés után minden kigyűjtött synset-hez megszámoltam, hogy milyen valószínűséggel tartozik pozitív jelölthöz. Ez alapján minden synsetet egy-egy valószínűséggel jellemeztem. Így a már ismertetett szózsák modell elvek alapján itt is minden jelöltet jellemeztem az ide tartozó szózsák-átlag és szózsák-legnagyobb jellemzőkkel.

9.5.2 Kiegészítő módszerek

A, Célszavak csoportosítása a vállalati vásárlások doménen

A következő részben arra kerestem a választ, hogy jobb eredményeket kapunk-e, ha a célszavakat csoportokba szervezzük egy doménen belül, vagy inkább együtt kezeljük azokat. A *vásárlásokkal* kapcsolatos mondatoknál a *vevő* és az *eladó* szerepek viselkedését meghatározza, hogy az adott célszónál az alany általában vevő vagy eladó, ezért a *célszavakat két csoportra bontottam* a következő egyszerű módszerrel. A *vevő-centrikus* csoportba azok a szavak kerültek, amelyeknél az alany általában a *vevő*: vesz, vásárol, szerez, bekebelez, gyarapít. Az *eladó-centrikus* csoportba pedig azok, amelyiknél az alany általában az *eladó*: ad, átruház, értékesít, forgalmaz. Egy harmadik esetben pedig nem végeztem csoportosítást. Azt tapasztaltam, hogy a csoportosítás csak a *Vevő* szerep detektálását segíti, a többit rontja. Ennek eredményét látjuk a (9.4. táblázat) táblázatban. **A célszavak csoportosítása nem keverendő össze a jelöltek csoportosításával,** amit a főnévi események detektálásánál alkalmaztam, itt pedig a célszavakat csoportosítottam.

B, Ritka jellemző-előfordulások elhagyása

A *vektortér méretét csökkentettem* a következő módszerrel a szöveges (string) adatoknál: A jellemzők kigyűjtése után csak azokat a *jellemző-előfordulásokat* vettem fel az osztályozáshoz, amelyek a tanító halmazon *legalább N-szer* szerepeltek (N= 0, 5, 10, 15, 20 eseteket vizsgálva). Például, ha egy jelölt *Szófaj-útvonal* jellemző-értéke kevesebbszer fordult elő a tanító halmazon, mint N, akkor nem vettem fel hozzá Szófaj-útvonal jellemzőt. Ezzel csak az osztályozás szempontjából jelentéktelen jellemző-előfordulásokat hagytam ki.

9.5.3 Baseline módszerek

Baseline módszereket a *vállalati vásárlások* doménen vizsgáltam.

Ennek során azokat a jelölteket vettem pozitívnak, amelyekre teljesülnek a következők:

- Az *Áru szerepnél* azokat, amelyek tárgy (OBJ) szintaktikai kapcsolatban vannak a célszóval.
- Az *Ár szerepnél* azokat, amelyeket egy előre elkészített pénznemek lista tartalmazott.
- Az *Idő szerepnél* azokat, amelyeket a következő lista tartalmazott: évszámok 1990-2014-ig, hónapnevek, napnevek, sorszámok 1-31-ig.
- *A vevő-centrikus célszavaknál a Vevő szerepnél és az eladó-centrikus célszavaknál az Eladó szerepnél* azokat, amelyek alany (SUBJ) kapcsolatban vannak a célszóval.
- *Az eladó-centrikus célszavaknál a Vevő szerepnél* azokat, amelyek részes eset (DAT) kapcsolatban vannak a célszóval.
- *A vevő-centrikus célszavaknál az Eladó szerepnél* azokat, amelyek végén a következő trigramok állnak: tól, től, ból, ből.

9.5.4 Statisztikai adatok

A vállalati vásárlások doménen: Mondatok száma: 1000 db
A szerepek száma a vizsgált mondatokban (9.1. táblázat):

Vevő	Eladó	Áru	Ár	Idő
783	579	1025	299	312

9.1. táblázat: Az adott szerepet tartalmazó mondatok száma a vállalati vásárlások doménen

A tőzsdei hírek doménen: Mondatok száma: 1000 db
A szerepek száma a vizsgált mondatokban (9.2. táblázat):

Instrumentum	Ár	Elmozdulás irány	Elmozdulás érték	Piac	Nap	Idő	Forgalom
787	530	431	683	485	436	109	302

9.2. táblázat: Az adott szerepet tartalmazó mondatok száma a tőzsdei hírek doménen

9.6 Eredmények

A kiértékelés során a pontosság, fedés és F-mérték metrikákat használtam.

A, Baseline mérések eredményei (9.3. táblázat):

A táblázatban a *Vevő- és eladó-centrikus célszavak átlaga* sorokat azért tüntettem fel, hogy könnyebben összehasonlítható legyen az azt követő táblázat eredményeivel.
A következő eredményeken látni fogjuk, *hogy a gépi tanulási modellem jóval felülteljesítette a Baseline mérések eredményeit*.

	Szerep	Pontosság	Fedés	F-mérték
Vevő-centrikus célszavak	Vevő	48,24	59,73	**53,37**
	Eladó	54,77	72,13	**62,26**
	Áru	73,25	73,25	**73,25**
	Ár	67,33	96,02	**79,16**
	Idő	34,74	57,89	**43,42**
Eladó-centrikus célszavak	Vevő	78,18	44,10	**56,39**
	Eladó	42,63	47,50	**44,93**
	Áru	77,47	72,97	**75,15**
	Ár	62,64	93,44	**75,00**
	Idő	23,95	46,51	**31,62**
Vevő- és eladó-centrikus célszavak F-mérték átlagok	Vevő			**54,88**
	Eladó			**53,59**
	Áru			**74,2**
	Ár			**77,08**
	Idő			**37,52**

9.3. táblázat: Baseline mérések eredményei

B, Eredmények a célszavak csoportosításával és anélkül

A modell a vevő-centrikus célszavaknál legjobban az *Ár* és az *Áru* szerepekre, leggyengébben pedig a *Vevő* szerepre teljesített. Az eladó-centrikus célszavak esetében legjobban az *Áru* és az *Ár* szerepekre, leggyengébben pedig az *Eladó* szerepre teljesített. A célszavak csoportosítása nélküli esetben legjobban az *Áru* és az *Ár* szerepekre, leggyengébben pedig az *Eladó* szerepre teljesített. A táblázat eredményein látjuk, hogy a célszavak csoportosítása csak a Vevő szerep megtalálását segítette, a többit inkább gátolta (9.4. táblázat), valamint azt, hogy ezen a doménen a vizsgált szerepek közül az Ár és az Áru szerepeket lehet legeredményesebben meghatározni.

Szerep	Vevő centrikus célszavak	Eladó centrikus célszavak	Csoportosítás nélkül
Vevő	65,16	67,95	63,81
Eladó	68,74	56,45	59,19
Áru	77,67	79,79	**81,85**
Ár	84,05	76,78	**84,91**
Idő	70,35	60,28	71,83
Átlag	**73,19**	**68,25**	**72,32**

9.4. táblázat: Eredmények a célszavak csoportosítására (F-mérték)

C, Eredmények a *Ritka jellemző-előfordulások elhagyása* módszerhez

Az eddigi eredményeket a legjobb N=10 esetben kaptam.

A Csoportosítás nélküli esetben az N változtatásával a következő átlag F-mértékeket kaptam (9.5. táblázat):

N (db)	0	5	**10**	15	20
Átlag	63,72	68,37	**72,32**	70,11	67,25

9.5. táblázat - Ritka jellemző-előfordulások elhagyása (F-mérték)

Az eredményeken látszik, hogy a jellemzők kigyűjtése után érdemes a jelentéktelen (a kis számban előforduló) jellemző előfordulásokat kihagyni. Ennél a problématípusnál az N=10 db esetén kaptam a legjobb eredményt.

D, Eredmények porlasztásos méréssel

Megvizsgáltam, hogy az egyes *jellemzőcsoportok* hogyan befolyásolják a gépi tanulórendszer eredményeit. Ehhez *porlasztásos mérést* végeztem mind az öt szerepre a célszavak csoportosítása nélküli esetben (9.6. táblázat). Ekkor a teljes jellemzőkészletből elhagytam az egyes jellemzőcsoportokat, majd a maradék jellemzőkre támaszkodva tanítottam. Az adatok azt mutatják, hogy az adott jellemzőcsoportot elhagyva hogyan változott az eredmény. A csökkenő (negatív) eredmény azt jelzi, hogy a vizsgált jellemzőcsoportnak pozitív hatása van az adott szerep felismerésében.

		Szerep					
		Vevő	**Eladó**	**Áru**	**Ár**	**Idő**	**Átlag**
Elhagyott jellemzők	**Felszíni**	-1,99	-1,31	-1,3	0,64	-0,13	**-0,82**
	Morfológiai	**-0,47**	**-4,67**	**-1,56**	**-2,35**	**-1,79**	**-2,17**
	Szintaktikai (függőségi)	**-6,54**	**-7,71**	**-5,43**	**-2,29**	**-3,36**	**-5,07**
	Szemantikai	0,17	-1,53	-0,03	-2,63	-0,58	**-0,92**

9.6. táblázat: A porlasztásos mérés eredményei (% változás)

Az eredményeken látható, hogy a legjobbnak a *szintaktikai és a morfológiai jellemzőcsoportok* bizonyultak. Minden szerepnél a hatásuk pozitív volt. A következő jellemzőcsoportoknak négy szerepnél pozitív hatása volt: felszíni, szemantikai (WordNet). A Szózsák jellemzőcsoportnak csak három szerepnél volt pozitív hatása.

A Szemantikai jellemzőknél *ha nem alkalmaztam a Lesk algoritmust*, akkor kissé gyengébb eredményt kaptam: a porlasztásos mérésnél a -0,92-ös átlag helyett csak -0,81-es értéket.

A *szózsák modellt* alkalmaztam a morfológiai, a szintaktikai és a szemantikai (WordNet) jellemzőcsoportoknál. A WordNet jellemzők mindegyikét erre alapoztam, de a morfológiai és a szintaktikai csoportnál is hasznosak voltak. Ha nem alkalmaztam volna a szózsák modellt ezekre a csoportokra, akkor a porlasztásos mérésnél a változások átlagára a morfológiai csoportnál a -2,17-es érték helyett csak -1,53-as értéket, a szintaktikai csoportnál a -5,07 érték helyett csak -3,24-es értéket kaptam volna.

E, A célszótól távolabbi szerepek keresése

A szerepek a legegyszerűbb esetekben a célszó szintaktikai kapcsolatai voltak, de voltak ettől eltérő esetek is. Sokszor a keresett szerep távolabb helyezkedett el a függőségi fában a célszótól. Megvizsgáltam, hogy milyen eredményt kaptam volna, ha csak azokkal a jelöltekkel foglalkozok, amelyek közvetlenül kapcsolódnak a célszóhoz az elemzőfában.

Láttuk, hogy a csoportosítás nélküli esetben 72,32-es átlag F-mértéket kaptam. Ha a célszónak csak a közvetlen kapcsolatait vizsgáltam az elemzőfában, akkor a pontosság növekedett, de az összes jelöltre vetítve a fedés viszont csökkent. Ebben az esetben az F-mértékre 69,58-as értéket kaptam. A két értékből látszik, hogy hasznos volt az elemzőfában a célszótól távolabbi jelöltek vizsgálata is a tanításnál és a kiértékelésnél.

F, Eredmények a tőzsdei hírek doménen

Az előző fejezetekben részletesen bemutatott modellt kipróbáltam a *tőzsdei hírek* domén nyolc szerepére is (9.7. táblázat). Az eredményeken látható, hogy *a modell jól teljesített ezen a doménen is*, valamint az, hogy a tőzsdei rövidhírek doménen a vizsgált szerepek közül az Ár és az Elmozdulás-irány szerepeket lehet legeredményesebben meghatározni.

Szerep	**Pontosság**	**Fedés**	**F-mérték**
Instrumentum	74,36	68,63	**71,12**
Ár	88,45	84,91	**86,54**
Elmozdulás-irány	82,13	81,98	**81,78**
Elmozdulás-érték	72,80	67,36	**69,64**
Piac	77,98	69,32	**72,07**
Nap	80,54	69,33	**72,47**
Idő	82,85	75,17	**78,15**
Forgalom	82,16	72,32	**76,14**
Átlag	**80,16**	**73,63**	**75,98**

9.7. táblázat: Eredmények a tőzsdei hírek doménen (%)

Az eredmények összehasonlítása a kapcsolódó munkákkal

Angol nyelvű szövegekre Gildea és társa (Gildea & Jurafsky, 2002) sok keretre és azokon belül sok szerepre végezték el a feladatot. Elsősorban igékkel foglalkoznak, de kerestek nem igei célszavakra is. Ezek átlagolt eredményére 63%-os F-mértéket kaptak. Eredményeim (72,32% és 75,98% F-mérték átlag) jónak számítanak annak ellenére, hogy én csak két keretet és ahhoz öt illetve nyolc fő-szerepet vizsgáltam, és csak igei és főnévi igenevekhez kerestem szerepeket.

