Szegedi Tudományegyetem
Nyelvtudományi Doktori Iskola
Elméleti nyelvészet program

Névmási anaforafeloldási kísérletek a magyar nyelvben

DOKTORI (PhD) ÉRTEKEZÉS

Kovács Viktória

Témavezető: Dr. Szécsényi Tibor

Szeged

2021

Tartalomjegyzék

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani mindenkinek, aki valamilyen módon hozzájárult a disszertációm létrejöttéhez, legyen az szakmai észrevétel vagy emberi támogatás.

Szeretnék köszönetet mondani a Szegedi Tudományegyetem Általános Nyelvészeti Tanszékének minden oktatójának az egyetemi éveim kezdete óta tartó folyamatos segítségért, a konstruktív kritikákért és a szakmai támogatásért, különösképpen a témavezetőmnek, Dr. Szécsényi Tibornak, aki az egyetemi és doktoranduszi éveimet is végig segítette tanácsaival, iránymutatásával.

Köszönettel tartozom a dolgozat előopponenseinek Dr. Vincze Veronikának és Dr. Sass Bálintnak a részletes véleményezésért, a rendkívül értékes ötletekért és a disszertációm műhelyvitáján folytatott szakmai eszmecseréért.

Szeretném megköszönni a PhD-képzésben résztvevő hallgatótársaimnak a közös élményeket, a hasznos információkat és a folyamatos kölcsönös támogatást, valamint a volt tanítványaimnak a közös munka során szerzett tapasztalatokat.

Végül, de nem utolsó sorban a legnagyobb köszönettel a családomnak és a páromnak tartozom, akik folyamatosan bíztattak, kiegyensúlyozott, nyugodt családi hátteret biztosítottak a doktori disszertációm megírása során.

1. Bevezetés

A disszertáció célja egy vizsgálat bemutatása, amely során az automatikus névmási anaforafeloldás lehetőségeit gépi tanulási kísérleteken keresztül veszem sorra a magyar nyelvben. Az anaforafeloldással kapcsolatban az elsődleges kérdés, hogy valójában mit is szeretnénk automatikusan kinyerni a szövegből, és itt nem feltétlenül az anafora definíciójára kell gondolnunk.

A számítógépes nyelvészet az anaforafeloldást hagyományosan az automatikus tartalomkinyerés vagy a gépi fordítás problémájának szemszögéből közelíti meg, ezért a kinyerés során kevéssé veszi figyelembe a szövegalkotót és a mentális állapotát. A hangsúly a *helyes* visszautalások kinyerésén van, azokon, amelyek megfelelnek a grammatikai és szintaktikai szabályoknak, éppen ezért gyakran ezek a szabályok az alapjai az automatikus feloldást célzó rendszereknek. A visszautalások azonban nem mindig ilyen egyszerűen azonosíthatók. Számos esetet képzelhetünk el, amelyekben a visszautalás annak ellenére sikeresnek mondható, tehát a befogadó helyesen azonosítja a visszautaló szó antecedensét, hogy a visszautalás a nyelvi szabályoknak megfelelt volna. Ilyen eset például, ha egy nem anyanyelvi beszélő helytelenül egyezteti nemben a névmást, vagy ha egy enyhe kognitív zavarban szenvedő személy az ’aktuális miniszterelnök’ kifejezéssel utal egy olyan személyre, akire nem illik valójában a leírás. Az enyhe kognitív zavarban szenvedők beszédátirataiban található koreferenciakapcsolatokról bővebben egy korábbi tanulmányomban (Kovács 2017) értekeztem. Tehát, ha az automatikus anaforafeloldás célja a szövegalkotó nyelvtudásának vagy mentális állapotának felmérése, teljesen más megközelítés szükséges a probléma megoldásához, hiszen ezekben az esetekben a rendszernek elsődlegesen a szövegalkotó szándékát kell felismernie, nem a helyes nyelvtani szabályokon alapuló visszautalásokat.

Abban az esetben, ha az anaforafeloldás nem egy komplexebb feladat része, amelyet meg szeretnénk oldani, akkor is el kell határozni azt, hogy az adott rendszer melyik aspektusra támaszkodik erőteljesebben, a szövegalkotó egyéni nyelvhasználatára, vagy a nyelvtani szabályokra. Ezt a döntést általában a szövegtípust figyelembe véve érdemes meghozni. Ha az anaforafeloldó rendszer hivatalos szövegekben keresi a visszautalásokat, akkor számíthatunk rá, hogy a szövegben a nyelvtani szabályoknak megfelelő visszautalásokat fogunk találni, ha azonban személyes levelezésekben, közösségi médiás bejegyzésekben, akkor nem tekinthetünk kizárólagos megszorításokként a grammatikai szabályokra. Ez azt jelenti, hogy nem érdemes

kiszűrni az antecedensjelöltek közül például azokat, amelyek nem felelnek meg az egyeztetési szabályoknak, mert lehetséges, hogy a szövegalkotó hibásan egyeztet.

Abban az esetben, ha a nyelvtani szabályokra és a *helyes* visszautalásokra megállapított szintaktikai vagy szemantikai szabályokra nem támaszkodhatunk a szövegtípus vagy maga a célnyelv felszíni szerkezete miatt, a funkcionális és kognitív nyelvészet mentális állapotra és figyelmi állapotra tett megállapításai mérvadók lehetnek. Éppen ezért érdemes megvizsgálni, hogy az egyes anaforafeloldással kapcsolatos mentális állapotra vonatkozó megállapítások hogyan és mekkora pontossággal implementálhatók számítógépes környezetbe, és hogy milyen hatással vannak az automatikus anaforafeloldás eredményességére.

A dolgozat elején meghatározom a kutatás célkitűzéseit, és az előzetes hipotéziseimet a gépi tanulási kísérletekkel kapcsolatban. Ezután egy szakirodalmi áttekintésen keresztül megvizsgálom a koreferencia és az anafora meghatározási lehetőségeit a nyelvészeti és a számítógépes nyelvészeti szakirodalomban, ennek során kitérek az automatikus feloldási lehetőségekre is. A dolgozat második felében először ismertetem az általam felhasznált korpuszok felépítését és a kísérletek végrehajtásához szükséges megelőző lépéseket, majd részletes bemutatom a tanulási kísérletek során felhasznált jellemzőket. Végül az általam elvégzett gépi tanulási kísérletek részletezése és az eredményekből származó konklúzió levonása található.

2. Célkitűzések

A disszertáció célja gépi tanulási kísérleteken keresztül megvizsgálni a jelenleg bevett, automatikus anaforafeloldást célzó statisztikai alapú felügyelt gépi tanulási kísérleti módszerek eredményeit a névmási anaforafeloldás tekintetében a magyar nyelvben, ezen belül is nagy hangsúlyt fektetve a tanulás alapjául szolgáló jellemzőkészlet összeállításának lehetőségeire. A kísérletekkel kapcsolatos programok elérhetők az alábbi linken: https://github.com/viktoria-kovacs/MentionPair. A nullhipotézisem az, hogy lehetséges az automatikus névmási anaforafeloldás a magyar nyelvben szemantikai információk nélkül is, pusztán morfológiai, szintaktikai és egyéb, a felszíni szerkezetből kinyerhető, kognitív nyelvészeti alapú jellemzők segítségével. Ennek bizonyításához több kísérletet végeztem el, amelyekben nem vettem figyelembe szemantikai információkat. A kísérletekhez először meg kell vizsgálni a nyelvészeti és számítógépes nyelvészeti szakirodalom jelenlegi álláspontját a névmási anafora definíciójáról és a névmáshoz tartozó antecedens azonosítási lehetőségeiről, valamint annak nehézségeiről. Ezután a két, névmási visszautalások tekintetében manuálisan annotált magyar nyelvű korpuszt, a Szeged Korpusz (Csendes–Csirik–Gyimóthy 2004; Csendes et al. 2005; Vincze et al. 2010) koreferenciaannotált alkorpuszát, a SzegedKoref korpuszt (Vincze et al. 2018) és a KorKorpuszt (Vadász 2020) kell megvizsgálni, hogy azonos típusú információk legyen kinyerhetők belőlük, majd meg kell határozni a gépi tanulási kísérletek során milyen és mennyi pozitív és negatív példát, illetve milyen jellemzőket veszek figyelembe.

A tanító és tesztelő fájlok felépítésének szempontjából a Mention-pair technikát (Aone–Benett 1995) alkalmazom minden esetben. A tesztelés során a tesztfájlokban megtalálható az adott korpuszrészletben előforduló összes névmás, és hozzá párként hozzárendelve az összes névmást megelőző főnévi csoport, mint lehetséges antecedensjelölt.

A gépi tanulás célja, hogy az épített modell felismerjen legalább egy antecedenst, amellyel a visszautaló névmás anaforikus kapcsolatban áll. Az anaforafeloldás tekintetében az egyik probléma a negatív és pozitív tanítópéldák kiegyensúlyozatlan eloszlása, ennek a problémának a kiküszöbölésére több módszer is létezik. A kísérletek alapjául szolgáló módszerből kifolyólag a tanulás során a negatív példák olyan szópárok lesznek, amelyek nem állnak anaforikus kapcsolatban, pozitív példák pedig olyan szópárok lesznek, amelyek anaforikus kapcsolatban állnak (kézzel annotált esetek a korpuszokban). Ennek következtében egy szövegből lényegesen több negatív, mint pozitív példa állítható elő, ami befolyásolja annak a valószínűségét, hogy az

osztályozó mennyire sikeresen ismeri fel a két csoport (anaforikus, nem anaforikus) tagjait. Az első hipotézisem szerint azok a modellek érik el a legjobb eredményeket a tesztelés során, amelyekben a pozitív és negatív példák eloszlása azonos a tesztfájlokban várható pozitív és negatív esetek eloszlásával, tehát sem a pozitív, sem a negatív példák számát nem csökkentjük a tanítófájlokban manuálisan. Fontos kiemelni ebben az esetben, hogy az általam összehasonlított, a tanítófájlokban megtalálható pozitív és negatív párok megoszlására vonatkozó módszerek pusztán elméleti jellegű kísérletek, hiszen egy valós, számítógépes nyelvészeti alkalmazás esetében nem határozható meg előre milyen lesz az adott szöveg, amelyen a feladatot el kell végezni, lehet akár egy egész regény vagy épp csak egy mondat, így a bennük található pozitív és negatív párok arányára sem lehet előjelzést tenni.

Mivel a kísérletek célja a névmáshoz tartozó egyetlen antecedens kiválasztása, a modell azonban névmás-antecedensjelölt párokat osztályoz, a második hipotézis arra vonatkozik, hogyan választhatunk a pozitívnak ítélt párok segítségével egyetlen antecedensjelöltet. Két módszert hasonlítok össze a tanulási kísérletek során, ezek a Best-first (Ng–Cardie 2002a) és a Closest-first (Soon–Ng–Lim 2001) módszerek. A Best-first módszer az osztályozó által legmagasabb valószínűségi értékkel ellátott névmás-antecedens párt jelöli meg a névmás antecedensének, a Closest-first módszer a szövegben a névmáshoz legközelebb eső, az osztályozó által antecedensnek ítélt főnévi csoportot jelöli a névmás antecedensének. A második hipotézisem szerint a legnagyobb valószínűségi értékkel ellátott névmás-antecedens pár kiválasztása nagyobb hatékonyságot eredményez, hiszen a névmások gyakran utalnak a szövegben messzebbre, így pusztán a lehetséges antecedensjelölt közelségének figyelembe vétele fals pozitív eredményt okozhat.

A kísérletek harmadik szempontja a kifejezéspárokhoz rendelhető jellemzők vizsgálata. A disszertáció célja megvizsgálni, hogy kizárólag morfológiai és szintaktikai jellemzők, valamint egyéb felszíni szerkezetből kinyerhető, kognitív alapú jellemzők segítségével is lehetséges automatikus névmási anaforafeloldást végezni a magyar nyelvben. A kísérletek során a harmadik kutatási kérdésem, hogy ezek közül a jellemzők közül melyek a leghatékonyabbak a modellépítés szempontjából. Először megvizsgálom, hogy a két kifejezés közötti távolság kiszámítására milyen lehetőségek merülnek fel, és ezek közül melyik a legeredményesebb, másrészt a korpuszból kinyerhető morfológiai és szintaktikai jellemzőket négy kognitív alapon megfogalmazott jellemzőcsomaggal egészítem ki egyesével, hogy megvizsgáljam, az általam megfogalmazott jellemzők hogyan módosítják a modellépítés sikerességét. A harmadik

hipotézisem, hogy a tanulási kísérlethez hozzáadott nem nyelvi jellemzők javítanak a modellépítés sikerességén.

A gépi tanulási kísérleteket külön végzem el az egyes névmási visszautalási típusok tekintetében (személyes névmás, mutató névmás, vonatkozó névmás), feltételezve, hogy egymástól eltérően viselkedhetnek a fent említett hipotézisek szempontjából. Az egyes névmástípusokkal elvégzett kísérletek végén megvizsgálom, hogy melyik a legsikeresebb módszer a modellépítés szempontjából, mind a pozitív és negatív példák aránya, mind az alkalmazott jellemzők szempontját figyelembe véve.

3. A névmási anafora különböző értelmezési lehetőségei a nyelvészeti és számítógépes nyelvészeti szakirodalomban

A névmási anafora fogalma az anafora kérdésköréhez tartozik, ezért elsődlegesen a különböző területek anaforadefinícióját tartom szem előtt, ami a nyelvészeti szakirodalomban merőben eltér a számítógépes szakirodalom meghatározásától, de még a nyelvészeti irányzatokon belül is számos eltérés tapasztalható. A következő fejezetben a különböző irányzatok anaforameghatározásait veszem sorra és hasonlítom össze, először a nyelvészeti szakirodalomban fennálló eltérésekre koncentrálva. A bemutatás során figyelembe veszem, hogy a legsarkalatosabb eltérés ebben a kérdéskörben is a szövegalkotó figyelembe vétele vagy figyelmen kívül hagyása a szöveg vizsgálata során, ezért a mondatértelmezés kérdéskörétől közelítem meg a probléma bemutatását. Mivel az anafora meghatározása a számítógépes nyelvészeti szakirodalomban szorosan kapcsolódik a koreferenciához, ezért a koreferencia definíciójára is kitérek az egyes megközelítések vizsgálata során. A különböző nyelvészeti megközelítési módok összehasonlítása után kitérek a számítógépes nyelvészet anaforadefiníciójára, és annak eltéréseire a nyelvészeti szakirodalom anaforadefinícióitól.

3.1. A mondatértelmezésben megjelenő különbségek a nyelvészeti szakirodalomban

A névmási anafora értelmezésének tekintetében két élesen elkülönülő csoportba sorolhatók a nyelvészeti irányzatok. Az elsőbe a generatív, strukturalista és formális nyelvészeti elméletek, a másodikba pedig a funkcionalista, kognitív-funkcionalista és kognitív nyelvészeti irányzatok. Ez nem azt jelenti, hogy két módja lehetséges az anaforák vizsgálatának, pusztán csak azt, hogy a két csoport között éles eltérés figyelhető meg a nyelvtanfelfogás és a mondatértelmezés tekintetében, ami maga után vonja az anaforaértelmezésében mutatkozó alapvető különbségeket is. A következőkben ezért két-két nagyobb fejezetre bontva fogom ismertetni a nyelvtan megközelítésének két fő módját, majd a referencia és az anafora megközelítési módjának lehetőségeit is.

3.1.1. A generatív, strukturalista és formális irányzatok

Az első csoport irányzatai autonóm rendszerként kezelik a nyelvtant: elkülönítik egymástól a nyelvi kompetenciát és a performanciát, tehát azt a nyelvtudást, amelyre elvben képes egy

anyanyelvi beszélő, a nyelvtudás adott szituációban való felhasználásától. A nyelv leírása során kizárólag a kompetenciát veszik figyelembe, így azt grammatikus mondatok összességeként fogják fel. A nyelvtan egy sor szabályból áll, amelyek fonológiai, szintaktikai és szemantikai komponensekből tevődnek össze, a szabályok segítségével pedig minden grammatikus mondat leírható a nyelvben (Bagha 2009).

3.1.2. A funkcionalista és kognitív nyelvészeti megközelítések

A generativista nyelvfelfogással szemben a funkcionalista és kognitív irányzatok használatalapú modellekkel dolgoznak (Barlow–Kemmer 2000), azaz a beszélő vagy szövegalkotó nézőpontjából közelítik meg a nyelvet és működését, és nem különítik el egymástól a kompetenciát a performanciától. A kognitív nyelvtan szerint a nyelv nem szabályok, hanem analógiák alapján működik, és a grammatikai struktúra három egység segítségével írható le: szemantikai, fonológiai és szimbolikus (Langacker 1987). A jelentés (szemantikai) egy hangalakkal (fonológiai) együtt alkot egy szimbolikus egységet, amely lehet egy morféma, de akár egy teljes mondat is. Konvencionális szimbolizációs struktúraként van jelen a szintaxis, azaz a ténylegesen előforduló szerkezetekből alkot a nyelvhasználó mintázatokat. Ezeket a sémákat használja fel új szerkezetek létrehozásakor, különálló szintaktikai kategóriák nincsenek jelen a nyelvtanban. Tehát a nyelvtant nem egy formális rendszernek tekinti, amely a jelentéstől elkülönülten működik (Langacker 1986).

A funkcionalista irányzatok nyelvfelfogásában a szöveg mondatok által realizálódik, de nem azokból áll (Halliday–Hasan 1976). Éppen ezért a mondatleírás során három szintet különítenek el: formális szinten az összetevős elemzést, funkcionális szinten a tematikus szerepekkel is azonosítható szerepeket, és a téma-réma szerkezetet(Tolcsvai Nagy 2000). A funkcionalista irányzatok nem csak a grammatikai szerkezetet és a formális szerkezetet veszik figyelembe, hanem az elemzés során fontos a kommunikációs szituáció, a kontextus, valamint a résztvevők is, hiszen ezek fogják meghatározni a megnyilatkozás tartalmát és formáját (Newmeyer 2000).

3.2. A névmási anafora értelmezésében megjelenő különbségek

A nyelvfelfogásban és a mondatértelmezésben megjelenő különbségek maguk után vonják, hogy a különböző irányzatok eltérő módon értelmezzék és magyarázzák a referencia, a koreferencia, ezáltal az anafora és ezen belül a névmási anafora szerepét is. Az anafora kifejezés leggyakrabban visszautalást, a koreferencia pedig együtt utalást jelent, ezért ahhoz, hogy a

különböző irányzatok visszautalásról alkotott képe összevethető legyen, először azt kell megvizsgálni, mit jelent az utalás a szakirodalomban, ezután pedig azt, hogy mi a különbség a visszautalás és az együtt utalás között. A nyelvtanfelfogásból kiindulva két élesen elkülönülő csoportba oszthatók a nyelvészeti irányzatok az anaforaértelmezés tekintetében, ennek megfelelően a következőkben két különálló fejezetben ismertetem a referenciával, koreferenciával és az anafora fogalmával kapcsolatos főbb megközelítési lehetőségeket.

3.2.1. A névmási anafora és a formális, generatív és strukturális megközelítések

A generatív irányzatok a nyelvleírás során kizárólag a nyelvi kompetenciát veszik figyelembe. A szintaktikai és grammatikai ítéletek mellett az irányzat már nem foglalkozik azzal, hogy az adott mondatnak vagy kifejezésnek mi a célja vagy hogyan illeszkedik a kontextusba, kizárólag azt vizsgálja, hogy az adott mondat grammatikus-e vagy sem. A nyelvhasználatot, a kifejezések kiválasztásának okait és módját már a performancia kérdéskörébe sorolja. Éppen ezért a generatív iskola a referencialitással és a koreferencia vizsgálatával sem foglalkozik, az anaforaértelmezése pedig erőteljesen eltér a funkcionális és kognitív irányzatok megközelítési módszereitől.

A generatív iskola a referenciáról annyi megállapítást tesz, hogy a különböző kifejezések különböző referenciális tulajdonságokkal rendelkeznek. Az úgynevezett R-kifejezéseknek, mint például a tulajdonneveknek, határozott leírásoknak (*Péter, Szeged, a fiú, a megyeszékhely*), az adott kontextusban rögzített a referenciájuk. Míg a neveket merev jelölőnek tartja, amelyeknek minden kontextusban azonos a referenciája, addig a határozott leírások referenciája kontextusról kontextusra változhat, de azon belül nem. A második kategóriába a névmások kerülnek, amelyek abban különböznek az előző kifejezésektől, hogy a referenciájuk a kontextuson belül is változhat (*ő, őt*). A harmadik kategóriába a kölcsönös és visszaható névmások kerülnek (*egymást, magukat*), melyeket a generatív szakirodalom egyszerűen anaforáknak vagy anaforikus névmásoknak nevez. A különbség az anafora és a névmás között a generatív szakirodalomban az antecedens megléte és helye. Az anaforának kötve kell lennie az antecedense által, a névmásnak nem (Hicks 2009), ez a kötés pedig automatikusnak tekinthető, a nyelvészeti vizsgálatok során adottnak tekinthetjük. Az, hogy az anafora kötve van, azt jelenti, hogy kötelező egy bizonyos lokális tartományon belül állnia a névmás antecedensének; ezeknek a kifejezéseknek anaforikus referenciájuk van, ezért ezek anaforák. Ezért a generatív irányzatok a mondatok nyelvtani viszonyaiból vezetik le az anaforát, a szöveggrammatika egyik lényegi elemének tartják. Az anafora a kohézió fontos elemévé válik azáltal, hogy az azonos utalás révén összefüggővé teszi a

szöveget, ennek megfelelően az anaforát relációként értelmezi. A generatív iskolán belül az anafora értelmezése a kötés definíciója alapján különbözhet. A magyar nyelvvel kapcsolatban például É. Kiss Katalin megállapítja, hogy a birtokos szerkezet egy különálló kötési tartományt alkot (É. Kiss 1987).

Chomsky transzformációs nyelvtanában az anafora értelmezése jelentősen leszűkült és kizárólag azokat a névmásokat tekinti anaforának, amelyek az antecedensük által kötelezően kötve vannak lokális mondattani tartományon belül, tehát a visszaható és kölcsönös névmásokat (*magát, egymást*). A nyelvtanában ezért foglalkozik az antecedens helyének kijelölésével is. A kötési elvek azt határozzák meg, hogy az egyes főnévi csoportok adott szintaktikai pozícióban kötelezően koreferensek, opcionálisan koreferensek, vagy nem lehetnek koreferensek (Chomsky 1981). Chomsky a következő szabályt fogalmazza meg: *(1) α akkor köti β-t, ha α és β azonos indexet visel, és α k-vezérli β-t.*

A kötési elvek a fent is meghatározott három kategória mentén fogalmazhatók meg. Az önállóan referáló kifejezésekre (nevek, határozott leírások) vonatkozó kötési elv kimondja, hogy a referáló kifejezések nem lehetnek kötve, se kormányzó kategórián belül, sem azon kívül (Chomsky 1981).

Azok a kifejezések, amelyek nem önállóan referálnak, az anaforák és a névmások. Az anaforákra vonatkozó kötési elv kimondja, hogy egy anaforát kormányzó kategórián belül kötni kell (Chomsky 1981), tehát az 1) példában a *magát* névmás egyértelműen Péterre utal, nem utalhat Marira.

1) Mari$_i$ azt gondolja, hogy Péter$_j$ meglátta magát$_{j/*i}$ a szemben lévő tükörben.

A névmások kategóriájába pedig az anaforikus névmásokon kívüli névmások sorolhatók. A névmásokra vonatkozó kötési elv kimondja, hogy a névmásoknak kormányzó kategórián belül szabadnak kell lenniük (Chomsky 1981), tehát a 2) példában az *őt* névmás nem utalhat Péterre, csak Marira.

2) Mari$_i$ azt gondolja, hogy Péter$_j$ meglátta őt$_{*j/i}$ a szemben lévő tükörben.

Az anafora definíciója tehát ebben a megközelítésben kimerül annyiban, hogy az anafora kötve kell legyen egy antecedens által, de nem cél a nyelv leírása során az antecedens kiválasztási módjainak magyarázata. Így a 3) és 4) példákban szereplő jelenségekre sem keres egyéb magyarázatot, azon felül, hogy mind a két lehetőség grammatikus.

3) Mari$_i$ azt gondolja, hogy kiválasztják őt$_i$ a következő sorsoláson.
4) Mari$_i$ azt gondolja, hogy kiválasztják őt$_j$ a következő sorsoláson.

A strukturalista és formális megközelítések hasonlóan foglalnak állást a referencia és koreferencia tekintetében. A nyelvi jelentést a nyelven belüli viszonyok összességéből számítják ki (Kiefer 2007), tehát szintén nem veszik figyelembe a nyelven kívüli tényezőket, mint a kontextus vagy a beszélő személye. Ezért szintén csak azokat a névmásokat tekintik anaforának, amelyek antecedense a nyelvtani viszonyok alapján kizárólagos és adott, az anafora vizsgálata pedig az anafora és antecedense jelentésviszonyaira szorítkozik. Így azonban bármely névmás tekinthető anaforának, amelynek adott az antecedense. Wasow két elemet akkor tekint anaforikusnak egy szövegben, ha a második elem értelmezéséhez az első elem teljes jelentésének ismerete szükséges, a második elem az első jelentésének egy részét ismétli meg. Ennek megfelelően anafora alatt egy szövegben előforduló kifejezéspár bináris relációja érthető (Wasow 1979). Huang értelmezésében már nem feltétlenül párról van szó, az ő definíciója alapján anaforikus kapcsolat két vagy több nyelvi elem között áll fenn, ahol az egyik elem interpretációja egy másik elem függvénye (Huang 2000).

A fent említett megközelítési módszerek tehát a nyelvet a forma szempontjából vizsgálják, nem céljuk a referenciáról összetettebb következtetéseket levonni, kizárólag grammatikalitási ítéleteket hozni szabályok segítségével a felszíni- és egyes esetekben a mélyszerkezet alapján, éppen ezért a kutatásom során sem használom fel ezeknek a területeknek az eredményeit.

3.2.2. A funkcionális és kognitív irányzatok értelmezése

Míg a generatív iskola irányzatai inkább azokra a szerkezeti jellemzőkre összpontosítanak, amelyek diszreferenciát mutathatják, addig a funkcionalista irányzatok azokat a különböző nyelvi szintekhez köthető jellemzőket keresik, amelyek egy referáló kifejezés értelmezéséhez segítenek hozzá. Ehhez felhasználják a nyelv strukturális leírását, de kiindulópontnak szerkezet helyett a jelentést tekintik. A funkcionális szemléletű nyelvészeti irányzatokban a nyelv és elemeinek funkció szempontjából történő vizsgálata valósul meg. A vizsgálat alapját tehát a forma vagy szerkezet helyett a jelentés vagy tartalom adja. Mivel a nyelv fő funkciójának a jelentésközvetítést tekintik, ezért nem a kifejezések jelentését keresik, hanem egy jelentés kifejezési módjait.

A referencialitással kapcsolatban ezeknek az irányzatoknak az a fő kérdése, hogy a beszélő vagy szövegalkotó hogyan választ referáló kifejezést, milyen formát választ az átadni kívánt

információ tekintetében. A választást pedig a beszélő birtokában lévő információk, valamint a hallgatóról kialakított képe fogja meghatározni. Ezekben a megközelítési módszerekben már az a kérdés is felmerülhet a kifejezések vizsgálata során, hogy ugyanazok a tényezők irányítják-e a produkciót, mint az értelmezést. Ebben a kérdésben eltérő vélemények olvashatók a szakirodalomban, az azonban egyértelműnek tűnik, hogy a kifejezések megválasztása nem lehet véletlenszerű, hiszen a beszélő vagy szövegalkotó célja az, hogy a címzett értelmezni tudja az adott információt.

A referencia fogalma a funkcionális nyelvészeti irányzatokban eltér a logikai pozitivista tradíciótól, ahol a referáló nyelvi kifejezés és a való világ entitásai közötti kapcsoló elemet jelenti.

Halliday szisztémikus funkcionális nyelvtanában (Halliday 1994) a jelentés vizsgálata során megállapítja, hogy három metafunkciót tölthet be: ideációs vagy reflektív (célja a környezet megértése), interperszonális vagy aktív (célja a környezetben található másokon végrehajtott cselekvés), textuális (célja a nyelvi kifejezések szövegbeli elhelyezése) (Halliday 1994; Tolcsvai Nagy 2005).

Halliday a referenciának két lehetséges értelmezését különíti el annak tekintetében, hogy hogyan utalunk az elemekre, amelyek részesei egy szituációnak. A szövegen kívülre, tehát a való világ egy elemére történő utalást exoforikus referenciának nevezi. Ilyen referenciával rendelkező kifejezések tipikusan az egyes szám első és második személyű személyes névmások, amelyek a mindenkori beszélőre és hallgatóra utalnak (*én*, *te*), vagy a helyet és időt meghatározó kifejezések (*itt, most*) (Halliday 1994; Halliday–Matthiessen 2004). A szövegen belül, a szöveg egy részére történő utalást nevezi anaforikus referenciának. Az anaforikus referencia azáltal válik a szöveg kohéziójának egyik fő elemévé, hogy a sorozatos visszautalás során láncot alkot a szövegben. Az anaforikus referenciával rendelkező kifejezéseket három kategóriába sorolja: személyre referáló elem, demonstratív referáló elem, összehasonlító referáló elem. A személyre referáló elemek közé a személyekre utaló kifejezések tartoznak pl. *ő*. A demonstratívok közé a mutató névmások tartoznak: *ez, az*. Az összehasonlító referencia kategória alá pedig azok a kifejezések tartoznak, amelyek nem koreferensek az antecedenssel, keretet teremtenek az értelmezéshez pl. *ugyanolyan, hasonló*.

Halliday nyelvtanában különös hangsúlyt kap a diskurzus információs struktúrájának és a nyelvtannak a viszonya. A diskurzus funkciójának vizsgálata során elkülöníti egymástól a pszichológiai alanyt (amire az üzenet vonatkozik), a nyelvtani alanyt (amiről valami állítódik) és a logikai alanyt (a cselekvés végrehajtója) (Halliday 1994: 31–32). Ebben a megközelítésben a

pszichológiai alany a téma (Theme), a nyelvtani alany az alany (Subject), a logikai alany a cselekvő (Actor).

Szoros kapcsolat van az információs szerkezet (adott- új) és a téma-réma szerkezet között. Egy klóz egy információs egység, így az adott-új skálán a téma az adott kategóriával, a réma pedig az új kategóriával van összhangban, azonban a kettő nem azonos. A téma- réma szerkezet beszélő központú, tehát a téma az, amit a beszélő választ. Az adott-új rendszer azonban hallgató központú, tehát az adott az az információ, ami a hallgató birtokában van. Az anaforikus elemek ebben a keretben azáltal interpretálhatók, hogy referálnak valamilyen a szövegben korábban említett dologra vagy a szituációra. Az anafora tehát a visszautalása révén lesz adott, a kifejezésnek nem szükséges további információt tartalmaznia, mivel interperszonálisan meghatározott a beszédszituációban (Halliday–Matthiessen 2004).

Dik értelmezésében az anafora olyan nyelvi elem, amely a korábbi szövegrészben közvetve vagy közvetlenül létrehozott (established) entitásra utal. Az a kifejezés az antecedens, amelynek a segítségével létrejött az entitás. Ha az antecedens és az anafora ugyanarra az entitásra utal, akkor koreferensek, de ez nem minden esetben van így. Például, ha a visszautalás egy halmaz egy elemére történik 5), vagy maga az antecedens nem referáló kifejezés 6).