9.7 Összegzés

Munkámban bemutattam gazdag jellemzőtéren alapuló gépi tanuló megközelítésemet, amely automatikusan képes magyar nyelvű szövegekben szemantikus szerepek címkézésére függőségi reprezentáció alkalmazásával. A *vállalati vásárlások és a tőzsdei hírek* kereteivel foglalkoztam, ezen a kereten belül 1000-1000 annotált mondatot dolgoztam fel. A vállalati vásárlások doménen öt, a tőzsdei hírek doménen pedig nyolc domén-specifikus szerepet címkéztem. Modellem teljesítményét a Szeged Dependency Treebank rövid hírek doménjén teszteltem. Gazdag jellemzőtéren alapuló osztályozót használtam. Kiemelt feladatomnak tekintettem olyan jellemzőcsoportok részletes kidolgozását, amelyek figyelembe veszik a magyar nyelv

sajátosságait. Ezek a *morfológiai* és a *függőségifa-alapú jellemzőcsoportok* voltak. *Jellemzőkészletemben* felszíni, morfológiai, szintaktikai (függőségi-alapú) elemzés alapján kinyert és WordNet jellemzőket használtam fel. Új jellemzőkkel is kibővítettem a jellemzőkészletemet, amelyeket a magyar szövegek tulajdonságai alapján választottam ki. Megvizsgáltam, hogy az egyes jellemzőcsoportok hogyan befolyásolják a modellem hatékonyságát. Bár munkámban a vizsgált szövegek kevesebb témát fedtek le, mint az angol nyelvű szövegekre bemutatott munkák, de eredményeim a magyar nyelvű szövegeken jónak számítanak a bemutatott angol munkák eredményeivel összehasonlítva. Szemantikus szerepek automatikus címkézésére domén-specifikus szerepekre, függőségi reprezentáció alkalmazásával, magyar nyelvű szövegekre ismereteim szerint ez az első kutatási eredmény.

9.8 A fejezet eredményei

A fejezet fő eredményeinek összefoglalása.

- Az **események szemantikus szerepeinek automatikus címkézésével** foglalkoztam.
- Az igei és főnévi igenévi célszavak szerepeit kerestem.
- A szemantikus szerepek címkézése területén a *vállalati vásárlások, tulajdonváltozások* és a *tőzsdei hírek* kereteit vizsgáltam, mindkét esetben több domén-specifikus szerepet címkéztem (5 és 8 szerep az egyes keretek esetén).
- Szintaktikai jellemzéshez felhasználtam a **függőségi reprezentációt**, szemantikai jellemzéshez a magyar **WordNet**et. A WordNetben az egyes jelentések között **egyértelműsítést** végeztem a **Lesk algoritmussal** (Jurafsky & Martin, 2009).
- Modellemben gazdag jellemzőtéren alapuló osztályozót használtam a következő **jellemzőcsoportokkal**: felszíni, morfológiai, szintaktikai (függőségifa-alapú reprezentáció) és szemantikai (WordNet) jellemzők.
- A WordNet jellemzőcsoportnál a modellt kipróbáltam a **Lesk algoritmus** alkalmazásával és nélküle is.
- A **szózsák modellt** alkalmaztam a morfológiai, szintaktikai és szemantikai jellemzőknél a következő szócsoportokra: szótő és toldalékok; egy részfa tokenei; az elemzőfában két csomópont közötti tokenek; a WordNet hipernim hierarchiájában két synset közötti synsetek.
- A szerepek a legegyszerűbb esetekben a **célszó szintaktikai kapcsolatai** voltak, de voltak ettől eltérő esetek is. Megvizsgáltam a modell eredményét arra az esetre, ha csak azokkal a jelöltekkel foglalkozok, amelyek közvetlenül kapcsolódnak a célszóhoz az elemzőfában.
- A vásárlásokkal kapcsolatos kereten belül megvizsgáltam modellem teljesítményét a célszavak **vevő-centrikus** és **eladó-centrikus** csoportokra bontása esetén is.
- Az osztályozás szempontjából **jelentéktelen (kis számban előforduló) jellemző-előfordulásokat kihagytam** az osztályozásnál, ezzel csökkentettem a vektortér méretét. Megvizsgáltam, hogy ennek a kihagyásnak milyen hatása van az eredményekre.
- Megvizsgáltam, hogy az egyes doméneken melyik **szerepeket** lehet legeredményesebben meghatározni.

Eredmények a tézispontokon:
Igazoltam a következőt az események szemantikus szerepeinek címkézésénél (3. tézispont):

- *Megmutattam, hogy az Igei események célszavaihoz hatékonyan lehet szerepeket keresni gépi tanulásos módszerekkel.*
- *Bizonyítottam, hogy ezen a területen a legjobban teljesítő jellemzőcsoport a szintaktikai és a morfológiai elemzés csoport, ezeknek a csoportoknak minden vizsgált szerepre pozitív hatása van.*
- *Igazoltam, hogy ezek mellett a szemantikai jellemzők használata is a legtöbb esetben javítja az eredményeket, ezért a WordNet használata javasolt ezen a területen is.*
- *Bizonyítottam, hogy a WordNet jellemzőcsoportnál a Lesk algoritmus alkalmazása javítja az eredményeket.*
- *Igazoltam, hogy a szózsák modell alkalmazása a morfológiai, a szintaktikai (függőségifa-alapú) és a szemantikai jellemzőknél javítja az eredményeket a következő jellemzőcsoportokra: szótő és toldalékok; egy részfa tokenei; az elemzőfában két csomópont közötti tokenek; a WordNet hipernim hierarchiájában két synset közötti synsetek.*
- *Megmutattam, hogy jobb eredményeket érek el, ha az elemzőfában a célszótól távolabbi jelöltekkel is foglalkozok a tanításnál és kiértékelésnél.*
- *Bizonyítottam, hogy ezen a területen, ha a kis előfordulású jellemző-*eseteket elhagyjuk *az osztályozó kialakításánál, akkor jobb eredményeket kapunk.*
- *Igazoltam, hogy a vállalati vásárlások doménen a vizsgált szerepek közül az Ár és az Áru szerepeket, a tőzsdei rövidhírek doménen az Ár és az Elmozdulás-irány szerepeket lehet legeredményesebben meghatározni.*

10 Összefoglalás

10.1 Magyar nyelvű összefoglalás

Az értekezés fő célkitűzése természetes szövegekben lévő események detektálása, osztályozása és szemantikus szerepeik címkézése volt. Vizsgálataim során figyelmet fordítottam az események sajátosságaira, külön tárgyalva az igei és főnévi igenévi, valamint a főnévi eseményeket. Új, gépi tanuláson alapuló eljárásokat implementáltam az események detektálására és osztályozására, valamint a szemantikus szerepeik címkézésére. Módszereimnél törekedtem a gazdag jellemzőtér alkalmazására, ahol sok fajta jellemzőt teszteltem és hasonlítottam össze.
A gépi tanulásos módszerek mellett alkalmaztam szabályalapú módszereket is.
Mindhárom fő kutatási résznél kiemelt feladatomnak tekintettem olyan jellemzőcsoportok részletes kidolgozását, amelyek figyelembe veszik a magyar nyelv sajátosságait. Ezek a *morfológiai* és a *függőségifa-alapú jellemzőcsoportok* voltak. Mivel a magyar morfológiailag gazdag nyelv, így a *morfológiai jellemzőcsoportra* kiemelt figyelmet fordítottam. És mivel a magyar nyelv szabad szórendű és a függőségi fákkal dolgozó reprezentáció különösen jól használható szabad szórendű nyelvek elemzésére, ezért a *függőségifa-alapú jellemzőcsoportot* is kiemelten kezeltem. Ezek a jellemzőcsoportok jelentősen hozzájárultak az angol nyelvre már alkalmazott jellemzők eredményeinek javításához a magyar nyelvű szövegeken.
A jelen értekezésben az elért főbb eredményeket foglaltam össze. Először ismertettem az események és a szemantikus szerepek jellegzetességeit, majd a vizsgált területeken használt korpuszokat, valamint az alkalmazott gépi tanulási technikákat ismertettem. Ezek után bemutattam az események detektálása, osztályozása és szemantikus szerepeik címkézése területén alkalmazott különböző módszereket.

Az értekezésben elért főbb eredményeket az alábbi pontokban foglalom össze.

Mindhárom témánál alkalmaztam a következő forrásokat és módszereket:

- A magyar szövegek szavainak morfológiai elemzésére, majd szófaji egyértelműsítésére és a mondatok függőségi nyelvtan szerinti szintaktikai reprezentációjára a Szeged Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) egy részét használtam fel, a következő területekről: üzleti rövidhírek, szépirodalom, jogi szövegek, újsághírek, fogalmazás. Az alkalmazás működését megvizsgáltam részkorpuszonként is.
- A függőségi elemzőfa eredményénél nem csak a közvetlen szintaktikai kapcsolatokat vizsgáltam, hanem a jelöltek és a fában tőle távolabbi igék kapcsolatát is.
- Rendszereimben a vizsgált főnevek szemantikai jellemzéséhez alkalmaztam a magyar WordNet-et (Miháltz, és mtsai., 2008), ahol a WordNet hiperním hierarchiájában található szemantikai kapcsolatokat használtam fel. Mivel egy szóalakhoz több jelentés is tartozhat a WordNet-ben, ezért az egyes jelentések között egyértelműsítést végeztem a Lesk algoritmussal (Jurafsky & Martin, 2009).

- Porlasztásos méréssel megvizsgáltam, hogy az egyes jellemzőcsoportok hogyan befolyásolják a gépi tanulórendszer eredményeit.
- Szintaktikai jellemzéshez felhasználtam a függőségifa-alapú reprezentációt. Ennek során nem csak az igéhez közvetlenül kapcsolódó szavakat vizsgáltam, hanem a jelölt főnév és a fában tőle távolabbi igék kapcsolatát is (csak a 2. és 3. témánál).

A *morfológiai* és a *függőségifa-alapú szintaktikai* jellemzőcsoportoknak több témánál is kiemelt szerepe volt, ami azt igazolja, hogy az angol nyelvű szövegekre már használt jellemzők mellett hasznos olyan *jellemzők definiálása* is a magyar nyelvű szövegek elemzésénél, amelyek *felhasználják a magyar nyelv sajátosságait.*

10.1.1 Igei és főnévi igenévi események detektálása és osztályozása természetes nyelvű szövegekben

A szövegekben a legtöbb esemény igékhez kapcsolódik, és az igék általában eseményeket jelölnek, ezért külön foglalkoztam az igei és főnévi igenévi események azonosításával és osztályozásával (Subecz Z. , 2014). Bemutattam gazdag jellemzőtéren alapuló gépi tanuló megközelítésemet, amely automatikusan képes igei és főnévi igenévi események detektálására és osztályozására.

A legtöbb munkában csak adott eseményekkel foglalkoznak (például üzleti), vagy még azon belül is csak kiemelt eseményekkel (például cégfelvásárlás). Én minden típusú igei és főnévi igenévi esemény detektálásával és osztályozásával foglalkoztam.

A feladatot három részre osztottam. A szövegekben először az egy- és többszavas főnév + igei és főnévi igenévi kifejezéseket válogattam ki, majd a kiválogatottak közül detektáltam az eseményeket. A megtalált eseményeket ezután osztályoztam. Az általam megvalósított megközelítés gépi tanuló módszer alapján detektálja és osztályozza az eseményeket, amit szabályalapú módszerrel is kiegészítettem a jogi korpuszon.

Modellemben gazdag jellemzőtéren alapuló osztályozót használtam a következő jellemzőcsoportokkal: felszíni, lexikai, morfológiai, szintaktikai (függőségifa-alapú reprezentáció) és szemantikai (WordNet) jellemzők.

A WordNet jellemzőnél egy külön modellt is készítettem, ami kiválogatja azokat a synseteket, amelyek alá jellemzően események tartoznak, majd a kiválogatott elemeket felhasználtam a fő osztályozónál. Ugyancsak a WordNet jellemzőnél kipróbáltam a Lesk algoritmus alkalmazásával és anélkül is a modellemet.

Morfológiai elemzéshez felhasználtam még a magyarlanc nyelvészeti programcsomag RFSA morfológiai elemzőjét (Zsibrita, Vincze, & Farkas, 2013).