5) Minden gyerek szereti a(z ő saját) játékát
6) A fal fehér volt és Mari utálta ezt a színt

Dik megközelítésében nem a kifejezések referálnak, hanem a beszélő, aki a kifejezést használja (Lyons 1977). Nem a megelőző szövegrészre utal vissza a beszélő, hanem a megelőző szövegrész által létrehozott entitásra. Dik is kiemeli, hogy az anaforikus visszautalások által láncok jönnek létre a szövegekben. Ezek a láncok hívhatók *"topik láncoknak*" (Dixon 1972), "*identification spans*"-nak (Grimes 2015), vagy "*anaforikus láncoknak*" (Chastain 1975).

Givón funkcionális nyelvtana (Givón 1983) nyelvtipológiai megközelítésű és erőteljesebben épít a pszicholingvisztikára és kognitív pszichológiára, ezt mutatja az is, hogy Givón (Givón 1993) szerint a nyelv legfontosabb funkciója a mentális reprezentáció és ezeknek a reprezentációknak a kommunikációja.

Givón megközelítésében a referencia nem a való világ dolgaira vonatkozik, de nem is a szöveg egy korábbi részére, hanem az úgynevezett diskurzusuniverzumra (*Universe of Discourse)*. A diskurzusuniverzumokat beszélők alkotják meg azáltal, hogy szándékosan hoznak benne létre entitásokat, amelyekre később vagy referálnak vagy nem. A beszélő referálási szándéka az, ami releváns a referencia nyelvtanának vizsgálata során. Amikor nem referál a

kifejezés, hanem jellemez, akkor beszélhetünk a Donellani értelemben vett attributív használatról (Donnellan 1966). Éppen ezért nem is különíti el egymástól élesen a referáló és nem referáló kifejezéseket, hanem skálaként képzeli el magát a referálásra való alkalmasságot, amelyen a kifejezések az alapján helyezkednek el, hogy mennyire valószínű, hogy a beszélő szándéka az általuk való referálás.

Givón nyelvtanában megjelenik és összekapcsolódik a határozottság (*definiteness*) és elérhetőség (*accesibility*) fogalma. Azokat a kifejezéseket nevezi határozottnak, amelyek azonosíthatók, azaz elérhetőek a hallgató tudatában a diskurzus egy adott pontján. Az, hogy elérhető, azt is jelenti, hogy már létezik a hallgató mentális állapotában, azaz visszakereshető onnan. Amikor határozott kifejezéssel utal vissza egy beszélő, akkor számos már korábban a beszélő által létrehozott mentális modellre alapozhat. Egyrészt alapozhat olyan információkra, amelyek egyaránt elérhetők a beszélő és a hallgató számára: Givón ezt a mindenkori beszédhelyzet mentális modelljének (*speech-situation*) nevezi, például ennek segítségével értelmezhető az egyes szám első és második személyű személyes névmás. Alapozható a visszautalás az általános lexikai tudás mentális modelljére (*permanent generic-lexical knowledge*), amennyiben a beszélő tisztában van a hallgató lexikai tudásával, illetve alapozható a visszautalás az adott szöveg mentális modelljére is (*the current text*). Utóbbi kettő általában egyszerre érvényesül, azaz a visszautalás duplán megalapozott (Givón 2001).

Givón megközelítésében a diskurzus folytonosságának három szintje van, a tematikus, cselekvés-, és szereplő- vagy topikfolytonosság. Tematikus folytonosság alatt a diskurzus szerkezetét érti, cselekvésfolytonosság alatt pedig a predikátumok koherenciáját (idő, ok-okozat). Topik alatt a diskurzus központi témája értendő. Az anafora ebben az értelmezési keretben a szereplő- vagy topikfolytonosság eszköze. A különböző anaforikus kifejezéseket hierarchiába szervezte az alapján, hogy mennyire erős topikfolytonosságot feltételeznek. A hierarchia alapelve az volt, hogy minél gyengébb a topikfolytonosság, annál több információra van szükség ahhoz, hogy koherens maradjon a diskurzus. A névmási anafora ebben az értelmezésben azt feltételezi, hogy erős a topikfolytonosság, hiszen minél nagyobb a távolság az anafora és az antecedense között, illetve minél több közbeékelt visszautalás történik, annál nehezebb azonosítani az anaforához tartozó antecedenst, ezáltal pedig gyengül a topikfolytonosság is (Givón 1983).

Tehát, egységesen elmondható a funkcionális irányzatok referenciaértelmezéséről, hogy merőben eltér a logikai pozitivista irányzatoktól. Míg Halliday és Givón esetében a nyelvi kifejezéseknek tulajdonítható a referálásra való képesség, addig Dik, átvéve Lyons

megközelítését, a referálás képességét a szövegalkotónak, és nem a kifejezésnek tulajdonítja. Az anafora Halliday esetében a szöveg egy korábbi részére utal, Dik esetében a szöveg egy korábbi része által létrehozott entitásra, Givón esetében pedig a diskurzusuniverzum egy entitására, de egyik esetben sem a való világ elemeire. A névmási anafora szerepe mindegyik megközelítésben az, hogy az ismétlés által frissíti az antecedens referensét, ami hozzájárul a diskurzus dinamikusságához. Tehát nem csupán kohéziós elem, hanem a diskurzus központi témáinak ismétlésére szolgáló, azokat a memóriában aktívvá tevő elemek, a címzett figyelmének irányítására szolgáló nyelvi elem (Ehlich 1982). Mivel a névmási anafora ebben az esetben alacsony szemantikai tartalmú nyelvi elem, amelynek tartalmi feltöltését a szövegelőzmény biztosítja, tehát a szövegvilágon belüli visszautalás, ezért a névmás és az antecedense nem feltétlenül koreferens egymással. A visszautalás alapja lehet: ismétlés, alá-fölérendelés, szinonima, zéró anafora, névmási anafora, epiteton, valószínű rész, szükségszerű rész, esetkeret (Pléh 1994).

A kognitív nyelvtanokban az anafora értelmezése kizárólag a szemantikai és pragmatikai alapelveken alapul. Egy kifejezés jelentése a beszélő elméjében aktiválódott konceptualizáció, ez különböző tudásrendszereket aktivál, amelyekkel a kialakult szituáció egyes aspektusai értelmezhetők. A konceptuális referenciapontok mutatják meg a kontextust, amelyben a koreferencia elfogadható és, amelyben nem, ezek rendszerezése főként szemantikai alapú. (Langacker 1987; van Hoek 1995).

3.3. Névmási anafora a számítógépes nyelvészet szakirodalomban

A névmási anaforafeloldás a számítógépes nyelvészeti gyakorlatban is az anaforafeloldás része, azonban gyakran (Desislava 2013) tekintenek rá a koreferenciafeloldás részfeladataként is, ezért az automatikus feloldást célzó rendszerek többsége nem pusztán a névmáshoz tartozó antecedens felismerésére összpontosít.

Azt, hogy az anaforafeloldás kizárólag a számítógépes szakirodalomban lehet része a koreferenciafeloldásnak, különösen fontos kiemelni. Egyrészt a terminológiai félreértések elkerülése miatt, hiszen sok automatikus feloldást célzó rendszer nyelvészeti megfontolásokon alapul, annak ellenére, hogy a nyelvészeti modellek által kitűzött célok nem fedik pontosan az automatizálást célzó adott rendszer céljait. Másrészt pedig fontos tisztában lenni az adott rendszer céljával a kiértékelése során, hiszen az ellenőrzésre használt korpusznak azonos megfontolások mentén ellátott annotációval kell rendelkeznie. Éppen ezért igen gyakori, hogy egy-egy rendszer dokumentációja nem kizárólag a rendszer leírását tartalmazza, hanem a feladat

pontos meghatározását is példákkal is szemléltetve. Az, hogy az anaforafeloldás több módon értelmezhető, főleg a kiértékelés és a rendszerek összehasonlítása során okoz gondot, de befolyással bír az adott rendszer további felhasználási lehetőségeire is.

A fő oka, hogy a nyelvészeti szakirodalomban nem része az anaforafeloldás a koreferenciafeloldásnak az a koreferencia mint reláció tulajdonságaival magyarázható. A nyelvészeti szakirodalomban két kifejezés akkor koreferens, ha azonos dologra referálnak, tehát a reláció szimmetrikus, tranzitív és reflexív. Az anaforának már magában a nyelvészeti szakirodalomban is tágabb az értelmezési köre. Az előző fejezetben ismertetett elméleti keretek némelyike a szöveg egy korábbi részére való utalásként értelmezte, más a szöveg egy korábbi része által létrehozott entitásra való utalásként vagy a diskurzusuniverzum egy entitására való utalásként is. Általánosságban azonban elmondható, hogy a visszautaló szavakat tekinthetjük anaforikusnak, azokat a kifejezéseket, amelyek koherenciát teremtenek a visszautalás által. A nyelvészeti szakirodalom ennek a visszautalási folyamatnak a szabályszerűségeit és körülményeit vizsgálja. Az anafora értelmezéséhez a nyelvészeti szakirodalomban az antecedens mellett a lexikon, a világtudás és a szövegkörnyezet is segítségül hívható. Az ilyen típusú visszautalás nem kizárólag akkor tekinthető sikeresnek, ha két kifejezés referense azonos, sőt tágabb értelemben véve az sem feltétele a visszautalásnak, hogy az adott kifejezés referáló legyen.

A számítógépes nyelvészeti szakirodalomban ettől eltérő módon az anaforikus kifejezés értelmezése kizárólag az antecedens által történhet, hiszen a szöveg feldolgozása során, kizárólag az anaforát megelőző szövegrészlet áll a rendelkezésre. Éppen ezért nem is a visszautalás tényén van a hangsúly, hanem az adott kifejezés értelmezéséhez szükséges korábbi információk azonosításán.

A számítógépes szakirodalomban nem csak az anafora, hanem a koreferencia fogalma is eltér a nyelvészeti szakirodalom definíciójától, ezért leggyakrabban kerülik a koreferencia kifejezést (van Deemter–Kibble 2000; Sidner 1979). Koreferenciafeloldás során a rendszer azokat a kifejezéseket keresi, amelyek tartalomkinyerés szempontjából összefüggők, ugyanarra az entitásra vonatkozó információt hordoznak. Szemléletes példa erre a MUC feladathoz (Message Understanding Conference - a feladat célja az információkinyerés hatékonyságának a növelése) megfogalmazott irányelvekben a *John is a fool* példa, amely alapján *John* és az *a fool* kifejezések jelölendők mint összetartozó kifejezések (van Deemter–Kibble 2000). Ebben az esetben nyelvészeti értelemben vett koreferenciáról nem beszélhetünk, hiszen az *a fool* kifejezés nem referál, így nem lehet azonos a két kifejezés referense. Ebből azonban arra lehet

következtetni, hogy a számítógépes nyelvészeti terminológiát figyelembe véve sem szerencsés az anaforafeloldást a koreferenciafeloldás részfeladatának tekinteni (Mitkov 2001), hiszen az anaforikus elemeknek kizárólag az értelmezéséhez használjuk az antecedenst. Ezért gyakran nem is antecedensnek, hanem horgonynak nevezi a szakirodalom (Schwarz 2000; Kocsány 2018), mert nem feltétlenül azonos entitásra vonatkoznak.

A számítógépes nyelvészeti értelemben vett anafora, mivel nem kötelező, hogy azonos objektumra vonatkozzon, mint a horgony, az azonos referencia mellett négyféle további kapcsolatban állhat a horgonnyal. Az első csoportba azok a kifejezések kerülnek, amelyek különböző objektumra utalnak, mint a horgony, de ugyanabból a típusból egy elemre. Ezek általában a határozatlan névmások, amelyek úgynevezett *identity of sense* kapcsolatban állnak a horgonnyal (Garnham 2001). A 7) példában a *zöld sálad* és az *egyet* kifejezés nem ugyanarra az objektumra utalnak, az *egyet* kifejezés értelmezéséhez segít hozzá a *zöld sálad* kifejezés, tehát a második tagmondat ugyanebből az objektumtípusból utal egy másik elemre.

7) Nagyon tetszik a zöld sálad, már régóta akarok egyet én is.

A második csoportba az úgynevezett *paycheck* névmások (Karttunen 1969) tartoznak, amelyek hasonló szerepet töltenek be, mint az első csoport, de határozott névmás segítségével azonosítják a csoport vagy kategória egy elemét. A 8) példában sem azonos a *pénztárcáját* és az *azt* zéró névmás referense, ebben az esetben a *pénztárcáját* kifejezés szintén a második tagmondat zéró névmásának értelmezésben segít, amely egy azonos kategóriájú elemre utal, nevezetesen annak a lánynak a pénztárcájára, aki a zsebében tartja azt.

8) A lány, aki a pénztárcáját a táskájában tartja, bölcsebb, mint aki a zsebében tartja (azt).

A harmadik csoportba az úgynevezett *bound* anaforák tartoznak Ezeknek az anaforáknak a horgonya egy kvantifikált kifejezés, a kifejezés referensének minden egyes tagjára külön-külön lesz igaz az állítás, tehát a névmás kötött változóként viselkedik. A 9) példában az *az* kifejezés egyesével teljesül azokra a diákokra, akik ötöst kaptak.

9) Minden diák, aki ötöst kapott, az kapott egy oklevelet is.

Az utolsó csoportba az asszociatív anaforák tartoznak, amelyeket *bridging anaforának* is szokott nevezni a szakirodalom (Hou–Markert–Strube 2013). Ezeknek az anaforáknak az értelmezéséhez következtetésre van szüksége a befogadónak, mert nem áll olyan szoros kapcsolatban a két kifejezés referense, mint a korábbi példákban. A magyar nyelvben az indirekt

anaforák viselkednek így, azonban ezeket a személyragokkal vagy a zéró névmásokkal szoktuk azonosítani.

10) Ameddig nem állítod le a mosogatógépet, nem fogod tudni kivenni azokat.

Az anafora azonosításához szükséges horgony, mivel az általa hordozott információra van szükség az anafora azonosításához, ami nem kizárólag névszói csoport 11) lehet, hanem mondat(ok) 12) vagy akár igei csoport 13) is.

11) Mari nagyon megörült Petinek$_i$, már rég nem látta őt$_i$.
12) [Mari és Peti már általános iskola óta barátok voltak]$_i$. Ez$_i$ tény.
13) Peti elköltözött$_i$ az egyetem miatt. Mari szomorú volt emiatt$_i$.

Sukthanker és munkatársai alapján a számítógépes nyelvészeti értelemben vett névmási anaforának három típusát különböztethetjük meg (Sukthanker et al. 2020), ezek a határozatlan névmási anafora 7), a határozott névmási anafora 11) és a melléknévi névmási anafora 14).

14) Peti vett Marinak egy zöld sálat$_i$, mert már rég akart ilyet$_i$.

De nem csak arról van szó, hogy az anaforikus kifejezések nem mindig koreferensek, éppen úgy az egymással koreferens kifejezések sem mindig anaforikusak. Például egy adott személy két különálló dokumentumban való említése koreferens egymással, hiszen ugyanarra a dologra utalnak, de nem anaforikusak, hiszen a két kifejezés egymástól függetlenül is értelmezhető. Ugyanez a helyzet azokban az esetekben is, amikor a névmást egy teljes főnévi csoporttal ismételjük meg 15). A következő példában a *Péter* kifejezés értelmezéséhez nem szükséges a megelőző mondat zéró névmása, ezért nem tekinthetjük a két kifejezés közötti kapcsolatot anaforikusnak. Ebben az esetben a *Péter* kifejezés kataforának tekinthető, hiszen ugyanarra az entitásra utalnak, ezért koreferensek is.

15) Már általános iskola óta ismerem (őt)$_i$, de még soha nem láttam ilyennek Pétert$_i$.

3.3.1. Anaforafeloldás és koreferenciafeloldás

A fentiekből tehát megállapítható, hogy a nyelvészeti szakirodalomban meghatározott fogalmak nem fedik pontosan a számítógépes nyelvészeti gyakorlatban bevett meghatározásokat, és a számítógépes nyelvészeti gyakorlaton belül is a rendszer további céljai határozzák meg, hogy pontosan mi is a feladat. Ennek az az oka, hogy egy magasabb szintű feladat megoldása,

mint például a tartalomkinyerés, vagy a gépi fordítás, más-más információtípusokat igényel. Leggyakrabban (de nem minden esetben) a számítógépes szakirodalom koreferenciafeloldásnak tartja egy entitás összes említésének azonosítását egy vagy több dokumentumban, anaforafeloldásnak pedig egy kontextusfüggő kifejezéshez tartozó egyetlen megelőző anchor, azaz horgony azonosítását, amely a kontextusfüggő kifejezés értelmezéséhez szükséges, így ezek párok lesznek. Az anaforafeloldás során tehát nem is igazán visszautalásokat azonosít a rendszer, hanem kontextusfüggő kifejezéseket és az azok értelmezéséhez szükséges horgonyt. Míg a koreferenciafeloldás során a cél a szövegben az összes azonos referenssel rendelkező kifejezés azonosítása, addig az anaforafeloldás során elegendő a visszautaló szóhoz tartozó egyetlen, leggyakrabban a szövegben az anaforát közvetlen megelőző antecedens azonosítása.

Az anaforák közül a határozott névmási anafora lesz az egyetlen, amely egyben koreferens is az antecedensével. Ezért is lehetséges az, hogy a névmási anaforafeloldást, amennyiben az kizárólag a határozott névmásokra vonatkozik, a számítógépes szakirodalom a koreferenciafeloldás alfeladatának is tekinti.

Annak ellenére, hogy a nyelvészeti szakirodalomban a probléma besorolása ettől merőben eltér, erőteljesen támaszkodik a számítógépes nyelvészeti módszertan a nyelvészeti eredményekre. Ezt azonban érdemes szem előtt tartani az egyes rendszerek eredményeinek értékelése során, valamint az ezekben a rendszerekben felhasznált elvek egy-egy feladatra való alkalmazhatóságának mérlegelése során.

4. Anaforafeloldás a nyelvészeti szakirodalomban

A következő fejezetben a nyelvészeti szakirodalomban meghatározott elveket és modelleket veszem sorra az anaforafeloldás szempontjából. Először ismertetem a feloldást célzó modellek főbb szempontjait, ezután pedig azokat a jellemzőket, amelyekre támaszkodva lehetővé válik az anaforafeloldás. A jellemzőkön belül kitérek a megszorítások és preferenciák közötti különbségekre, majd ismertetek két modellt, amelyek egymástól erőteljesen különböznek a jellemzők felhasználásának tekintetében. Az utolsó alfejezetben a magyar nyelvvel kapcsolatos megfigyeléseket közlöm, emellett pedig kitérek a már ismertetett jellemzők sajátosságaira a magyar nyelv tekintetében.

4.1. Az anaforafeloldást modellező elméletek alapjai

A névmáshoz tartozó antecedens azonosításának folyamatát modellező elméletek egy része figyelembe veszi mind a nyelvtani szabályokat, mind a teljes diskurzust, valamint a szövegalkotót vagy beszélőt, esetleg a befogadót is.

Az antecedens azonosításához használható jellemzőket csoportosíthatjuk aszerint, hogy mire vonatkoznak (az antecedensre, az anaforára vagy a két kifejezés közötti kapcsolatra), milyen jellegűek (nyelvi vagy nem nyelvi), illetve hogy kötelezőek-e (szabály vagy heurisztika).

Az antecedens azonosításához a jellemzők három forrásból származhatnak: 1) a névmásra vonatkozó jellemzők, 2) az antecedensjelöltre vonatkozó jellemzők és 3) a két kifejezés közötti kapcsolatra vonatkozó jellemzők.

A névmási anaforafeloldás során az anaforára vonatkozó információk adottak, ezekből az információkból vonhatók le további következtetések az antecedens természetére nézve.

A pszicholingvisztikai szakirodalomban a névmáshasználat a referensnek a beszélő mentális állapotában levő magas aktivitását mutatja. Erre a szakirodalom különböző módokon utal: *prominent, salient, accesible, in focus, center of attention* (Ariel 1990; Grosz–Joshi–Weinstein 1995; Gundel–Hedberg–Zacharski 1993; Arnold 2001; Arnold 2010). Vannak jellemzők, amelyek ezt a magas aktivitást előidézik, és számos jellemző kizárólag csak utólag mutatja, ezzel is segítve a beszélő számára az értelmezést. Tehát az anaforafeloldás során ezeket a jellemzőket kell keresni, megvizsgálni.

A három forrásból származó jellemzők csoportosíthatók nyelvi és nem nyelvi jellemzőkre. Nyelvi jellemzők azok, amelyek a kifejezések morfológiai, szintaktikai és szemantikai elemzése

után állnak rendelkezésünkre. Ilyenek például a kifejezés száma és személye, szintaktikai funkciója vagy szemantikai kategóriája. Nem nyelvi jellemzőknek tekintjük azokat, amelyek a szöveg felszíni szerkezetének segítségével határozhatók meg, illetve amelyekhez következtetés vagy a világtudás felhasználása révén juthatunk. Ilyen például a kifejezés említésének gyakorisága vagy a két kifejezés közötti távolság a szövegben.

A jellemzőket csoportosíthatjuk felhasználásuk alapján szabályokra és heurisztikákra. Szabályokról akkor beszélhetünk, ha az állításnak, feltételezve hogy a szöveg helyesen megszerkesztett, minden visszautalásra teljesülnie kell. A két kifejezés közötti kapcsolatot mutató jellemzők többsége nyelvtani szabályokon alapuló jellemző, ezeket gyakran megszorításoknak nevezi a szakirodalom. Ilyenek például az egyeztetési megszorítások az angol nyelvben. Heurisztikáknak pedig azokat az állításokat tekintjük, amelyek nem teljesülnek minden egyes esetben, ezeket preferenciáknak is nevezhetjük. A heurisztikák abban különböznek a szabályoktól, hogy valószínűségek alapján működnek, tehát az állítás nem minden esetben lesz igaz. A heurisztikák általában korpuszokon végzett vizsgálatok alapján megállapított, az esetek többségére teljesülő állítások.

A jellemzők segítségével nem alkothatók szükséges és elégséges feltételek az antecedens kiválasztására nézve még a nyelvtani szabályok segítségével sem, a már a bevezetésben is említett esetekből kiindulva. Leggyakrabban a jellemzők összessége teszi valószínűvé azt, hogy egy kifejezés az adott névmás antecedense-e vagy sem, azonban az egyes csoportok elnevezéséből is látható, hogy nem egységes a szerepük súlya az antecedens kiválasztása során. Leggyakrabban a megszorításoknak van erőteljes szerepük, a preferenciák pedig akkor segítenek, ha a megszorítások nem csökkentik le egy darabra a lehetséges antecedensjelöltek számát.

A következőkben sorra veszem a névmási anaforafeloldás során figyelembe vehető jellemzőket, majd ezután néhány olyan nyelvészeti modellt ismertetek, amelyek a jellemzők segítségével modellezik a névmási anaforafeloldást.

4.1.1. Megszorítások

Az anaforafeloldás során alkalmazható megszorítások kizárólag a helyes visszautalások tulajdonságait írják le, ezért ezek mind nyelvi információkat használnak fel. A nyelvi jellemzők a morfológia, a szintaxis és a szemantika területeinek elemzéseiből adódnak (Caramazza et al. 1977; Clark–Clark 1977; Garvey–Caramazza–Yates 1974; Grober–Beardsley–Caramazza 1978; Chomsky 1981). A névmáshoz tartozó antecedens azonosításához szintaktikai elemzés

segítségével fel kell ismerni a mondat szerkezeti egységeit, majd a névmásnak és az antecedensjelölteknek mint összetevőknek az egymáshoz való viszonyát. A morfológiai elemzés során a névmásnak és az antecedensjelölteknek a grammatikai jegyeit állapíthatjuk meg, ezekre vonatkoznak az egyeztetési megszorítások. Az angol nyelvben például a névmásnak és az antecedensének egyeztetve kell lennie számban és nemben (Gordon et al. 1999; Clifton–Ferreira 1987) és nemben (Arnold et al. 2000; Ehrlich–Rayner 1983; Garnham et al. 1995), tehát azok a főnévi csoportok, amelyek ennek a két kritériumnak nem felelnek meg, nem tekinthetők potenciális antecedensnek. A magyar nyelvben nincs nembeli egyeztetés, így erre nem hagyatkozhatunk, a számban történő egyeztetés kérdéskörére a későbbiekben térek ki.

Szintaktikai elemzés alapján meghatározott megszorításként az antecedens keresési hatókörével kapcsolatos megállapítások említhetők meg. A névmások és az antecedensjelöltjeik elhelyezkedésére vonatkozó állításokat a már ismertetett kötési elvek (Chomsky 1981; Lasnik 1976; Reinhart 1976) mondanak ki.

Szemantikai elemzés után megszorításként a kvantorok hatóköri megszorításai említhetők meg (Karttunen 1969).

A nyelvi jellemzőkön alapuló megszorítások tehát a helyes visszautalásokat írják le, éppen ezért megfigyelhető, hogy ezeknek az elveknek a többsége a generatív vagy strukturalista nyelvészeti irányzatok eredményeiből jöttek létre.

4.1.2. Heurisztikák és következtetések

Heurisztikák megállapíthatók nyelvi és nem nyelvi jellemzők segítségével is. A következőkben sorra veszem azokat a nyelvi jellemzőkön alapuló heurisztikákat, amelyek megállapításához elegendő az antecedenst tartalmazó tagmondat vizsgálata, majd azokat, amelyekhez a teljes szöveg vizsgálata szükséges. Ezek után röviden áttekintem azokat a jellemzőket, amelyek nem nyelvi jellemzők segítségével megállapítható heurisztikák, illetve következtetési folyamatok vagy a világtudás felhasználásának eredményeiként jönnek létre.

4.1.2.1. Nyelvi jellemzőkön alapuló heurisztikák

Az anaforafeloldás során nyelvi jellemzők segítségével alkalmazható heurisztikák két csoportra bonthatók, mondat és szöveg alapú megközelítésekre. A mondat alapú heurisztikák megállapításához elegendő az antecedensjelöltet tartalmazó tagmondat vizsgálata. A szöveg alapú heurisztikák megállapításához a teljes szöveg vizsgálata szükséges.

Mondat alapú megközelítések a *subject-assignment strategy* (Crawley–Stevenson 1990; Crawley–Stevenson–Kleinman 1990), a *parallel grammatical role preference* (Chambers–Smyth 1998) vagy másnéven a *paralell function strategy* (Caramazza–Gupta 1979; Grober–Beardsley–Caramazza 1978).

A *subject-assignment strategy* alapján a címzett a névmáshoz a megelőző legközelebbi alanyi pozícióban levő főnévi csoportot azonosítja antecedensként, míg a *parallel function strategy* alapján a megelőző legközelebbi azonos pozícióban levő, a *parallel grammatical role* esetében pedig az feltételezhető, hogy a névmás és az antecedense azonos grammatikai szerepben van. A következő bekezdésekben a stratégiákat alátámasztó kísérleteket fogom részletezni.

Groeber, Beardsley és Caramazza mondatkiegészítéses tesztjének hipotézise az volt, hogy az alárendelt tagmondatban, alanyi pozícióban szereplő névmást az adatközlők koreferensnek fogják tekinteni a főmondat alanyi pozíciójában levő főnévi csoporttal (Grober–Beardsley–Caramazza 1978). Az esetek többségében a hipotézisük helytálló volt, azokban az esetekben pedig, ahol nem, ott további szemantikai tényezők befolyásolták a döntést. Mivel az alanyi pozícióban szereplő névmásokat vizsgálták, nem határozható meg ebben az esetben, hogy melyik stratégia érvényesült a döntés során. Caramazza és Gupta három kísérletben vizsgálta a *parallel function* stratégia hatását a névmáshoz tartozó antecedens kiválasztása során (Caramazza–Gupta 1979). A kísérleti alanyoknak arról kellett döntést hozniuk, hogy egy adott mondatban szereplő névmás melyik megelőző mondatban szereplő személyre utal, a válaszaikon kívül pedig a reakcióidőt is rögzítették. Az eredményeik alapján a kísérleti alanyok a grammatikai alanyt választották az esetek többségében antecedensként.

Chambers és Smyth három elvégzett kísérlete közül kettő kötődik a párhuzamos grammatikai szerep hatásához, ezeket a diskurzus koherenciájára összpontosítva végezték el (Chambers–Smyth 1998). Az első kísérletük eredményei alapján arra a megállapításra jutottak, hogy a névmáshoz tartozó antecedensnek azonos grammatikai szerepben kell lennie, mint a névmásnak. A második kísérletük eredményei alapján pedig azt a következtetést vonták le, hogy a diskurzus csak akkor mondható koherensnek, ha a névmás és az antecedense azonos grammatikai szerepben van.

Crawley és munkatársai több olvasásértési, hozzárendelési és mondatkiegészítéses kísérletet is elvégeztek annak érdekében, hogy megállapítsák, a névmáshoz tartozó antecedens kiválasztása során az esetek többségében a *subject-assignement* vagy a *parallel-function* stratégia érvényesül-e gyakrabban. Ennek eldöntéséhez tárgyi pozícióban szereplő névmásokat vizsgáltak, mind olvasásértési, mind a hozzárendelési feladat segítségével (Crawley–Stevenson–Kleinman 1990).

Az eredmények a *subject assignement* stratégiát támasztották alá, azokban az esetekben, ahol a névmáshoz több antecedens is azonosítható volt. Azokban az esetekben, ahol a nembeli egyeztetési megszorítás figyelembe vehető volt, nem játszott szerepet a kiválasztás során a két stratégia. Crawley és Stevenson mondatkiegészítéses tesztjei alapján (Crawley–Stevenson 1990) szintén az a következtetés vonható le az eredmények alapján, hogy általánosságban szívesebben referálnak az adatközlők az alanyi pozícióban szereplő objektumokra, mint a tárgyi pozícióban szereplőkre a nembeli egyeztetéstől függetlenül.

Szöveg alapú heurisztika a *topic assignement strategy* (Sanford–Garrod 1981), ami alapján a névmáshoz tartozó antecedens azonosítása során a pronominalizáció ténye az, ami a leginkább érvényesül. Mivel névmással utalunk vissza, ezért az adott kifejezés referense a központi téma. Ennek az entitásnak a kiemelése a diskurzusból, mint diskurzus topik, tehát a diskurzus fő témája, befolyásolja leginkább a névmások értelmezését (Sanford–Garrod 1981). Sanford és Garrod kutatásaik alapján nem a megelőző mondat alanyi pozícióban szereplő főnévi csoportját vagy a megelőző mondat azonos grammatikai szerepű főnévi csoportját, hanem abban az adott pillanatban fennálló diskurzustopikot tartja a legvalószínűbb antecedensnek a mondatkezdő névmás számára. Crawley és munkatársai kísérletükben (Crawley–Stevenson 1990) azt találták, hogy az értelmezés során nem érvényesül a topik előnye, de ez annak is betudható, hogy a kísérleteikben nem voltak igazán jelöltek a topikok. Ezzel szemben voltak olyan kísérletek is, ahol épp ennek ellenkezője bizonyult igaznak (Sanford–Moar–Garrod 1988). Sanford Moar és Garrod kimutatták, hogy gyorsabb az olvasási sebesség, ha a névmás a topikra utal.

A nyelvi jellemzőkön alapuló heurisztikák vizsgálatai alapján elmondható, hogy a kísérleti módszerek és a választott névmások tulajdonságai nagyban befolyásolják a kísérletek eredményeit. Az alárendelt tagmondatban szereplő névmás a kísérletek alapján leggyakrabban a főmondat alanyával koreferens, azonban a szerzők nem vizsgálták, mi történik abban az esetben, ha az alárendelt mondatban több névmás is található. A főmondat kezdő pozíciójában szereplő névmás a megelőző tagmondatban szereplő főnévi csoportokkal más kapcsolatban van információs struktúra szempontjából, mint az alárendelő tagmondat névmása a főmondat főnévi csoportjaival. Ezt alátámasztják a fent ismertetett kísérletek is, hiszen más eredmények születtek a névmás pozícióját figyelembe véve. Ezek alapján az a következtetés vonható le, hogy akár az összes heurisztika is teljesülhet, és az, hogy melyik heurisztikát szükséges figyelembe vennünk, leginkább a névmás tulajdonságain múlik.