A morfológiai és a szintaktikai (függőségifa-alapú) jellemzőknél alkalmaztam a szózsák modellt szócsoportok jellemzésére a következő szócsoportokra: a szó töve és toldalékai; a kapcsolatok címkéi és a kapcsolatban lévő szavak lemmája a függőségi reprezentációnál.
A detektálásnál megvizsgáltam külön az igékre és külön a főnévi igenevekre.
Domének közötti keresztmérést is végeztem, ennek során a forráskorpuszon tanított modellt értékeltem ki a célkorpuszon. A domének közötti hasonlóságot gráfban ábrázoltam.
Mérésekkel megvizsgáltam, hogy a korpusz méretének változtatása hogyan befolyásolja az eredményeket.
Az igei események detektálása után *osztályoztam* azokat. Az osztályozást több szempont szerint is elvégeztem. Az első csoportnál az igék alapkategóriáit vizsgáltam meg: cselekvés, történés, létezés, állapot. Ezek közül az eseményeknél a *cselekvésnek* és a *történésnek* van fő szerepe, így ezt a két kategóriát emeltem ki. Modellemet két kisebb, de még gyakori kategórián is megvizsgáltam: a mozgás és a kommunikáció kategóriákon.

Igazoltam a következőket az igei és főnévi igenévi események detektálásánál és osztályozásánál (1. tézispont):

- *Bizonyítottam, hogy ezen a területen a legjobban teljesítő jellemzőcsoportok a morfológiai, a függőségifa-alapú szintaktikai és a szemantikai csoportok.*
- *Igazoltam, hogy a szabályalapú módszer alkalmazása a jogi korpuszon javítja a gépi tanulási rendszer eredményeit.*
- *Megmutattam, hogy a WordNet jellemzőcsoportnál a Lesk algoritmus alkalmazása javítja az eredményeket.*
- *Megmutattam, hogy a morfológiai és a szintaktikai (függőségifa-alapú) jellemzőknél a szózsák modellt hatékonyan lehet alkalmazni a következő szócsoportokra: a szó töve és toldalékai; függőségi reprezentációnál a kapcsolatok címkéi és a kapcsolatban lévő szavak lemmája.*
- *Igazoltam, hogy a detektálásnál az igékre jobb eredményt ad a modell, mint a főnévi igenevekre.*
- *Megmutattam, hogy a detektálás és az osztályozás szempontjából a Fogalmazás, Szépirodalom, Üzleti rövidhírek és az Újsághírek domének hasonlítottak legjobban egymásra, ezektől jelentősen eltért a Jogi domén.*
- *Bizonyítottam, hogy a Detektálásnál és osztályozásnál is a korpusz méretének növelése javítja az eredményeket, de a hozzáadott érték folyamatosan csökken.*

A tézispontban elért eredményeket a következő publikációkban ismertettem: (Subecz Z. , 2014), (Subecz & Csák, 2014). Az utóbbi munka társszerzője a nyelvészeti háttér biztosításában vett részt.

10.1.2 Főnévi események automatikus detektálása magyar nyelvű szövegekben függőségifa- és konstituensfa-alapú szintaktikai reprezentációval és WordNettel

Az igéken kívül lehetnek események más szófajú szavak is pl. főnevek, igenevek stb. Az igék mellett a főnévi események a leggyakoribbak, ezért a főnévi események detektálásával külön foglalkoztam (általános főnevek és igéből képzett főnevek) (Subecz Z. , 2016).

Bemutattam gazdag jellemzőtéren alapuló gépi tanuló megközelítésemet, amely automatikusan képes főnévi események detektálására függőségifa- és konstituensfa-alapú reprezentáció és WordNet alkalmazásával.

Modellemben gazdag jellemzőtéren alapuló osztályozót használtam a következő jellemzőcsoportokkal: felszíni, morfológiai, függőségifa-alapú, konstituensfa-alapú, szemantikai (WordNet), szózsák, lista és kombinált jellemzők.

A főnévi események detektálásához Névelem-felismerő alkalmazást (Szarvas, Farkas, & Kocsor, 2006) is implementáltam.

Szintaktikai jellemzéshez függőségifa- és konstituensfa-alapú reprezentációt is alkalmaztam és azok hatékonyságát összehasonlítottam.

Modellem teljesítményét megvizsgáltam az igéből képzett főnévi eseményekre és a nem igéből képzett főnévi eseményekre is.

A feladathoz több jellemző esetén felhasználtam a szózsák modellt szócsoportok jellemzéséhez.

A morfológiai elemzéshez két morfológiai elemzőt használtam fel. Két adatbányászati algoritmust implementáltam és hasonlítottam össze (Döntési fa, SVM).

Megvizsgáltam, hogy a jelöltek csoportosításával vagy anélkül lehet jobb eredményeket elérni.

Az alapjellemzők mellé a következő kiegészítő módszereket is alkalmaztam, amelyek javították az eredményeket: jelöltek csoportosítása; jellemzők súlyozása.

Bizonyítottam a következőket a főnévi események detektálásánál (2. tézispont):

- *Igazoltam, hogy a legjobban teljesítő jellemzőcsoportok a szemantikai és a szózsák csoportok.*

- *Megmutattam, hogy ezen a területen a szózsák modellt hatékonyan lehet alkalmazni a következő szócsoportok esetében: egy részfa tokenei; az elemzőfában két csomópont közötti tokenek; a WordNet hiperním hierarchiájában két synset közötti synsetek; a jelöltek környezetében lévő szavak a mondatban.*
- *Igazoltam, hogy ezen a területen jobb eredményt lehet elérni a függőségifa-alapú szintaktikai reprezentáció használatával, mint a konstituensfa-alapú reprezentáció használatával.*
- *Bizonyítottam, hogy ha a főnévi események detektálását a jelöltek csoportosításával végezzük el, akkor jobb eredményeket lehet elérni, mintha egy csoportban kezelnénk minden jelöltet.*

10.1.3 Események szemantikus szerepeinek automatikus címkézése

Az események detektálása mellett fontos azok szemantikus kapcsolatainak, *szemantikus szerepeinek* megtalálása is (*szemantikus szerepek címkézése*). Az események és azok szemantikus szerepeinek detektálását a természetesnyelv-feldolgozás sok területén lehet hasznosítani, például az összegzéskészítés, gépi fordítás és a válasz-keresés területein.

Ismertettem gazdag jellemzőtéren alapuló gépi tanuló megközelítésemet, amely automatikusan képes események szemantikus szerepeinek (Subecz Z. , 2015a). Az igei és főnévi igenévi célszavak szerepeit kerestem.

A szemantikus szerepek címkézése területén a *vállalati vásárlások, tulajdonváltozások* és a *tőzsdei hírek* kereteit vizsgáltam, mindkét esetben több domén-specifikus szerepet címkéztem (5 és 8 szerep az egyes keretek esetén).

Modellemben gazdag jellemzőtéren alapuló osztályozót használtam a következő jellemzőcsoportokkal: felszíni, morfológiai, szintaktikai (függőségifa-alapú reprezentáció) és szemantikai (WordNet) jellemzők.

A *WordNet jellemzőcsoportnál* a modellt kipróbáltam a Lesk algoritmus alkalmazásával és nélküle is.

A *szózsák modellt* alkalmaztam a morfológiai, szintaktikai és szemantikai jellemzőknél a következő szócsoportokra: szótő és toldalékok; egy részfa tokenei; az elemzőfában két csomópont közötti tokenek; a WordNet hipernim hierarchiájában két synset közötti synsetek.

A szerepek a legegyszerűbb esetekben a célszó szintaktikai kapcsolatai voltak, de voltak ettől eltérő esetek is. Megvizsgáltam a modell eredményét arra az esetre, ha csak azokkal a jelöltekkel foglalkozok, amelyek közvetlenül kapcsolódnak a célszóhoz az elemzőfában.

A *vásárlásokkal* kapcsolatos kereten belül megvizsgáltam modellem teljesítményét a célszavak vevő-centrikus és eladó-centrikus csoportokra bontása esetén is.
Az osztályozás szempontjából *jelentéktelen (kis számban előforduló) jellemző-előfordulásokat kihagytam* az osztályozásnál, ezzel csökkentettem a vektortér méretét. Megvizsgáltam, hogy ennek a kihagyásnak milyen hatása van az eredményekre.
Megvizsgáltam, hogy az egyes doméneken melyik **szerepeket** lehet **legeredményesebben** meghatározni.

Igazoltam a következőt az események szemantikus szerepeinek címkézésénél (3. tézispont):

- *Megmutattam, hogy az Igei események célszavaihoz hatékonyan lehet szerepeket keresni gépi tanulásos módszerekkel.*
- *Bizonyítottam, hogy ezen a területen a legjobban teljesítő jellemzőcsoport a szintaktikai és a morfológiai elemzés csoport, ezeknek a csoportoknak minden vizsgált szerepre pozitív hatása van.*
- *Igazoltam, hogy ezek mellett a szemantikai jellemzők használata is a legtöbb esetben javítja az eredményeket, ezért a WordNet használata javasolt ezen a területen is.*
- *Bizonyítottam, hogy a WordNet jellemzőcsoportnál a Lesk algoritmus alkalmazása javítja az eredményeket.*
- *Igazoltam, hogy a szózsák modell alkalmazása a morfológiai, a szintaktikai (függőségifa-alapú) és a szemantikai jellemzőknél javítja az eredményeket a következő jellemzőcsoportokra: szótő és toldalékok; egy részfa tokenei; az elemzőfában két csomópont közötti tokenek; a WordNet hipernim hierarchiájában két synset közötti synsetek.*
- *Megmutattam, hogy jobb eredményeket érek el, ha az elemzőfában a célszótól távolabbi jelöltekkel is foglalkozok a tanításnál és kiértékelésnél.*
- *Bizonyítottam, hogy ezen a területen, ha a kis előfordulású jellemző-eseteket elhagyjuk az osztályozó kialakításánál, akkor jobb eredményeket kapunk.*
- *Igazoltam, hogy a vállalati vásárlások doménen a vizsgált szerepek közül az Ár és az Áru szerepeket, a tőzsdei rövidhírek doménen az Ár és az Elmozdulás-irány szerepeket lehet legeredményesebben meghatározni.*

10.1.4 Jövőbeli tervek

Az eseményi információk kinyerése egyre időszerűbbé vált sok NLP alkalmazás számára, mint például a válaszkeresés, az automatikus összegzés, az információ vissza-keresés és az információkinyerés. A válaszkeresési kutatások szerint a legtöbb webes kereső kérdés eseményekkel kapcsolatos. Az automatikus összegzés szintén igényli az eseményinformációkat, felhasználva az események egymáshoz viszonyított sorrendjét.

A jövőben szeretném rendszereimet továbbfejleszteni az egyes jellemzők hatásainak részletesebb elemzésével, valamint azokat kidolgozni a magyartól eltérő más nyelvek esetére, nyelvspecifikus jellemzők megvalósításával. Emellett szeretném a jellemzőket általánosítani, hogy rendszerem alkalmas legyen nyelv-független eseményi információkinyerésre. Továbbá tervezem az eseménydetektáló, eseményosztályozó és a szerepfelismerő alrendszerek összekapcsolását egy összetett rendszerbe.

Véleményem szerint az értekezésben ismertetett módszereim, amelyeket az események detektálására, osztályozására és szemantikus szerepeik címkézésére dolgoztam ki, jól hasznosíthatóak más számítógépes nyelvészeti feladat megoldására is.

10.2 Summary in English

This dissertation is concerned with computer processing of events expressed in natural languages. Its main tasks are *event detection, event classification and the labelling of their semantic roles.*

The achievements of the thesis

The main achievements of this dissertation are summarized below along with the related publications.

I considered my *main task* in all three research areas to develop in detail feature groups that *take into account the characteristics of the Hungarian language*. These were the morphological and the dependency-tree based feature groups**.** Since the Hungarian language is a morphologically rich language, I paid great attention to the *morphological feature group.* And since the Hungarian language has free word order and dependency-tree based representations are well-suited for the analysis of languages with free word order, I paid particular attention to the *dependency-tree based feature group* as well. These feature groups significantly improved the results for the Hungarian texts.

In all three themes I used the following sources and methods:

- In my applications one part of the Hungarian Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) was used from the following areas: business and financial news, fictions, legal texts, newspaper articles, compositions of pupils. I tested the model's performance on each subcorpus.
- In my systems I used the Hungarian WordNet (Miháltz at al. 2008) for the semantic characterization of the words examined, where the semantic relations of the WordNet hypernym hierarchy were used. Since a word form may have more than one sense in the WordNet, I performed word sense disambiguation (WSD) of the particular senses using the Lesk algorithm (Jurafsky & Martin, Speech and Language Processing, 2009).
- I examined the efficacy of the particular feature groups using ablation analysis.
- For syntactic characterization the dependency-tree based representation was used. In this case not only the directly connected words were examined for the candidate verb, but also the relation between the candidate and the words farther away in the tree was analyzed (theme 2 and 3).

The key role the *morphological* and the *dependency-tree based syntactic* feature groups have in more cases confirms that besides the features used for English texts, it is useful to define *features for Hungarian* text analysis which take into account the *characteristics of the Hungarian language.*

10.2.1 The detection and classification of verbal and infinitival events in natural language texts.

Most events belong to verbs in texts and verbs usually denote events. That is why I dealt with the detection and classification of verbal and infinitival events separately (Subecz Z. , 2014). I introduced my rich feature set based machine-learning approach that can automatically detect and classify verbal and infinitival events.

Most works deal only with certain events (e.g. business) or more specific ones (e.g. acquisitions). I dealt with all types of verbal and infinitival event detection and classification.

I divided the task into three parts. First, the multiword noun + verb and noun + infinitive expressions were identified, then, the events were detected from them. Afterwards, the events found were classified. My approach detects and classifies events with machine learning techniques and has been expanded with a rule based method on the Legal Corpus.

I used a rich feature set based classifier in my model with the following feature groups: surface, lexical, morphological, syntactic (dependency-tree based representation) and semantic (WordNet) features.

A new model has been created for the WordNet feature group, which picks out the synsets that are typically found in the hypernym chains of events, then, the synsets picked out have been used in the main classifier. Also, in the WordNet feature group I tested my model with and without the Lesk algorithm.

For morphological analysis I also used the RFSA morphological parser of the magyarlanc linguistic toolkit (Zsibrita, Vincze, & Farkas, 2013).

For the morphological and syntactical (dependency-tree based representation) features I applied the bag of words model for the characterization of word groups.

I tested the model's performance separately for verbs and infinitives.

Besides the main examinations cross-domain measurements were carried out. In this case, the model trained on the source corpus was evaluated on the target corpus. I represented the similarity between domains in graph.

I examined using measurements how changes in corpus size modifies results.

After the detection of verbal and infinitival events I *classified* them. The classification was performed according to several criteria. First, I investigated the main verb types: actions, occurrences, existence and states. Out of them the *action* and *occurrence* categories are mostly related to events, therefore I focused on these two categories. I tested my model on smaller, but frequent categories: movement and communication.

The following points have been verified regarding the detection and classification of verbal and infinitival events (Thesis 1):

- *I have showed that in this area the best performing feature groups are the morphological, dependency-tree based syntactic and semantic groups.*
- *I have justified that applying the rule based method to the Legal Corpus improves the results of the machine learning system.*
- *I have showed that applying the Lesk algorithm to the WordNet feature group improves the results.*
- *I have proved that applying the bag of words model at the morphological and syntactic (dependency-tree based representation) features improves the results for the following word groups: word stem, prefixes and suffixes; relations and relation-lemmas at the parse tree.*

- *I have showed that regarding detection the model performs better for verbs than for infinitives.*
- *I have justified that regarding detection and classification Compositions of pupils, Fictions, Financial news and Newspaper articles were very similar to each other; legal texts were significantly different.*
- *I have proved that regarding detection and classification, increasing the corpus size improves the results, but the added value constantly declines.*

I presented the results of this Thesis in the following publications: (Subecz Z. , 2014), (Subecz & Csák, 2014). The co-author of the last publication provided the linguistic background.

10.2.2 Automatic detection of nominal events in Hungarian texts with dependency-tree and constituency-tree based representations and the WordNet

Besides verbs other parts of speech (e.g. nouns, participles) can also denote events. Among them nominal events are the most frequent, therefore I dealt with nominal event detection in detail (Subecz Z. , 2016). I introduced my machine learning approach based upon a rich feature set, which can detect nominal events in Hungarian texts using dependency-tree and constituency-tree based representations and the WordNet.

I used a classifier based upon a rich feature set with the following feature groups: Surface, Morphological, Dependency-tree based, Constituency-tree based, Semantic (WordNet), Bag of words, List and Combined features. For nominal event detection I also implemented a Named Entity Recognition System (Szarvas, Farkas, & Kocsor, 2006).

For syntactic characterization I used dependency-tree based and constituency-tree based representations and compared their efficacy.

The performance of the model was tested for deverbal and non-deverbal nominal events too.

The bag of words model was used for the characterization of word groups for these features: tokens of subtrees; tokens between two nodes in the parse-tree; the synsets between two nodes in the WordNet hypernym hierarchy; the words around the candidate in the sentence.

Two morphological parsing were used for morphological analysis. Two data mining algorithms were implemented and compared (Decision Tree and Support Vector Machine (SVM) algorithms).

I compared the results by forming groups from the candidates and without groups.

Besides the main features the following additional methods were used, which improved the results: forming candidate groups; feature weighting.

The following points have been verified for nominal event detection (Thesis 2):

- *I have showed that in this area the best performing feature groups are the semantic and the bag of words groups.*
- *I have proved that the bag of words model can be applied effectively in the case of the following word groups: tokens of subtrees; tokens between two nodes in the parse-tree; the synsets between two nodes in the WordNet hypernym hierarchy; the words around the candidate in the sentence.*
- *I have showed that, in this area, the dependency-tree based representation produces better results than the constituency-tree based one.*
- *I have justified that nominal event detection produces better results by forming groups from the candidates than by handling all candidates in one group only.*

I presented the results of this Thesis in the following publications: (Subecz Z. , 2016), (Subecz Z. , 2017a), (Subecz Z. , 2017b).

10.2.3 Automatic semantic role labelling of events

Besides event detection, the *labelling of the semantic relations* of events is an important task (Semantic Role Labelling, SRL). The detection of the events and their semantic roles can be utilized in several areas of natural language processing, for example, in summarization, machine translation and question answering.

I introduced my machine learning approach based upon a rich feature set, which can automatically label semantic roles of events (Subecz Z. , 2015a). I searched for roles for target words of verbal and infinitival events.

In semantic role labelling I dealt with the frames of *company purchases* and *stock market news*. In both cases, several domain-specific roles were labelled (5 and 8 roles in the particular frames).

In my model I used a classifier based upon a rich feature set with the following feature groups: Surface, Morphological, Syntactic (dependency-tree based representation) and Semantic (WordNet) features.

In the WordNet feature group I tested my model with and without the *Lesk algorithm*.

The *bag of words model* was used for the morphological, syntactic and semantic (WordNet) feature groups for the following word groups: word stem, prefixes and suffixes; tokens of subtrees; tokens between two nodes in the parse-tree; the synsets between two nodes in the WordNet hypernym hierarchy;

In the *company purchases* domain I examined my model's performance sorting out the targets into two groups: customer-centric and seller-centric groups.
I *left out the infrequent feature-entities* from the classifier, thus the vectorspace-size was reduced. I have investigated the effect of this leaving out.
In simpler cases the roles have direct syntactic relationship with the target word. I examined the model's performance if it dealt only with these directly connected candidates.
On each domain I have **searched for the roles** that the model can detect most **efficiently**.

The following points have been verified for the semantic role labelling of events (Thesis 3):

- *I have showed that for the target words of verbal events we can effectively search roles with machine learning techniques.*
- *I have justified that in this area, the best performing feature groups are the syntactic and the morphological groups. These groups have had positive impact on all examined roles.*
- *I have showed that in addition to the above, using semantic features also improves the results therefore it is recommended that the WordNet be applied here too.*
- *I have proved that applying the Lesk algorithm to the WordNet feature group improves the results.*
- *I have justified that applying the bag of words model to the morphological, syntactic and semantic features improves the results for the following word groups: word stem, prefixes and suffixes; tokens of subtrees; tokens between two nodes in the parse-tree; the synsets between two nodes in the WordNet hypernym hierarchy.*
- *I have showed that the model achieves better results if it deals with farther candidates in the dependency tree as well.*
- *I have proved that leaving out the infrequent feature-entities from the classifier improves the results.*
- *I have showed that on the purchases of companies domain the Price (Ár) and the Item (Árú) and on the e news from stock markets domain the Price (Ár) and the Shift-direction (Elmozdulás-irány) roles can be detect with most efficiency.*

10.2.4 Future Work

In the future, I would like to improve my systems by conducting a detailed analysis of the effects of the features included and developing systems for other languages as well by adapt-

ing language specific features. Later, I would like to generalize the features to achieve a language-independent event extraction system. Moreover, I plan to integrate my event detection and classification, and semantic role labelling applications into one complex system.

I believe that my research on automatic event detection and classification, and semantic role labelling can be successfully exploited in several NLP tasks and it will contribute to developing novel approaches in many areas of natural language processing.

Irodalomjegyzék

Ahn, D. (2006). The stages of event extraction. *Proceedings of Workshop on Annotating and Reasoning about Time and Events* (pp. 1-8). Sydney, Australia: Association for Computational Linguistics.

Allan, J., Lavrenko, V., & Jin, H. (2000). First Story Detection In TDT Is Hard. *CIKM '00, Proceedings of the ninth international conference on Information and knowledge management*, 374–381.

Allan, J., Lavrenko, V., Malin, D., & Swan, R. (2000). Detections, Bounds, and Timelines: UMass and TDT-3. *Information Retrieval*, 167–174.

Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of ACM, 26 (11), ACM New York, NY, USA* , 832–843.

Allen, J. F. (1995). *Natural language understanding (2nd ed.).* Redwood City, CA, USA. ISBN:0-8053-0334-0: Benjamin-Cummings Publishing Co., Inc.,.

Alonso, O., Gertz, M., & Baeza-Yates, R. (2007). On the Value of Temporal Information in Information Retrieval. *ACM SIGIR Forum, ACM New York, NY, USA*, 35-41.

Aone, C., & Ramos-Santacruz, M. (2000). REES: A Large-Scale Relation and Event Extraction System. *6th Applied Natural Language Processing Conference (ANLP 2000)* (pp. 76–83). Association for Computational Linguistics.

Atkinson, M., Du, M., Piskorski, J., Tanev, H., Yangarber, R., & Zavarella, V. (2013). Techniques for Multilingual Security-Related Event Extraction from Online News. *Computational Linguistics*, 163-186.

Atkinson, M., Piskorski, J., Tanev, H., Goot, E., Yangarber, R., & Zavarella, V. (2009). Automated Event Extraction in the Domain of Border Security. *User Centric Media*, 321–326.

Bach, E. (1981). On time, tense and aspect. In *Radical Pragmatics* (pp. 63-81). New York: Academic Press.

Bach, E. (1986). The Algebra of Events. *Linguistics and Philosophy, Tense and Aspect in Discourse, volume 9, Springer*, 5-16.

Baker, C. F., Fillmore, C. J., Lowe, & B., J. (1998). The Berkeley framenet project. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics* (pp. 86–90). Montreal, Canada: ACL.

Baker, M. (1988). *Incorporation: A Theory of Grammatical Function Changing.* Chicago, ISBN: 0226035417: University of Chicago Press.

Bańczerowski, J. (1994). A kommunikációs kompetencia és összetevői. *Magyar Nyelvőr. 118.*, 277-286.

Bańczerowski, J. (1999). A kognitív nyelvészet alapelvei. *Magyar Nyelvőr 123*, 78–87.

Belletti, A., & Rizzi, A. (1988). Psych-verbs and theta-theory. *Natural Language and Linguistic Theory, 6, Springer*, 291–352.

Bernard, C. (1976). *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems.* Cambridge Textbooks in Linguistics, ISBN 0521290457: Cambridge University Press.

Bethard, S. (2002). *Temporal and Causal Structure in Text: A Machine Learning Approach.* Boulder, CO, USA: Doctoral Dissertation, University of Colorado at Boulder, ISBN: 978-0-549-31488-2.

Bethard, S. (2007). *Finding Event, Temporal and Causal Structure in Text: A Machine Learning Approach.* University of Colorado at Boulder, ISBN: 978-0-549-31488-2: Doctoral Dissertation, Boulder, CO, USA.

Bethard, S., & Martin, J. (2006). Identification of Event Mentions and their Semantic Class. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 146-154). Sydney, Australia: Association for Computational Linguistics.

Bittar, A. (2009). Annotation of events and temporal expressions in French. *Proceeding ACL-IJCNLP '09 Proceedings of the Third Linguistic Annotation Workshop* (pp. 48–51). Stroudsburg, PA, USA: Association for Computational Linguistics.

Björne, J., Ginter, F., Pyysalo, S., Tsujii, J., & Salakoski, T. (2010). Complex event extraction at PubMed scale. *Bioinformatics 26 (12):i382-90, PMID: 20529932, DOI: 10.1093/bioinformatics/btq180*, 382-390.

Blaheta, D., & Charniak, E. (2000). Assigning function tags to parsed text. *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*, (pp. 234–240). Seattle, Washington.

Boas, H. C. (2002). Bilingual framenet dictionaries for machine translation. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), Anthology ID: L02-1052.* Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA).

Boguraev, B. J., Pustejovsky, R., Ando, & Verhagen, M. (2007). TimeBank Evolution as a Community Resource for TimeML Parsing. *Language Resources and Evaluation, February 2007, Volume 41, Issue 1, Online ISSN: 1572-8412, DOI: https://doi.org/10.1007/s10579-007-9018-8, Springer Netherlands*, 91-115.