4.1.2.2. Nem nyelvi jellemzőkön alapuló heurisztikák

A nem nyelvi jellemzők közé azok a szöveg felszíni szerkezete vagy előelemzés segítségével kinyerhető tulajdonságok kerülnek, amelyek nem nyelvi információt hordoznak. Ezek egy része pragmatikai vagy világtudásból származik, más részük pedig kognitív elveken alapul.

Hobbs 1979-es (Hobbs 1979) munkájában a diskurzus szintjén folyó következtetési folyamatok „melléktermékeként" értelmezi a névmás referenshez való kapcsolódását. A világtudás, mint központi jellemző az anaforafeloldás során megjelenik például Hirst és Brill munkájában (Hirst–Brill 1980). Arra a következtetésre jutnak, hogy a névmás megjelenése a szövegben nem feltétlenül keresési folyamatot indít el a megelőző szövegrészben, inkább arra ösztönzi a címzettet, hogy a névmást tartalmazó tagmondatban szereplő információkat összekösse a megelőző információkkal. Tipikusan következtetést és világtudást igényel az indirekt anaforához tartozó antecedens, illetve referens azonosítása, mivel ezekben az esetekben az anafora és az antecedense nem koreferensek. Az antecedens itt valójában egy horgony, „Anchor" lesz, amelynek a segítségével értelmezhetővé válik a névmás (Schwarz 2000; Kocsány 2018). A magyar nyelvben nagyon ritkán fordul elő névmás indirekt anaforaként, leginkább az igei személyragok tekinthetők annak, mint ahogyan a 16) példa is mutatja:

16) Hónapok óta írok$_i$, mégsem készül**nek**$_i$ el (a fejezetek).

A kognitív jellemzők és kognitív heurisztikák közé általában azok kerülnek, amelyek azt mutatják, hogy az egyes referensek a beszélő, illetve a hallgató mentális állapotában adott pillanatban mennyire lehetnek aktívak, elérhetőek. Egy referens elérhetősége a memóriában azon múlik, hogy hogyan és hol lett megemlítve korábban a diskurzusban (Clark–Sengul 1979).

Az elérhetőségre vonatkozónak a látszólag egymással szemben álló *advantage of the first-mentioned participant* (Gernsbacher–Hargreaves 1988) és *advantage of the most recent clause* (Gernsbacher–Hargreaves–Beeman 1989) heurisztikák.

Az *advantage of the first-mentioned participant* heurisztika azt mondja ki, hogy a mondatban szereplő első entitás könnyebben elérhető, mint a többi. Az elmélet alapját több kísérlet is alátámasztja, amelyek eredményei alapján kijelenthető, hogy az értelmezés során a befogadó a kezdő szavak alapján építi fel a nagyobb struktúrákat, így a tagmondatokat és a mondatokat is (Cairns–Kamerman 1975; Aaronson–Ferres 1983), valamint hogy az első szó a legalkalmasabb arra, hogy a segítségével a teljes mondat visszahívható legyen (Bock–Irwin 1980). Gernsbacher és Hargreaves hét kísérletben vizsgálta a heurisztika helytállóságát. A kísérleti alanyok sokkal

gyorsabban reagáltak az elsőként említett szavakra, mint a másodikként említettekre. A kísérletekben azt vizsgálták, hogy az eredményre hatással voltak-e nyelvi jellemzők, és azt a konklúziót vonták le, hogy sokkal inkább a kognitív folyamatok eredményei, mint a nyelvi jellemzők befolyása az, ami ezeket az eredményeket okozta. Ezek az eredmények azt mutatják, hogy a mondat első szava az, amelyhez a befogadó a későbbi információkat hozzákapcsolja (Gernsbacher–Hargreaves 1988).

Az *advantage of the most recent clause* heurisztika a megelőző tagmondatban szereplő entitások könnyebb elérhetőségét állítja a korábbiakkal szemben. Ez az előny abból az egyszerű elvből indul ki, hogy a legutóbb elhangzott információk könnyebben előhívhatók, mint a korábbiak. Gernsbacher, Hargreaves és Beeman munkájában a két heurisztikát vizsgálták úgy, hogy megmérték, milyen gyorsan reagál a befogadó egy összetett mondat első, illetve második tagmondatának entitásaira (Gernsbacher–Hargreaves–Beeman 1989). Eredményeik alapján a structure building framework (Gernsbacher 1985) segítségével mind a két heurisztika figyelembe vehető. Az *advantage of the most recent clause* heurisztika két mondat között, az *advantage of the first mention* heurisztika pedig a tagmondatok között érvényesül.

Számos kognitív alapú heurisztika a szövegtopik, azaz a diskurzus fő témájának meghatározására irányul. Mivel a szövegtopik kitüntetett eleme a szövegnek, igen könnyen elérhető referens a befogadó számára a mentális állapotban, a pronominalizáció pedig szintén ezt a könnyű elérhetőséget mutatja. Ezért a névmási anaforafeloldás során számos kognitív nyelvészeti alapú modell nem is kifejezetten a névmást szem előtt tartva keresi az antecedenst, hanem egyszerűen a névmás megjelenésének pillanatában éppen fennálló diskurzustopikot keresi, és a névmást automatikusan a diskurzustopikra való utalásnak tartja. A diskurzustopik referensének azonosítására elsősorban szintén kognitív alapú jellemzők alkalmazhatók. A szöveg feldolgozása előtt a cím az, ami először kijelölheti egy diskurzus fő témáját, tehát a diskurzustopikot (Kozminsky 1977).

Kiemelt pozícióban található a mentális állapotban egy referens közvetlenül a bevezetése, tehát az első említése után (Ariel 1990), hiszen maga a referens bevezetésének ténye is mutatja, hogy abban az adott pillanatban arról az entitásról lesz szó.

Azok a kifejezések, amelyek egy egység (bekezdés, mondat) elején találhatóak, sokkal figyelemfelkeltőbbek, illetve általában ezek hordozzák a lényeges információkat, a diskurzustopik is általában elöl található, a jellemzőként való felhasználása során ezt nevezi a szakirodalom *initial mention*-nek (Kieras 1980; Sanford–Garrod 1981).

Azokat az információkat, amelyek fontosak, gyakran megemlítjük egy szövegben, mivel a diskurzustopik a diskurzus fő témája, és ezért ez a legfontosabb információ, valószínűsíthető, hogy az említése gyakori lesz: *frequency of mention* (Perfetti–Goldman 1974).

A következő fejezetben két olyan modellt mutatok be, amelyeknek célja a névmási anaforafeloldás, és a fent említett megszorításokat és heurisztikákat alkalmazza eltérő arányban és módszerekkel. A két modell az angol nyelv sajátosságait figyelembe véve készült, így más nyelvekre történő alkalmazásukhoz az elvek módosítása szükséges.

4.2. A központiság elmélet és az elérhetőségi elmélet

A névmási anaforát a következő két modell a teljes diskurzus és a bekezdések alapján vizsgálja. A teljes diskurzus koherenciájához hozzájárul az, hogy a diskurzuson belüli kisebb lokális szegmensek, tehát a bekezdések koherensek, ennek a koherenciának pedig az egyik fő eszköze a központi téma. Grosz, Weinstein és Joshi (Grosz–Joshi–Weinstein 1995) alapján a központi téma az az entitás, amely a figyelmi állapot központjában van, tehát a leginkább szembetűnő elem a diskurzus adott pontján. A diskurzuson belül megkülönböztetnek globális és lokális központot az alapján, hogy a teljes diskurzus kommunikációs célját jelöli, vagy a kisebb lokális szegmensek céljait, vagyis a beszélő pillanatnyi intencióit.

A kisebb lokális diskurzusszegmensek központi témáját a már korábban is említett jellemzők segítségével tudjuk utólag a szövegből azonosítani. A központiság elmélet alapján pedig a figyelmi állapot központjában lévő kifejezésre meghatározott típusú referáló kifejezéssel tudunk visszautalni. Tehát a kifejezések formájából következtetni tudunk arra, hogy adott pillanatban a diskurzus melyik objektuma van a figyelmi állapot központjában, és arra is, ha ez az objektum kikerül a központból, és egy másik objektum veszi át a helyét. Mivel a diskurzusban a központ az az elem, amely a beszélő és a hallgató figyelmi állapotának a központjában van, olyan kifejezéssel utalunk rá vissza, amely mutatja, hogy a beszélő és a hallgató is könnyedén eléri a referenst a mentális állapotából.

Grosz, Weinstein és Joshi központiság elmélete (Grosz–Joshi–Weinstein 1995) alapján minden megnyilatkozás tartalmaz úgynevezett *forward-looking center*-eket *(Cf),* amelyek valójában a megnyilatkozás referáló kifejezései, tehát a lehetséges antecedensjelöltjei a következő megnyilatkozás úgynevezett *backward-looking center*-ének *(Cb),* ami a megnyilatkozás központi témája, tehát az anafora. A diskurzusszegmens első megnyilatkozásában még nincs *Cb*, hiszen eben a megnyilatkozásban vezetjük be a diskurzus objektumait először, és csak a következő megnyilatkozás alapján határozhatjuk meg, hogy ezek

közül melyik a téma. A *Cf* lista részben rendezett bizonyos tulajdonságaik mentén, amelyek közül a grammatikai szerep, a pronominalizáció és az elhangzás sorrendje a legprominensebb, azonban ezek között a tulajdonságok között is van eltérés annak tekintetében, hogy mennyiben befolyásolják a sorrendet. Ezek alapján az a kifejezés lesz *Cb* a következő megnyilatkozásban, amely a legutóbb hangzott el, névmás és alany vagy zéró névmás és alany.

17) a) Odaérkezett [a kávézóhoz] [a lány]$_j$ …().
b) Bement [(ő)]$_j$,
c) hogy lefoglalja [(ő)]$_j$ [az asztalt],
d) ameddig [a fiúra] vár [(ő)]$_j$.

A 17) példa alapján a diskurzusszegmens a) megnyilatkozásának *Cf* listáján két darab objektum található, amelyek *a lány* kifejezés és *a kávézóhoz* kifejezés referensei. A b) megnyilatkozásban a *Cf* lista megegyezik a *Cb*-vel, mivel egy darab referens található a megnyilatkozásban. A megnyilatkozás *Cb*-je azonos kell hogy legyen az a) megnyilatkozás *Cf* listáján első helyre rendezett elemével. Tehát a b) megnyilatkozásban a zéró névmás a lányra utal vissza, mivel ez az alany és a legközelebbi referens. A c) megnyilatkozás *Cf* listáján két objektum található, a zéró névmás és az *az asztalt* kifejezések referensei, az első helyre sorolt elem a pronominalizáció miatt a zéró névmás referense lesz, tehát ez lesz az adott megnyilatkozás *Cb*-je, ami azonos a b) megnyilatkozás egyetlen referensével. A d) megnyilatkozás *Cf* listáján szintén két referens található, a zéró névmás és az *a fiú* referensei, ezért itt szintén a zéró névmás referense lesz az első helyre sorolt elem, azaz a *Cb,* ami azonos lesz a c) megnyilatkozásban szereplő zéró névmás referensével.

A központiság elmélet célja a lokális koherencia modellezése, tehát a diskurzusszegmensek meghatározása a bennük szereplő témák segítségével. Ennek modellezését egy preferencia-sorrend alapján képzelik el, amely a figyelmi állapot központjában lévő objektum megtartására és leváltására vonatkozik. Ez alapján a legpreferáltabb a központi téma folytonossága, ezután a központi téma megtartása, a legkevésbé preferált pedig a központi téma leváltása.

A központi téma folytonossága azt jelenti, hogy az U_n megnyilatkozásban levő *Cb* azonos az U_{n+1} megnyilatkozásban levő *Cb*-vel, és ezzel egyidejűleg a legmagasabbra rangsorolt az U_{n+1} megnyilatkozás *Cf* listáján, tehát a legalkalmasabb jelöltje az U_{n+2} megnyilatkozás *Cb*-jének. A téma folytonosság során tehát a központi objektum három megnyilatkozáson keresztül tudja tartani a *Cb,* jelen esetben az anafora, szerepet.

A központi téma megtartása során az U_n megnyilatkozásban levő *Cb* azonos az U_{n+1} *megnyilatkozás Cb*-jével, azonban nem ez az objektum a legmagasabbra sorolt *a Cf* listán, tehát nem ez az objektum a legvalószínűbb jelölt az U_{n+2} megnyilatkozás *Cb*-jének. Az U_{n+1} megnyilatkozásban tehát megtartjuk lehetséges jelöltnek, de nem ő lesz az U_{n+2} *Cb*-je, ezáltal két megnyilatkozáson keresztül tudja betölteni a *Cb* szerepet.

A harmadik típus a központi téma váltása, amely azt jelenti, hogy az U_{n+1} megnyilatkozás *Cb*-je nem azonos az U_n megnyilatkozás *Cb*-jével.

Ez a prominencia-sorrend is azt mutatja, hogy a diskurzusszegmensen belül a központi téma lesz az, aminek a leggyakoribb az említése. A gyakori említés szintén a könnyű elérhetőséget biztosítja, minél gyakoribb egy referens említése, annál valószínűbb, hogy az ő elérhetősége lesz a legmagasabb a lehetséges antecedensek közül.

18) központi téma folytonosság
   a) [A szomszéd embernek] volt [egy olyan nyolcévesforma huncut fia]. (Cf1 = fiú)
   b) [Annak a gyereknek] mindig valami komiszságon járt az esze. (Cf1=fiú, Cb = fiú)
   c) Legjobban szúrta a szemét (neki), hogy mennyi körte van [az öregember körtefáján]. (Cf1=fiú, Cb = fiú)
   d) Egyszer, amikor átlesett a kerítésen [a gyerek]… (Cb = fiú)

19) központi téma megtartás
   a) Na de [a kötél] vásott volt, s [a gyerek] is megnehezedett [a sok körtétől].(Cf1 = a gyerek)
   b) (…) éppen odaérkezett [a körtefához] [(ő)], (Cb= a gyerek, Cf1 = a gyerek)
   c) hát elszakadt [a kötél], s - zsuppsz! – [a gyerek] leesett. Éppen rá [az öregemberre]. (Cb = a gyerek, Cf1 = az öregember)
   d) [Az öregember] felszökött, s látta, hogy [a gyereknek] tele [a zsebe] körtével. (Cb = az öregember)

20) központi téma váltás
   a) volt [egy öreg gazdaember]. (Cf1 = gazdaember)
   b) Annak [az öreg gazdaembernek] a kertjében volt [egy körtefa], (Cb = gazdaember, Cf1 = körtefa)
   c) s azon [a körtefán] szép nagy körték termettek (Cb = körtefa)

A központiság elmélet tehát a lokális koherencia modellezését tűzte ki célul, de ezáltal alkalmas a névmási anaforához tartozó antecedens azonosítására is. A pronominalizációt automatikusan a központi témává válás jelének tekinti, ezen keresztül pedig a megelőző megnyilatkozásban a legkönnyebben elérhető objektummal azonosítja. Számos probléma merülhet fel azonban a módszer alkalmazása során. Elsősorban a szöveg megnyilatkozásokra való szegmentálása az, ami gondot okozhat, hiszen az elmélet nem tér ki arra, mit tekint megnyilatkozásnak. Arra sem tér ki az elmélet, hogy több névmás megjelenése esetén, hogyan azonosítható antecedens a nem *Cb* szerepű névmáshoz, illetve hogy hogyan különböztethetők meg azok a névmások, amelyek nem a megelőző megnyilatkozás központi témájára, hanem a globális topikra, azaz a teljes diskurzus fő témájára utalnak. Szintén érdemes szem előtt tartani, hogy a *Cf* listán történő rendezéshez felhasznált jellemzők minden egyes nyelv esetében eltérőek lehetnek.

Az *elérhetőségi elmélet* (Ariel 1990; Ariel 2014; Ariel 2001) a kifejezések formáját hozza összefüggésbe a referensük mentális elérhetőségével. Az elérhetőségi elmélet szerint a beszélő olyan kifejezést választ, amiről feltételezi, hogy a hallgató értelmezni tudja, tehát a kiválasztás során figyelembe veszi, hogy a hallgató mentális állapotában adott pillanatban mennyire van központi pozícióban az adott objektum, amire utalni akar. Az elméletben a különböző fokú mentális elérhetőséget mutató kifejezéseket Ariel egy skálán helyezi el, így a kifejezések a formájuk és nyelvtani tulajdonságaik alapján összehasonlíthatók az elérhetőség szempontjából. Azok a kifejezések, amelyek kevés nyelvi információt tartalmaznak, mint amilyen a névmás is, magas vagy könnyű elérhetőséget mutatnak, azaz a referensük a hallgató mentális állapotának középpontjában helyezkedik el. Az Ariel által legfontosabbnak vélt jellemzők az informativitás mértéke, a kifejezés rigidsége és a hossza. Az informativitás mértéke arra vonatkozik, hogy az adott kifejezés mennyire jellemzi hiánytalanul az adott dolgot, amire utal. A rigidség azt mutatja, hogy az adott kifejezés mennyire merev jelölő, a hossz pedig a nyelvi forma hossza, amely a kifejezés írott méretére és fonológiai méretére is vonatkozik. Minél informatívabb, rigidebb és hosszabb egy nyelvi kifejezés, annál alacsonyabb az értéke az elérhetőségi skálán, tehát feltételezhetően a referens nehezen elérhető.

Ariel korábbi kognitív nyelvészeti kutatások kísérleteinek eredményeiből kiindulva az elérhetőség alapján három nagyobb csoportba sorolja a kifejezéseket: alacsony, közepes és magas elérhetőséget kódolóba. Az alacsony elérhetőséget kódoló csoportba a tulajdonneveket és a határozott leírásokat, a közepes elérhetőségűbe a személyes és mutató névmást, a magas elérhetőségűbe pedig a zérókat sorolja. Fontos azonban kiemelni, hogy az elérhetőség egy

skaláris tulajdonság, a különböző elérhetőségi fokokat kódoló kifejezések pedig különböző módon viselkednek a nyelvekben. Az Ariel által meghatározott elérhetőségi skálát a következő ábra mutatja:

Alacsony elérhetőségi érték

a Módosító + teljes név

b Teljes név

c Hosszú határozott leírás

d Rövid határozott leírás

e Vezetéknév

f Keresztnév

g Módosító + távolra mutató demonstratívot tartalmazó kifejezés

h Módosító + közelre mutató demonstratívot tartalmazó kifejezés

i Távolra mutató demonstratívot tartalmazó kifejezés

j Közelre mutató demonstratívot tartalmazó kifejezés

k Hangsúlyos névmás + gesztikuláció

l Hangsúlyos névmás

m Hangsúlytalan névmás

n Klitikum

o Extrém magas elérhetőséget mutató jelölők
(gap, pro, PRO, kérdő kifejezés nyoma, reflexívek, egyeztetés)

Magas elérhetőségi érték

1) ábra Elérhetőségi skála (Ariel 2014)

Ez azt jelenti, hogy ha a beszélő azt feltételezi, hogy a hallgató mentális állapotában nehezen elérhető a referens, akkor nagyon specifikusnak kell lennie a kifejezésnek ahhoz, hogy a hallgató értelmezni tudja az információt. Ezzel szemben, ha könnyen elérhető a címzett számára a referens, elegendő kevésbé rigid vagy rövidebb kifejezéssel utalni rá, mivel kevesebb erőfeszítésre van szüksége a hallgatónak ahhoz, hogy azonosítsa a kifejezés referensét. Az, hogy egy objektum a beszélgetésben a figyelmi állapot középpontjába kerüljön, és elegendő legyen névmással utalni rá, számos módon kiváltható. Elérhetőbbé teszi az objektumot az, ha jelen van a fizikai kontextusban, vagy ha korábban már szó volt róla. Szintén magas az elérhetősége azoknak az objektumoknak, amelyek szorosan illeszkednek a diskurzusuniverzumba, vagy épp feltűnően nem illeszkednek oda. További tényező még, hogy az adott objektum élőlény vagy

tárgy, az elmélet szerint ugyanis az élőlények egyszerűbben elérhetőek. Tehát amikor a visszautaló névmáshoz tartozó antecedenst keressük, olyan kifejezést keresünk, amely már önmagán hordozza az elérhetőség jeleit.

A névmás és az anafora közötti kapcsolatot Ariel, nem túl szerencsés módon, szintén az elérhetőség fogalmával jellemzi, azonban itt nem teljesen ugyanarról a kapcsolatról van szó, mint a referens mentális állapotban való elérhetősége. A két kifejezés közötti elérhetőség egy külön reláció, melyben szerepet játszik a két kifejezés által mutatott, a referensükre vonatkozó mentális állapotban való elérhetőség, de ezen kívül még a két kifejezés közötti kapcsolatok is. Ilyen például a kifejezések közötti távolság is. Ariel azt a megállapítást tette, hogy a két kifejezés közötti elérhetőség, azaz kapcsolat, annál szorosabb, minél közelebb vannak egymáshoz. Szorosabb a kapcsolat, ha a névmás az antecedenst tartalmazó mondat alárendelő tagmondatában található, mintha mellérendelt tagmondatokban lennének, és akkor a leggyengébb, ha külön mondatokban helyezkednek el. Abban az esetben, ha a két kifejezés közötti kapcsolat a távolság miatt gyengébb, arra következtethetünk, hogy a szövegalkotó egyéb nyelvi jelekkel fog segíteni a befogadónak a névmás értelmezésében. Ebben az esetben vizsgálhatjuk meg például az antecedensjelöltek formája által mutatott, a referensükre vonatkozó elérhetőséget. Az antecedensjelölt formáján kívül vizsgálható még az említésének gyakorisága, a mondaton belüli pozíciója, hogy új diskurzusreferenst vezet-e be, tehát azok a tulajdonságok, amelyek a szalienciát mutatják.

Az elérhetőségi elmélet tehát a kifejezések formáját veszi elsősorban figyelembe, ezen kívül pedig az egyéb korábban tárgyalt szalienciát mutató tényezőket. A névmáshoz tartozó antecedens keresése során nem határozza meg a keresés hatókörét, így a megelőző tagmondatnál messzebbre történő visszautalásokat is képes felismerni. Az elmélet nem tér ki arra, hogy a szalienciát és az elérhetőséget mutató tényezők az anaforafeloldásban egymáshoz képest milyen jelentőséggel játszanak szerepet, illetve hogy az egyes esetekben mely jellemzőket szükséges figyelembe venni az antecedens sikeres azonosításához.

A fent ismertetett két modell a névmási anaforához tartozó antecedens azonosítására is alkalmas, azonban erőteljesen eltér a két megközelítés egymástól. A központiság elmélet a névmást a központi témával azonosítja, és feltételezve, hogy a szöveg tartalmilag összefüggő egységet alkot, az antecedenst a megelőző tagmondat kifejezései között keresi. Az elmélet mindössze három jellemző alapján rangsorolja a potenciális antecedensjelölteket, leginkább megszorításokra támaszkodik, és ezek alapján próbál egy antecedenst azonosítani a névmáshoz, azonban nem képes kezelni a messzebbre történő visszautalásokat. Az elérhetőségi elmélet a

kifejezések formájából következtet a referenseik elérhetőségére, és azt feltételezi, hogy az elérhetőség önmagában elegendő ahhoz, hogy a névmáshoz tartozó antecedenst azonosítsuk. A modell előnye, hogy a messzebbre történő visszautalások is azonosíthatók a segítségével, abban az esetben, ha a közelebbi antecedensjelölteket kizárjuk, hátránya azonban, hogy az elérhetőséget okozó és mutató jellemzőket nem határolja el egymástól élesen, illetve nem határozza meg a jellemzők anaforafeloldásban betöltött szerepének súlyát, így nem alkot meg egy egységes szabályrendszert, amely alapján az összes névmáshoz tartozó antecedens egységesen módon kereshetővé válik.

4.3. A magyar névmáskutatás eredményei

A következő fejezetben a már ismertetett, az anaforafeloldás során felhasználható jellemzőket fogom megvizsgálni a magyar nyelvvel kapcsolatban. A jellemzők ismertetése során sorra veszem a magyar névmáskutatás főbb eredményeit, ami többnyire szintén a funkcionalista, funkcionalista-kognitív szemléletmódot követi.

A magyar nyelvvel kapcsolatban a névmások vizsgálata gyakrabban terjed ki a névmások csoportosíthatóságának szempontrendszereire (Laczkó 2004; Laczkó 2005), vagy arra, hogy az egyes kontextusokban milyen típusú névmás használata a kötelező vagy lehetséges (Laczkó–Tátrai 2012; Kocsány 1996). Pszicholingvisztikai megközelítéssel a névmás értelmezésének folyamatáról készült több tanulmány (Pléh 1998; Pléh–Radics 1976). A névmási deixis kérdésköréről is szintén az adott kontextusokban való használat szempontjából olvashatunk (Boronkai 2010; Laczkó 2008; Tátrai 2010; Laczkó–Tátrai 2012). A névmási anaforáról és az antecedens azonosításának modellezéséről kevesebb szó esik a magyar szakirodalomban.

A névmási anaforafeloldás során felhasználható megszorításokkal kapcsolatban kevesebb információra támaszkodhatunk, mint az angol nyelvvel kapcsolatban. A morfológiai elemzés során megállapíthatók a szám és személy jegyek, de a nembeli egyeztetés hiányában ezek a megszorítások általában nem zárnak ki elegendő antecedensjelöltet a potenciális jelöltek közül. Azt is érdemes megfontolni, hogy a megszorításokat mekkora jelentőséggel vesszük figyelembe. Ha az anaforafeloldásra tágabb értelemben tekintünk, vagy a számítógépes nyelvészeti megközelítést vesszük figyelembe, akkor a szám szerinti egyeztetés sem feltétlenül szükséges a visszautalás sikerességéhez. Ilyen eset például, ha a visszautaló szó és az antecedense rész-egész viszonyban vannak egymással:

21) Mi$_i$ nyertünk annak ellenére, hogy én$_i$ megsérültem az egyik gyakorlat közben.

A kötési elvek és a kvantorok hatóköri elvei a magyar nyelvben is érvényesülnek, ezek szintén az anafora tágabb definícióját figyelembe véve használhatók fel, de ezek sem zárnak ki elegendő antecedensjelöltet a legtöbb alkalommal, ezek az elvek leggyakrabban a keresés hatókörét képesek meghatározni, korlátozni. Tehát a visszaható 22) és kölcsönös névmásnak 23) a magyar nyelvben is tagmondaton belül keresendő az antecedense.

22) Mari$_i$ látja magát$_i$ a tükörben.
23) [Mari és Peti]$_i$ már általános iskola óta ismerik egymást$_i$.

Tágabb értelemben véve hatókörnek azt a tartományt nevezzük, amelynek értelmezése a kvantortól függ. Abban az esetben, ha egy névmás egy kvantor hatókörében van, a névmás értelmezése is a kvantor segítségével történik.

24) Minden diák$_i$ átment a vizsgán és kapott (ő)$_i$ egy oklevelet.

A kifejezések mondaton belüli pozíciói, mivel a magyar diskurzuskonfigurációs nyelv, nem a grammatikai funkciót mutatják, hanem a mondat szavai által kifejezett információs szerkezetre következtethetünk belőlük. Az ige előtt sorrendben a topik, a kvantor és a fókusz helyezked(het)nek el, amelyek közül a topik egy már ismert információt ismétel meg, a fókusz pedig általában új információt hordoz. Az esetrag segítségével következtethetünk az alany, tárgy és további argumentumok grammatikai funkciójára. Az angol nyelvvel kapcsolatban a *subject assignement* stratégia és *a parallel function* stratégia alkalmazása közötti különbséget a névmás és az antecedens egymáshoz viszonyított pozíciója határozta meg leggyakrabban. A magyar nyelvvel kapcsolatban Pléh és munkatársai (Pléh–Radics 1976; Pléh 1994; Pléh 1998) arra a megállapításra jutottak, hogy a névmás típusa mutatja utólag a szövegben, hogy a két stratégia közül melyik érvényesül.

Pléh 1994-es munkájában a következő megállapításokat tette (Pléh 1994 : 297):

> 1) a jelöletlen, prototipikus alapesetben, amikor az előzmény alanya ismétlődik, a magyarban elhagyjuk az alanyt a második mondatban. 2) Az **az** mutató névmás az alanyváltás jele. Azt mutatja, hogy a korábban nem alanyi főnévi csoport vált alannyá. 3) Nyomatékosítás esetén az ismétlődő alany is személyes névmásként jelenik meg, ha élő lexikai jegyű. 4) Az előzmény-mondat 'másik' főnevével koreferens nem alanyi főnévi csoport (vagyis nem alanyból lesz nem alany) személyes névmás formában realizálódik, ha nem tárgy. Ha tárgy, akkor elhagyjuk.

Ezek alapján Pléh és Radics egy két szabályon alapuló nyelvtani modellt dolgoztak ki (Pléh–Radics 1976):

> A. Törlési szabály: az alanyismétlés törléshez vezet, ugyanakkor ez nem a felszíni szerkezeti alanyeseten, hanem alapvetőbb nyelvtani viszonyokon alapul (akkori terminológiánkban: a megismételt mélyszerkezeti alany törlődik).
> B. Alanyváltás szabály: A mutató névmás (az) viszont az alany- (vagy pespektíva-) váltás jele. (Pléh 1994: 278)

Kísérleteik során egy diskurzusérzékeny feldolgozási modell megalkotását tűzték ki célul. Megállapítják, hogy az anaforaértelmezés során a befogadónak úgy kell integrálnia a különböző forrású információkat, hogy először egy előzetes nyelvtani elemzést végez, majd ennek eredményeire támaszkodva vizsgálja részletesebben a lehetséges antecedensjelölteket. A modell megalkotásához kísérletek segítségével egy sor heurisztikát állapítottak meg (Pléh 1994: 315–316):

> 1) A tematikus szerepek fontosabbak, mint a morfológiai jelölés és a szintaktikai keret, ha az érintettebb szereplő az előzményben elöl áll a mondatban. 2) A tematikus szerep és a topikalizációs tagolás együttesen egy feldolgozási perspektívát adnak. Ez kiindulópontként kezeli az aktívabb szereplőt. Az aktivitásba azonban belejátszik a lineáris sorrend is. 3) A szereplők kognitív reprezentációja ebből a szempontból nem lényeges. Az egyedi propozíciók kognitív reprezentációja (pl. annak leképezése, hogy akik találkoznak vagy akik sakkoznak, azok egyenértékűek) csak egy későbbi fázisban tekintődik. 4) Az egyedi propozíció feldolgozásnak van egy beépített interpropozíciós oldala is. Nevezetesen a perspektívafenntartás vagy párhuzamos funkció hipotézis. Mindig úgy tartjuk, hogy a perspektíva ugyanaz marad, hacsak valami nem bírálja felül ezt. E felülbírálásnak több párhuzamosan működő módja van, a referenciaváltó névmástól a propozíciók közti viszonyokig.

Kocsány 2016-os munkájában az *ő* névmás szerepeit, illetve a mondatközi anaforát vizsgálta. (Kocsány 2016). Eredményei megerősítették a törlési és alanyváltási szabályokat, kiegészítve azzal a megállapítással, hogy a zéró névmás nem kizárólag az alanyismétlést jelöli, tehát a mutató névmás is törölhető. A zéró névmással való alanyváltás összefügg a beszélő intenciójának megváltozásával. A névmás típusa nem függ össze a +/- élő jeggyel, a mutató névmás használata alanyváltás során például utalhat előre (Kocsány 2016: 147):

> Ha az antecedens és a visszautaló anafora közé egyéb elemek ékelődnek, a [+élő] jegyét aktiváló személyes névmással [+élő] jegyű alanyokra vissza tudunk utalni, a [–élő] jegyű mutató névmással azonban [–élő] jegyű antecedensekre sem utalhatunk vissza, muszáj névszóval megneveznünk az antecedenst. Ez nyilvánvalóan a [+élő] jegy mentális elsődlegességével függ össze.