Boguraev, B., & Ando, R. (2005). *TimeBank-driven TimeML analysis.* Annotating, Extracting and Reasoning about Time and Events: Dagstuhl Seminars. German ResearchFoundation.

Boguraev, B., & Ando, R. (2007). Effective Use of TimeBank for TimeML Analysis. *Annotating, Extracting and Reasoning about Time and Events, Springer, Berlin, Heidelberg, ISBN: 978-3-540-75988-1*, 41-58.

Capet, P., Delavallade, T., Nakamura, T., Sandor, A., Tarsitano, C., & Voyatzi, S. (2008). A Risk Assessment System with Automatic Extraction of Event Types. *Intelligent Information Processing IV* (pp. 220-229). Springer Boston.

Carreras, X. (2005). Learning and Inference in Phrase Recognition. *Doctoral thesis in Universitat Politècnica de Catalunya (UPC)*, 11.

Carreras, X., & Marquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. *Proceeding CONLL '05 Proceedings of the Ninth Conference on Computational Natural Language Learning* (pp. 152-164). Stroudsburg, PA, USA: Association for Computational Linguistics.

Carreras, X., & Marquez., a. L. (2004). Introduction to the conll-2004 shared task: Semantic role labeling. *Proceeding CONLL '05 Proceedings of the Ninth Conference on Computational Natural Language Learning* (pp. 152-164). Stroudsburg, PA, USA: Association for Computational Linguistics.

Caselli, T. (2009). Time, Events and Temporal Relations: an Empirical Model for Temporal Processing of Italian Texts. *Ph.D. thesis*. Pisa, Italy: Istituto Di Linguistica Computazionale, Consiglio Nazionale delle Ricerche.

Caselli, T., dell'Orletta, F., & Prodanof, I. (2009). TETI: a TimeML compliant TimEx tagger for Italian. *Proceedings of the International Multiconference on Computer Science and Information Technology* (pp. 185–192). IEEE.

Caselli, T., Russo, I., & Rubino, F. (2011). Recognizing deverbal events in context. *International Journal of Computational Linguistics and Applications*, 91.

Charniak, E. (2000). A maximum-entropy-inspired parser. *Proceeding NAACL 2000 Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference* (pp. 132-139). Stroudsburg, PA, USA: Association for Computational Linguistics.

Chen, C., & NG, V. (2012). Joint modeling for chinese event extraction with rich linguistic features. *Proceedings of COLING 2012* (pp. 529-544). Mumbai, India: The COLING 2012 Organizing Committee.

Chen, J., & Rambow, O. (2003). Use of deep linguistic features for the recognition and labeling of semantic arguments. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 41-48). Sapporo, Japan: SIGDAT.

Chen, M., Zhang, C., & Chen, S. (2007). Semantic Event Extraction Using Neural NetworkEnsembles. *1st IEEE International Conference on Semantic Computing (ICSC 2007)* (pp. 575–580). IEEE Computer Society.

Chen, Z., & Ji, H. (2009). Language specific issue and feature exploration in chinese event extraction. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* (pp. 209-212). Boulder, Colorado: Association for Computational Linguistics.

Chiticariu, L., Li, Y., & Reiss, F. (2013). Rule-based information extraction is dead! long live rule-based information extraction systems! (pp. 827-832). Seattle, WA: Proceedings Conference on Empirical Methods in Natural Language Process.

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton and Co.

Christensen, J., Mausam, S. S., & Etzioni, O. (2010). Semantic role labeling for open information extraction. *Proceeding FAM-LbR '10 Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading* (pp. 52-60). Stroudsburg, PA, USA: Association for Computational Linguistics.

Chun, H.-W., Hwang, Y.-S., & Rim, H.-C. (2004). Unsupervised Event Extraction from Biomedical Literature Using Co-occurrence Information and Basic Patterns. *1st International Joint Conference on Natural Language Processing (IJCNLP 2004)* (pp. 777–786). Springer Berlin Heidelberg.

Cohen, K., Verspoor, K., Johnson, H., Roeder, C., Ogren, P., & Baumgartner, W. (2009). High-Precision Biological Event Extraction with a Concept Recognizer. *Workshop on BioNLP: Shared Task at 47th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 50–58). Association for Computational Linguistics.

Cohn, T., & Blunsom, P. (2005). Semantic role labelling with tree conditional random fields. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)* (pp. 169-172). Stroudsburg, PA, USA: Association for Computational Linguistics.

Collins, M. (2003). Head-Driven Statistical Models for Natural Language Parsing. *Computational Linguistics, Volume 29 Issue 4, December 2003, MIT Press Cambridge, MA, USA, ISSN: 0891-2017*, 589-637.

Comrie, B. (1985). *Tense.* Cambridge, ISBN: 0521281385: Cambridge University Press, Cambridge Textbooks in Linguistics.

Copestake, A., & Flickinger, D. (2000). An open-source grammar development environment and broad-coverage english grammar using hpsg. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), Anthology ID: L00-1276.* Athens, Greece: European Language Resources Association (ELRA).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, September 1995, Volume 20, Issue 3, Kluwer Academic Publishers Hingham, MA, USA, doi: 10.1023/A:1022627411411*, 273–297.

Cowie, J., & Lehnert, W. (1996). Information Extraction. *Communications of the ACM, Volume 39 Issue 1, Jan. 1996, ACM New York, NY, USA, doi: 10.1145/234173.234209*, 80-91.

Crystal, D. (2011). *A Dictionary of Linguistics and Phonetics.* ISBN 1444356755: John Wiley & Sons, The Language Library.

Cybulska, A., & Vossen, P. (2011). Historical Event Extraction from Text. *5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011) at 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)* (pp. 39–43). Association for Computational Linguistics.

Csendes, D., Csirik, J., & Gyimothy, T. (2004). The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. *Seventh International Conference on Text, Speech and Dialogue (TSD 2004)*, (pp. 41-49). Brno, Czech Republic.

Csendes, D., Csirik, J., Gyimóthy, T., & Kocsor, A. (2005). The Szeged Treebank. *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)* (pp. 123-131). Karlovy Vary, Czech Republic: Springer LNAI.

Dahiya, Y. V. (1995). *Panini as a linguist: Ideas and Patterns.* Delhi, India, ASIN: B002A9QYEU: Eastern Book Linkers.

Dahl, D. A., Palmer, M. S., & Passonneau, R. J. (1987). Nominalizations in pundit. *Proceedings of the 25th annual meeting on Association for Computational Linguistics* (pp. 131–139). Morristown, NJ, USA: Association for Computational Linguistics.

Deribas, V. (1968). Ustojchivye glagolno-imennye slovosochetanija v obshchectvenno-politicheskih tekstah. In B. A. Anisimov, *Iz opyta perpodavanija russkogo jazyka nerusskim. 4. kiadás.* Moszkva: Akadémiai Kiadó.

Dorr, B. J., Farwell, D., Green, R., Habash, N., & Helmreich, S. (2004). Interlingual annotation of parallel text corpora: A new framework for annotation and evaluation. *Journal of Natural Language Engineering*.

Dowty, D. (1979). *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ.* Springer Netherlands, ISBN 9027710090: Springer Science & Business Media.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language, 67(3)*, 547-619.

Dowty, D. R. (1986). The effects of aspectual class on the temporal structure of discourse: Semantics or pragmatics?. *Linguistics and Philosophy, 9,*, 37–61.

Dowty, D. R. (1989). On the semantic content of the notion thematic role. *Properties, Types, and Meanings, volume 2*, 69–130.

Ehmann, B., Lendvai, P., Miháltz, M., Vincze, O., & László, J. (2013). Szemantikus szerepek a narrative kategoriális elemzés (NARRCAT) rendszerében. *IX.Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 121-123). Szeged.

Engelen, B. (1968). Zum System der Funktionsverbgefüge. *Wirkendes Wort 18*, 289–303.

Farkas, R., Konczer, K., & Szarvas, G. (2004). Szemantikus keretillesztés és az IE-rendszer automatikus kiértékelése. *II. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 49-53). Szeged.

Feng, X., Qin, B., & Liu, T. (2018). A language-independent neural network for event detection. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 66-71). Berlin, Germany: Association for Computational Linguistics, DOI: 10.18653/v1/P16-2011.

Fillmore, C. (1968). *The case for case.* New York, NY: Holt, Rinehart, and Winston.: Universals in Linguistic Theory.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *New York Academy of Sciences* (pp. 20–32). New York: Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech volume 280.

Fillmore, C. J., Narayanan, S., & Baker, C. F. (2006). What can linguistics contribute to event extraction? *Aaai Workshop - Technical Report*, 18-23.

Forascu, C. (2008). GMT to +2 or how can TimeML be used in Romanian. *Proceedings of the Sixth International Language Resources* (pp. 3238–3242). Marrakech, Morocco: ELRA.

Gábor, K., & Héja, E. (2007). Clustering Hungarian verbs on the basis of complementation patterns. *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop* (pp. 91–96). Prague, Czech Republic: Association for Computational Linguistics.

Gerber, M., Chai, J., & Meyers, A. (2009). The role of implicit argumentation in nominal SRL. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 146–154). Boulder, Colorado: Association for Computational Linguistics.

Gildea, D., & Hockenmaier, J. (2003). Identifying semantic roles using Combinatory Categorial Grammar. *2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 57–64). Sapporo, Japan.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics, MIT Press Cambridge, MA, USA, DOI: 10.1162/089120102760275983*, 245–288.

Gildea, D., & Palmer, M. (2002). The necessity of syntactic parsing for predicate argument recognition. *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02)*, (pp. 239–246). Philadelphia, PA.

Gorzitze, S., & Pado, S. (2012). Corpus-based acquisition of German event-and object-denoting nouns. *Proceedings of KONVENS 2012* (pp. 259-263). Vienna, Austria: KONVENS.

Grétsy, L., & Kemény, G. (1996). *Nyelvművelő kéziszótár.* Budapest: Auktor Könyvkiadó.

Grishman, R., Huttunen, S., & & Yangarber, R. (2002). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics, Volume 35 Issue 4, August 2002, Elsevier Science San Diego, USA*, 236-246.

Grishman, R., Huttunen, S., & & Yangarber, R. (2002). Real-time event extraction for infectious disease outbreaks. *Proceedings of the second international conference on Human Language Technology Research* (pp. 366-369). Association for Computational Linguistics.

Grishman, R., Westbrook, D., & Meyers, A. (2005). *Nyu english ace 2005 system description.* Gaithersburg, Maryland: Department of Computer Science, New York University.

Gruber, J. (1967). *Studies in Lexical Relations.* North Holland: MIT Linguistics Dissertations, MIT Working Papers in Linguistics, .

Gruber, J. S. (1965). *Studies in lexical relations.* Massachusetts Institute of Technology.: MIT Linguistics Dissertations, MIT Working Papers in Linguistics.

Hacioglu, K., Pradhan, S., Ward, W., Martin, J. H., & Jurafsky, D. (2004). Shallow semantic parsing using support vector machines. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004* (pp. 233–240). Boston, Massachusetts, USA: Association for Computational Linguistics.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: an update. *SIGKDD Explorations, ACM SIGKDD Explorations Newsletter, Volume 11 Issue 1, ACM New York, NY, USA, DOI: 10.1145/1656274.1656278*, 10–18.

He, Q., Chang, K., & Lim, E. (2007). Analyzing feature trajectories for event detection. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 207-214). Amsterdam, Netherland: ACM New York, NY, USA.

Hemphill, C., Godfrey, J., & Doddington, G. (1990). The atis spoken language systems pilot corpus. *HLT '90 Proceedings of the workshop on Speech and Natural Language* (pp. 96-101). Hidden Valley, Pennsylvania: Association for Computational Linguistics.

Heringer, H.-J. (1968). *Die Opposition von 'kommen' und 'bringen' als Funktionsverben. Untersuchungen zur grammatischen Wertigkeit und Aktionsart. Sprache der Gegenwart 3.* Düsseldorf, Germany: Schwann, Sprache der Gegenwart.

Hirst, G. (1987). *Semantic Interpretation and the Resolution of Ambiguity.* New York, NY, USA, ISBN:0-521-32203-0: Cambridge University Press.

Hovav, M., & Levin, B. (1998). The Projection of Arguments: Lexical and Compositional Factors. In *Building Verb Meanings* (pp. 97-134). Stanford University ISBN: 1575861100 : CSLI Publications, Stanford, CA.

Hull, R., & Gomez, F. (1996). Semantic interpretation of nominalizations. *AAAI'96 Proceedings of the thirteenth national conference on Artificial intelligence - Volume 2, ISBN:0-262-51091-X* (pp. 1062-1068). Portland, Oregon: AAAI Press.