Tolcsvai kutatásában (Tolcsvai Nagy 2000) kognitív keretben vizsgálja az *ő* és az *az* névmások megoszlását, a szövegvilág, szövegtopik és szövegfókusz fogalmak segítségével. Kutatásai során Pléh és munkatársaival azonos következtetésekre jutott, azaz azonos alanyú mondatok esetén a magyarban kötelező a második mondatban törölni az alanyt, a távolra mutató névmás pedig az alanyváltást jelöli.

Kocsány kutatásaiban az *ez* közelre mutató névmást és a komplex anafora szerepét is vizsgálja a magyar nyelvben (Kocsány 2018; Kocsány 2011). Komplex anaforának tekinti azokat az anaforákat, amelyek megelőző mondatértékű struktúrát vezetnek be, tehát eseményre, folyamatra, állapotra utalnak. A magyar nyelvben a leggyakrabban az *ez* névmás tekinthető komplex anaforának. Kutatásai során az alábbi következtetésekre jutott: A közelre mutató névmás antecedense a megelőző mondatban leggyakrabban, de nem minden esetben, nem a topik lesz. Az anafora szerepe a nem topikra való visszautalás esetében a beszélő kommentárjának bevezetése, a topikra való visszautalás pedig a témaváltást jelzi előre.

Tehát a kutatások alapján, hasonlóan a *subject assignement* és *parallel grammatical role* stratégiákhoz, a *topic assignement* stratégia használata is a névmás típusától függ, és nem a pozíciójától.

A szövegtopik azonosítására a korábban ismertetett nem nyelvi jellemzők hasonlóan alkalmazhatók, tehát az elhangzás gyakoriságának, a bekezdésben elfoglalt pozíciónak és a címnek a vizsgálata a magyar nyelvben is támpontok lehetnek az azonosítás során. Az új entitás bevezetése a diskurzusba, valamint az elérhetőségi elméletben megfogalmazott, a kifejezés formájára és hosszára vonatkozó elvek is hasonlóan érvényesülnek. A pronominalizáció mint a szövegtopikot automatikusan mutató jellemző azonban a magyar nyelvvel kapcsolatban nem teljesül a törlés miatt. A magyar nyelvben a könnyen elérhető, mentális állapotban központi pozícióban található entitásokra utaló kifejezések a felszíni szerkezetből automatikusan törlődnek.

A fenti kutatásokból is látszik, hogy a magyar nyelvvel kapcsolatban főleg a tagmondatok közötti visszautalások vizsgálata a jellemző. Ennek az az oka, hogy a felszíni szerkezetben megjelenő információk mennyisége sokkal kevesebb, mint az angolban, mind a morfológiai

tulajdonságok, mind a szintaktikai funkciók tekintetében. Mindebből arra következtethetünk, hogy az anaforafeloldás során sokkal erőteljesebben támaszkodik a nyelvhasználó a kognitív tényezőkre, a pragmatikai információkra, azaz a következtetésre és a világtudásra (Laczkó–Tátrai 2012).

5. Az automatikus anaforafeloldás a számítógépes nyelvészeti szakirodalomban

Az automatikus anaforafeloldásnak két módja lehetséges: szabály- vagy tudásalapú, illetve statisztikai alapú, azaz tudásszegény.

A szabályalapú kinyerés során elméleti szempontok figyelembevételével megalkotott szabályokra támaszkodunk, tehát a szabályok megalkotójának tudása beépül a rendszerbe. Ilyen rendszerek például a központiság elmélet eredményein alapuló BFC algoritmus (Brennan–Friedman–Pollard 1987) vagy a szintaktikai tudást felhasználó, szélességi keresést végző Hobbs algoritmus (Hobbs 1978). A névmáshoz tartozó antecedens kinyerése komplex feladat, minden nyelvi szint részt vesz a folyamatban. Szintén probléma a szabályok hierarchiába rendezése, tehát az egymáshoz való viszonyuk meghatározása.

A tudásszegény rendszereken belül is két módszer különíthető el, a felügyelt (supervised) és a felügyelet nélküli (unsupervised) rendszerek. A felügyelt rendszerek a szövegben található manuálisan vagy előelemzés során hozzáadott információk alapján hoznak döntést (pl. az anafora és az antecedense egyeztetve vannak-e számban, illetve személyben, azonos tagmondatban találhatóak-e stb.), tehát a szöveg előfeldolgozáson, rendszerezésen esik át. Ilyen rendszer például Soon és munkatársai Mention-pair módszeren alapuló rendszere (Soon–Ng–Lim 2001). A felügyelt rendszerek fő problémája, hogy az előelemzésekből származó hibák további hibákat okoznak a célfeladat megoldása során. A másik nehézség pedig azokat a nyelveket érinti, amelyek esetében a szöveg felszíni szerkezetében az anaforafeloldást segítő információk száma csekély. Ez a probléma a szabályalapú feloldási módszereket is érinti, hiszen a szabályokat is ezeknek az információknak a segítségével fogalmazhatók meg. Ezekben az esetekben részletes manuális annotáció szükséges, ami erőforrás- és időigényes.

A felügyelet nélküli rendszerek tanító adatbázisában nincsenek rendszerezve az információk. Magának az algoritmusnak kell felismernie a mintázatokat a nyers szövegben, éppen ezért nagyszámú pozitív előfordulásra van szükség a modellépítéshez. Ennek a feltételnek köszönhetően ezeket a rendszereket gyakrabban koreferenciafeloldásra szokták alkalmazni, ilyen például Poon és Domingos Markov Logic-on alapuló kísérlete (Hoifung–Pedro 2008).

A következő részben áttekintem a névmási anaforafeloldást célzó rendszerek típusait és eredményeit, habár a fentiekhez igazodva, némely esetben a névmási anaforafeloldás csak részfeladata az adott rendszernek. A felügyelet nélküli rendszerekre egyrészt a fent ismertetett

okokból kifolyólag, másrészt a disszertáció eltérő céljai miatt nem térek ki. Az áttekintés során figyelembe veszem a megoldási kísérletek mögött meghúzódó elméleti háttereket, elképzeléseket, így jól látható lesz, hogy milyen irányban változnak, fejlődnek ezek a rendszerek.

A rendszerek áttekintése négy fő részből fog állni, az első három rész igazodik a módszertani elvek első három csoportjához, majd egy külön alfejezetben a magyar nyelvvel kapcsolatos eredményeket ismertetem. Az első névmási anaforafeloldást célzó rendszerek szabályalapú tudásgazdag, azaz nyelvészeti tudáson alapuló szabályok segítségével működő rendszerek voltak. Ezeket váltották fel a nagyobb korpuszokon meghatározott heurisztikákon, statisztikai eredményeken alapuló rendszerek. A harmadik egységbe pedig a különböző felügyelt gépi tanulási kísérleteken alapuló modellek és kiértékelési lehetőségeik tartoznak.

5.1. Szabályalapú rendszerek

5.1.1. Hobbs algoritmus

Az első kifejezetten névmási anaforafeloldást célzó rendszer Hobbs nevéhez fűződik (Hobbs 1976a; Hobbs 1976b). Az általa kidolgozott módszer nyelvészeti megfontolású elveken alapult, így kizárólag morfológiai és szintaktikai információkat vett figyelembe, és szabályok segítségével működött. A módszer lényege, hogy az antecedensjelölteket az őket tartalmazó mondatok szintaktikai fái alapján meghatározott sorrendben vizsgálja meg az algoritmus, ezen belül pedig nem és szám szerinti egyezést keres az anafora és a főnévi csoport között. Az algoritmus a szintaktikailag és morfológiailag előelemzett szövegen, kilenc szabály segítségével a következő módon működik (a szintaktikai fában a főnévi csoport NP, a mondat S rövidítéssel szerepel) (Hobbs 1976a: 9):

1. Először azt az NP csomópontot vizsgáljuk meg, amelynek közvetlen dominanciája van a névmás felett.
2. Feljebb lépünk a fán a legközelebbi NP vagy S csomópontig, nevezzük ezt a csomópontot X-nek, az utat, amit megteszünk, pedig p-nek.
3. Haladjunk végig az X csomópont alatti összes ágon balról jobbra szélességi keresést alkalmazva. Antecedensjelölt lesz bármely NP, amely az NP vagy S csomópont és X között előfordul.

4. Ha az X csomópont a legmagasabb S csomópont a mondatban, akkor térjünk át a szövegben közvetlen megelőző mondat szintaktikai fájához a legközelebbi utolsó csomóponttól visszafelé haladva. Minden fán balról jobbra szélességi keresést végzünk, és ha NP csomópontot találunk, akkor azt lehetséges antecedensnek tekintjük. Ha X nem a legmagasabb S csomópont a mondatban, akkor az 5. lépéssel folytatjuk.
5. Az X csomóponttól az első NP vagy S csomópontig megyünk és ezt új X csomópontnak nevezzük, az utat pedig hozzáadjuk p-hez.
6. Ha az X csomópont kategóriája NP, és a p X-hez vezető útvonal nem halad át az X által közvetlenül dominált főnévi csomóponton, akkor X antecedensjelölt.
7. Az X alatti összes ágon haladjunk keresztül balról jobbra szélességi keresést végezve. Minden NP csomópontot vizsgáljunk meg, és tippeljük meg, hogy antecedens-e.
8. Ha az X csomópont kategóriája S, akkor X jobb oldalán lévő összes ágát vizsgáljuk meg balról jobbra szélességi keresést végezve, de ne haladjunk egy NP vagy S csomópont alá se, amelyet találunk. Vizsgáljunk meg minden NP-t antecedensnek, amelyet találunk.
9. folytassuk a 4.-től.

A fenti módszer tehát a névmástól visszafelé halad a szintaktikai fán szélességi keresést végezve balról jobbra, és sorra vizsgálja a főnévi csoportokat a legközelebbitől egyre távolodva. Az az antecedensjelölt lesz a potenciális antecedens, amely számban és nemben is egyeztetve van a névmással, és a legközelebb található a névmáshoz a jelöltek közül. Hobbs nem automatizálta algoritmusát, így a kezdetekben az általa véletlenszerűen három angol nyelvű szövegből választott száz névmáson vizsgálta meg kézzel a módszert. A vizsgálathoz feltételezte a tökéletes szintaktikai és morfológiai elemzést. A vizsgált névmások az angol egyes szám harmadik személyű *he, she, it* és a *they* voltak. Az algoritmus előnye, hogy nem szükségesek szemantikai vagy világtudásból származó információk hozzá, így viszonylag kevés előelemzési lépéssel is működik, viszont annak a kevés előelemzési lépésnek tökéletesnek kell lennie, hiszen ez az alapja. Hátránya a rendszernek viszont az, hogy a közelebbi antecedensjelölteket preferálja, így a távolabbra történő visszautalásokat nem ismeri fel, ha már a névmáshoz közelebb is található egy számban, személyben és nemben is egyeztetett antecedensjelölt.

5.1.2. BFP algoritmus

A másik széles körben elterjedt, szabályalapú módszer a 4.2 szakaszban már ismertetett központiság elmélet számítógépes implementációja. A diskurzus modellezésére készült központiság elmélet a lokális koherencia megállapítását tűzte ki célul a megnyilatkozások központi témájának figyelembevételével, ezen keresztül azonban alkalmas a névmási anaforához tartozó antecedens azonosítására is. Ennek a modellnek a formalizált számítógépes implementációja a BFP algoritmus (Brennan–Friedman–Pollard 1987). A központiság elmélet egy sor szabályt és megszorítást tartalmaz, amely meghatározza a kapcsolatot a diskurzus résztémái között, ezáltal pedig a diskurzus kisebb lokális összefüggő egységeit. Ilyen szabályok a nyelvtani funkciók megválasztása, a szintaktikai szerkezet vagy a referáló kifejezés típusa: tulajdonnév, határozott leírás, névmás. A modell azáltal válik alkalmassá a névmási anafora kezelésére, hogy a pronominalizációt automatikusan, a központi témára való figyelemirányítás eszközének tekinti.

A központiság elméleten alapuló algoritmusnak három fázisa van. Az első fázis a lehetséges referáló kifejezések kilistázása, amelyek ebben az esetben automatikusan az adott megnyilatkozás *forward looking center* (Cf)-ei lesznek. Ezek azok a kifejezések, amelyek a későbbiekben potenciális antecedensei lehetnek majd egy névmásnak. Szintén az első lépés része a kifejezések grammatikai szerep szerinti sorrendbe tétele. A legelső pozícióba rendezett névmás lesz a megnyilatkozásban a *backward looking center* (Cb), tehát az a visszautaló névmás, ami az adott megnyilatkozás központi témája. Ezután megvizsgáljuk, hogy a jelenlegi megnyilatkozás névmásai egyeztetve vannak-e. Ha nem, akkor ezek biztosan eltérő dologra utalnak. A második fázis a kifejezések párokba rendezése. Minden Cf listán szereplő kifejezést párba rendezünk a következő megnyilatkozás Cb-jével. Ezután a következő fázisban a generált párokat szűrjük. Erős megszorítások alapján kiszűrjük azokat a párokat, amelyek nem felelnek meg például a kötési elveknek vagy az egyeztetési paradigmáknak. Ha egy megnyilatkozás több névmást is tartalmaz, akkor a további párokat a diskurzus témájának preferált folytonossága alapján rendezzük: a legvalószínűbb a folytonosság, ezután a megtartás, majd a váltás.

5.1.3. A Hobbs és a BFP algoritmusok összehasonlítása és más nyelvekre való implementálása

A szerzők ugyan nem értékelték ki az algoritmust, de Walker (Walker 1989) tanulmányában sorra vizsgálja a különböző szövegtípusok tekintetében a BFP algoritmust összehasonlítva a Hobbs algoritmussal. A kiértékelés során pusztán a valóban visszautaló egyes és többes számú

harmadik személyű személyes névmásokat vette figyelembe, így csak az állapítható meg, hogy az antecedens azonosítása sikeres volt-e a módszer segítségével vagy sem. Az elbeszéléseken végzett kísérletben 100 esetet vizsgáltak meg: míg a Hobbs algoritmus 88, addig a BFP algoritmus 90 névmáshoz azonosította a helyes antecedenst. Az újságcikkekből gyűjtött, szintén 100 előfordulás tekintetében, illetve egy párbeszédekből álló korpusz 81 előfordulását tekintve a Hobbs algoritmus teljesített jobban. Míg a BFP algoritmus az újságcikkekben szereplő 100 esetből 79-et, addig a Hobbs algoritmus 89-et azonosított helyesen. A párbeszédek tekintetében a BFP 81-ből 49, a Hobbs algoritmus 51 esetet talált el. Az eltérés egyedül az újságcikkek tekintetében nagyobb, ez pedig valószínűleg abból fakad, hogy a Hobbs algoritmus a mondaton belüli antecedens azonosításában eredményesebb, a BFP pedig a mondaton kívüliben, hiszen kifejezetten a mondatok közötti kohéziós viszonyok azonosítását tűzte ki célul.

Ezek a megoldási kísérletek mind az angol nyelvben szereplő névmásokhoz tartozó antecedensek azonosítását tűzték ki célul, és a bennük megfogalmazott szabályok és elvek is az angol nyelvhez igazodtak. Az évek során a két algoritmust számos nyelvre implementálták eltérő sikerességgel, mely erősen függ az adott nyelv szerkezetétől és a benne szereplő visszautalások jellemzőitől.

A Hobbs algoritmus kínai nyelvre történő implementációja (Converse 2005) sikeresnek mondható, mivel az angolhoz hasonlóan a kínai is kötött szórenddel rendelkezik. A harmadik személyű személyes névmásokra 77,6%, a zéró névmásokra 73,3%-os pontosságot ért el. A spanyol nyelvre implementált változat (Palomar et al. 2001) 62,7%-os eredményt ért el, a portugál nyelvre (Santos–Carvalho 2007) a magazinokból összeállított korpuszon 61,22%-ot, az irodalmi szövegekből készült korpuszon 49,68%-ot, a jogi szövegekből készült korpuszon pedig 40,4%-ot ért el.

A BFP algoritmus portugálra implementált változatát (Aires et al. 2004) 16 jogi szövegen értékelték ki és 51%-os eredményt ért el. A japán (Manabu–Kouji 1996), illetve thai (Aroonmanakun 2000) nyelvekben azt találták, hogy a központiság elmélet a zéró névmásokhoz tartozó antecedens azonosításában ér el jobb eredményeket. A thai nyelvre 87,32%-os pontosságot ért el, egy kibővített változata pedig 90,67%-ot, a japán nyelvre pedig 78%-ot.

Ezek az eredmények rávilágítanak arra, hogy a névmási anaforafeloldás problémájának megoldása során a módszer kiválasztása erőteljesen függ nem csak az adott nyelvtől, de a szövegtípustól is, amire alkalmazni szeretnénk az adott módszert, illetve arra, hogy a megoldás során eredményes lehet az adott nyelvben, illetve adott szövegtípusban előforduló esetek figyelembevétele és az ezek az esetek alapján megfogalmazott szabályok alkalmazása is.

5.2. Heurisztika alapú rendszerek

Nem sokkal a szabályalapú rendszerek megjelenése után elkezdődött olyan korpuszok létrehozása, amelyeken az automatizálást célzó rendszerek ellenőrizhetők, illetve a visszautalások nagyobb mennyiségben vizsgálhatók. Ezzel párhuzamosan a statisztikai alapú vizsgálatok is elkezdődtek a területen, amelyek ugyanúgy felhasználtak nyelvészeti megközelítésű szabályokat, de már erőteljesebben koncentráltak a korpuszokban gyakrabban előforduló jelenségek leírására: ezek az úgynevezett tudásszegény rendszerek.

5.2.1. Dagan–Itai 1990 módszere

Dagan és Itai statisztikai megközelítésű rendszere (Dagan–Itai 1990) már nagyobb korpuszon számított együtt-előfordulási mintázatokon alapul. A módszer célja az volt, hogy szelekciós megszorítások segítségével az angol *it* egyes szám harmadik személyű névmáshoz antecedenst azonosítsanak. A megközelítés két fázisból állt, az elsőben egy nagy korpusz alapján statisztikai adatbázist építettek, a másodikban pedig a megépített adatbázis felhasználása történt meg. A statisztikai adatbázis szintaktikai kapcsolatok együtt-előfordulási mintázatait tartalmazta. A módszer ellenőrzéséhez véletlenszerűen választottak mondatokat, amelyek tartalmazták az egyes szám harmadik személyű személyes névmást, majd kézzel kiszűrték azokat az eseteket, amelyekben a kifejezés nem referált, vagy nem volt a szövegben megelőző antecedense, illetve azokat az eseteket, amelyekben olyan szerkezet szerepelt, amire nem volt példa az adatbázisban. Így a megmaradt 59 előforduláson értékeltek ki. Minden esetben azokat az antecedensjelölteket vették figyelembe, amelyek kielégítették a szám-, személy- és nembeli egyeztetésre vonatkozó megszorításokat. 21 esetben az algoritmus egyik jelöltet sem választotta ki, tehát a statisztikai adatbázis alapján nem volt képes döntést hozni. A maradék 38 példa közül 33-hoz azonosította az algoritmus a valódi antecedensjelöltet, ami 87%-os pontosságot jelent.

5.2.2. Resolution of Anaphora Procedure (RAP)

Lappin és Leass RAP (Resolution of Anaphora Procedure) algoritmusa (Lappin–Leass 1994) a McCord's Slot Grammar elemző szintaktikai elemzésén alapult, de már a figyelmi állapotot is megkísérelte figyelembe venni. Munkájuk célja volt a névmáshoz tartozó mondaton belüli és a mondaton kívüli antecedenseket is azonosítani. Az algoritmus három fő komponensből áll: előszűrésből, a megmaradt jelöltek értékeléséből és a kiválasztásból. Az értékelés alapja a szaliencia, amelyet a szintaktikai szerkezet segítségével határoztak meg.

A RAP első komponense az előszűrés, amely több lépésből áll. Az előszűrés során szintaktikai és morfológiai információk segítségével zárják ki azokat a kifejezéseket, amelyek biztosan nem lehetnek antecedensek. A szintaktikai szűrés a Lappin és McCord által bemutatott szűrő segítségével történik (Lappin–McCord 1990a), a morfológiai szűrés pedig egyszerűen kizárja azokat a főnévi csoportokat, amelyek nincsenek számban és nemben egyeztetve a névmással. Ezen kívül szűrik még a pleonasztikus *it*-et, azaz a szemantikailag üres névmásokat. A szűrés utolsó lépése a kölcsönös és visszaható névmásokhoz tartozó antecedensek azonosítása, ezzel a lépéssel ezek az antecedensek lekerülnek a potenciális antecedenseket tartalmazó listáról. Ezt az utolsó lépést szintén Lappin és McCord munkája alapján végzik el (Lappin–McCord 1990b).

A RAP második komponense az antecedensjelöltek rangsorolása szalienciájuk alapján. A rangsorolás alapja egy érték, amelyet számos, szalienciát mutató jellemző alapján számol ki a rendszer, ilyenek pl. szemantikai szerep, a grammatikai szerepek párhuzamossága, említések gyakorisága, közelség, mondat közelség. A nyelvtani szerepekre felállított hierarchia a következő módon alakul: a) az alany valószínűbb antecedensjelölt, mint a nem alany, b) a közvetlen tárgy, mint más komplementum, c) az igei argumentumok valószínűbb antecedensek, mint az adjunktumok, d) a fejek pedig valószínűbb antecedensek, mint a fejek komplementumai.

Az utolsó fázis a lehetséges jelöltek közül való választás, amely a morfológiai információk és szalienciát mutató érték alapján történik.

A módszert először egy előelemzett, 560 névmásból álló mintán értékelték ki, amely 475 névmáshoz azonosított antecedenst sikeresen, ami 85%-ot jelent. Az összes névmási visszautalásból 89 volt mondatok közötti, ebből 72-höz azonosított sikeresen antecedenst, ami 81%-ot jelent, 471 volt mondaton belüli, amelyből 403-hoz azonosított antecedenst sikeresen, ez 86%-ot jelent. A vakteszt 360 névmási visszautalásból állt, ebből 70 mondat közötti és 290 mondaton belüli visszautalás. Az összes eset 86%-át azonosította sikeresen az algoritmus, ami a mondatok közötti visszautalás tekintetében 74%-ot, a mondaton belüli visszautalások tekintetében pedig 89%-ot jelent.

A portugálra implementált változat (Coelho–Carvalho 2005) magazinok szövegein 43,56%-ot, irodalmi szövegeken 32,61%-ot, jogi szövegeken pedig 35,15%-ot ért el. A német nyelvre Wunsch (Wunsch 2006) tesztelte le az algoritmus eredményességét. Kísérletéhez a “Tübingen Treebank of Written German” (TüBa-D/Z) (Telljohann–Hinrichs–Kübler 2004) korpuszt használta fel, amely újságcikkekből készült, és a Hinrichs és munkatársai által hozzáadott

koreferenciaannotációt tartalmazza (Hinrichs et al. 2004). A TüBa-D/Z újságcikkeiben szereplő visszautalásokra 76,6-os F-mértéket ért el.

A RAP algoritmus előnye, hogy nem pusztán szintaktikai és morfológiai információk alapján hoz döntést, hanem az ezekből az információkból következtethető szalienciát is figyelembe veszi, hátránya azonban, hogy hasonlóan a Hobbs algoritmushoz részletes szintaktikai előelemzést igényel a rendszer. Kennedy és Boguraev nevéhez fűződik a RAP algoritmus egy módosított változata (Kennedy–Boguraev 1996), amely szintaktikai előelemzés nélkül működik. Az algoritmus egy egyszerűsített morfoszintaktikai elemzésen alapul, ezáltal többféle szövegen futtatható. Az elemzőt 27 véletlenszerűen választott szövegen tesztelték változatos műfajban. A szövegekben 306 harmadik személyű személyes névmás volt, amiből 231-nek helyesen azonosított antecedenst a rendszer.

Egy másik statisztikai alapú megoldás Ge, Hale és Charniak nevéhez fűződik (Ge–Hale–Charniak 1998). Három fő jellemző statisztikái alapján: távolság, szelekciós megszorítások, említések száma. A módszerrel 84,2%-os pontosságot értek el.

A különböző statisztikai alapú megoldásoknak, valamint a RAP algoritmus és továbbfejlesztett változatainak újdonsága, hogy nem csak a konkrét nyelvi előelemzésből származó információkra támaszkodik, hanem ezekből további következtetéseket von le, valószínűségeket is figyelembe vesz az antecedensre vonatkozóan. Ezek a valószínűségek abban térnek el a szabályoktól, hogy nem lesznek minden esetben igazak, de az esetek többségében igen. Az anaforafeloldás problémakörének megoldása ezzel elmozdult abba az irányba, amely nem az összes lehetséges visszautalás leírására és feloldására törekszik, hanem a leggyakrabban előforduló, legjellemzőbb mintázatok felismerésére. A kizárólag morfológiai és szintaktikai szabályokon alapuló rendszerektől való távolodás azt is lehetővé tette, hogy olyan nyelvekben előforduló visszautalásokra is alkalmazhatóak legyenek a módszerek, amelyeknek felszíni szerkezetében kevesebb morfológiai információ jelenik meg, illetve nem kötött szórenddel rendelkeznek.

5.3. Gépi tanuláson alapuló rendszerek és kiértékelési lehetőségeik

A 2000-es évek elejére a korpuszalapú technikák váltak népszerűbbé a természetesnyelv-feldolgozás szempontjából, hiszen nem a lehetséges esetekre koncentráltak, hanem a valóban nagyobb számban is előforduló példákra, így jobb eredményeket produkáltak a tesztelések során. A MUC-6 és MUC-7 koreferenciakorpuszok (Linguistic Data Consortium 2003; Linguistic Data

Consortium 2001) elkészülésével adott volt a lehetőség, hogy a koreferenciafeloldás is a korpuszalapú megoldások felé mozduljon. A koreferenciakorpusz segítségével kiértékelhetővé váltak már meglévő algoritmusok, technikák, megvizsgálhatóvá váltak heurisztikák. Emellett a gépi tanulás is egyre népszerűbbé vált.

A koreferencia- és anaforafeloldás során különösen fontos kitérni a kiértékelési metrika kérdéskörére, hiszen egyik feladat esetében sem triviális a gold standard korpusszal való összehasonlítás menete. Mást tekintünk sikeres feloldásnak, ha a cél anaforikus kifejezés-antecedens párok azonítása, és mást, ha a feladat referencialitási ekvivalenciaosztályok azonosítása egy szövegben, így a metrika kiválasztása során az egyik legfontosabb szempont maga az elvégzendő feladat.

A MUC feladat a kiértékelés során (Vilain et al. 1995) a standard IR metrikákat alkalmazza (F=F-mérték, P=Pontosság, R=Fedés), ezek kiszámításához pedig a gold standard korpuszban jelölt, valamint a modell által azonosított koreferens kapcsolatok számát használja. A pontosság (precision) a mind a két esetben megfigyelhető kapcsolatok száma osztva a modell kimenetében azonosított kapcsolatok számával. A fedés (recall) a mind a két esetben megfigyelhető kapcsolatok száma osztva a gold standard korpuszban előforduló kapcsolatok számával. Az F-mérték, ami alapján végül a modellek összehasonlíthatók, a két érték harmonikus közepe. A metrika tehát kifejezéspárok közötti kapcsolatok felismerését méri. A koreferenciafeloldás szempontjából két probléma merül fel a metrikával kapcsolatban. Egyrészt nem veszi figyelembe azokat az entitásokat, amelyek kizárólag egyszer fordulnak elő a szövegben, vagyis antecedens nélküliek, ami a probléma, ha az összes referens azonosítása, majd azok ekvivalenciaosztályokba való csoportosítása a feladat. Másrészt a számítási módszerből fakadóan a kevesebb kapcsolatot azonosító modellek esetében jobb eredményeket mutat, hiszen a magasabb pontosság magasabb F-mértéket is fog eredményezni a gyengébben teljesítő modellek esetében is.

Bagga és Baldwin módszere, a B-cubed metrika (Bagga–Baldwin 1998) már nem a kifejezések közötti kapcsolatokra, hanem a kifejezések egyes klaszterekhez való tartozására támaszkodik, minden egyes előforduláshoz kiszámolja a pontosságot és a fedést, majd ezeknek az értékeknek veszi a súlyozott átlagát. A módszer az egyes kifejezések jelenlétét, illetve hiányát értékeli ki a gold standard korpusz ekvivalenciaosztályaiban. Ezzel a módszerrel figyelembe vehetők az egyszeri említések is, illetve mérhetővé válnak az azonosított ekvivalencia osztályok méreti különbségeiből fakadó eltérések is. A módszerből fakadó probléma, hogy az adott

előfordulás pontossága és fedése a koreferencialáncokkal való összehasonlításából számítandó ki, így maguk a láncok többször lesznek kiértékelve.

Az ACE program célja azonosítani a szövegben az említett entitásokat, a közöttük lévő kapcsolatokat, és az eseményeket, amelyekben ezek az entitások részt vesznek. Tehát a koreferenciafeloldás szempontjából a cél egy entitás említésének összes azonosítása, legyen az tulajdonnév, határozott leírás vagy névmás, és ezek ekvivalencia osztályokba sorolása. A feladat komplexitása maga után vonja a saját kiértékelési metrika alkalmazását: minden egyes kifejezéshez egy saját ACE-értéket (NIST 2003) számítanak ki, a modell eredményessége ezek összességéből számítandó. Az ACE-értékek kiszámítása különbözik az entitás típusa (hely, személy) és az említés szintje (tulajdonnév, névszó, névmás) tekintetében is. A tökéletes koreferenciafeloldó rendszer ACE-értéke 100%, míg egy olyan rendszeré, amelynek a kimenete üres, 0. Ha csak hibás azonosításai vannak egy rendszernek, akkor negatív ACE-értéke lesz.

A CEAF mutató (Luo 2005) már nem a kapcsolatokra vagy a kifejezésekre koncentrál, hanem az entitásokra. Az érték meghatározásához egy az egyhez azonosítja a kifejezésekből alkotott koreferencialáncokat az entitásokhoz. A cél kinyerni az entitásokat (szereplők, objektumok), ezután meghatározni az egy lánchoz tartozó kifejezéseket, majd összekapcsolni az entitásokat a láncokkal úgy, hogy egy lánc csak egy entitáshoz tartozhat és fordítva. Mivel teljes azonosítás történik, így pontosan meghatározható a pontosság, a fedés és az F-mérték is. Nem szerepelhet egy kifejezés (referens) több láncban, és több lánc sem tartozhat ugyanahhoz a referenshez, így az F-mérték pontosan mutatja a modell sikerességét. Abban az esetben, ha túl sok entitást azonosít a rendszer, akkor arányosan romlik a pontosság, ha túl keveset akkor pedig a fedés. A tökéletes rendszer F-mértéke 1, míg azé, amelyik nem azonosít egy entitást sem 0.

A MELA kiértékelési metrika Denis és Baldridge (Denis–Baldridge 2009) nevéhez köthető, és három korábban említett metrikán alapul. Az érték a MUC, a B-CUBED, és a CEAF értékek súlyozott átlaga, így egyszerre figyelembe vehető az összes korábban említett mérési módszer sajátossága.

5.3.1. A Mention-pair modell

Az egyik leggyakrabban alkalmazott koreferenciafeloldást célzó modell a Mention-pair modell. A módszert Aone és Bennet (Aone–Benett 1995), valamint McCarthy (McCarthy 1996) dolgozta ki, de a végleges számítógépes környezetbe implementált változat, amelyen gépi tanulási kísérlet végezhető, Soon és munkatársai (Soon–Ng–Lim 2001), valamint Ng és Cardie (Ng–Cardie

2002b) nevéhez fűződik. A Soon és munkatársai által alkotott modell és jellemzőkészlet, valamint tanulóalgoritmus és kiértékelési metrika vált a standard baseline-ná a 2000-es évek elején.

A modell a szövegből előzetesen kinyert főnévi csoportokból alkotott párokból áll. Ezek közül a párok közül az egyik a visszautaló szó, a másik pedig a lehetséges antecedensjelölt. A páron kívül tartalmazza még a visszautaló szóra vonatkozó jellemzőket, a lehetséges antecedensjelöltre vonatkozó jellemzőket és a két kifejezés közötti lehetséges kapcsolatot mutató jellemzőket. Ebből a három forrásból származó információk összessége az a jellemzővektor, ami a párt reprezentálja a modellben. A tanulóalgoritmus így egyszerű klasszifikációs feladatot végez, azt állapítja meg, hogy az adott jellemzők alapján a két főnévi csoport koreferens-e vagy sem. Az osztályozás után a koreferensnek ítélt párok közül minden egyes kifejezés tekintetében egyet veszünk figyelembe, hiszen a cél egy antecedens azonosítása. A teljes folyamat során négy helyen befolyásolhatjuk az eredményt: 1) a párok generálása során, 2) a párokat jellemző tulajdonságok meghatározáskor és kinyeréskor, 3) tanuló algoritmus kiválasztásánál, és 4) egyetlen jelölt azonosítása során.

A párok generálására számos példát találunk a szakirodalomban, a következőkben ezek közül tekintek át néhányat.

5.3.1.1. A párok kiválasztása

A kiinduló állapot, amely minden lehetőséget megvizsgál, a következő: a pozitív példák generálásához minden anaforikus főnévi csoportot párba rendezünk minden őt a koreferencialáncban megelőző főnévi csoporttal. A negatív példák generálásához szintén az anaforikus főnévi csoportokat használjuk fel, de azokkal a főnévi csoportokkal rendezzük őket párba, amelyek nem voltak a koreferencialánc tagjai, de megelőzték az anaforát a szövegben. Ennek a megoldásnak a hátránya, hogy a negatív példák jelentős túlsúlyba kerülhetnek a pozitív példákkal szemben. Ez a tanulás során azt eredményezheti, hogy az algoritmus minden párt a nem koreferensek kategóriába sorol, ami magas pontosságot eredményez, hiszen több az egymással nem koreferens kapcsolatban lévő pár, mint a koreferens, de egy visszautalást sem azonosít sikeresen.

Ezt a problémát Soon és munkatársai (Soon–Ng–Lim 2001) úgy küszöbölték ki, hogy a negatív példák számát előszűrő lépésekkel csökkentették. A negatív példák generálásához csak azokat a főnévi csoportokat használták fel, amelyek az anaforikus főnévi csoport és a hozzá tartozó legközelebbi antecedense között voltak megtalálhatók. Ezen kívül még számos előszűrő

lépés alkalmazható, amely a negatív párok számát csökkenti, ezzel pedig elősegíthető a későbbiekben a tanuló algoritmus sikeressége. Például Strube, Rapp és Müller (Strube–Rapp–Müller 2002) nyelvészeti alapú szűrőket alkalmazott még a gépi tanulás előtt, ami a negatív példák számát 50%-kal csökkentette. Többek között kiszűrték a határozatlan és a beágyazott főnévi csoportokat, a névmások tekintetében pedig az egyeztetési jegyek alapján szűrtek. Yang és munkatársai (Yang et al. 2003) a negatív példák számát nem azzal csökkentették, hogy a kézzel is annotált antecedensig generálták őket, hanem azzal, hogy a szöveg eleje helyett egy általuk meghatározott távolságig rendelték hozzá a főnévi csoportokat párként az anaforához. Az anaforát tartalmazó mondat és az azt megelőző két mondat főnévi csoportjait, névmás tekintetében a nemben számban és személyben is egyeztetett főnévi csoportokat vették figyelembe. Mivel az anaforafeloldás célja a Mention-pair technikával az anaforához tartozó egyetlen antecedenssel való kapcsolat azonosítása, így nem csak a negatív, de a pozitív példák szűrése is hozzájárulhat a tanulás sikerességéhez. Ng és Cardie (Ng–Cardie 2002b) nem csak a negatív, de a pozitív példákat is szűrte, a negatív példákat Soon és munkatársai módszerével, kizárólag a legközelebbi antecedensig vették figyelembe. Mivel egy antecedens azonosítása is elegendő az anaforafeloldás során, ezért a pozitív példákat is szűrték, Harabagiu és munkatársai módszerét alkalmazva (Harabagiu–Bunescu–Maiorano 2001) kizárólag az anaforához tartozó legközelebbi antecedens azonosítását tűzték ki célul. Uryupina (Uryupina 2004) a szűrés során az anafora típusát is figyelembe vette. A főnévi csoportokat négy csoportba sorolta és a különböző csoportokat különböző módon szűrte (névmás, határozott leírás, tulajdonnév, összes többi NP). Az ezeknek az előszűréseknek a segítségével létrejött rendszerek sokkal jobb eredményeket értek el, mint a kizárólag tanuláson alapulók.

5.3.1.2. A párok jellemzőinek meghatározása

A második feladat a párok generálása után a párokat jellemző tulajdonságok meghatározása és kinyerése. A jellemzőket általában két csoportba sorolhatjuk: nyelvi és nem nyelvi jellemzők. A nyelvi jellemzők a morfológiai, szintaktikai és szemantikai információkból származnak, a két kifejezés grammatikai tulajdonságaira (szám, személy, nem) és azok egyeztetésére, szintaktikai információkra (szintaktikai funkció), szemantikai kategóriáikra (élő vagy élettelen, személy, hely, dátum vagy intézmény) vonatkoznak. A nem nyelvi jellemzők általában a két kifejezés közti távolságot, a kifejezések említésének gyakoriságát mutatják. Szintén gyakori még annak vizsgálata, hogy a két kifejezésben szereplő szavak között van-e azonos, ami az ismétlést feltételezheti. Ezeknek a jellemzőknek a megállapításához számos előelemzési lépésre van

szükség, ezért nem jellemző, hogy az összes megállapításra kerül egy-egy tanulási kísérlet során. Soon és munkatársai (Soon–Ng–Lim 2001) 12 jellemző segítségével építettek osztályozót, amelyek között szemantikai információk nem szerepeltek. Ng és Cardie ezt a jellemzőkészletet egészítette ki további 41 jellemzővel, de a jellemzőkészlet méretének növelése csökkentett a pontosságon, ezért végül a 41 új jellemző közül 18 jellemzőt választottak ki, és ezzel a 30 jellemzővel építettek osztályozót.

5.3.1.3. A tanuló algoritmusok

A harmadik kérdés az algoritmus kiválasztása, amelyre szintén számos példa hozható. A tapasztalatok alapján a választott jellemzőkészlet és a párok előszűrésének mértéke alapján érdemes választani. Soon és munkatársai a C4.5 (C5.0) döntési fát alkalmazták (Quinlan 1993). De található Hoste (Hoste 2005), Recasens és Hovy (Hoste 2005; Recasens–Hovy 2009) munkáiban példa memória alapú tanulóra (Daelemans–van den Bosch 2005), maximum entrópia tanulóra (Berger–Della Pietra–Della Pietra 1996; Le 2004) Yang és munkatársai (Yang et al. 2003), Hendrick és munkatársai (Hendrickx–Hoste–Daelemans 2007) vagy Kehler és munkatársai kísérleteiben (Kehler et al. 2004). A RIPPER szabályalapú tanulót (Cohen 1995) alkalmazzák Ng és munkatársai (Ng–Cardie 2002b; Ng–Cardie 2002a; Ng–Cardie 2002c), Hoste (Hoste 2005) és Uryupina (Uryupina 2006) is. Bengtson és munkatársai (Bengtson–Roth 2008) a voted perceptrons algoritmust (Freund–Shapire 1999) alkalmazza. Uryuppina (Uryupina 2006) munkájában és Versley és munkatársai munkájában (Versley et al. 2008) pedig az SVM implementációját alkalmazza (Vapnik 1995).

5.3.1.4. Egyetlen jelölt azonosítása

A negyedik kérdés végül az, milyen módszer alapján válasszon az osztályozó egyetlen jelöltet. Mivel az anaforához több főnévi csoportot is hozzárendelünk a modell építése során párként, ezért egy anaforához több antecedenst is pozitívnak ítélhet az osztályozó. Az osztályozó döntése szempontjából két bevett módszer követhető, a Closest-first és a Best-first megközelítési módszerek. A Closes-first módszer során az anaforához pozitív párként rendelt legközelebbi főnévi csoportot tekintjük az anafora antecedensének, ezt a módszert alkalmazták Soon és munkatársai is (Soon–Ng–Lim 2001). A másik módszer a Best-first, amelynek során azt a főnévi csoportot tekintjük antecedensnek, amelyet a legnagyobb valószínűséggel értékelt annak az

osztályozó. A Best-first módszert alkalmazta Ng és Cardie is kísérletük során (Ng–Cardie 2002a).

Eredmények tekintetében elmondható, hogy a Soon és munkatársai nevéhez köthető módszer igen jó eredményeket ért el a MUC-6 és MUC-7 korpuszokon. A MUC-6 korpuszon 62,6-os F-mértéket, a MUC-7 korpuszon pedig 60,4-os F-mértéket értek el. Ng és Cardie nagyobb jellemzőkészletet használó osztályozója a MUC-6 korpuszon 70,4-os F-mértéket, a MUC-7 korpuszon pedig 63,4-os F-mértéket ért el.

5.3.2. Automatikus koreferenciafeloldásra alkalmas rendszerek

A koreferenciafeloldás szempontjából problémát jelent a Mention-pair modellel kapcsolatban, hogy kevéssé veszi figyelembe a kontextusból származó információkat, illetve, hogy főleg a lokális koherenciát képes felismerni, hiszen minden anaforához egy antecedenst azonosítunk, a legközelebbit vagy a legbiztosabbra értékeltet. Ez a módszer tehát az anaforafeloldásra alkalmasabb, mint a koreferenciafeloldásra, mivel koreferenciafeloldás során általában egy entitás összes említését keressük. Cardie és Wagstaff nem kizárólag a névmásokkal foglalkozott, céljuk a főnévi csoportok koreferenciakapcsolatainak felismerése volt egy felügyelet nélküli algoritmus segítségével (Cardie–Wagstaff 1999). A koreferenciafeloldásra klaszterezési feladatként tekintettek, azaz nem kizárólag a visszautaló szóból és az antecedenséből álló párokat, hanem teljes láncokat szerettek volna azonosítani. Az ötlet abból a kiindulópontból származik, hogy a koreferens főnévi csoportok ekvivalencia osztályokat képeznek, mivel a koreferencia egy szimmetrikus, reflexív és tranzitív reláció. A megközelítésük lényege, hogy a szövegben minden főnévi csoportot egy jellemzővektorként képzeltek el, és ezen jellemzők alapján csoportosította az algoritmus a főnévi csoportokat osztályokba. Az azonos osztályokba csoportosított főnévi csoportok tekinthetők koreferensnek. A kontextusfüggetlen jellemzők két főnévi csoport eltérő osztályba való sorolását, a kontextusfüggők pedig az azonos osztályba sorolást segítették elő. A módszerüket a MUC-6 koreferencia korpuszon tesztelték, és 53,6-os F-mértéket értek el.

5.3.2.1. Entity-mention modell

A fent említett problémának a megoldására jött létre az Entity-mention koreferencia modell (Luo et al. 2004; Yang et al. 2004), amely nem csak egy antecedens, hanem a teljes koreferencialánc azonosítását tűzte ki célul. A módszer célja, hogy az osztályozó eldöntse, az adott anafora tagja-e

egy már létező koreferencialáncnak. Ennek a módszernek a segítségével nem csak egy kifejezés jellemzői alapján kíséreljük meg azonosítani a referenst, hanem a lánc összes korábbi tagja segítségével. Ebben az esetben a modell építésére szolgáló jellemzők a névmás tulajdonságaiból, a névmást megelőző összes azonosított koreferencialánc-tag tulajdonságaiból és a két csoport tulajdonságainak összehasonlításából származnak. A névmást megelőző koreferencialánc-tagok az egyes tulajdonságok tekintetében pedig mindig négy értéket vehetnek fel: 1 minden tagra igaz, 2 a legtöbb tagra igaz, 3 néhány tagra igaz, 4 egyik tagra sem igaz.

A fent ismertetett felügyelt gépi tanulási rendszerek előnye, hogy kisebb korpuszokon is felhasználhatók, így olyan nyelvek esetében is alkalmazhatók, amelyekkel kapcsolatban nem rendelkezünk több tízezer pozitív tanítópéldával a modell építéshez. Ezzel szemben a felügyelet nélküli rendszerek létrehozásához nagyobb mennyiségű adat szükséges.

5.3.2.2. CoNLL-2012 Shared Task

A CoNLL-2012 Shared Task (Pradhan et al. 2012) célja az OntoNotes többnyelvű korpuszon történő koreferenciafeloldás volt. A korpuszban angol, kínai és arab nyelvű szövegek is voltak az annotáció során pedig nem kizárólag főnévi csoportokat jelöltek, így lehetővé vált, az entitás láncokon túl az esemény láncok azonosítása is. A verseny során a MELA kiértékelési metrikát alkalmazták.

Fernandes fa reprezentáción alapuló rendszere (Fernandes–Santos–Milidiú 2012) érte el a legjobb eredményt összességében a három nyelv tekintetében. Soon és munkatársai megközelítésében (Soon–Ng–Lim 2001) a kifejezés párok egy lineáris rendszert modelleznek, amely nem elágazó faként is értelmezhető, ezzel szemben Fernandes és munkatársai megközelítési módszerének lényege, hogy a koreferencia osztályokat elágazó faként reprezentálja. A szöveg minden egyes kifejezése egy csomópont a fában a kifejezések közötti élek pedig a koreferens kapcsolatot mutatják. A gyökér csomópontból kiinduló alfák az egymással koreferens kifejezések klaszterei. A módszer az angol és az arab nyelv tekintetében is a legjobb eredményt érte el a versenyben, az angolra 63,37-es F-mértéket, az arabra pedig 54,22-es F-mértéket ért el.

Björkelund és Farkas rendszere (Björkelund–Farkas 2012) a korábban is említett, kifejezés párokból álló reprezentációt követte verem módszerrel kiegészítve. A módszer az angol nyelvre 61,24, a kínai nyelvre 59,97, az arab nyelvre pedig 53,55-ös F-mértéket ért el.

Martschat rendszere (Martschat et al. 2012) a szövegben szereplő koreferens kapcsolatokat több gráf segítségével modellezte. A gráfokban a csomópontok a kifejezések, az élek pedig a

kapcsolatok. A klasztereket mohó algoritmussal határozták meg. A módszer az angol nyelvre 61,31-es F-mértéket ért el.

A kínai nyelv tekintetében fontos még megemlíteni Chen és Ng munkáját (Chen–Ng 2012), amely szabály alapú és tanulás alapú rendszerek előnyeit ötvözve 62,24-es F-mértéket ért el, valamint Yuan és munkatársai módszerét, amely a feladatot több gépi tanulási módszerrel illetve szabály alapú döntésekkel megoldott alfeladatra osztotta, így 60,69-es F-mértéket ért el.

A versenyen azok a módszerek érték el a legjobb eredményeket, amelyek ötvözték a szabály alapú és gépi tanulási módszereket valamint részletesen és aprólékosan megtervezték, hogy a kifejezések mely jellemzőit veszik figyelembe a feloldás során.

Az OntoNotes többnyelvű korpusz és a CoNLL-2012 Shared Task részletesen kidolgozott kiértékelési módszertana lehetővé tette későbbi rendszerek kiértékelését is, valamint a Shared Taskban résztvevő rendszerekkel való összehasonlításukat is.

Durett és Klein 2013-as tanulás alapú rendszere (Durrett–Klein 2013) egyszerű, felszíni szerkezetből kinyerhető, kevés jellemző segítségével 60,3 F értéket ért el. Következő rendszerük (Durrett–Klein 2014) célja már nem kizárólag koreferencia feloldás, hanem teljes entitás vizsgálat volt. A dokumentumon belüli koreferencia feloldáson túl névelem-felismerés és entitáskapcsolás, valamint ezeknek a feladatoknak az összeegyeztetése is megtörtént. Koreferencia feloldás tekintetében a rendszer 61,7 F értéket ért el.

Björkelund és Kuhn (Björkelund–Kuhn 2014) a tanítóadatbázis teljes kihasználása érdekében Daumé és Marcu rendszerét (Daumé–Marcu 2005) implementálták, így 61,6-os F értéket értek el.

Clark és Manning rendszere (Clark–Manning 2015) a Mention-pair modellből indult ki, kiegészítve azt entitásra vonatkozó információkkal úgy, hogy az egyes említésekre vonatkozó információkat összesítették. A tanulás alapjául használt jellemzők az egyes Mention-pair modellek lesznek, így a több modell a koreferenciát több aspektusból is jellemezni tudja. A módszer segítségével nem az egyes említések, hanem maguk az entitások válnak fontos tényezővé a tanulás során, ennek segítségével 63-as F mértéket értek el.

5.3.2.3. Neurális hálók

Az utóbbi fél évtizedben több kutatás is neurális hálók segítségével készült el (Clark–Manning 2016; Lee et al. 2017; Wiseman–Rush–Shieber 2016). A módszer lényege, hogy a reprezentáció tanulása automatikusan történik, a klaszterezéshez nem szükséges előzetes kézi annotáció az azonos klaszterekhez tartozó kifejezések meghatározásához. A neurális hálózatoknak több rétege

van: bementi réteg, rejtett rétegek és a kimeneti réteg. A bementi réteg jelen esetben a nyers, elemzést nem tartalmazó szöveg, a kimeneti réteg pedig a meghatározott klasztereket tartalmazza.

Wiseman, Rush és Shieber rekurrens neurális háló alkalmazásával az egyes említések alapján entitás klaszterek meghatározására törekedtek. A módszer a névmási említések előjelzésére különösen alkalmasnak bizonyult, 64,2-es F mértéket ért el. Clark és Manning koreferencia klaszter párok vektor reprezentációján alapuló rendszere 65,3-as F mértéket ért el. Lee és munkatársai kutatásának célja end-to-end koreferenciafeloldás volt. Ehhez kizárólag a gold standard korpuszt használták, szintaktikai és kézi előelemzés nélkül, a kifejezéseket pedig vektor beágyazásokkal reprezentálták. A rendszer 68,8-as F mértéket ért el.

5.4. Automatikus anaforafeloldás a magyar nyelvben

A magyar nyelv az anaforafeloldás szempontjából különösen problémás, hiszen a legtöbb automatikus azonosításra törekvő rendszer az angol nyelvet alapul véve morfológiai és szintaktikai tulajdonságokon alapul, és ezen információk segítségével is viszonylag jó eredményeket ér el, a magyar nyelvben azonban a felszíni szerkezetből kinyerhető nyelvi információk mennyisége alacsonyabb, mint az angol nyelvben (Mitkov 2009). A morfológiai tulajdonságok közül a magyar nyelvben nem tudjuk figyelembe venni a nembeli egyeztetést, amely az angol nyelvben például erősen hozzájárul az automatikus rendszerek sikerességének növeléséhez. Mivel a magyar diskurzuskonfigurációs nyelv, ezért a szintaktikai információk is csak korlátozott mértékben állnak a rendelkezésünkre. Ugyan a topik-komment szerkezet, valamint a tematikus szerepek is segíthetnék az automatikus anaforafeloldást, ezek az információk még nem nyerhetők ki automatikusan a felszíni szerkezetből, csak manuálisan adhatók hozzá az antecedens-névmás párokhoz. Tehát a felügyelt tanulási módszerek használatát a jellemzők alacsony száma korlátozza. Ennek ellenére a magyar nyelvvel kapcsolatban számos kezdeményezésről olvashatunk, amelyek egyre jobb eredményeket produkálnak a koreferencia- és anaforafeloldás tekintetében (Lejtovicz–Kardkovács 2006; Miháltz et al. 2007; Miháltz 2012), azonban egyik kutatásnak sem kifejezetten a névmási anaforafeloldás áll a fókuszában.

Lejtovicz és Kardkovács 2006-os munkájában a *Centering Theory,* azaz központiság elméleten alapuló tudásalapú algoritmust mutat be. A központiság elmélet alapelve, hogy a diskurzusban minden egyes megnyilatkozásban egy központi téma van, a következő megnyilatkozásban szereplő anafora pedig valószínűleg erre a központi témára fog visszautalni a topikfolytonosság miatt. Ezen az alapelven működő algoritmus a BFP algoritmus (Brennan–

Friedman–Pollard 1987), amelyet a szerzők a magyar nyelvre adaptáltak. A magyarra megvalósított BFP algoritmus a Szeged Treebank 2.0-n 39,6%- os találati arányt ért el az összes anafora tekintetében. A szerzők nem tértek ki arra, hogy ezt az eredményt pontosan milyen annotációs módszerrel azonosított gold standard adatokon érték el.

Miháltz és munkatársai munkájukban tudásalapú anafora-feloldó rendszert mutattak be. Az általuk létrehozott algoritmus 4 lépésben működik: előszűrés, az antecedensek listájának előállítása, a jelöltek szűrése, antecedens kiválasztása a fennmaradó jelöltek közül. Minden egyes visszautalási típus esetében eltérő a feloldási folyamat. A névmások tekintetében az előszűrés során a zéró névmásokat, személyes névmásokat, valamint az *az* mutató névmást szűrték ki, ezeknek keresték a szövegben korábban előforduló antecedensét. A jelöltek kiválasztásának során az anafora mondata előtti második mondattól kezdve választja ki az összes NP-t az anaforát tartalmazó mondat határáig. Ezután az anafora és az antecedensjelölt számának, személyének, és két szemantikai jegynek (+/- élő, +/- ember) egyezését vizsgálták. Az utolsó lépésben mindig az alanyi szerepű névmási anaforát oldják fel először, és utána a többit, így a már korábban kötött antecedensek kiesnek a vizsgálatból. Az alanyi szerepű névmáshoz tartozó antecedens azonosításához a szerkezeti párhuzamosság heurisztikát alkalmazták, ezt felülbírálhatja az alanyi szerepben álló *az* névmás, amely alanyváltást jelöl. Amennyiben több nem alanyi szerepű antecedensjelölt NP is található, két szabály alkalmazandó: hozzáférhetőség (az oblikuszi hierarchiában), távolság (a névmáshoz közelebb eső NP a preferált). A kiértékeléshez létrehoztak egy kiértékelő-korpuszt, amely 5 általános iskolai történelemkönyvekből származó szövegből áll, az előelemzésből származó hibák miatt a korpusz a fedés kiértékelésére azonban nem alkalmas, így kizárólag a pontosságot vették figyelembe. A szövegekben a MetaMorpho segítségével először a maximális NP-ket azonosították, majd egy annotátor jelölte közöttük a koreferencia-viszonyokat. Az annotáció után a korpusz 338 antecedenssel rendelkező főnévi csoportot tartalmazott. A fenti módszer segítségével a névmások tekintetében a rendszer 68%-os pontosságot ért el. Ezt a kutatást vitte tovább Miháltz 2012-es munkája, amelyet tíz általános iskolai történelemkönyvekből kiemelt szövegrészleten értékeltek ki. A rendszer a névmások tekintetében a korábbihoz igen hasonló, 71,4%-os pontosságot ért el.

2016-ban Munkácsy és Farkas bemutatta az első statisztikai alapú módszert, amely magyar nyelvű szövegeken végzett koreferenciafeloldást (Munkácsy–Farkas 2016). Munkájukban a SzegedKoref korpuszon a HOTCoref rendszert (Björkelund–Kuhn 2014) tanították, majd a rendszer egyes moduljait a magyar korpusznak megfelelően átalakították. Az egy entitásra

vonatkozó láncok azonosításához a HOTCoref először szabályok alapján kiválasztja a lehetséges anaforajelölteket, majd felügyelt gépi tanulási módszertant követve, látens faszerkezettel reprezentálva, említési láncokat alakít ki. Négy különböző kiértékelés metrikát (MUC, BCUC, CEAFM, CEAFE) használtak a feloldó rendszer kiértékeléséhez, ezek átlaga a fedés tekintetében 37,37, a pontosság tekintetében 50,815, az F1 pedig 43,05 lett.

6. A kutatások alapjául szolgáló adatok, módszerek

A következő fejezetben az általam végzett kísérletek alapjául szolgáló két korpuszt, az ezekből készült tanító és tesztfájlok felépítését és tartalmát, valamint a tanulási kísérletekhez felhasznált jellemzők és hozzájuk tartozó címkekészleteket ismertetem, ezek mellett pedig bemutatom a jelenlegi kutatást megelőző korábbi kísérleteimet.

A gépi tanulási kísérletekhez a Mention-pair (Soon és mtsai, 2001) (lásd az 5.3.1 szakaszt) modellt használtam, amelyhez a lehetséges visszautaló névmások és a hozzájuk tartozó lehetséges antecedensjelöltekből álló párokat kellett kinyerni a korpuszból. Mivel a két korpuszban nem csak a koreferens kapcsolatok találhatók meg, hanem a kifejezésekhez tartozó morfológiai és szintaktikai elemzések is, így ezek a párok és a hozzájuk rendelt morfológiai és szintaktikai jellemzők adták a tanító és tesztfájlokat. A modell előnye az, hogy segítségével azoknak a jellemzőknek a tanulásra gyakorolt hatása egymástól függetlenül vizsgálható, amelyek a korábban ismertetett különböző elméletekben megfogalmazott elvek szerint hatással vannak az anaforafeloldásra. Ezek a jellemzők expliciten nem szerepelnek a korpuszokban, viszont automatikus eszközökkel meghatározhatók. A konstituens elemzés segítségével a főnévi csoportok, valamint a frázisok fejéhez tartozó morfológiai elemzések kinyerhetők a fájlokból. A párok első eleme olyan főnévi csoport, amely a korpuszban a morfológiai elemzés oszlopban PronType attribútummal rendelkezik, ehhez pedig a prs, dem vagy rel címkét kapta. Az antecedensjelöltek pedig a névmásokat a szövegben megelőző főnévi csoportok (NP).

A gépi tanulás során a lehetséges anafora-antecedens párok közül kell kiválasztani a korpuszban ténylegesen koreferensként jelölt párokat. Ez a kiválasztás vonatkozhat az összes koreferens pár azonosítására, vagy az anaforához legalább egy kézzel is annotált antecedens azonosítására, amely lehet a legközelebbi vagy az osztályozó által a legnagyobb valószínűségi értékkel ellátott antecedens. Kísérleteim során a névmáshoz egyetlen kézzel is annotált antecedensjelölt azonosítását tűztem ki célul, ennek során pedig megvizsgáltam az osztályozó által a legközelebbi és a legnagyobb valószínűségi értékkel ellátott antecedensként értékelt főnévi csoportot is.

6.1. A felhasznált korpuszok

A kísérlet során két korpuszt használok fel: a Szeged Koreferencia Korpuszt, amely a Szeged Korpusz (Csendes et al. 2005) koreferenciaannotált alkorpusza, valamint összehasonlításként a

KorKorpuszt (Vadász 2020). A Szeged Treebank és alkorpuszai az alábbi linken: https://rgai.inf.u-szeged.hu/node/113, a KorKorpusz pedig a következő linken: https://github.com/vadno/korkor_pilot érhető el.

A SzegedKoref Korpusz a Szeged Korpusz egy alkorpusza, ezért részletes morfológiai és szintaktikai annotációt is tartalmaz. Az eredeti Szeged Korpuszból a 8. és 10. osztályosok fogalmazásait, valamint hvg-s cikkeket tartalmazza. Mivel ezek a szövegek megtalálhatók a Szeged Korpusz egy másik alkorpuszában, a Szeged Dependencia Korpuszban (Vincze et al. 2010) is, így a függőségi elemzések is a rendelkezésemre álltak a szövegekhez. A szavak eredeti alakja mellett megtalálható a korpuszokban a lemma, a szófaj, a morfológiai elemzés, valamint a függőségi elemzés kimenete: élek, él címke, szófaji címke, továbbá a konstituens elemzés kimenete. A korpuszban minden szóhoz tartozik egy azonosítószám, ami azt mutatja, hogy az adott szó a mondat hányadik szava, ezt az értéket használja fel a függőségi elemzés is az élek megállapításához. A mondatokat üres sor választja el egymástól. Mivel a korpuszban az összes elemzett fogalmazás és cikk egy fájlban található, van egy plusz azonosító is, ami azt mutatja, hogy az adott sor melyik szöveghez tartozik.

A SzegedKoref koreferencia korpuszban az eredeti Szeged Korpuszban található információkon túl egy további oszlop is található, amely azt mutatja meg, hogy az adott szó, illetve az a frázis, amelynek a szó része, a szöveg melyik koreferencialáncába tartozik. A koreferens szavak, frázisok a korpuszban ugyanazt az azonosítót kapják, vagyis nem az anafora–antecedens párok vannak jelölve, hanem ekvivalencia osztályok.

A KorKorpuszban, hasonlóan a SzegedKoref koreferencia korpuszhoz, minden szónak van egy azonosítója, ami a mondatban elfoglalt pozícióját mutatja. Szintén megtalálható a szó eredeti formája, a lemma, a szófaji címke, morfológiai elemzés és függőségi elemzés kimenete. A koreferenciaannotáció két oszlopban található, az első oszlop azt mutatja, hogy az adott szónak hol található az antecedense, ez két szám érték: hányadik mondat, hányadik szó, kettősponttal elválasztva. A koreferencia jelölése során tehát nem a teljes antecedens van kijelölve, hanem a frázisok fejei. Ezenkívül még a kapcsolat típusa is jelölve van. Az anaforikus kapcsolatoknál a névmás típusával egyezett meg a jelölés. A korpuszban a következő névmások szerepelnek anaforikus kapcsolatban: személyes (prs), mutató (dem), kölcsönös (recip), visszaható (refl), vonatkozó (rel), birtokos (poss).

6.2. A kutatás alapjául szolgáló korábbi kísérletek és a felmerülő kérdések

Jelen fejezetben a korábban a témakörben végzett kutatásaimat és a kutatásaim során felvetett kérdéseket és az általam levont következtetéseket tárgyalom.

Az első tanulási kísérleteket egy interneten található blogbejegyzésekből, rövid cikkekből álló saját korpuszon végeztem. Ezek a rövid cikkek a következő blogokról származnak: https://prohardver.hu/fooldal/rovat/fujitsu_blog, http://webisztan.blog.hu, https://www.egyedikutya.hu/egyedi-kutya-blog, http://otthonedes.blog.hu/, http://neszeszer.blog.hu , http://konyvkritikak.blog.hu/, http://filmvilag.blog.hu , http://jateknaplo.blog.hu/, http://varosikonyha.blog.hu A korpusz 60 db szöveget tartalmazott, összesen 430 névmási visszautalást, ebből 216 vonatkozó névmási, 126 személyes névmási, 88 mutató névmási visszautalás volt.

Mivel ez egy pilot kutatás volt a későbbi kísérleteimhez, ezért a manuális annotációt egyedül végeztem el, amely során kizárólag a névmási visszautalásokat jelöltem az MMAX2 annotációs szoftver (Müller – Strube 2006) segítségével. Emellett a szövegeket a magyarlánc (Zsibrita–Vincze–Farkas 2013) parse moduljával elemeztem le, így a tanítás során felhasználhattam az MSD kódot, a Szófaji címkét, a morfológiai információkat, valamint a függőségi és konstituens elemzést. Az előelemzés és a kézi annotáció segítségével a Mention-pair technikával (Soon–Ng–Lim 2001) tanító és tesztfájlokat generáltam a korpuszban található névmás és antecedensjelölt párokból. A kísérlet célja az elérhetőségi elméletben megfogalmazott kognitív alapú jellemzők hatásának vizsgálata volt a gépi tanulás során, ezért egy pusztán morfológiai és szintaktikai jellemzőkön alapuló jellemzőkészlet eredményességét hasonlítottam össze, egy kibővített jellemzőkészlettel. A pozitív és negatív példák alapján a tanítófájlon a Random Forest tanuló algoritmussal (Breiman 2001) két osztályozót építettem. A két osztályozó teszteléséhez az alacsony számú visszautalás miatt a keresztvalidálás módszerét alkalmaztam. A korpuszt a szövegek alapján tíz részre osztottam, kilenc részből készült el a tanítófájl, és egy részből a tesztfájl, ezt a módszert pedig tízszer megismételtem, a végleges kiértékeléshez pedig az egyes tesztek átlagát használtam fel.