Hwang, C. H., & Schubert, L. K. (1992). Tense trees as the "fine structure" of discourse. *Proceedings of the 30th annual meeting on Association for Computational Linguistics* (pp. 232–240). Morristown, NJ, USA: ACL.

IJntema, W., Sangers, J., Hogenboom, F., & Frasincar, F. (2012). A Lexico-Semantic Pattern Language for Learning Ontology Instances from Text. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 15(1)*, 37–50.

Im, S., You, H., Jang, H., Nam, S., & Shin, H. (2009). KTimeML: Specification of temporal and event expressions in Korean text. *Proceedings of the 7th Workshop on Asian Language Resources* (pp. 115–122). ACL.

Jackendoff, R. (1975). Semantic Interpretation in Generative Grammar. In R. Freidin, *Language - Review* (pp. 189-205). Massachusetts: Linguistic Society of America, DOI: 10.2307/413161.

Jackendoff, R. (1987). The status of thematic relations in linguistic theory. *Linguistic Inquiry, Vol. 18, No. 3 (Summer, 1987), The MIT Press*, 369–411.

Jackendoff, R. (1990). *Semantic Structures.* Cambridge, MA : The MIT Press, Current Studies in Linguistics, ISBN: 9780262100434.

Jacobs, G., Lefever, E., & Hoste, V. (2018). Economic event detection in company-specific news text. *Proceedings of the First Workshop (ECONLP 2018) - Economics and Natural Language Processing* (pp. 1-10). Melbourne, Australia: Association for Computational Linguistics.

Jeong, Y., & Myaeng, S. (2012). Using Syntactic Dependencies and WordNet Classes for Noun Event Recognition. *The 2nd Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web in Conjunction with the 11th International Semantic Web Conference,*, (pp. 40-50).

Ji, H., & Grishman, R. (2008). Refining Event Extraction through Cross-Document Inference. *Proceedings of ACL08 HLT, Anthology ID: P08-1030* (pp. 254-262). Columbus, Ohio: Association for Computational Linguistics.

Johnson, C. R., Fillmore, C. J., Petruck, M. R., & Baker, C. F. (2002). *FrameNet: Theroy and Practice.* ICSI Technical Report TR-02-009: International Computer Science Institute, Technical Report.

Jungermann, F., & Morik, K. (2008). Enhanced Services for Targeted Information Retrieval by Event Extraction and Data Mining. *13th International Conference on Natural Language and Information ystems: Applications of Natural Language to Information Systems* (pp. 335–336). Springer Berlin Heidelberg.

Jurafsky, D., & Martin, J. (2009). *Speech and Language Processing.* New Jersey, ISBN-10: 9780131873216: Prentice Hall, Upper Saddle River.

Kakkonen., E., & Arendarenko, T. (2012). Ontology-Based Information and Event Extraction for Business Intelligence. *15th International Conference on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA 2012)* (pp. 89–102). Springer Berlin Heidelberg.

Kamijo, S., Matsushita, Y., Ikeuchi, K., & Sakauchi, M. (2000). Traffic Monitoring and Accident Detection at Intersections. *IEEE Transactions on Intelligent Transportation Systems, 1(2)* (pp. 108–118). IEEE.

Kamp, H., & Reyle, U. (1983). *From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Dordrecht, The Netherlands: Studies in Linguistics and Philosophy 42.

Kearns, K. (1998). Extraction from make the claim constructions. *Journal of Linguistics, Vol. 34, No. 1 (Mar., 1998), Cambridge University Press*, 53-72.

Kiefer, F. (1992). *Strukturális magyar nyelvtan 1.-Mondattan.* Budapest, ISBN: 0469001307707: Akadémiai Kiadó.

Kiefer, F. (2006). *Aspektus és akcióminőség, különös tekintettel a magyar nyelvre.* Buapest: Akadémiai Kiadó, ISBN: 9630583887 .

Kiefer, F. (2007). *Jelentéselmélet.* Budapest: Corvina Kiadó, ISBN: 9789631356823.

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., & Tsujii, J. (2002). Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web. *Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM 2002),*, (pp. 1-6).

King, G., & Lowe, W. (dátum nélk.). An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders. *International Organization, 57(03)*, 617-642.

Kiparsky, P. (2002). On the architecture of panini's grammar. *Three lectures delivered at the Hyderabad Conference on the architecture of grammar.* Hyderabad: UCLA.

Kipper, K., Dang, H. T., & Palmer, M. (2000). Classbased construction of a verb lexicon. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence* (pp. 691-696). Austin, United States: AAAI Press, ISBN:0-262-51112-6.

Kong, F., Li, Y., Zhou, G., Zhu, Q., & Qian, P. (2008). Using semantic roles for coreference resolution. *International Conference on Advanced Language Processing and Web Information Technology, ISBN: 978-0-7695-3273-8* (pp. 150–155). Dalian Liaoning, China: Advanced Language Processing and Web Information Technology, International Conference on Advanced Language.

Kuti, J., Varasdi, K., Cziczelszki, J., Gyarmati, Á., Nagy, A., Tóth, M., és mtsai. (2006). Igei wordnet és igei eseményszerkezet ábrázolása. *IV. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 97-108). Szeged: SZTE.

Kwak, H., Lee, C., Park, H., & Sue, M. (2010). What is Twitter , a Social Network or a News Media ? *ACM* (pp. 591–600). Raleigh, North Carolina, USA: Proceeding of WWW '10 Proceedings of the 19th international conference on World wide web.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning, ISBN:1-55860-778-1* (pp. 282-289). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Lagos, N., Segond, F., Castellani, S., & O'Neill, J. (2010). Event Extraction for Legal Case Building and Reasoning. *Springer, Intelligent Information Processing, ISBN: 978-3-642-16326-5* (pp. 92–101). Berlin, Germany: Springer Berlin Heidelberg.

Lakoff, G., & Thompson, H. (1975). Dative Questions in Cognitive Grammar. In R. Grossman, J. San, & T. Vance, *Papers from the Parasession on Functionalism* (pp. 337-350). Chicago: Chicago Linguistics Society.

Landeghem, S., Björne, J., Wei, C., Hakala, K., Pyysalo, S., & Ananiadou, S. (2013). Large-Scale Event Extraction from Literature with Multi-Level Gene Normalization. *PLoS One. 2013; 8(4): e55814, doi: 10.1371/journal.pone.0055814*.

Langacker, R. W. (1987). *Foundation of cognitive grammar. Volume I. , Theoretical Prerequisites.* California: Stanford University Press.

Langer, S. (2004). A Linguistic Test Battery for Support Verb Constructions. *Lingvisticae Investigationes, 27 (2)*, 171–184.

Lascarides, A., & Asher, N. (1993). A semantics and pragmatics for the pluperfect. *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics* (pp. 250–259). Morristown, NJ, USA: ACL.

Lee, C., Chen, Y., & Z., J. (2003). Ontology-Based Fuzzy Event Extraction Agent for Chinese E-News Summarization. *Expert Systems with Applications, 25(3)*, 431–447.

Lei, Z., L., W., Zhang, Y., & Liu, Y. (2005). A System for Detecting and Tracking Internet News Event. *6th Pacific-Rim Conference on Multimedia (PCM 2005)* (pp. 754–764). Springer Berlin Heidelberg.

Levin, B. (1995). English Verb Classes and Alternations: A Preliminary Investigation. *Language - Review, Vol. 71, No. 1 (Mar., 1995), Linguistic Society of America, DOI: 10.2307/415968*, 144-146.

Levin, B., & Rappaport, M. (1986). The formation of adjective passives. *Linguistic Inquiry 17*, 623–662.

Li, C., Sun, A., & Datta, A. (2012). Twevent: segment-based event detection from tweets. *Proceeding of the ICIKM '12 Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 155-164). Maui, Hawaii, USA, ISBN: 978-1-4503-1156-4: ACM.

Li, F., Sheng, H., & Zhang, D. (2002). Event Pattern Discovery from the Stock Market Bulletin. *5th International Conference on Discovery Science (DS 2002)* (pp. 35–49). Springer Berlin Heidelberg.

Litkowski, K. (2004). Senseval-3 task: Automatic labeling of semantic roles. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* (pp. 9-12). Barcelona, Spain: Association for Computational Linguistics.

Liu, M., Liu, Y., Xiang, L., Chen, X., & Yang, Q. (2008). Extracting Key Entities and Significant Events from Online Daily News. *9th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2008)* (pp. 201–209). Springer Berlin Heidelberg.

Liu, R.-L., & Soo, V.-W. (1993). An Empirical Study on Thematic Knowledge Acquisitin based on Syntactic Clues and Heuristics. *ACL* (pp. 243-250). Columbus, Ohio: Processing of 31st Annual Meeting of the ACL.

Llorens, H., Saquete, E., & Navarro-Colorado, B. (2010). TimeML Events Recognition and Classification: Learning CRF Models with Semantic Roles. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, (pp. 725-733). Beijing.

Makrai, M. (2015). Mélyesetek a 4lang fogalmi szótárban. (pp. 50-57). Szeged, Szegedi Tudományegyetem: XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015).

Mani, I., & Shiffman, B. (2003). Inferring temporal ordering of events in news. *NAACL-Short '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL* (pp. 55-57). Edmonton, Canada: Association for Computational Linguistics, DOI: 10.3115/1073483.1073502.

Mani, I., & Wilson, G. (2000). Robust temporal processing of news. *ACL '00 Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, DOI: 10.3115/1075218.1075228* (pp. 69-76). Hong Kong: Association for Computational Linguistics.

Mani, I., Verhagen, M., Wellner, B., Lee, C. M., & & Pustejovsky, J. (2006). Machine learning of temporal relations. *ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 753–760). Sydney, Australia: Association for Computational Linguistics, DOI: 10.3115/1220175.1220270.

Marantz, A. P. (1984). *On the Nature of Grammatical Relations.* Cambridge, Massachusetts: MIT Press, Linguistic Inquiry Monographs, ISBN: 9780262131933.

March, O., & Baldwin, T. (2008). Automatic event reference identi¯cation. *Proceedings of the 2008 Australasian Language Technology Workshop*, (pp. 79-87). Hobart, Australia.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1994). Building a large annotated corpus of english : The penn treebank. *Computational Linguistics - Special issue on using large corpora: II archive, MIT Press Cambridge, MA, USA*, 313–330.

Marquez, L., Comas, P., Gimenez, J., & Catala, N. (2005). Semantic role labeling as sequential tagging. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)* (pp. 193–196). Ann Arbor, Michigan: Association for Computational Linguistics.

Marquez, L., Recasens, M., & Sapena, E. (2013). Coreference resolution: An empirical study based on semeval-2010 shared task. *Language Resources and Evaluation Volume 47 Issue 3, September 2013, Springer-Verlag New York, Inc, DOI: 10.1007/s10579-012-9194-z*, 661–694.

Marsic, G. (2011). *Temporal processing of news: annotation of temporal expressions, verbal events and temporal relations.* Wolverhampton: PhD thesis, University of Wolverhampton, LAP LAMBERT Academic Publishing, ISBN: 9783659497612.

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. Forrás: http://mallet.cs.umass.edu.

McClosky, D., & Surdeanu, M. M. (2011). Event Extraction as Dependency Parsing. *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (pp. 1626-1635). Portland, Oregon: Association for Computational Linguistics.

Melli, G., Shi, Z., Wang, Y., Liu, Y., Sarkar, A., & Popowich, F. (2006). Description of squash, the sfu question answering summary handler for the duc-2006 summarization task. *Proceedings of the HLT/EMNLP Document Understanding Workshop (DUC).* Vancouver, Canada: DUC-2006 Summarization Task.

Meulen, A. (1995). *Representing time in natural language: the dynamic interpretation of tense and aspect.* Cambridge, DOI:10.7551/mitpress/5897.001.0001: Cambridge University Press.

Meyers, A., Macleod, C., Yangarber, R., Grishman, R., Barrett, L., & Reeves, R. (1998). Using nomlex to produce nominalization patterns for information extraction. *Proceedings of the COLING-ACL Workshop on the Computational Treatment of Nominals.* Montreal, Canada: ACL.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., & Young, B. G. (2004). The nombank project: An interim report. *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation* (pp. 24-31). Boston, Massachusetts, USA: Association for Computational Linguistics.

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Gábor, P., és mtsai. (2008). Methods and Results of the Hungarian WordNet Project. (pp. 311–320). Szeged. University of Szeged: Proceedings of the Fourth GlobalWordNet Conference (GWC 2008).

Miháltz, M., Indig, B., & Prószéky, G. (2015). Igei vonzatkeretek és tematikus szerepek felismerése nyelvi erőforrások összekapcsolásával egy kereslet-kínálat elvű mondatelemzőben. (pp. 298-302). SZTE, Szeged: XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015).

Miller George. (1995). WordNet: A Lexical Database for English. *Communications of the ACM Volume 38 Issue 11, Nov. 1995* (pp. 39-41). New York, NY, USA: ACM.