Az első osztályozó (Base) a következő jellemzők segítségével épült fel: 1 a névmás és antecedensjelölt közötti tagmondati távolság, 2 a névmás és antecedensjelölt közötti főnévi csoportok száma, 3 a névmás esete, száma és személye, 4 az antecedensjelölt esete, száma és személye, 5 a 3 és 4 jellemzők egyeztetésére vonatkozó értékek (0-nem, 1-igen), 6 a névmás típusa (mutató-, személyes-, vonatkozó névmás), 7 Az antecedensjelölt típusa (Np, Cp), 8 Az antecedensjelölt szófaji címkéje, kiemelt jellemzőként, ha tulajdonnév vagy névmás.

A második osztályozó (withAcc) az alábbi jellemzőkkel kiegészítve épült:

1. az antecedensjelölt hossza (szavak száma),
2. az antecedensjelölt három vagy annál több szavas (igen, nem),
3. az antecedensjelölt határozottsága (Def, Indef),
4. az antecedensjelölt alany esetű-e (igen, nem),
5. a névmás és antecedensjelölt közötti szavak száma.

A kísérlet eredményei alapján az elérhetőségi elméletben megfogalmazott elvekből generált jellemzők javítottak a tanulás sikerességén a puszta morfológiai és szintaktikai információkon alapuló tanuláshoz képest (Kovács 2019).

Base			***withAcc***		
P	**R**	**F**	**P**	**R**	**F**
31,58	13,33	18,75	38,89	15,56	22,23
52,38	18,33	27,16	68,75	18,33	28,94
55,00	22,92	32,36	63,13	25,00	35,82
53,58	25,00	34,09	63,63	32,56	43,08
75,00	39,13	51,43	75,00	39,13	51,43
50,00	27,50	35,48	61,11	27,50	37,93
70,59	28,57	40,68	68,75	26,19	37,93
57,90	32,35	41,51	71,43	29,41	41,67
55,56	32,26	40,82	57,90	35,48	44,00
54,55	23,53	32,88	68,42	25,49	37,14
55,61	**26,29**	**35,51**	**63,70**	**27,47**	**38,02**

1. táblázat A pilot kutatás eredménye

Mivel az általam készített korpusz nem tartalmazott elegendő visszautalást ahhoz, hogy tényleges következtetéseket vonjak le a kognitív elveken alapuló jellemzők hatásáról, ezért a második kísérletben a Szeged Korpusz koreferenciaannotált alkorpuszát (Vincze et al. 2018) használtam fel. A SzegedKoref koreferencia korpusz névmási visszautalásai közül, kizárólag a PRON szófaji címkével ellátott névmásokat vizsgáltam meg. A szűrés után 725 visszautalást azonosítottam, amelyek segítségével a tagmondati távolság, mint jellemző, meghatározási módjainak hatását vizsgáltam (Kovács 2020).

A kísérletekhez használt jellemzőkészlet az alábbi értékeket tartalmazta:

1. a névmás és antecedensjelölt közötti tagmondatok száma,
2. a névmás és antecedensjelölt közötti főnévi csoportok száma,
3. az antecedensjelölthöz rendelt szófaji címke
4. a névmás típusa
5. az antecedensjelölt esete, száma, személye,
6. a névmás esete, száma, személye,
7. az 5 és 6 értékeinek egyeztetésére vonatkozó információk.

A három kísérlet között a két kifejezés közötti tagmondatok számának meghatározása között volt különbség. A Baseline tesztelése során nem tettem különbséget a tagmondatok között, ezek az eredmények láthatók a 2. táblázat Baseline oszlopában. Az első tesztelésnél már figyelembe vettem a közbeékelődéseket és az alá- és mellérendelő mondatok közötti különbségeket, ezt mutatja a táblázatban az Exp1 oszlop. A második teszt során már a nagy hatókörű anaforák alapján megfogalmazott elveket is figyelembe vettem, ezt mutatja az Exp2 oszlop.

	Baseline			**Exp1**			**Exp2**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TEST1	22,41	35,14	27,37	22,31	36,49	27,69	23,53	37,84	29,02
TEST2	28,07	45,71	34,78	29,66	50,00	37,23	32,14	51,43	39,56
TEST3	29,20	43,42	34,92	28,57	42,11	34,04	30,63	44,74	36,36
TEST4	37,50	45,21	40,99	34,83	42,47	38,27	38,46	47,95	42,68
TEST5	40,19	55,13	46,49	39,62	53,85	45,65	41,18	53,85	46,67
TEST6	31,65	39,68	35,21	35,62	41,27	38,24	35,82	38,10	36,92
TEST7	36,61	61,19	45,81	41,84	61,19	49,70	39,60	59,70	47,62
TEST8	39,02	47,76	42,95	38,55	47,76	42,67	40,74	49,25	44,59
TEST9	30,85	39,19	34,52	34,04	43,24	38,10	34,02	44,59	38,6
TEST10	41,75	51,81	46,24	37,72	51,81	43,65	51,81	51,81	51,81
ÁTLAG	**33,73**	**46,42**	**38,93**	**34,28**	**47,02**	**39,52**	**36,79**	**47,92**	**41,38**

2. táblázat A tagmondatok számának meghatározását vizsgáló kísérlet eredménye

A korábbi kísérletek során az egyik gond a potenciális visszautaló névmások azonosítása volt a szövegekben, ami leginkább a tanítófájlokban megjelenő pozitív példák számára van hatással. Abban az esetben, ha kizárólag a PRON szófaji címkét vesszük figyelembe, sok olyan névmás is kizárásra kerül, ami ADV, azaz határozószói szófaji címkét kap, ezek kihagyásával csökken a pozitív tanítópéldák száma. Eredményesebbnek mutatkozik, ha azokat a szavakat keresem, amelyek rendelkeznek PronType attribútummal a morfológiai elemzésben. Ezután azt kell megvizsgálni, hogy melyek azok a névmástípusok, amelyeknek lehet antecedense a szövegben. Azokban az esetekben, ahol a morfológiai elemzés során például a PronType attribútum a Neg, v. Default címkéket kapta, biztosan nem rendelkeznek antecedenssel a szövegben.

Mivel a SzegedKoref Korpuszban ezer és kétezer közötti a névmási visszautalások száma az általam végzett szűrés után, a keresztvalidálás módszerét alkalmaztam a validálás során, azaz öt részre osztottam a teljes korpuszt, és mindig csak egy részből készítettem a teszt-, a többiből pedig a tanító fájlt.

További problémát jelent a pozitív névmás- antecedensjelölt párok száma a tesztelés során. Az első Szeged Korpuszon végzett tanulási kísérletek során a tesztfájlokban a névmáshoz antecedensként kézzel annotált első, legközelebbi főnévi csoportot kerestem. Ez azt eredményezte, hogy ha az algoritmus egy a szövegben korábban előforduló, tehát a névmástól távolabbi főnévi csoportot azonosított antecedensként, azt szintén fals pozitív találatnak kellett tekintenem annak ellenére, hogy azonos koreferencialáncban szerepeltek a korpuszban. Ez a probléma kiküszöbölhető, ha a névmáshoz az őt tartalmazó koreferencialáncban megelőző összes főnévi csoportot lehetséges antecedensének tekintem, azonban a kiértékelés során ezek közül már mindössze egy azonosítását is elegendőnek értékelem. Ez a megközelítési módszer vezet el a disszertáció elején megfogalmazott, a Best-first és Closest-first módszerek összehasonlítására vonatkozó kutatási kérdésemhez.

Megvizsgálható a tanulás sikeressége úgy, hogy az összes névmásból egységesen építek tanító és teszt fájlt. Ezzel az a probléma, hogy a vonatkozó névmási visszautalások száma nagyon magas, de ezek a visszautalások erőteljesen eltérnek a mutató vagy személyes névmási visszautalásoktól. Tehát feltételezhetően a vonatkozó névmások nagy száma miatt túláltalánosít az osztályozó, és minden visszautaláshoz a megelőző két-három főnévi csoport közül választ antecedenst. Ha a korábbi tanítási kísérleteket vesszük alapul, akkor ez a megállapítás helytállónak tűnik. Mind a tíz tesztben a mutató és személyes névmási visszautalások azonosítása volt a legkevésbé eredményes. Ha nem csak a szófaji címke alapján szűröm a névmásokat, akkor magasabb számú visszautalást kapunk, hiszen az első és második személyű személyes névmások és sokkal több mutató névmás is a fájlokba kerül, azaz nagyobb lesz a tanító és a tesztfájl is. Mivel az osztályozónak azt a bináris döntést kell meghoznia, hogy egy pár anaforikus-e vagy sem, azaz két csoportba kell sorolnia a párokat, ezért kapcsolatot keres a vonatkozó névmási visszautalás és a mutató és személyes névmási visszautalás között. Erre a problémára két megoldási lehetőség van: 1 Nem bináris osztályozást végez az algoritmus, hanem több csoportot adunk meg neki, amit fel kell ismernie. 2 Külön tanító és tesztfájlokat generálunk visszautalási típus szerint. Ezek közül a lehetőségek közül a második módszert alkalmaztam. Ezt indokolta továbbá a kognitív jellemzők hatására feltett kutatási kérdésem is, hiszen feltételezhető, hogy az egyes névmás típusok esetében eltérő lesz a jellemzők hatása.

A névmások típusonként való tanítását és tesztelését mutatja célszerűbbnek a második hipotézisem is, mely szerint a legnagyobb valószínűségi értékkel ellátott névmás-antecedens pár kiválasztása lesz a legcélravezetőbb módszer, Best-first (Ng–Cardie 2002a), a névmáshoz tartozó antecedens kiválasztása során, hiszen a névmások gyakran utalnak a szövegben messzebbre, így

pusztán a lehetséges antecedensjelölt közelségének figyelembe vétele, Closest-first (Soon–Ng–Lim 2001), fals pozitív eredményt okozhat. Az is előfordulhat azonban, hogy az egyes visszautalási típusok eltérő módon fognak viselkedni, és míg a vonatkozó névmási visszautalás esetében a Closest-first, addig a mutató és személyes névmási visszautalás esetében a Best-first módszer lehet a kiértékelésnél a célravezetőbb, mivel a személyes névmási és mutató névmási visszautalások esetében a két kifejezés közötti távolság nagyobb lehet, mint a vonatkozó névmási visszautalás esetében.

További kérdés, hogy a modellek kiértékelése során mely mérőszámokat vegyem figyelembe. Mivel a kutatás célja kifejezéspárok azonosítása, tehát kizárólag két kifejezés közötti kapcsolat keresése, így a MUC kiértékelési metrikáit alkalmaztam (lásd az 5.3. szakaszt).

A harmadik hipotézisem, hogy a tanulási kísérlethez hozzáadott nem nyelvi jellemzők javítanak a modellépítés sikerességén. A kognitív jellemzők célja a névmásokhoz tartozó antecedensek pontosabb azonosítása a tanulási kísérletek során, tehát az várható, hogy a recall, azaz fedés értékeken javítanak a jellemzők. Ez azt jelenti, hogy azokhoz a névmásokhoz, amelyekről egyébként is tudtuk, hogy szükséges antecedenst keresni hozzájuk, nagyobb valószínűséggel azonosítja a megfelelő antecedenst. Abban az esetben, ha a tesztfájlokban olyan névmások is szerepelnek, amelyekhez nem szükséges antecedenst azonosítani, a precision, azaz pontosság értékek két módon is csökkenhetnek. Egyrészt ronthat a modell eredményén, ha olyan névmáshoz azonosít antecedenst, amely nem volt visszautaló névmás, másrészt, ha olyan névmáshoz azonosít antecedenst, amely visszautaló volt, de hozzá helytelenül. Mindemellett azt is figyelembe kell venni, hogy a kognitív jellemzők a korábban ismertetett elméletek alapján nem járulnak hozzá ahhoz, hogy eldönthessük, egy névmás visszautaló-e vagy sem.

6.3. A korpuszok egységesítése

Ahhoz, hogy a két korpuszt egységesen fel tudjam használni, meg kellett vizsgálnom, milyen információk hogyan vannak jelölve bennük, az eltéréseket pedig vagy egységesítenem kellett vagy a későbbiekben nem használhattam fel a tanulás és tesztelés során.

Az olyan információkat tartalmazó oszlopokat, amelyek hiányoztak valamelyik korpuszból nem vettem figyelembe mivel egységes tanító és tesztfájlokat szerettem volna generálni. Ilyen információ például a KorKorpuszba annotált visszautalási típusok, ez azonban nem okozott gondot, hiszen a célom kizárólag a névmáshoz tartozó antecedens azonosítása, így a visszautalási típusa minden esetben névmási lenne.

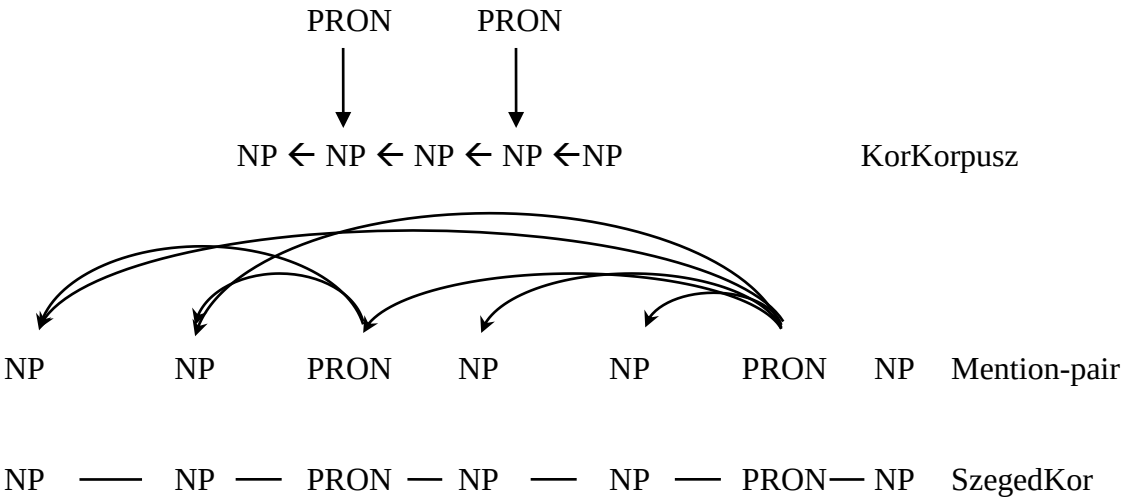
A két korpuszban használatos szófaji címkék megegyeztek az ’X’ és ’Y’ kategóriákat kivéve. Az ’X’ a Szeged Koreferencia Korpuszban a hibát, míg a Korkorpuszban a különböző írásjeleket, illetve a zérónévmásokat jelöli. Az ’Y’ a Szeged Koreferencia Korpuszban a rövidítéseket jelöli, a Korkorpuszban pedig nincs ilyen szófaji címke. Ezt a két címkét tehát a tanulás során nem vehettem figyelembe, ezeket a feldolgozás során hiányzó értékként értelmezte az algoritmus („?”). Erre azért is volt szükség, mert attól, hogy egy szó nyelvtanilag hibásan lett leírva, még lehet antecedens, de ha ’X’ marad, akkor közös csoportot generálunk ezekből a szavakból, és az elemző összefüggést próbál majd keresni közöttük. A SzegedKoref Korpuszban megtalálható még egy INTJ szófaji címke, ami a *nos, sajnos* típusú szavakat jelölte, és a Korkorpuszban nem volt megtalálható (ott ezek leginkább ADV-nak vannak jelölve). Ezeken nem változtattam, mivel a névmási anaforafeloldás szempontjából nem befolyásolják az eredményeket.

A kutatás egyik célja a SzegedKoref Korpuszon épített modell kiértékelése a KorKorpuszon, ezért egységes tanító és tesztfájlokat generáltam a korpuszokból. Mivel a kutatás egyik kérdése a felszíni szerkezetből kinyerhető kognitív alapú jellemzők hatása a tanítás sikerességére, így a felszíni szerkezetben nem megtalálható zérónévmásokat nem vizsgáltam a kísérletek során. Zérónévmások hiányában a KorKorpuszban lényegesen kevesebb névmási visszautalás maradt, ezért a SzegedKoref korpuszhoz igazodtam az előfeldolgozás során. A KorKorpuszban található dependenciaannotáció kimenete alapján meghatároztam a frázisokat: NP, AdvP, PRONP, CP stb. A CP címkét megtartottam, az összes többit pedig összevontam egy NP címkébe, mivel a Szeged Koreferencia Korpuszban is csak ez a két kategória volt jelölve a konstituens elemzésnél. A fejekhez rendelt annotációkat kiterjesztettem az őket tartalmazó teljes frázisra, így ugyanolyan intervallumokat kaptam, mint amilyenek a Szeged Koreferencia Korpuszban találhatóak. Majd ezeket az azonosítókat rendeltem az antecedensükhöz is.

Oszlop száma	Oszlop funkciója
1	szövegre utaló ID
2	szóra utaló ID
3	szóalak
4	lemma
5	Szófaji címke (ADJ, ADP, ADV, AUX, CONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, VERB)
6	morfológiai elemzés
7	dependencia él
8	él címke
9	konstituens elemzés (NP, CP)
10	koreferenciaannotáció

3. táblázat A két korpuszban megjelenő információk és elhelyezkedésük

A két korpuszban a névmási visszautalások sem egységesen vannak jelölve, ennek oka, hogy a SzegedKoref korpusz egy koreferencia korpusz, amelyben az ekvivalencia osztályokat azonos azonosítóval látják el, a KorKorpuszban azonban a névmási visszautalások külön vannak a koreferencialáncokhoz annotálva. Ezek az eltérések azonban nem okoznak gondot, hiszen a Mention-par technika alapján párokat generálunk. Az eltéréseket a következő ábrák mutatják.



2) ábra Anaforikus kapcsolatok jelölése a korpuszokban és a Mention-pair technikában

6.4. A névmások azonosítása

A korpuszok egységesítése után az első feladat a korpuszban a névmások azonosítása volt. A használt korpuszokban a névmások nem azonos módon voltak jelölve, valamint az egyes korpuszokon belül is többféle jelöléssel lettek ellátva.

A SzegedKoref korpuszban a névmások PRON vagy ADV szófaji címkéket kaphattak. PRON szófaji címkét kaptak azok a névmások, amelyek a főnevekhez hasonlóan viselkednek, helyettesíthetik azokat. Nagyon sok névmás azonban, mint például az *ekkor*, *azóta*, nem PRON-ként jelenik meg a korpuszban, hanem ADV szófaji címkével. Hogy az anaforafeloldás minél szélesebb körű legyen, ezeket is figyelembe vettem. A lehetséges visszautaló névmások azonosításához tehát nem a szófaji címkét, hanem a morfológiai elemzésben megtalálható PronType attribútumot vettem figyelembe. A PronType attribútum a következő értékeket veheti föl: Prs, Dem, Rel, Rcp, Ind, Int, Tot, Neg, Default, Art, v. Ezzel kapcsolatban három további gond merül fel.

1. A SzegedKoref korpuszban nem csak Prs, hanem PrsPron címke is megtalálható. Ezek személyragozott névutók vagy esettel ellátott névmások. Mivel ezek az esetek visszautalás szempontjából nem térnek el a személyes névmástól, egységesen kezeltem őket.

2. A visszaható névmások nem kaptak PronType címkét a morfológiai elemzésben. A *maga* kifejezés lehet az E/3 személyes névmás is, de ebben az esetben a PronType=Prs jegyet kapja, ha viszont visszaható névmás, akkor Reflex= yes vagy Reflexive= yes címkét. Mivel a két korpuszban nem voltak nagy számban visszaható névmások, és mivel ezeket különböző módon kezeli a két korpusz, ezért nem a lehetséges visszautaló névmások kigyűjtés során szabtam meg ezt a plusz kitételt, hanem a két korpuszban cseréltem le ezt a két jegyet egységesen PronType=Refl címkére.

3. További gondot jelentett a PronType címkét kapó szavak szűrése, hiszen nem mindegyik névmástípus lehet visszautaló. A nyelvészeti szakirodalom alapján öt típust megkérdőjelezhetetlenül potenciális visszautaló névmásként kell kezelnem:

	Kód	Szeged Koreferencia	KorKorpusz
személyes névmás	Prs és PrsPron	982	90
mutató névmás	Dem	743	114
vonatkozó névmás	Rel	825	325
kölcsönös névmás	Rcp	14	8
visszaható névmás	Refl	25	15

4. táblázat A potenciális visszautalások 5 típusa

Azonban a számok alapján megfigyelhető, hogy a kölcsönös és visszaható névmási visszautalásokból a két korpuszban nincs elegendő példa a tanuláshoz és a teszteléshez, így ezeket végül nem vettem figyelembe.

A személyes névmásokkal kapcsolatban a nyelvészeti szakirodalom általános álláspontja, hogy az első személyű és második személyű alakokhoz szintén nem szükséges antecedenst keresni, mivel azok a szövegvilágon kívülre, a beszélőre vagy a beszélőt bennfoglaló csoportra, illetve a hallgatóra vagy a hallgatót bennfoglaló csoportra utalnak. Mivel azonban a SzegedKoref Korpuszban is és a KorKorpuszban is jelölve vannak az ilyen típusú visszautalások, én is figyelembe vettem őket. A következőkben a példák a Szeged Korpusz koreferenciaannotált alkorpuszából származnak.

25) És **én** örömmel hagytam el a stadiont, a rendőrök elválasztották a két szurkoló tábort, mi az UTE táborral mentünk. És az úton végig hazafelé énekeltük a Fradi indulókat. És mikor hazaértem, elmeséltem szüleinek az élményeimet. Hát ez volt az **én** legérdekesebb napom.

26) Na márpedig én kitaláltam a **te** nevedet, most találd ki az enyémet. Találgassa, mi is lehet, Niki, nem, Kati, nem, Linda, nem, Petra, nem, Melinda, nem, Zsanett, hasonló, de nem Éva, nem, Betti, nem, Zsuzsi, igen, végre kitaláltad, könnyebb volt, mint a **tiéd** Nick, ha hamarabb tudom, minden ajándékot megcímezek.

A következőkben azokat a névmástípusokat vizsgálom meg, amelyek a nyelvészeti szakirodalom szerint nem referálnak, azonban a két korpusz valamelyikében vagy mind a kettőben koreferencialáncban szerepelnek. Azt, hogy az adott névmástípust végül potenciális visszautaló névmásnak tekintem-e, az alapján döntöm el, hogy a korpuszokban milyen arányban van az előfordulások száma a koreferencialáncban való előfordulások számával, a sikeres tanuláshoz és teszteléshez úgy vélem, legalább az előfordulások negyedének visszautalónak kellene lennie. Azt is figyelembe kell vennem, hogy a koreferencialáncban való előfordulás nem garancia a visszautalásra, az is előfordulhat, hogy az adott névmás antecedens, illetve azt, hogy a keresztvalidálás során ez a szám tovább csökken, hiszen a későbbiekben öt részre osztom majd a korpuszt.

6.4.1. Határozatlan névmások

Ha a PronType attribútum az Ind címkét kapja, az a határozatlanságot mutatja pl. *néhány, valamilyen*. A teljes SzegedKoref korpuszban 621 ilyen címkével ellátott névmás található,

amely teljes fráziskén, NP-nek jelölve fordul elő, tehát potenciálisan visszautalhat. Az összes közül 36 tagja koreferencialáncnak, de ezek között előfordulnak olyanok, amelyek nem önállóan szerepelnek a koreferencialáncban, hanem egy szerkezet tagjaként, pl. *mások életéről szóló könyveket* kifejezés esetében a *mások* önálló NP, de az őt bennfoglaló teljes NP lesz koreferencialánc tagja. A teljes korpuszban 18 darab határozatlan névmás található, ami nem szerkezetben, hanem önmagában fordul elő, és mind a 18 visszautal vagy antecedense egy visszautalásnak. A KorKorpuszban 14 darab Ind címkével rendelkező névmás található, amiből 6 darab szerepel koreferencialáncban, ebből 3 szerepel önmagában, tehát nem egy nagyobb szerkezet tagjaként. Ez azt jelenti, hogy a két korpuszban összesen 635 darab potenciális visszautalás között, azaz az összes olyan kifejezés, amely NP és PronType=Ind, 21 tényleges visszautalás található, tehát azok a névmások, amelyek nem szerkezetben fordulnak elő és megtalálhatók koreferencialáncokban, ami mind tanítás, mind tesztelés tekintetében igen kevésnek mondható, ezért az Ind címkével ellátott névmásokat nem tekintettem potenciális visszautaló névmásnak.

27) Vége lett az első órának, odajött **néhány csaj** hozzám és elkezdtek velem beszélgetni. **Némelyik** úgy bánt velem, mint ugyanolyan lánnyal, mint ők, de a többiek kimutatták, hogy mennyire gyűlölnek. (8oelb.33)

6.4.2. Kérdő névmások

Int címkét a kérdő névmások kaptak a morfológiai elemzés során. A SzegedKoref korpuszban összesen 512 ilyen jeggyel ellátott névmás található, amely teljes frázist alkot, ebből 5 darab található koreferencialáncban, és kizárólag 1 olyan van közöttük, amelyik nem szerkezetben, ez pedig egy mutató névmási visszautalás antecedense. A KorKorpuszban 32 Int címkével ellátott névmás közül 6 darab található koreferencialáncban, és ezek közül egyik sem önállóan, tehát minden esetben a névmást bennfoglaló teljes szerkezet az, amelyik visszautal vagy antecedens. Mivel a két korpuszban egy darab Int címkével rendelkező névmás sem található, amely önállóan, tehát nem egy nagyobb frázis tagjaként, visszautalna, ezért az Int címkével elemzett névmásokat nem vettem potenciális visszautaló névmásnak.

28) 1997 nyarán néhány barátommal megbeszéltük, hogy elmegyünk valahova biciklizni. Már csak az volt a kérdés, hogy **hova**. **Azt** is gyorsan eldöntöttük.

6.4.3. Általános névmások

A Tot címkét az általános névmások kapják. Ebből 990 olyan található a SzegedKoref korpuszban, amely frázist alkot, tehát NP-ként van annotálva, és ebből 111 tagja koreferencialáncnak. A 111 előfordulásból 95 olyan van, amelyik nem egy nagyobb szerkezet tagjaként, hanem önálló fráziskénti utal vissza vagy antecedens. A KorKorpuszban 18 NP található, amely névmás és Tot címkével rendelkezik, ezek közül 3 tagja koreferencialáncnak, és ebből 2 olyan van, amelyik nem szerkezetben, hanem önálló fráziskénti utal vissza vagy antecedens. Tehát a Tot címkével ellátott névmások közül a két korpuszban 1008 potenciális visszautaló névmás található, amelyből 97 önálló fráziskénti is koreferenciálánc tagja. Mivel ebben az esetben is kisebb mennyiségű visszautalásról lehet szó, hiszen a 97 előfordulás között azok az esetek is ott vannak, amelyek antecedensek, de nem utalnak vissza, a Tot címkével ellátott névmásokat sem kezeltem potenciális visszautaló névmásként.

29) Másnap szép napos délelőtt volt, s mindenki megérkezett a találkozóhelyre. Elindultunk a kiszemelt hely felé, a Karancsra. Szép, nyugodt tempóban bicikliztünk. **Senki** sem sietett, mert időnk volt, s a tájat is jól szemügyre vette **mindenki**.

6.4.4. Tagadó névmások

Neg címkével a tagadó névmások rendelkeznek. A SzegedKoref korpuszban 1472 Neg címkével ellátott NP található, amelyből 6 szerepel koreferencialáncban, de ezek közül mindegyik egy nagyobb frázis tagja. A KorKorpuszban 153 Neg címkével szereplő NP-ből 8 tagja koreferencialáncnak, és szintén mindegyik egy nagyobb, őt bennfoglaló frázis tagja. Tehát a Neg címkével ellátott névmások nem potenciális visszautaló névmás jelöltek.

30) A fiúk horgásztak, aki meg nem, az sétált valahol vagy még a szobájában volt.

6.4.5. Igekötőszerű névmások

A PronType attribútum a Default címkét kapta az igekötőszerű szavak esetében, ilyenek például: *haza, be, le, fel*. A teljes korpuszban 1343 Default névmási címkével ellátott NP található, ebből 34 tagja koreferencia láncnak, 28 teljes fráziskénti. A KorKorpuszban nem található ilyen címkével ellátott névmás. A fenti két okból kifolyólag a Default névmási címkével ellátott frázisokat nem tekintem potenciális visszautaló névmásnak.

31) Éjjel indultunk **haza** és hajnali 3 h fele értünk **haza**.

6.4.6. Határozói igenevek

A Szeged Koreferencia Korpuszban v címkét határozói igenevek kaptak. 332 NP közül 3 darab van koreferencialáncban, és ebből 2 darab önállóan, tehát nem egy bennfoglaló szerkezet tagjaként.. A KorKorpuszban nem található ilyen címkével ellátott névmás. A fenti két okból kifolyólag a v és PronType címkével egyaránt rendelkező kifejezéseket sem tekintem potenciális visszautaló névmás jelölteknek.

32) Amikor elindultunk, kiszámoltuk, hogy 6 óra körül kényelmesen **hazaérkezünk**. Csakhogy a nagy számolgatás közben véletlenül letértünk az útról és az új utunkon haladtunk tovább. 3 óra volt és még semmi ismerőset nem láttunk. Megálltunk, körülnéztünk és csak fákat láttunk mindenhol. Nagyon megijedtünk és megpróbáltunk visszajutni a tóhoz. 2 óra barangolás után ismét megpillantottuk a nagy vizet. Megint elindultunk a helyes úton és most már semmi másra nem figyeltünk, csak arra, hogy megmaradjunk a keskeny úton. Amikor már észrevettük a házakat, nagyon megörültünk és elkezdtünk szaladni. Hazaérkezve szüleinket nagy félelemben találtuk, mivel 6 óra helyett 11 órakor érkeztünk **haza**.

6.5. Az antecedensjelöltek azonosítása

A potenciális antecedensjelöltek tekintetében el kell választani egymástól a tanító és tesztfájlokat a felépítésük tekintetében. Az első kísérletek során minden az adott potenciális visszautaló névmást szövegben megelőző főnévi csoport potenciális antecedensjelölt volt, tehát a jelölteken nem szűrtem morfológiai vagy szintaktikai szabályok segítségével. A későbbi kísérletek során azonban két szűrési feltételt is megfogalmaztam az antecedensek tekintetében Az első a személyes névmások esetében a személyjegy alapján történő egyeztetésre vonatkozott, erről részletesebben írok a későbbiekben. A második szűrési feltétel mind a három névmástípusra vonatkozott, ehhez létrehoztam egy listát azokból a kifejezésekből, amelyek főnévi csoportok, de biztosan nem antecedensek, ezt a listát figyelembe véve néhány kifejezést kizártam a jelöltek közül, ezek a kifejezések a következők: *és, is, csak, még, mégis, de, -e, hát, kb., pl., hogy, már*.

A tanító fájlok összetételével kapcsolatban kísérleteket végeztem, ezért az ezekre vonatkozó információkat a későbbiekben, a kísérleteknél közlöm.

6.6. Tanító és tesztfájlok létrehozása

A tanító és tesztfájlok több módon is létrejöhetnek az alapján, hogy milyen típusú példákat szeretnénk, hogy tartalmazzanak. A fájlokhoz az első lépés minden esetben kilistázni a potenciális visszautaló névmásokat a korpuszból.

A potenciális visszautaló névmások morfológiai elemzésében (a korpuszok 5. oszlopa) megtalálható a PronType= Prs, Dem, Rel címke, és a konstituens elemzés alapján teljes frázis, vagyis van nyitó és csukó zárójel is a konstituens elemzés oszlopban, ami lehet NP vagy ADVP is.

Ezen a ponton két lehetőség van a tanító és a tesztfájlok felépítésére: 1 Tartalmaz minden potenciális visszautaló névmást, így olyat is, amelynek egyáltalán nem lesz antecedense a szövegben. Ezzel a módszerrel az osztályozónak azt is fel kell ismernie, ha egy névmás nem utal vissza. 2 Csak olyan visszautaló névmásokat tartalmaznak a fájlok, amelyeknek kézzel is annotált antecedense van. Ebben az esetben a modellnek kizárólag a helyes antecedenst kell felismernie. Az első módszer előnye, hogy pusztán morfológiai és szintaktikai előelemzéssel elvégezhető, hiszen egyedül azt kell felismerni a fájl létrehozásához, hogy az adott kifejezés névmás, hátránya viszont, hogy az épített osztályozónak nem csak antecedenst kell azonosítani, hanem azt is fel kell ismernie, ha egy névmáshoz egyáltalán nem szükséges antecedenst azonosítani. A második módszer előnye, hogy az osztályozónak kizárólag a helyes antecedenst kell azonosítania, viszont hátránya, hogy ehhez egy automatikus vagy manuális előelemzés szükséges, amely előzetesen kizárja a nem visszautaló névmásokat. Éppen ezért is a későbbiekben a kísérletek során én az első módszert alkalmazom, és a tesztfájlok felépítése során nem veszem figyelembe a manuális koreferenciaannotációt, azaz minden névmást potenciális visszautaló névmásnak tekintek.

A tesztfájlok felépítéséhez a potenciális névmást megelőző összes főnévi csoportot kilistázom mint antecedensjelöltet, és egyesével a névmáshoz rendelem.

Abban az esetben, ha a korpuszban kézzel össze voltak indexelve, pozitív pár volt, ha nem, akkor negatív. Később a Closest-first (Soon–Ng–Lim 2001) vagy a Best-first (Ng–Cardie 2002a) módszerrel az egy névmáshoz rendelt összes pozitív pár közül egyet választok ki antecedensként.

A tanítófájlok esetében a névmásokhoz tartozó antecedensjelöltek hozzárendelésére pozitív példák esetében szintén két mód van: 1 Mivel a névmási anaforafeloldás során a cél kizárólag egy antecedens azonosítása, így egy antecedens van hozzá pozitív párként rendelve, jellemzően a legközelebbi. 2 Az összes névmást láncban megelőző kifejezése pozitív példaként van

feltüntetve a fájlokban. Ezzel a módszerrel növelhető a tanítófájlokban a pozitív példák száma, illetve valószínűbbé válik, hogy az osztályozó távolabbi visszautalásokat felismer.

Abban az esetben, ha a tanítás során a névmáshoz csak egy antecedenst jelölünk pozitív példaként, kezelni kell azokat az eseteket, amelyekben a névmás több antecedensre is visszautal. Ez jelen esetben két módon okozhat gondot. Egyrészt mivel a KorKorpuszban a koreferenciaannotációt kiterjesztettem a frázisokra, ezért ebben az esetben egymásba ágyazott visszautalások keletkeztek, ami automatikusan növelte a pozitív példák számát. A következő példa a KorKorpuszból származik.

33) Három hónap telt el az újságíró házaspár, Sagar Sarwar és (felesége (Meherun Runi)) meggyilkolása óta.

A fenti esetben mind *a felesége Meherun Runi*, mind a *Meherun Runi* NP-k pozitív párként lettek kilistázva. Ha azonban csak egy pozitív példát szeretnénk felhasználni a tanítás során, akkor ezek közül a lehetőségek közül választani kell valamilyen szempont alapján. Ebben az esetben mindig a névmáshoz legközelebbi és legnagyobb szerkezetet vettem figyelembe, a fenti példában tehát az antecedens a *felesége Meherun Runi* lenne.

A másik eset, ha a névmás több koreferencialáncban is megtalálható, például a többes számú visszautaló névmások esetében a Szeged Koreferencia Korpuszban '|' jellel van elválasztva egymástól a két (vagy több) ekvivalencia osztály azonosítószáma. Ebben az esetben, ha |-al elválasztva több azonosító szám van az utolsó oszlopban, akkor csak azt kell figyelembe venni, ami zárt (), tehát nem egy bennfoglaló szerkezet tagja. Ha pedig mind a kettő zárt, akkor minden esetben az annotáció alapján elsőt vettem figyelembe. A Mention-pair módszer egyik hátránya az anaforafeloldás szempontjából, hogy egy antecedens azonosítása esetén nem határozható meg előre mely névmások utalhatnak vissza több antecedensre is. Ilyen eset, amikor például az *ők* külön-külön visszautal Petire is és Marira is, de a tesztfájlban nem határozható meg előre, hogy melyik névmásokhoz szükséges a modellnek két vagy több antecedenst azonsítania.

34) **Péter** megérkezett, de **Mari** még nem, bár **ők** mindig külön érkeznek.

Megadható lenne, hogy feltételezze ezt minden többes számú névmásról, de az sem lenne helytálló, mivel előfordul, hogy a többes számú névmásnak is csak egy antecedense van: *ők – a diákok*, de az is előfordulhat, hogy a két antecedense *és*-sel összekapcsolva közösen is egy NP-t alkot: *ők - a diákok és a tanárok*. Egy többes számú névmás tehát visszautalhat egy többes számú

antecedensre, vagy két többes számú antecedensre vagy két egyes számú antecedensre vagy egy többes számú és egy egyes számú antecedensre.

Tehát ha a többes számú visszautaló névmás visszautal egy másik többes számú szerkezetre *ők – a diákok*, azt lehetséges, hogy azonosítani tudja az osztályozó, ha azonban két antecedenshez van hozzárendelve azonos azonosítóval, akkor csak az elsőt, ha pedig különböző azonosítóval, akkor is csak az elsőt tudja azonosítani.

A negatív példák a tanítófájlokban a névmás és a hozzá kézzel is annotált antecedense között elhelyezkedő főnévi csoportokból képződnek. Abban az esetben, ha csak egy antecedens van pozitív példaként feltüntetve a tanítófájlban, a negatív példák is sokkal kisebb arányban fordulnak elő. Abban az esetben, ha azokat a névmásokat is figyelembe szeretnénk venni, amelyeknek nincs antecedense, meg kell szabni egy hatókört, amiben hozzárendelhetők a megelőző főnévi csoportok negatív példaként.

Miután meghatároztuk az anafora-antecedensjelölt párokat, a következő lépés a párokat meghatározó jellemzők kinyerése a korpuszokból. Ezek a jellemzők pedig a szintaktikai és morfológiai elemzések címkéiből, valamint a felszíni szerkezetből származnak.

6.7. Tanítás során felhasználható jellemzők

A tanulási kísérletek során meg kellett határoznom egy alap jellemzőkészletet, amelyhez viszonyítottam a későbbiekben a kognitív alapú jellemzők hozzáadásával végzett tanulás eredményeit. A következő két fejezetben ezeket a jellemzőket ismertetem úgy, hogy először külön bemutatom az alap jellemzőkészlet elemeit, majd pedig a további, kognitív alapon megfogalmazott jellemzőket. A bemutatás során kitérek arra, hogy a korpusz elemzéseiből származik-e az információ, vagy további következtetési szabályok segítségével jöttek létre, emellett azt is, hogy milyen elméleti alapja van a jellemző relevanciájának, és hogy mennyire pontosan implementálható az adott jellemző számítógépes környezetbe.

6.7.1. Alap jellemzőkészlet

Felügyelt gépi tanulási kísérletek során az irányadó elv a jellemzőkészletre nézve, hogy legyen informatív, és ne tartalmazzon túl sok jellemzőt, mert az maga után vonja a túltanulás lehetőségét. Már korábban kitértem arra is, hogy a jellemzők funkciójukat tekintve három csoportba oszthatók: 1) azokra, amelyek a visszautaló szót jellemzik 2), azokra, amelyek az antecedensjelöltet jellemzik és 3) azokra, amelyek a két kifejezés közötti kapcsolatot. A magyar

nyelvvel kapcsolatban leginkább a morfológiai előelemzés kimenetéből lehet kiindulni. A névmások minden esetben egyszavas kifejezések, tehát a hozzájuk rendelt információ egyértelmű, az antecedensjelöltek azonban lehetnek többszavas kifejezések is, ezekben az esetekben az antecedensjelölt fejéhez rendelt morfológiai és szintaktikai információkat vettem figyelembe. Az antecedensjelöltek fejét a függőségi elemzés segítségével határoztam meg.

Mindenekelőtt a névmással kapcsolatban a morfológiai elemzés során megállapítható a 'PronType' attribútum, ami az adott névmás típusát mutatja. Erre az attribútumra a lehetséges visszautalások kigyűjtése, azaz a tanuló és tesztfájlok generálása során van szükség. Abban az esetben, ha az összes névmási visszautalásra egy tanulási kísérlet keretein belül építünk osztályozót, érdemes lehet megadni a névmás típusát is címkeként, azonban ha csak kizárólag valamelyik névmástípushoz tartozó antecedens azonosítása a cél, akkor minden visszautaló névmás azonos címkével rendelkezne, így nem szükséges ezt az attribútumot figyelembe venni. A következő példákban a *neki* PronType=Prs, azaz személyes névmási címkét, az *ott* PronType=Dem, azaz mutató névmási címkét, az *ami* pedig PronType=Rel, azaz vonatkozó névmási címkét kap a morfológiai elemzés során.

35) [Egy 45 cm-es pontyot]$_i$ fogtam. Nagyon örültem neki$_i$

36) Egy-két órán keresztül csak kis halakat fogtam, de tudtam, hogy [ahol kis halak vannak]$_i$, ott$_i$ nagyobbak is.

37) Elindultam otthonról [a tó]$_i$ felé, ami$_i$ két-három kilométerre volt.

A következő alfejezetekben mind a visszautaló szó, mind az antecedensjelöltek esetében a morfológiai és szintaktikai elemzéséből kinyerhető információkat mutatom be. Ezekről a jellemzőkről továbbá megállapítható még az is, hogy azonosak-e a két kifejezés esetében vagy sem. Ezek az egyeztetési attribútumok, amelyek mindig két értéket vehetnek fel: igen, nem. Abban az esetben, ha valamelyik attribútummal valamelyik kifejezés nem rendelkezik, akkor a tanulás során hiányzó információként jelöljük egy kérdőjellel (?), ami azt eredményezi, hogy az egyeztetési attribútuma is '?' címkét kap. A (38) példa esetében az *a filmeket* és az *őket* kifejezések például egyaránt tárgyesetűek és többes számúak, ezért ezek az egyeztetési jegyeik az 1 értéket kapták, azonban amíg az *őket* kifejezésről tudjuk, hogy harmadik személyű személyes névmás, addig az *a filmeket* kifejezésnek nincs személyjegye, ezért '?' értéket kap, ahogy a személyjegy alapján történő egyeztetésre utaló attribútum címkéje is '?' lesz.

6.7.1.1. Case, SameCase

A ’Case’ attribútum jelöli a kifejezés esetét, ami lehet alany, tárgy, birtokos, eszköz… Mivel a magyarban nem kötött a szórend, ezért az esetrag az egyik kiindulópont, amelynek a segítségével megragadható az adott kifejezés mondatban betöltött szerepe. Az eset jellemző a következő értékeket veheti fel: *Ine, Nom, Acc, Sup, Ins, Sub, Dat, Tra, Ill, Abs, Gen, Ela, Abl, Ade, All, Del, Ter, Ess, Cau, Tem, Dis*. A leggyakoribb ezek közül *a Nom*, azaz az alanyeset, mint az *én és a barátnőm* kifejezés esetében, és az *Acc*, azaz a tárgyeset, mint az *a filmeket* kifejezés esetében.

38) Elindultunk haza miután kivettük [**a filmeket**]$_i$ és meg is néztük **őket**$_i$.
39) Már nagyon vártuk, hogy felérjünk a helyre, de [**én és a barátnőm**] lemaradtunk, a többiek pedig elhagytak [**bennünket**].

6.7.1.2. Number, SameNumber

A ’Number’ attribútum a kifejezés számát jelöli, tehát azt, hogy egyes számú vagy többes számú a kifejezés. Az attribútum így két címkével rendelkezhet, amelyek közül a *Sing* jelöli az egyes számot (40), a *Plur* pedig a többes számot (38). Ebben az esetben kizárólag a morfológiai számot tudtam figyelembe venni, tehát *a két kutya* kifejezés is Sing értéket venne fel. Szintén megvizsgáltam, hogy a névmásra és az antecedensre vonatkozó értékek azonosak-e vagy sem.

40) **Gábor**$_i$ a sziget mellé dobott, **ő**$_i$ csukázott.

6.7.1.3. Person, SamePerson

A ’Person’ attribútum a személyjegyre vonatkozik, első-, második- vagy harmadik személyű lehet az adott kifejezés. Az attribútumhoz rendelhető címkék az 1, 2 és 3. A (41) példában a *nekem* és *engem* kifejezések első személyűek, az *akivel* és *ő* pedig harmadik személyűek.

41) Egy napon olyaskivel társalogtam, **akivel**$_i$ eddig nem lehetett. Sokáig beszélgettünk, mikor azt mondta **nekem**$_j$, hogy már régóta ismer **engem**$_j$ és hogy már a születésem napján **engem**$_j$ köszöntött. Ez a meglepetés tényleg meglepett **engem**$_j$. Miért pont **ő**$_i$ köszöntött, kit$_i$ most ismertem meg.

6.7.1.4. PosTag, Pron, Propn, SamePosTag

Mind a két kifejezésre megállapítható attribútum még a POS Tag, ami azt mutatja, hogy az adott token melyik szófajba tartozik. Az attribútumhoz rendelhető címkék: ADJ, ADP, ADV, CONJ, DET, NOUN, NUM, PRON, PROPN, VERB, SCONJ, PUNCT, INTJ, AUX. A POS Tag-ek segítségével külön kiemelt, bináris jellemzők is megfogalmazhatók, például a 'PRON' címke arra utal, hogy a kifejezés névmás, például a (41) példában mind a két esetben az antecedens és a visszautaló szó is PRON címkét kap. Ha már az antecedensjelölt is névmás, az utalhat arra, hogy a kifejezés referense a szöveg fő témája. A 'PROPN' címke arra, hogy az antecedens tulajdonnév, ilyen például a (40) példában a *Gábor*. A visszautalás során hasznos lehet külön jelölni azt is, ha egy antecedensjelölt tulajdonnév, hiszen akkor személyt, intézményt vagy helyet jelöl, a tulajdonnévvel, azaz a specifikus kifejezéssel való utalás pedig kognitív alapú jellemző lehet. Egyes POS Tageket kitüntetett jegyként kiemelni érdemes lehet még a tanuló algoritmus működése miatt is, erre a későbbiekben külön ki fogok térni.

6.7.1.5. Subj, Obj, AgrSubj, AgrObj

A dependenciaelemzés során az élekhez rendelt címkék közül szintén bináris jellemző képezhető a 'SUBJ', illetve az 'OBJ' jegyekből, amelyek segítségével az esetragokkal együtt megragadhatók a már korábban bemutatott magyar nyelvre vonatkozó alanyváltással kapcsolatos szabályok. A (42) példában az *én* alanyesetű és SUBJ jegyet kap, a második tagmondatban azonban alanyváltás történik, és az *enyémet* visszautaló névmás tárgyesetű és OBJ jegyet kap. Az AgrSubj és AgrObj bináris jegyek a két kifejezés SUBJ és OBJ jegyeinek egyeztetésére vonatkoznak.

42) Na márpedig **én**$_i$ kitaláltam a te nevedet, most találd ki az **enyémet**$_i$.

6.7.2. Kognitív alapon megfogalmazott jellemzők implementálása

Mivel a tanulás alapját képező jellemzőket kizárólag utólag, a felszíni szerkezet és az előelemzés segítségével tudjuk meghatározni, a kognitív alapú jellemzők többsége nem tükrözi pontosan a különböző elméletekben megfogalmazott elveket. Mivel a célom az, hogy az eredetileg megállapított elveket a lehető legpontosabban implementáljam számítógépes környezetbe, a következő fejezetben ezeket a jellemzőket veszem sorra, úgy, hogy ismertetem a jellemzők alapjául szolgáló elméleti megfontolásokat, illetve empirikus vizsgálatokat, majd kitérek arra is,

hogy a korpuszok melyik részének segítségével és milyen pontossággal használhatók fel a jellemzők.

6.7.2.1. Távolság

A szöveg felszíni szerkezetéből kinyerhető egyik kognitív alapú jellemző az anafora és az antecedens(jelölt) közötti távolság. Minél nagyobb a távolság, annál nehezebb a befogadónak azonosítani az antecedenst, hiszen a közbeékelt főnévi csoportok említésével, különösen az először említett entitásokkal, az antecedens a mentális állapotban a központi pozícióból egyre inkább perifériára kerül, így egyre nehezebben ismerhető fel a kapcsolat a két kifejezés között. A két kifejezés közötti távolságból tehát következtethetünk arra az erőfeszítésre, amelyet a címzettnek ki kell fejtenie ahhoz, hogy azonosítsa az anaforához tartozó antecedenst. A távolság megadása több módon is történhet, az érték kiszámításához pedig számos tényező figyelembe vehető.

A távolságot két mérőszám alapján számolhatjuk a szövegben: főnévi csoportok szerint és tagmondatok szerint. Főnévi csoport szerinti távolságszámítás során a Hobbs-távolság (Hobbs 1978) a bevett mérőszám. A Hobbs-távolság a két kifejezés közötti főnévi csoportok számát mutatja, azaz azoknak a lehetséges antecedensjelölteknek a számát, amelyeket el kell vetnünk, mint a kifejezés antecedense. A mérőszám megadása során el kell dönteni, hogy az összes főnévi csoportot figyelembe vegyük, vagy ezek közül kizárjuk a beágyazott főnévi csoportokat. Ez utóbbi esetben csak azokat számoljuk, amelyeket nem tartalmaz másik főnévi csoport. Erre az esetre jó példa a (43) mondatban a *Mari és Peti* kifejezés, amely tartalmaz két további főnévi csoportot, *Mari*-t és *Peti*-t, tehát ha ez a kifejezés közbeékelődik egy névmás és az antecedense közé, akkor számolható 3-nak és 1-nek is. Szintén kérdés még a felhasznált szintaktikai előelemzés tekintetében, hogy kizárólag NP-ket vagy az ADVP-ket is beleszámoljuk-e ebbe a számba, hiszen, mint ahogyan a (43) példa is mutatja, az anaforafeloldás során antecedens lehet ADV is.

43) Mari és Peti régen$_i$ minden nap együtt játszottak. Ekkor$_i$ még (ők) szomszédok voltak.

Ha a (43) példában a zéró (*ők*) névmáshoz keresnénk antecedenst, és kizárólag a főnévi csoportokat számolnánk, akkor csak a *minden nap* és a *Mari és Peti* vehetők figyelembe, azaz a *Mari és Peti* távolsága a zéró névmástól 2 lesz. Ha viszont az ADVP-ket is beleszámoljuk, hiszen a visszautaló névmások között találhatunk olyat, amelyiknek ilyen típusú antecedense

lesz, akkor figyelembe kell vennünk az *ekkor* és a *régen* kifejezést is, így pedig 4 lesz a versengő antecedensek száma.

Az alap jellemzőkészlet minden esetben tartalmazta a Hobbs-távolságot, vagyis azt az információt, hogy az adott főnévi csoport a névmástól számított hányadik. Ehhez figyelembe vettem a beágyazott főnévi csoportokat is, mégpedig úgy, hogy a teljes szerkezetet részesítettem előnyben, majd a szerkezetben szereplő további főnévi csoportokat. A (43) példában a *Mari és Peti* főnévi csoport lenne először a névmáshoz rendelve a legkisebb *NP distance* értékkel, ezután a *Peti* eggyel magasabb NP distance értékkel, ezután pedig *Mari* még eggyel magasabb értékkel.

A második távolságmérték a két kifejezés közötti tagmondatok száma. A tagmondatok számára hagyományos módon úgy tekintünk, mint egy hatókörre, amelyben az antecedens keresendő. A leggyakrabban egyszerűen a két kifejezés közötti tagmondatok számát mutatja a jellemző. Az érték megadása során itt is el kell dönteni, hogy a beágyazott tagmondatok is növeljék-e az értéket, vagy kizárólag azokat a mondatokat vegyük figyelembe, amelyek nem tartalmaznak más mondatokat. Ha az érték meghatározásának célja, hogy a kognitív erőfeszítést mutassa, akkor a kognitív nyelvészeti kutatások alapján nem pusztán a névmás és az antecedense közötti tagmondatok száma, de a tagmondatok egymáshoz való viszonya is mérvadó. A következő példában szögletes zárójelekkel ([]) azok a tagmondathatárok vannak jelölve, amelyek a SzegedKoref korpuszban is jelölve vannak.

44) [[Az elején még nem tudtunk **mi**$_i$ se támadni], [meg ők se.]] [[A felénél már rúgtunk egy gólt], [már 3-2 volt az eredmény.]] [[Ekkor még ők is rúgtak egy gólt]és [már **mi**$_i$ is azt hittük, [hogy a meccset már elveszítettük], [[mert már csak 11 perc volt hátra]és[4-2-re ki voltunk kapva.]]]]

A fenti példában a *mi* névmással ismétléssel utal vissza a szövegalkotó. Ha a teljes mondatok száma határozza meg a két kifejezés közötti távolságot, akkor 1 az érték. Ha a teljes mondathatár-átlépések számát vesszük, akkor 2. Ha csak azokat a mondatokat vesszük figyelembe, amelyeket más mondat tartalmaz, vagyis a tagmondatokat, akkor a két kifejezés közötti tagmondatok száma 4, a határátlépések száma 5.

A nyelvfeldolgozás során a feladat a szavak értelmezése és szerkezetbe való beépítése. Ezt a feldolgozást azonban kognitív szempontból nehezíti, amikor egy szerkezetbe egy újabb szerkezet közbeékelődik, és a korábbi szerkezetet a címzettnek hiányos állapotában tárolnia kell mindaddig, amíg a közbeékelt szerkezet teljessé nem válik. Minél több ilyen közbeékelődés

található egy mondatban, annál nehezebb a címzettnek feldolgoznia az információt. Ebből az a következtetés vonható le, hogy az ilyen típusú mondatok értelmezése során a címzettnek nagyobb erőfeszítésre van szüksége az értelmezéshez, mint az egyszerű tagmondatok értelmezése során, és ez a különbség hatással van az antecedens azonosításához szükséges erőfeszítésre is (Gibson 2000).

Az elérhetőségi elmélet (Ariel 1990; Ariel 2001; Ariel 2014) kitér az anafora és az antecedense közötti távolság anaforafeloldásra gyakorolt hatására is. Az elmélet szerint az alárendelő tagmondatból kisebb erőfeszítéssel érhető el az anaforához tartozó antecedens, mint a mellérendelő tagmondatból, a legnagyobb erőfeszítés pedig a teljes mondathatár átlépéséhez szükséges.

A tagmondati távolsággal kapcsolatban két jellemzőt definiáltam, és ezekkel kísérleteztem. Az első jellemző kizárólag a tagmondatzáró határátlépéseket vette figyelembe, amelyet a konstituens elemzés segítségével határoztam meg. Abban az esetben, ha több tagmondat is ugyanabban a pozícióban záródott le, az érték csak eggyel nőtt, mivel egy határátlépés történt. Ez a jellemző tehát egy numerikus érték, amely a későbbiekben a CP1 névvel fog szerepelni. A CP1 értéke a (45) és (46) Szeged Korpuszból származó példában szereplő visszautalás esetében így 5 lett.

45) [[Amíg vártuk **Petit**, [mert úgy hívják a kocsis haveromat], elmentünk fagyizni], [ott meg találkoztunk a barátom haverjaival.]] [Ők is épp fagyiztak.] [[Velük elbeszélgettünk], [aztán jött **ő**] és [mentünk Tófaluba]].

46) [[Amíg vártuk **Petit**, [mert úgy hívják a kocsis haveromat], elmentünk fagyizni], [ott meg találkoztunk a barátom haverjaival.]] [Ők is épp fagyiztak.] [[Velük elbeszélgettünk, [aztán jött **ő** is .]]

A második mondatszintű távolsági jellemző meghatározásának esetében figyelembe vettem közbeékelődéseket és az elérhetőségi elméletben felállított tagmondatokra vonatkozó megállapításokat. A szakirodalom alapján közbeékelődéseknek azokat az eseteket tekintettem, ahol a közbeékelt mondatnak sem a kezdete, sem a vége nem esik egybe az őt tartalmazó mondat kezdetével vagy végével. A (45) példában a *mert úgy hívják a kocsis haveromat* tagmondat megszakítja az *Amíg vártuk Petit (...), elmentünk fagyizni*. teljes tagmondatot. Tehát azt az egységet, hogy *Amíg vártuk Petit,* ebben a formában, hiányosan kell tárolnia a hallgatónak, mindaddig, amíg a közbeékelt mondat végéhez nem ér. Ezért a visszautaló névmástól számítva a

tagmondati határátlépések számát úgy vettem figyelembe, hogy a közbeékelődött mondat esetében egy belépési és egy kilépési értéket is számításba vettem.

Alárendelésnek tekintettem azokat az eseteket, ahol a beágyazott mondat kezdete vagy vége egybeesett az őt tartalmazó mondat kezdetével vagy végével, ezekben az esetekben a határátlépés egy pontot ért. Mellérendelésnek pedig azokat a szerkezeteket tekintettem, amelyeket más mondat tartalmazott, és ahol a megelőző mondat vége és a soron következő mondat eleje egybeesett vagy egymás után következett (írásjel vagy *és* kötőszó esetén nem minden esetben követik egymást közvetlenül): itt a határátlépés két ponttal növelte a *CP2* jellemző értékét. Ezeknél a szerkezeteknél is egy határátlépésnek számítottak az egybeeső mondatkezdő vagy egybeeső mondatzáró határok. A teljes mondat határátlépések, tehát azok a mondatok, amelyeket nem tartalmaz más mondat, nem egy, hanem három pontot értek, ezzel a nagy hatókörű anaforák esetét igyekeztem pontosítani. Ez esetben a (45) példa a 12 értéket vette fel, a (46) példa pedig 11-et.

A Szeged Korpuszban a konstituens elemzés, tehát a CP-k jelölése a szavak szövegen belüli pozíciójával történik, tehát minden CP-t két index jellemez: a CP első szavának szövegen belüli pozícióját jelölő index és a CP utolsó szavát vagy a mondatvégi írásjelet a szövegen belül jelölő index. A függőségi elemzésben a teljes mondatok továbbá nem CP, hanem ROOT címkét kaptak, így abban az esetben, ha a CP-ket kizárólag két számmal, a kezdő és utolsó szó pozíciójának sorszámával jellemezzük, akkor is megkülönböztethető egymástól a két mondattípus.

A CP2 érték tehát a következő módon került kiszámolásra:

1 Meghatározzuk a névmás szövegen belüli pozícióját, tehát az egyedi azonosító indexét.

2 Meghatározzuk az antecedens utolsó szavának (amennyiben többszavas) egyedi azonosító indexét.

3 Kilistázzuk a két index közötti kezdő CP indexeket úgy, hogy a névmás egyedi azonosítóját tartalmazza, de az antecedensét nem. Tehát, ha az antecedens mondatkezdő szó, abban az esetben ezt a határt már nem kell átlépni, hiszen elértük az antecedenst.

4 Kilistázzuk külön a CP kezdő indexek közül azokat, amelyek ROOT címkét kaptak.

5 Kilistázzuk a két kifejezés közötti záró CP indexeket úgy, hogy az antecedens egyedi azonosítóját tartalmazza, de a névmásét nem. Tehát, ha a névmással épp véget ér egy tagmondat, azt a határt nem kell átlépnünk.

6 Minden, a két kifejezés közötti egyedi (tehát ha több tagmondat is ugyanabban a pozícióban kezdődik, attól az még csak egy határátlépésnek számít) CP kezdő index kap 1 pontot, ezzel

növeljük a teljes mondatkezdésnél és a beágyazott alárendelő tagmondatoknál is a távolság értékét.

7 Minden egyedi, a két kifejezés közötti CP záróindex kap 1 pontot, ezzel növeljük egy ponttal a teljes mondat záró határok átlépését, valamint a mellérendelő mondatok esetében a záróhatárátlépést egy kezdő határátlépés fogja követni, így ez 2 pontot fog érni.

8 Minden a két kifejezés közé eső ROOT címkével ellátott CP kezdő intervallum, tehát a külön kilistázott ROOT-ok további növelik egy ponttal az értéket, így a teljes mondathatár átlépés 3 pontot fog érni.

```
function CpBetween(cpIntervals, antacedentInterval(head), anaphoraInterval(foot))
            relevantInterval := [antacedentInterval.start, anaphoraInterval.end]
            relevanCpIntervals := filterRelevantIntervals(cpIntevals, relevantInterval)
            rootCpInterval := filterRoots(cpIntervals)

            boundaries := []
            for cpInterval in relevantCpIntervals
              if relevantInterval.contains(cpInterval.start) &&
!boundaries.contains(cpInterval.start)
                boundaries.add(cpInterval.start)
              if relevantInterval.contains(cpInterval.end) &&
!boundaries.contains(cpInterval.end)
                boundaries.add(cpInterval.end)
            boundaryCount := boundaries.size() / 2

            lowerEndpoints := []
            for cpInterval in relevantCpIntervals
              lowerEndpoint := cpInterval.start
              if lowerEndpoint != antacedentInterval.start &&
!lowerEndpoints.contains(lowerEndpoint)
                lowerEndpoints.add(lowerEndpoint)
            countLower := 0
            for lowerEndpoint in lowerEndpoints
              if !lowerEndpoints.contains(lowerEndpoint + 1) &&
relevntInterval.contains(lowerEndpoint)
                countLower++

            upperEndpoints := []
            for cpInterval in relevantCpIntervals
              upperEndpoint := cpInterval.end
              if upperEndpoint != anaphoraInterval.end &&
!upperEndpoints.contains(upperEndpoint)
                upperEndpoints.add(upperEndpoint)
            countUpper := 0
            for upperEndpoint in upperEndpoints
              if !upperEndpoints.contains(upperEndpoint + 1) &&
relevntInterval.contains(upperEndpoint)
                countUpper++

            return boundaryCount + countLower + countUpper

function filterRelevantIntervals(cpIntevals, relevantInterval)
            relevanIntervals := []
            for interval in cpIntervals
              if interval.isConnected(relevantInterval)
                relevantIntervals.add(interval)

            return relevantIntervals

function filterRoots(intervals)
            roots := []
            for interval in intervals
              root := true
              for otherInterval in intervals
                if interval != otherInterval && otherInterval.encloses(interval)
                  root := false
                  break
              if root
                roots.append(interval)

            return roots
```

6.7.2.2. Hossz

Ariel munkáiban (Ariel 1990; Ariel 2001; Ariel 2014) hangsúlyozza, hogy minél nagyobb a távolság a két kifejezés között, annál gyengébb közöttük a kapcsolat, ezért egyéb mutatókat kell keresnie a címzettnek, amely segítségével azonosítani tudja az antecedest, és ezen keresztül a referenst. Az egyik ilyen mutató lehet a kifejezés hossza, amely Ariel munkáiban a kifejezés által jelölt referens elérhetőségét mutatja. A névmási visszautalás során az anafora hossza adott, egy szóból áll, ez magas, vagy más szóval könnyű elérhetőséget mutat. Ez azt jelenti, hogy a névmás egy olyan entitásra utal, amely a diskurzus éppen aktuális pontján a mentális állapot középpontjában van, tehát olyan antecedenst kell keresnünk, amely már magán hordozza ennek a központi pozíciónak a jeleit.