Miller, S., Stallard, D., Bobrow, R., & Schwartz, R. (1996). A fully statistical approach to natural language interfaces. *Proceedings of the 34th Annual Meeting of the ACL* (pp. 55–61). Santa Cruz, California: ACL.

Misra, & Niwas, V. (1966). *The Descriptive Technique of Panini: an introduction.* The Hague and Paris: Mouton: Janua Linguarum: Studia Memoriae Nicolai Van Wijk.

Mitchell, M., & Mulherin, J. (1994). The Impact of Public Information on the Stock Market. *Journal of Finance 49, Working paper. Chicago and Clemson, S.C.: University*, 923-950.

Miwa, M., Saetre, R., Kim, J., & Tsjujii, J. (2010). Event Extraction With Complex Event Classification Using Rich Features. *Journal of Bioinformatics and Computational Biology, 8(1), DOI: /10.1142/S0219720010004586*, 131–146.

Moens, M., & Steedman, M. (1988). Temporal ontology and temporal reference. *Computational Linguistics - Special issue on tense and aspect, Volume 14 Issue 2, June 1988, MIT Press Cambridge, MA, USA* , 15-28.

Moldovan, D., Clark, C., & Harabagiu, S. (2005). Temporal Context Representation and Reasoning. (pp. 1099-1104). Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-2005): Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Moreau, E., & Tellier, I. (2009). The crotal srl system : a generic tool based on tree-structured crf. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task* (pp. 91–96). Boulder, Colorado: Association for Computational Linguistics.

Moreda, P., Llorens, H., Saquete, E., & Palomar, M. (2011). Combining semantic information in question answering systems. *Information Processing and Management, Volume 47, Issue 6, November 2011, DOI: /10.1016/j.ipm.2010.03.008*, 870-885.

Moreda, P., Navarro, B., & & Palomar, M. (2007). Corpus-based semantic role approach in information retrieval. *Data and Knowledge Engineering, Volume 61, Issue 3, June 2007, DOI: /10.1016/j.datak.2006.06.010*, 467–483.

Moreno, L., Palomar, M., Molina, A., & Ferrandez, A. (1999). *Introduccion al Procesamiento del Lenguaje Natural.* Alicante, Spain: Servicio de Publicaciones de la Universidad de Alicante.

Moschitti, A. (2004). A study on convolution kernel for shallow semantic parsing. *Proceedings of the 42-th Conference on Association for Computational Linguistic (ACL-2004), Article No. 335, DOI: 10.3115/1218955.1218998.* Barcelona, Spain: Association for Computational Linguistics.

Moulin, B. (1997). Temporal contexts for discourse representation: An extension of the conceptual graph approach. *Applied Intelligence, July 1997, Volume 7, Issue 3, ISSN: 0924-669X*, 227–255.

Nagy, I., Vincze, V., & Zsibrita, J. (2013). Félig kompozicionális szerkezetek automatikus felismerése doménadaptációs technikák segítségével a Szeged Korpuszon. *IX. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 47-58). Szeged.

Naradowsky, J., Riedel, S., & David, S. (2012). Improving nlp through marginalization of hidden syntactic structure. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 810–820). Jeju Island, Korea: Association for Computational Linguistics.

Narayanan, S., & Harabagiu, S. (2004). Question answering based on semantic structures. *COLING '04 Proceedings of the 20th international conference on Computational Linguistics, Article No. 693, DOI: 10.3115/1220355.1220455* (pp. 184–191). Geneva, Switzerland: Association for Computational Linguistics.

Naughton, M., Kushmerick, N., & Carthy, J. (2006). Event Extraction from Heterogeneous News Sources. *Proceedings Workshop Event Extraction.* Menlo Park, California, USA: American National Conference Artificial.

Nguyen, T. H., & Grishman, R. (2015). Event Detection and Domain Adaptation with Convolutional Neural Networks. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics* (pp. 365–371). Beijing, China: Association for Computational Linguistics.

Nishigauchi., T. (1984). Control and the thematic domain. *Language, Vol. 60, No. 2 (Jun., 1984), Linguistic Society of America, DOI: 10.2307/413640*, 215-250.

Nuij, W., Milea, V., Hogenboom, F., Frasincar, F., & Kaymak, U. (2014). An Automated Framework for Incorporating News into Stock Trading Strategies. *IEEE Transactions on Knowledge and Data Engineering, 26(4)*, 823–835.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition bank:An annotated corpus of semantic roles. *Association for Computational Linguistics, Volume 31 | Issue 1, DOI: 10.1162/0891201053630264* , 71-106.

Parsons, T. (1990). *Events in the Semantics of English: A Study of Subatomic Semantics.* Cambridge: The MIT Press.

Peris, A., Taule, M., Boleda, G., & Rodriguez, H. (2010). ADN-Classifier: Automatically assigning denotation types to nominalizations. *Proceedings of the seventh LREC conference*, (pp. 1422 - 1428). Valetta, Malta.

Petrovic, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to twitter. *HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 181-189). Los Angeles, California: Association for Computational Linguistics, ISBN:1-932432-65-5.

Piskorski, J., Tanev, H., & Wennerberg, P. (287–300). Extracting Violent Events From On-Line News for Ontology Population. *10th International Conference on Business Information Systems (BIS 2007)* (pp. 2007). Springer Berlin Heidelberg.

Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J. H., & Jurafsky, D. (2003). Support vector learning for semantic argument classification. *Machine Learning, September 2005, Volume 60, Issue 1–3, ISSN: 0885-6125, Kluwer Academic Publishers*, 11–39.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J. H., & Jurafsky, D. (2004). Shallow Semantic Parsing Using Support Vector Machines. *Conference of the North American Chapter of the Association for Computational Linguistics & Human Language Technologies (NAACL-HLT)*, (pp. 233–240). Boston, MA.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J., & Jurafsky, D. (2005). Semantic role labeling using different syntactic views. *Proceedings of the Association for Computational Linguistics 43rd annual meeting (ACL-2005).* Ann Arbor, MI.

Prószéky, G. (2003). NewsPro: automatikus információszerzés gazdasági rövidhírekből. *Szegedi Tudományegyetem* (pp. 161–166). Egyetemi nyomda, Szeged: Magyar Számítógépes Nyelvészeti Konferencia.

Prószéky, G., & Kis, B. (2003). Mire jó a NewsPro? (Rövidhírek automatikus elemzése – magyarul). *Boss Magazin, 2003. október*, 40–41.

Punyakanok, V., Roth, D., & Yih, W. (2005). The necessity of syntactic parsing for semantic role labeling. *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, (pp. 1117–1123). Edinburgh, UK.

Punyakanok, V., Roth, D., & Yih, W. (2008). The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics, Volume 34 | Issue 2, DOI: 10.1162/coli.2008.34.2.257* , 257-287.

Punyakanok, V., Roth, D., Yih, W., Zimak, D., & Tu, Y. (2004). Semantic role labeling via generalized inference over classifiers. *Proceedings of the 8th Conference on Natural Language Learning (CoNLL-2004)* (pp. 130–133). Boston, MA: ACL.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics, Volume 17 Issue 4, MIT Press Cambridge, MA, USA* , 409–441.

Pustejovsky, J. (1991). The syntax of event structure. *Cognition, 41*, 47–81.

Pustejovsky, J., Casta, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., és mtsai. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. Proceedings of IWCS-5: Fifth International Workshop on Computational Semantics.

Pustejovsky, J., Verhagen, M., Sauri, R., Littman, J., Gaizauskas, R., Katz, G., és mtsai. (2006). *TimeBank 1.2.* Linguistic Data Consortium LDC2006T08.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Francisco, CA, USA: Morgan Kaufmann Publishers, Inc.

Rappaport, M., & Levin, B. (1988). What to do with theta-roles. In W. Wilkins, *Syntax and semantics* (pp. 7–37). New York, USA: Academic Press.

Reichenbach, H. (1966). Elements of Symbolic Logic. In *The Tenses of Verbs* (pp. 287–298). New York: The Macmillan Company.

Reuter, T., & Cimiano, P. (2012). Event-based classification of social media streams. *ICMR '12 Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Article No. 22, ISBN: 978-1-4503-1329-2* (pp. 22-29). Hong Kong, China: ACM.

Reuter, T., Cimiano, P., Drumond, L., Buza, K., & Schmidt-Thieme, L. (2011). Scalable event-based clustering of social media via record linkage techniques. *Fifth International AAAI Conference on Weblogs and Social Media* , (pp. 313-320). Barcelona, Spain.

Riedel, S., Chun, H., Takagi, T., & Tsujii, J. (2009). A Markov Logic Approach to Bio-Molecular Event Extraction. *BioNLP '09 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task* (pp. 41–49). Boulder, Colorado: Association for Computational Linguistics, ISBN: 978-1-932432-44-2.

Riemsdijk, H. V., & Williams, E. (1986). *Introduction to the Theory of Grammar.* Cambridge: MIT Press.

Rijsbergen, v. (1979). *Information retrieval.* London: Butterworths, Information Retrieval Group, University of Glasgow.

Riloff, E. (1993). Automatically constructing a dictionary for information extraction tasks. *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI)*, (pp. 811–816). Washington, D.C.

Riloff, E., & Schmelzenbach, M. (1998). An empirical approach to conceptual case frame acquisition. *Proceedings of the Sixth Workshop on Very Large Corpora*, (pp. 49–56). Montreal, Canada.

Ritter, A., Etzioni, O., & Clark, S. (2012). Open domain event extraction from twitter. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1104-1112). Beijing, China: ACM, ISBN: 978-1-4503-1462-6.

Robaldo, L., Caselli, T., Russo, I., & Grella, M. (2011). From Italian Text To TimeML Document Via Dependency Parsing. *International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2011, ISBN: 978-3-642-19436-8* (pp. 177–187). Springer, Berlin, Heidelberg.

Rocher, R. (1964). "Agent" et "Objet" chez Panini. *Journal of the American Oriental Society, Vol. 84, No. 1 (Jan. - Mar., 1964), DOI: 10.2307/597061*, 44–54.

Romeo, L., Lebani, G., Bel, N., & Lenci, A. (2014). Choosing which to use? A study of distributional models for nominal lexical semantic classification. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC2014)*, (pp. 4366-4373).

Rosa, J. L., & Francozo, E. (1999). Hybrid Thematic Role Processor: Symbolic Linguistic Relations Revised by Connectionist Learning. (pp. 852-857). Stockholm, Sweden: Proceedings of the 16th International Joint Confrerence on Artificial Intelligence-IJCAI'99.

Rozwadowska, B. (1988). Thematic restricitons on derived nominals. In W. Wilkins, *Syntax and Semantics 21* (pp. 147–66). New York, USA: Academic Press.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., & Johnson, C. R. (2005). *FrameNet II: Extended theory and practice.* Berkeley, CA: International Computer Science Institute.

Saeed, J. (1977). *Semantics.* Oxford: Wiley-Blackwell Publishers.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. (pp. 851–860). New York, NY, USA: Proceedings of the 19th international conference on World wide web.

Sauri, R. (2010). *Tempeval 2. spanish data release.* Barcelona: Barcelona Media Technical Report.

Sauri, R., Knippen, R., Verhagen, M., & Pustejovsky, J. (2005). Evita: A Robust Event Recognizer for QA Systems., (pp. 700-707). Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing.

Schuler, K. K. (2005). VerbNet: A Broad-coverage, Comprehensive Verb Lexicon. *Ph.D. thesis, Paper AAI3179808*. Philadelphia, PA, USA: University of Pennsylvania, Computer and Information Science Dept.

Sciullo, A., & E., W. (1987). On the definition of word. *Machine Translation, December 1990, Volume 4, Issue 4, ISBN 0-262-04091-3*, 313–317.

Seker, S. E., & Diri, B. (2010). TimeML and Turkish Temporal Logic. *International Conferenc on Artificial Intelligence, Vol 10*, (pp. 881–887).

Shen, D., & Lapata, M. (2007). Using semantic roles to improve question answering. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 12–21). Prague, Czech Republic: Association for Computational Linguistics.

Siegel, E. V., & McKeown, K. (2000). *Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights.* MIT Press Cambridge, MA, USA: Computational Linguistics, Volume 26 Issue 4, December 2000, DOI: 10.1162/089120100750105957.

Siklósi, B., & Novák, A. (2014). Identifying and Clustering Relevant Terms in Clinical Records Using Unsupervised Methods. *2nd International Conference on Statistical Language and Speech Processing* (pp. 233–243). Springer International Publishing.

Siklósi, B., & Novák, A. (2015). Restoring the Intended Structure of Hungarian Ophthalmology Documents. *The Association for Computational Linguistics* (pp. 152–157). Beijing, China: Proceedings of the BioNLP 2015 Workshop on Biomedical Natural Language Processing.

Siklósi, B., Novák, A., & Prószéky, G. (2014). Resolving abbreviations in clinical texts without pre-existing structured resources. *LREC.* Reykjavik, Iceland: Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing.