Központi pozícióban vannak például az éppen először említett entitások, amelyekre valószínűsíthető, hogy egy hosszabb, specifikusabb kifejezéssel utalunk, mivel az első említés maga után vonja, hogy a referens a címzett számára még egyáltalán nem elérhető. Ez persze nem valósul meg, ha az entitás, amire utalunk, a fizikai térben található meg, de ezeket az eseteket a szövegből utólag nem tudjuk felismerni. Ráadásul az eredetileg is írott szövegekben történő, azaz nem átiraton való anaforafeloldás során az ilyen esetek nem valószínűek.

Szintén könnyű elérhetőségre utal, ha a kifejezés a központi témája a szövegnek. Ebben az esetben lehetséges, hogy már az antecedens is egy könnyű elérhetőséget mutató, rövid kifejezés.

Az előbbi két esetből azt a következtetést vonhatjuk le, hogy az antecedens hossza, azaz specifikussága kapcsolatban állhat a névmástól való távolságával. Ha a szöveg központi témájára utalunk, akkor az az entitás annyira egyszerűen elérhető, hogy akármekkora távolságból utalhatunk rá névmással (49). Ha viszont frissen bevezetett, időszakosan a mentális állapot központjában levő entitásra utalunk vissza névmással, akkor valószínűleg az antecedens hosszabb, specifikusabb lesz, a távolság pedig kisebb (47)–(48), hiszen a közbeékelődő főnévi csoportok mind csökkentik a helyes azonosítás valószínűségét.

47) Ezért ez a nap a legérdekesebb számomra, mert még csak számításba sem volt [egy új bicikli]$_i$ (hát még egy ilyen). De ettől függetlenül nagyon örültem neki$_i$ és meglepett voltam az úton hazafelé is.

48) Mikor előkerültek [a sütemények és a hatalmas szép torták]$_i$, muszáj volt [őket]$_i$ lefényképezni.

49) Ekkor Jocónak$_i$ haza kellett mennie, ezért elszaladt az autóhoz a táskájáért. Nemsokára azt vettük észre, hogy kétségbeesetten siet vissza. Az idegességtől csak annyit tudott mondani, hogy feltörték az autót. Ő$_i$ hazament és telefonált a rendőrségre.

A gépi tanulás során több, összesen négy, az antecedensjelölt hosszára vonatkozó jellemzőt fogalmaztam meg, majd ezeket egy csomagként *Length* néven adtam hozzá a tanulási kísérlethez. Az első jellemző egy numerikus érték volt, és azt mutatta meg, hogy az antecedensjelölt milyen hosszú, azaz hány szóból áll. Ebben az esetben az egyszerűség kedvéért nem kizárólag a tartalmas szavakat vettem figyelembe, hanem a névelőket és a számokat is, azaz minden olyan elemet, amely a korpuszban külön sorba került az írásjeleket kivéve, így csak meg kellett számolni, hogy az adott főnévi csoport hány sorból állt a korpuszban.

A másik jellemző azt vizsgálta meg, hogy az adott főnévi csoport három szónál hosszabb-e, és ennek megfelelően bináris értéket vehetett fel. Ez az érték az Ariel elérhetőségi elméletében megfogalmazott „hosszú határozott leírások" meghatározásából származik. Mivel a puszta köznevek vagy az egyszerűen egy névelővel szereplő köznevek magasabb elérhetőséget mutatnak, mint a jelzővel is ellátott hosszabban körülírt kifejezések, ezért a három vagy annál több szóból álló kifejezéseket így ezzel a jellemzővel egy külön csoportba soroltam. A (48) példában a *a sütemények és a hatalmas szép torták* a hossz jellemző alapján a 7 értéket venné fel, valamint mivel három szónál hosszabb, az 1 értéket. További két jellemzőt is megfogalmaztam, amelyek azt biztosították, hogy ne kizárólag az adott antecedensjelöltet vegyem figyelembe, hanem a többi antecedensjelölthöz is hasonlítsam az adott NP-t. Az első jellemző azt mutatta meg, hogy az éppen vizsgált antecedensjelölt az antecedensjelöltek közül a leghosszabb-e, a másik pedig azt, hogy a legrövidebb-e. Mind a két jellemző bináris értéket vett fel.

6.7.2.3. Pozíció

A topik, vagy diskurzus topik, a kognitív-funkcionális keretekben a szöveg legkiemelkedőbb eleme, a szöveg fő témája. Egyes elméleti keretekben, mint Grosz és munkatársai (Grosz–Joshi–Weinstein 1995) munkáiban megkülönböztetik egymástól a globális topikot, azaz a teljes diskurzus fő témáját a lokális topiktól, azaz a kisebb lokális egységek fő témájától. A topikot utólag a szövegből azonosíthatjuk, például a kifejezések említésének száma alapján, mivel a topik a fő téma, ezért gyakran ismétlődik a szövegben. Mivel az entitás ismétlésének gyakorisága és száma éppen a koreferenciafeloldás segítségével lenne meghatározható, így csak az vizsgálható meg, hogy az egyes kifejezések milyen gyakran szerepelnek a szövegben, és

feltételezhetjük, hogy ezek a kifejezések egy szövegen, azaz fő témán belül nem változtatják a referenseiket.

A kifejezések pozíciójának vizsgálatával is következtethetünk a diskurzustopikra. Ariel munkáiban (Ariel 2001) arra a következtetésre jut, hogy a bekezdések elején és végén szereplő kifejezések referensei könnyebben elérhetőek, figyelemfelkeltőbbek, mint a bekezdések közepén szereplő kifejezések referensei. Mivel a vizsgált korpuszok nincsenek bekezdésekre tagolva, így ebből a szempontból nem vizsgálhatók a kifejezések, azonban a mondaton belüli pozíció igen. A magyar nyelv diskurzuskonfigurációs mivoltából következik, hogy semleges mondatban, a mondaton belül az ige előtti főnévi csoportok sorrendjében a topik szerepel az első helyen. Topik alatt itt az az entitás értendő, amiről már volt szó a szövegben. A második pozícióban a fókusz áll, azaz olyan új entitás, amiről eddig nem volt szó. A tagmondaton belül tehát az is megvizsgálható, hogy az adott antecedensjelölt a mondat első vagy második főnévi csoportja-e. A korpusz elemzése alapján automatikusan nehezen lehetne eldönteni, hogy az adott kifejezés valóban topik, illetve fókusz-e, azonban a kognitív nyelvészeti kutatások alapján a fontos információkat általában a mondat elejére helyezzük, tehát általánosságban ez az információ mégis segíthet az anaforafeloldás során is (Szécsényi–Kovács 2020).

50) Mi [az arra járó autókat]$_i$ megállítottuk és megkértük őket$_i$, hogy telefonáljanak a rendőröknek, ugyanis a közelben nem volt telefonfülke és a pénzünket is elrabolták.

51) [A ruháinkat]$_i$ [a hátsó ülésen]$_j$ helyeztük el, abban bízva, hogy [ott]$_j$ semmi baj nem történhet [velük]$_i$.

A jellemzők közé ezért egy Position jellemző csomagba felvettem két jellemzőt. Az egyik azt mutatja meg, hogy az adott főnévi csoport a tagmondat első főnévi csoportja-e, a másik jellemző pedig azt, hogy a tagmondat második főnévi csoportja-e. Ezekben az esetekben tehát kevéssé tudtam pontosan megragadni az elméletben megfogalmazott elveket, hiszen kizárólag a konstituens elemzésből indultam ki a két jellemző meghatározása során, amelyek szintén bináris értéket vehettek fel. A (50) példában a *Mi* első NP-nek *az arra járó autókat* pedig második NP-nek számítana.

6.7.2.4. Új diskurzusreferens

Az új diskurzusreferens bevezetése pillanatában az egyik legelérhetőbb entitás, mivel magának a bevezetésnek a ténye felhívja rá a figyelmet. A bevezetés tényéből arra is következtethetünk, hogy a téma szempontjából fontos, ezért még fog rá visszautalni a szövegalkotó. A szövegből

utólag szintén, mint az említések gyakoriságánál, itt is a már azonosított koreferencialáncok segítenének az első említés azonosításában. Ennek hiányában csak következtetéseket vonhatunk le a szöveg felszíni szerkezetéből, valószínűségek alapján fogalmazhatjuk meg a diskurzusreferens bevezetését mutató jellemzőt.

Az új diskurzusreferens bevezetését mutathatja például a határozatlan névelő használata, amit a morfológiai elemzés alapján, a korpuszon is ellenőrizhetünk. Ha a 'Definite' attribútuma az NP-hez tartozó névelőnek az 'Ind' értéket kapja, akkor biztosan határozatlan a kifejezés.

Két további jellemzőt határoztam még meg ezek alapján. A Pron jellemző azt mutatja meg, hogy az adott antecedensjelölt névmás-e, tehát bináris értéket vehet fel, és az antecedensjelölthöz rendelt szófaji címke alapján határozza meg az értéket. Abban az esetben, ha az antecedensjelölt névmás, magas az elérhetősége. A másik jellemző a Def jellemző, amely azt mutatja meg, hogy az adott antecedensjelölt határozott leírás-e. Ehhez a teljes NP morfológiai elemzésében keresem, hogy tartalmaz-e Definite attribútumot. Ha igen, akkor az attribútumhoz rendelt Def vagy Indef értéket kapja az antecedensjelölt, ha viszont nem tartalmaz ilyet, vagy az összetett szerkezet többet is tartalmaz, akkor az érték hiányzó adat lesz.

6.7.2.5. Jellemzők összefoglalása

Az általam vizsgált kognitív elvek alapján megfogalmazott jellemzőket tehát 4 csoportba soroltam Pron, Length, Def, Pos, és külön-külön hozzáadtam a Base jellemzőkészlethez. Ezeket a kiegészített jellemzőlistákat hasonlítottam össze a Base jellemzőkészlettel. Mind az öt csoporthoz külön-külön hozzárendeltem a CP1, valamint a CP2 tagmondati távolságszámítási módszert is, harmadik esetként pedig azt az esetet is figyelembe vettem, ha mind a két tagmondati távolságszámítási módszert is alkalmazom: CP12.

6.8. A tanításhoz használt algoritmus

A gépi tanulási kísérleteket a Weka szoftver (Eibe–Hall–Witten 2016) segítségével végeztem el, és a Random forest algoritmust (Breiman 2001) alkalmaztam.

A Random forest több különálló döntési fából áll, amelyek külön, egyesével hoznak döntést arról, hogy az adott minta melyik osztályba tartozhat, végül a legtöbb szavazatot kapó osztály lesz a modell előrejelzése. Ennek előnye, hogy az egyes fák hibáit kiküszöböli a többség döntése, mindaddig, ameddig nem hibás a többség döntése. Éppen ezért minél több nem korreláló fát tartalmaz a modell, annál biztosabb az eredmény. Tehát a hatékonyság a fák

számán, minőségén és eltérőségén múlik. Az eltérő fákat két módon biztosítja az algoritmus: 1 Minden egyes fa véletlenszerűen vesz mintát az adatkészletből, így biztosan eltérőek lesznek egymástól az egyes fák. Ez azt jelenti, hogy az eredeti tanítóadatokból vesz egy véletlenszerű mintát, majd ezt a mintát a teljes tanító adatkészlet méretére növeli véletlenszerű helyettesítésekkel. Így minden tanítóadat szerepel valamely fa létrehozásában, és az egyes fákat sem kevesebb adat segítségével kell létrehozni; ez a módszer a zsákolás, vagy *bagging*. 2. A másik diverzitást biztosító tényező, hogy minden egyes fa csak a jellemzők egy részhalmazát használja fel a modellépítéshez, így az egyes fák biztosan eltérőek lesznek egymástól. Ezért sem feltétlenül tekinthető redundánsnak jellemzők egyes címkéinek kitüntetettként való kezelése és bináris jellemzőként való újrafelvétele.

A negatív hozadéka a véletlenszerű mintavételnek, hogy kiegyenlítetlen tanítóadat-halmaz esetén könnyen a többségi osztályra eshet a fák szavazata, mivel ezekből a példákból többet választhat ki egy-egy fa építéséhez. Az is előfordulhat, hogy a kevesebb példát felmutató osztályból egyet sem tartalmaz a véletlenszerűen kiválasztott minta. Ennek az eredménye, hogy az előrejelzésében annak az osztálynak, amelyből kevesebbet tartalmaz a tanító minta, gyengébben teljesít az algoritmus. Ezen a problémán a tanítóadat-halmazban szereplő egyes osztályokba sorolt példák számának közelítésével lehet javítani, azaz a többségi csoport alulreprezentálásával vagy a kisebbségi csoport felülreprezentálásával.

A több döntési fa segítségével történő osztályozás mellett előnye még az algoritmusnak, hogy mind nominális, mind skaláris értékeket képes a tanulás alapjául szolgáló jellemzőként kezelni, így az anaforafeloldás során a két kifejezés közötti távolságot mint számszerűsített értéket is a jellemzőlistához lehet adni.

A legfontosabb jellemzők kiválasztására számos módszer létezik. Az egyik ilyen statisztikai adat, amely segítségünkre lehet a gépi tanulás során, a gini index (Yadolah 2008), amely a jellemzők értékeinek statisztikai eloszlásának egyenlőtlenségét méri, tehát biztosítani tudja a nem korreláló adatok kiválasztását. Az anaforafeloldásnál azonban a tanító és tesztfájlok egymástól eltérő eloszlásúak, így lehetséges, hogy az a jellemző, amely a tanító adathalmazon még nagy eltérést mutat a két osztály között, a tesztadatokon már nem rendelkezik ezzel a tulajdonsággal, és fordítva. A jellemzőkészletek összehasonlítására másik módszer, hogy az egyes jellemzőkészletekkel külön modelleket építünk, és ezeket a modelleket külön-külön kiértékeljük ugyanazon a kisebb adathalmazon, majd a legjobb eredményt elérő jellemzőkészlet segítségével épített modellt teszteljük egy nagyobb adathalmazon. Ennek a módszernek szintén vannak hátrányai, hiszen a kisebb adathalmazon elért eredmények nagyban függenek az adott

adathalmaz minőségétől, reprezentativitás és hibák, valamint kiugró értékek számától, és méretétől. Kísérleteim során az utóbbi módszert választottam, egyrészt, mert a kutatási kérdéseim nem kizárólag az általam megfogalmazott jellemzők befolyásoló erejére vonatkoztak, így a külön modellek építése elengedhetetlen volt, másrészt a fent említett, tanító és tesztfájlok eltérő összetételéből fakadó nehézségek miatt. Ezáltal kisebb mintán épített modellek eredményeit kellett összehasonlítanom, így ahhoz hogy a levont konklúziók biztosabbak legyenek, nem kizárólag a Szeged Korpuszon, de a KorKorpuszon is ellenőriztem az épített modellek sikerességét.

6.9. Kiértékelési metrika

Az általam végzett kísérletekben párokról hoz döntéseket az osztályozó, mégpedig azt, hogy anaforikus kapcsolatban állnak egymással vagy sem, így a MUC feladatban alkalmazott, a kifejezések közötti kapcsolatokra összpontosító kiértékelésből indulhatunk ki. A feladatból következik, hogy a standard IR metrikák is megfelelően mutatják majd a modellek sikerességét, hiszen, ha a párok többségét anaforikusnak ítéli a rendszer, akkor romlik a pontosság, ha viszont semmit sem ítél anaforikusnak, akkor romlik a fedés értéke. Mindemellett azt is figyelembe kell venni, hogy a tesztek során sokkal több párt kell az osztályozónak negatívnak értékelnie, mint pozitívnak, így abban az esetben, ha az osztályozó mindent negatívnak ítél, tehát nem talál egy anaforikus párt sem a tesztben, az F-mérték sokkal jobb lesz, mint abban az esetben, ha minden párt pozitívnak ítél. Ennek ellenére a tesztek egymással összehasonlíthatók lesznek, tehát a kognitív alapon megfogalmazott jellemzők pozitív vagy negatív hatása megfigyelhető lesz a párok anaforikus, illetve nem anaforikus csoportba sorolását tekintve.

7. Kísérletek

A gépi tanulási kísérletek során elsősorban azt vizsgálom, hogy lehetséges-e automatikus névmási anaforafeloldást végezni pusztán morfológiai és szintaktikai jellemzők segítségével. Az első hipotézisem, hogy a tanulófájlok létrehozása során az a módszer eredményezi a legsikeresebb modellt, ha a tesztfájlhoz hasonlóan a tanulófájl is tartalmaz minden névmást és hozzá minden, a névmást megelőző főnévi csoportot (Exp1 kísérletek). A második hipotézisem, hogy a névmáshoz tartozó pozitívnak ítélt antecedensjelöltek közül a legnagyobb valószínűségi értékkel antecedensnek tekintett főnév valódi antecedensként való azonosítása lesz a legsikeresebb stratégia (Best-first). A harmadik hipotézisem pedig az, hogy az általam, kognitív nyelvészeti elvek alapján meghatározott jellemzők pozitívan befolyásolják a tanulás sikerességét.

A hipotéziseim alátámasztásához hat kísérletet végzek el minden egyes névmástípus tekintetében. Az első két kísérlet a távolságszámítás különböző módjainak hatását vizsgálja, a maradék négy kísérlet pedig a tanítófájokban szereplő pozitív és negatív példák befolyását valamint az egyetlen antecedensjelölt azonosítására vonatkozó hipotézisemet fogja ellenőrizni. Az összes kísérletben megvizsgálom az általam meghatározott jellemzők hatását az épített modellekre, így ezt a hipotézist minden kísérleten keresztül ellenőrizni tudom.

A pilot kísérletek során a keresztvalidáció módszerét alkalmaztam, úgy, hogy a teljes korpuszt 10 részre osztottam. Ezekből a kísérletekből három fontos következtetést vontam le: 1 A teljes korpusz tizedében meglehetősen kevés számú visszautalást találunk, így a kiértékelés során már egy visszautalás helyes vagy helytelen módon történő azonosítása is nagy különbségeket okoz, abban az esetben ha a tesztelést csak a korpusz 10%-án végzem el. Ez kiküszöbölhető, ha a fájlokat öt részre osztom és így a teljes korpusznak nem 10%-ból, hanem 20%-ból készül a tesztfájl, ezzel azonban egyaránt csökkentem a tanító fájlok méretét is, ezért rosszabb eredmények születhetnek a kevesebb pozitív példa miatt. 2 A tesztfájlokban úgy is növelhető a pozitív példák száma, ha az első körben a validálás során az összes kézzel is annotált antecedensét megvizsgálom a névmásnak. 3 A teljes korpuszban a visszautalások egyenetlenül oszlanak el, így ha a korpusz felosztása során nem veszem figyelembe az egyes szövegekben megtalálható visszautalások mennyiségét, az egyes tesztekben eltérő mennyiségű visszautalás lesz megtalálható, ami a kiértékelés során szintén nagy különbségeket okozhat, ha azokat csak egy részen végzem el.

Az első kísérletek tapasztalatai alapján a teljes korpuszt a későbbiekben mindig öt részre osztottam úgy, hogy figyelembe vettem a szövegek csoportosítása során a bennük található koreferencialáncokban szereplő névmások számát is (ezt háromszor kellett megismételni, minden névmástípus tekintetében). Az első két kísérletben (Exp1v, Exp2v) a tesztfájlokban az összes névmáshoz kézzel is annotált antecedens pozitív példaként szerepelt. A további kísérletekben (Exp1, Exp2, Exp3, Exp4) a tesztelést már kizárólag csak a korpusz ötödén végeztem el és mindegyik esetben kizárólag egy antecedens azonosítása volt a cél.

Az első kísérletben (Exp1v) a tanító fájlban megtalálható a korpusz 80%-ában előforduló összes adott típusú névmás, legyen az visszautaló vagy sem, lehetséges antecedensjelölt pedig az összes névmást megelőző főnévi csoport, amelyeket a névmáshoz rendeltem párként. Tehát a tanító fájlban sem a pozitív, sem pedig a negatív párok nem lettek szűrve. A tesztfájlban a pozitív és negatív példák aránya a tanítófájllal azonos és a korpusz maradék 20%-ból készül. Ezt a módszert ötször ismétlem a tesztelés során mindig a korpusz különböző 20%-án. A jellemzők közül mind az 5 csoportot megvizsgálom (Base, Pron, Def, Length, Pos) úgy, hogy a távolságszámítás jellemzőkre gyakorolt hatását is figyelembe veszem, tehát minden egyes esetben három különböző jellemző lista jön létre: egy amelyikben kizárólag a tagmondati határátlépések számát vizsgálom (Cp1), egy amelyikben kizárólag az általam meghatározott tagmondatok számából kinyert értékeket adom hozzá (Cp2) a jellemzőkhöz, és egy, amelyikben mind a két érték szerepel (Cp12).

A második kísérletben (Exp2v) a tanító és tesztfájlok hasonlóan az elsőhöz a korpusz négyötöd és egyötöd részéből jönnek létre és változatlanul ötször ismétlem a tanulás és tesztelési folyamatot. A tanulás alapját képező jellemzők összetétele is megegyezik az első kísérletnél olvasottakkal. Az eltérés a tanítófájlokban szereplő pozitív és negatív példák mennyisége. A második kísérletnél kizárólag azokat a névmásokat veszem figyelembe a tanulás során, amelyekhez a korpuszban kézzel is annotált antecedens található, az antecedensek közül pedig az első kézzel is annotált lesz az egyetlen pozitív tanító példa. A negatív példák a névmás és az antecedense között elhelyezkedő főnévi csoportokból tevődnek össze. Tehát ezzel a módszerrel a tanító fájlban mind a pozitív, mind a negatív példák szűrésre kerülnek.

A harmadik és negyedik kísérletben az első két kísérlet kimeneteit vizsgálom tovább (Exp1, Exp2). Mivel ezekben az esetekben a tesztfájlokban a névmásokhoz kézzel is annotált összes antecedens pozitív párként van jelölve, de a névmási anaforafeloldás során a cél egy antecedens azonosítása, ezért a jelölteket két módszer segítségével szűkítem le egyre. Az első módszer a Best-first technika (Ng–Cardie 2002a), amely során az osztályozó által legmagasabb

valószínűségi értékkel antecedensnek ítélt főnévi csoportot tekintem a névmás antecedensének. Abban az esetben, ha több főnévi csoport is azonos valószínűségi értéket kap, akkor ezek közül a névmáshoz legközelebb eső főnévi csoport lesz az antecedens. A második módszer a Closest-first technika (Soon–Ng–Lim 2001), amely során az osztályozó által pozitívnak ítélt a névmáshoz legközelebb eső főnévi csoport lesz az antecedens. Ezt a két kísérletet 5-ször végzem el minden névmás esetében, minden jellemzőcsoportból az összességében legeredményesebbet kiválasztva egy darab teszt fájlon.

A Random Forest algoritmus (Breiman 2001) tulajdonságaiból kiindulva elvégeztem két további kísérletet is, amelyben a célom a pozitív tanítópéldák számának növelése volt a negatív tanítópéldák számával szemben, ezzel is növelve az esélyt, hogy a modellépítésben több pozitív példa vegyen részt. A tanítófájlok az első esetben (Exp3) úgy jöttek létre, hogy a visszautaló névmáshoz antecedensként hozzárendeltem minden a szövegben őt megelőző főnévi csoportot az első kézzel is annotált főnévi csoportig, ezek lettek a negatív példák, plusz a kézzel is jelölt főnévi csoport az első pozitív példa. Ezek után a további, szövegben névmást megelőző kézzel is annotált főnévi csoportokat szintén hozzárendeltem párként a névmáshoz, ebben az esetben már a két kifejezés közötti főnévi csoportokat nem vettem figyelembe. Ezzel a módszerrel tehát kizárólag a negatív példák számát csökkentettem. A tanítófájl segítségével modellt építettem a korpusz 80%-án és a tesztelés során szintén megvizsgáltam a Best-first és a Closest-first technikákat is az összes jellemző tekintetében. A második esetben (Exp4) arra törekedtem, hogy a pozitív és negatív példák a tanítófájlokban egyenlő arányban jelenjenek meg. Ehhez megvizsgáltam, hogy hány darab pozitív példa található a tanítófájlban és a maximális távolságot is. Ezután megvizsgáltam, hogy ezen a maximális távolságon belül a negatív példák milyen arányban oszlanak el, a CP1 távolságszámítási módszer alapján, majd ezt az arányt megtartva a pozitív példák számához csökkentettem a negatív példák számát is. A következő táblázat az egyes tanítófájlok közötti eltéréseket szemlélteti.

	NP	**NP**	**NP**	**NP**	**NP**	**NP**	**NP**	**NP**	**NP**	
EXP1	0	1	0	1	0	0	1	0	0	←PRON
EXP2	_	_	_	_	_	_	1	0	0	←PRON
EXP3	_	1	_	1	_	_	1	0	0	←PRON
EXP4	_	1	_	1	_	0	1	0	0	←PRON

5. táblázat Az egyes tanítófájlok közötti eltérések

A tesztelés során minden esetben megvizsgáltam minden névmást, és párként hozzárendeltem az összes főnévi csoportot lehetséges antecedensként egészen a szöveg elejéig. Az első két kísérletben (EXP1v, EXP2v) a cél az összes névmáshoz pozitív jelöltként rendelhető főnévi csoport azonosítása, tehát NP2, NP5, NP7 a további négy kísérletben már kizárólag egy antecedens azonosítása volt a cél, tehát a három közül bármelyik. A tesztfájlok tartalmát a következő táblázat mutatja.

	NP8	NP7	NP6	NP5	NP4	NP3	NP2	NP1	
BOF	0	1	0	1	0	0	1	0	←ANA_PRON
BOF	0	0	0	0	0	0	0	0	←NOT-ANA_PRON

6. táblázat A tesztfájlok felépítése (BOF= szöveg eleje, ANA_PRON= visszautaló névmás, NOT-ANA_PRON= nem visszautaló névmás)

7.1. Személyes névmás

A személyes névmási visszautalások tekintetében a SzegedKoref korpuszról elmondható, hogy nem csak a harmadik személyű, de a második és első személyű személyes névmások antecedensei is jelölve vannak a szövegekben. Ennek következtében egyrészt több a pozitív példa a gépi tanulás során, azonban másrészt általánosan jobb eredmények érhetők el, hiszen a morfológiai egyeztetések által minden egyes szám első személyű névmás koreferens lesz, illetve minden többes szám első személyű névmás is, és valószínűleg ugyanez teljesül a második személyű névmásokra is.

Az első általam végzett kísérletek alapján két fontos következtetést vontam le.

1 A gépi tanulás során a morfológiai információk és az egyeztetési jegyek nem kizáró megszorítások, ezért olyan antecedenseket is azonosíthat az algoritmus, amelyek személyben nincsenek egyeztetve a névmással. Ezek az antecedensjelöltek még a tanulási folyamat előtt kiszűrhetők mind a tanító, mind a tesztfájlokból. Erre azért van szükség, mert míg az előfordulhat, hogy számban, esetben vagy akár más nyelvekben hibásan nemben sincs egyeztetve a névmás és az antecedense, addig a személybeli egyeztetés minden esetben megtörténik.

52) Mi$_i$ elmentünk a boltba, én$_i$ vettem egy csokit. rész-egész viszony

53) Gabi$_i$ mindig elkésik, legközelebb leszidom őt$_i$. alanyváltás

54) Ti$_i$ elmentetek a boltba én$_i$* vettem egy csokit.

55) Gabi$_i$ elesett a biciklivel, én/mi,te/ti$_i$* eltörtem a kezem

A személyjegy alapján történő egyeztetést tehát kizáró szűrőként használhatom, ezzel pedig csökkentem a tesztfájlokban a negatív párokat, amelyek fals pozitív eredményt adhatnak.

2 A tesztelés során számos esetben kellett fals pozitív találatként értékelnem egymással anaforikus kapcsolatban álló párokat a kézi annotációból fakadó következetlenségek, apróbb hibák miatt. Ezek főleg az egyes szám első személyű, illetve a többes szám első személyű névmások esetében fordultak elő. Az ilyen típusú visszautalások gyakran nem láncként szerepeltek a szövegekben, tehát csak az első, legközelebbi antecedens volt a fájlokban jelölve, ez a későbbiekben nem feltétlenül fog gondot okozni, hiszen a fő cél egy darab antecedens azonosítása, a validációnál azonban sokat ront az eredményeken.

Az első kísérletekből fakadó tapasztalatok alapján két nagy változtatást hajtottam végre. A fals pozitív párok csökkentésére két lehetőség közül választhattam, vagy egyáltalán nem veszem figyelembe a mindenkori beszélőre utaló névmásokat sem a tanulás, sem a tesztelés során, vagy kézzel utólag azonosítom a korpuszban az összes ilyen típusú visszautalást. Először kiszűrtem a tanító- és tesztfájlokból az összes első, illetve második személyű személyes névmást, ezzel kizártam a mindenkori beszélőre és hallgatóra utaló névmásokat, és kizárólag a harmadik személyű visszautalásokat vettem figyelembe, mind a tanulás, mind a tesztelés során (PrsPer3). Az első táblázatban a teljes tanító fájl segítségével épített modell eredményei, a második táblázatban pedig a mind pozitív, mind negatív példákra szűrt tanító fájlok segítségével épített modellek eredményei láthatóak.

PrsPer3Exp1v	**RandomForest**		
	P	**R**	**F**
cp1_Base	34,40	12,50	18,30
cp2_Base	38,10	14,80	21,30
cp12_Base	39,30	13,30	19,80
cp1_Pron	35,20	13,10	19,10
cp2_Pron	36,70	15,20	21,50
cp12_Pron	37,60	13,30	19,60
cp1_Def	34,60	12,00	17,90
cp2_Def	38,30	14,30	20,80
cp12_Def	40,00	13,30	19,90
cp1_Length	30,70	10,50	15,70
cp2_Length	37,90	13,30	19,60
cp12_Length	40,20	12,00	18,50
cp1_Pos	39,50	12,50	19,00
cp2_Pos	41,50	13,30	20,10
cp12_Pos	43,10	11,70	18,50

7. táblázat A teljes tanítófájl segítségével épített modell eredményei

PrsPer3Exp2v	RandomForest		
	P	R	F
cp1_Base	04,00	37,38	07,20
cp2_Base	05,76	38,08	09,88
cp12_Base	03,26	35,30	05,98
cp1_Pron	04,24	37,62	07,56
cp2_Pron	05,78	38,70	09,96
cp12_Pron	03,38	35,54	06,14
cp1_Def	04,80	36,78	08,40
cp2_Def	06,28	36,66	10,54
cp12_Def	04,06	36,02	07,22
cp1_Length	04,00	35,68	07,18
cp2_Length	07,36	37,42	12,14
cp12_Length	03,80	35,28	06,88
cp1_Pos	03,24	14,84	05,30
cp2_Pos	03,36	16,16	05,56
cp12_Pos	02,82	16,22	04,84

8. táblázat A szűrt tanítófájl segítségével épített modell eredményei

A két kísérlet alapján összességében elmondható, hogy a pozitív tanítópéldák kis mennyisége miatt általánosan rosszak az eredmények. Ezekhez hozzájárul az is, hogy a tesztelés során a névmáshoz összes antecedensként azonosított főnévi csoportot vizsgáltam, tehát a sikeres találathoz nem volt elég egy antecedens azonosítása. A teljes tanítófájlt felhasználva jobb pontosság érhető el, mivel a névmástól távolabb eső főnévi csoportokról is találhatóak információk, azonban ez azt is eredményezi, hogy sokkal kevesebb főnévi csoportot ítél anaforikusnak a rendszer. A szűrt tanítófájloknál a fedés mutat jobb eredményeket, ezekben az esetekben azonban a pontosságon sokat ronthat