Smith, C. (1991). *The Parameter of Aspect.* Dordrecht: Kluwer Academic Press.

Song, F., & Cohen, R. (1991). Tense interpretation in the context of narrative. *Proceeding of AAAI'91 Proceedings of the ninth National conference on Artificial intelligence - Volume 1* (pp. 131–136). Anaheim, California, USA: AAAI Press, ISBN:0-262-51059-6.

Sprugnoli, R., & Tonelli, S. (2019). Novel Event Detection and Classification for Historical Texts. *Computational Linguistics - Association for Computational Linguistics - ISSN: 0891-2017*, 1-38.

Stallard, D. (2000). Talk'n'travel: A conversational system for air travel planning. *Proceedings of the Sixth Applied Natural Language Processing Conference* (pp. 68–75). Seattle, Washington, USA: Association for Computational Linguistics.

Steedman, M. (2000). *The Syntactic Process.* Cambridge, MA. USA: The MIT Press, ISBN:0-262-19420-1.

Strubell, E., Verga, P., & Andor, D. (2018). Linguistically-Informed Self-Attention for Semantic Role Labeling. *In Conference on Empirical Methods in Natural Language Processing (EMNLP).* Brussels, Belgium: Association for Computational Linguistics.

Subecz, Z. (2014). Detection and Classification of Events in Hungarian Natural Language Texts. *Proceedings of the 17th International Conference, TSD 2014,* (pp. 68-75). Brno, Czech Republic: Springer Lecture Notes in Computer Science Volume 8655.

Subecz, Z. (2015a). Automatic Labeling of Semantic Roles with a Dependency Parser in Hungarian Economic Texts. *18th International Conference on Text, Speech and Dialogue, TSD 2015* (pp. 261-272). Brno, Czech Republic: Springer.

Subecz, Z. (2015b). Szemantikus szerepek automatikus címkézése függőségi elemző alkalmazásával magyar nyelvű gazdasági szövegeken. *XI. MAGYAR SZÁMÍTÓGÉPES NYELVÉSZETI KONFERENCIA* (pp. 95-106). Szeged: Szegedi Tudományegyetem.

Subecz, Z. (2016). Automatic Detection of Nominal Events in Hungarian Texts with Dependency Parsing and WordNet. *Information and Software Technologies, 22nd International Conference, ICIST 2016* (pp. 580-592). Druskininkai, Lithuania: Springer.

Subecz, Z. (2017a). Event Detection in Hungarian Texts with Dependency and Constituency Parsing and WordNet. *Informatics 2017, IEEE 14th International Scientific Conference on Informatics* (pp. 365-371). Poprad Slovakia: IEEE Xplore.

Subecz, Z. (2017b). Főnévi események automatikus detektálása függőségi elemző és WordNet alkalmazásával magyar nyelvű szövegeken. *XIII. MAGYAR SZÁMÍTÓGÉPES NYELVÉSZETI KONFERENCIA* (pp. 13-24). Szeged: Szegedi Tudományegyetem.

Subecz, Z., & Csák, É. (2014). Igei események detektálása és osztályozása magyar nyelvű szövegekben. *X. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 237–247). Szeged.

Surdeanu, M., Harabagiu, S. M., Williams, J., & Aarseth, P. (2003). Using predicateargument structures for information extraction. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 8–15). Sapporo, Japan: Association for Computational Linguistics.

Surdeanu, M., Johansson, R., Meyers, A., Marquez, L., & Nivre, J. (2008). The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies. *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning* (pp. 159–177). Manchester, England: Coling 2008 Organizing Committee.

Surdeanu, M., Marquez, L., Carreras, X., & Comas, P. (2007). Combination strategies for semantic role labeling. *Journal of Artificial Intelligence Research 29, AI Access Foundation, DOI: 10.1613/jair.2088*, 105-151.

Szarvas, G., Farkas, R., & Kocsor, A. (2006). A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. *DS'06 Proceedings of the 9th international conference on Discovery Science* (pp. 267-278). Barcelona, Spain: Springer-Verlag Berlin, Heidelberg, LNAI 4265, ISBN:3-540-46491-3.

Szarvas, G., Farkas, R., & Kocsor, A. (2006). A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. *The Ninth International Conference on Discovery Science on Discovery Science* (pp. 267-278). Barcelona, Spain: Springer Verlag Berlin, Heidelberg, LNAI 4265, ISBN:3-540-46491-3.

Szőts, M., Csirik, J., Gergely, T., & Karvalics, L. (2010). MASZEKER: projekt szemantikus keresőtechnológia kidolgozására. (pp. 159-167). Szeged, Szegedi Tudományegyetem: VII. Magyar Számítógépes Nyelvészeti Konferencia.

Tackstrom, O., Ganchev, K., & Das, D. (2015). Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics, Volume 3, DOI: 10.1162/tacl_a_00120*, 29-41.

Talmy, L. (1985). Figure and ground as thematic roles. *Symposium on Thematic Relations.* Seattle: Proceedings of Annual Meeting of the Linguistic Society of America.

Tanev, H., Piskorski, J., & Atkinson, M. (2008). Real-Time News Event Extraction for Global Crisis Monitoring. *13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems (NLDB 2008)* (pp. 207–218). London: Springer Berlin Heidelberg.

Tanev, H., Zavarella, V., Linge, J., Kabadjov, M., Piskorski, J., Atkinson, M., és mtsai. (2009). Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish. *Linguamática, 1(2), ISSN: 1647–0818*, 55-66.

Tenny, C., & Pustejovsky, J. (2000). *Events as Grammatical Objects.* Stanford University, Stanford, CA: CSLI Publications, Series: CSLI Lecture Notes, ISBN: 1575862069.

Thompson, C. A., Levy, R., & Manning, C. (2003). A generative model for FrameNet semantic role labeling. *Machine Learning: Proceedings of the Fourteenth European Conference on Machine Learning, ECML 2003* (pp. 397-408). Cavtat-Dubrovnik, Croatia: Springer, Proceedings, volume 2837.

Tikk, D. (2007). *Szövegbányászat.* Budapest: Typotex kiadó.

Tolcsvai, N. G. (2009). A magyar segédige + igenév szerkezet szemantikája. *Magyar Nyelvőr. 133. évf. 4. sz.* , 373-393.

Toutanova, K., Haghighi, A., & D., C. (2005). Joint learning improves semantic role labeling. *Proceedings of the Association for Computational Linguistics 43rd annual meeting (ACL-2005).* Ann Arbor, MI.

Toutanova, K., Haghighi, A., & Manning, C. (2008). A global joint model for semantic role labeling. *Computational Linguistics, Volume 34 Issue 2, MIT Press Cambridge, MA, USA*, 161-191.

Tran, M., Nguyen, M., Nguyen, S., Nguyen, M., & Phan, X. (2012). A Real – Time News Event Extraction Framework for Vietnamese. *4th International Conference on Knowledge and Systems Engineering (KSE 2012)* (pp. 161–166). IEEE Computer Society.

Tron, V., Kornai, A., Gyepesi, G., Németh, L., Halácsy, P., & Varga, G. (2005). Hunmorph: Open source word analysis. *Proceedings of the Workshop on Software, Software '05,* (pp. 77–85). Stroudsburg, PA, USA.: Association for Computational Linguistics.

Vargas-Vera, M., & Celjuska, D. (2004). Event Recognition on News Stories and Semi-Automatic Population of an Ontology. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), ISBN:0-7695-2100-2* (pp. 615–618). Beijing, China: IEEE Computer Society.

Vendler, Z. (1957). Verbs and times. *Philosophical Review, 56 (2), DOI: 10.2307/2182371*, 143–160.

Vendler, Z. (1967). *Linguistics in Philosophy.* Ithaca, New York, USA: Cornell University Press, ISBN: 0801404363.

Verhagen, M., Gaizauskas, R., Hepple, M., Schilder, F., Katz, G., & Pustejovsky, J. (2007). Semeval-2007 task 15: Tempeval temporal relation identification. *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 75–80). Prague: ACL.

Verhagen, M., Mani, I., Sauri, R., Knippen, R., Jang, S. B., & Littman, J. (2005). Automating temporal annotation with TARSQI. *ACLdemo '05 Proceedings of the ACL 2005* (pp. 81–84). Ann Arbor, Michigan: Association for Computational Linguistics, DOI: 10.3115/1225753.1225774.

Verhagen, M., Sauri, R., Caselli, T., & Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 57–62). Uppsala, Sweden: ACL.

Verkuyl, H. (1993). A Theory of Aspectuality. The Interaction between Temporal and Atemporal Structure. *Journal of Linguistics, Get access Volume 31, Issue 1, Cambridge Studies in Linguistics, volume 64, Cambridge University Press*, 177-181.

Verkuyl, H. J. (1972). *On the Compositional Nature of the Aspects.* Dordrecht, ISBN: 9027702276: D. Reidel Publishing Company, Foundations of language. Supplementary series ; v. 15.

Vickrey, D., & Koller, D. (2008). Applying sentence simplification to the CoNLL-2008 Shared Task. *CoNLL '08 Proceedings of the Twelfth Conference on Computational Natural Language Learning* (pp. 268-272). Manchester, United Kingdom: Association for Computational Linguistics, ISBN: 978-1-905593-48-4.

Vincze, V. (2009). Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban. *VI. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 390–393). Szeged.

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., & Csirik, J. (2010). Hungarian Dependency Treebank. *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (pp. 1855-1862). Valletta, Malta: Springer.

Vincze, V., Zsibrita, J., & Nagy, T. (2013). Dependency Parsing for Identifying Hungarian Light Verb Constructions. *Proceedings of International Joint Conference on Natural Language Processing*, (pp. 207–215).

Webber, B. L. (1988). Tense as discourse anaphora. *Computational Linguistics, vol. 14, no. 2,*, 61–73.

Wei, C.-P., & Lee, Y.-H. (2004). Event detection from Online News Documents for Supporting Environmental Scanning. *Decision Support Systems - Special issue: Knowledge management technique, Volume 36 Issue 4, Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands, DOI: 10.1016/S0167-9236(03)00028-9*, 385–401.

Weng, J., & Lee, B. (2011). Event detection in twitter. *Proceedings of the AAAI conference on weblogs and social media (ICWSM-11), ISBN: 978-3-642-38561-2* (pp. 401-408). Barcelona, Catalonia, Spain: Springer, Lecture Notes in Computer Science 7923, .

Williams, E. (1981). Argument structure and morphology. *Linguistic Review, Volume 1, Issue 1, ISSN 0167-6318*, 81–114.

Xu, F., Uszkoreit, H., & Li, H. (2006). Automatic Event and Relation Detection with Seeds of Varying Complexity. *2006 AAAI Workshop on Event Extraction and Synthesis (W8) at 21st National Conference on Artificial Intelligence (AAAI 2006).*

Xue, N., & Palmer, M. (2004). Calibrating features for semantic role labeling. *Proceedings of the 2004 Conference on Empirical Methods on Natural Language Processing (EMNLP-2004)* (pp. 88-94). Barcelona, Spain: Association for Computational Linguistics.

Xue, N., & Zhou, Y. (2010). Applying Syntactic, Semantic and Discourse Constraints in Chinese Temporal Annotation. *COLING '10 Proceedings of the 23rd International Conference, Computational Linguistics* (pp. 1363–1372). Beijing, China: Association for Computational Linguistics.

Y. Nishihara, K. S. (2009). Event Extraction and Visualization for Obtaining Personal Experiences from Blogs. *Symposium on Human Interface 2009 on Human Interface and the Management of Information. Information and Interaction* (pp. 315–324). Springer Berlin Heidelberg.

Yakushiji, A., Tateisi, Y., & Miyao, Y. (2001). Event Extraction from Biomedical Papers using a Full Parser. *World Scientific Publishing, DOI: 10.1142/9789814447362_0040* (pp. 408–419). River Edge, New Jersey, USA: 6th Pacific Symposium on Biocomputing (PSB 2001).

Yang, Y., Pierce, T., & Carbonell, J. (1998). A Study of Retrospective and On-Line Event Detection. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 28–36). Melbourne, Australia: New York, USA.

Yangarber, R. (2005). Extracting Information about Outbreaks of Infectious Epidemics. *Proceedings of HLT/EMNLP, Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing* (pp. 22-23). Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Yi, S.-t., & Palmer, M. (2004). Pushing the boundaries of semantic role labeling with SVM. *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics; .* Hyderabad, India.

Zubizarreta, M. L. (1987). *Levels of Representation in the Lexicon and in the Syntax.* Dordrecht, Holland: Foris, ISBN: 9067652865.

Zsibrita, J., Vincze, V., & Farkas, R. (2013). magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. *Proceedings of RANLP-2013, International Conference on Recent Advances in Natural Language Processing* (pp. 763–771). Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA.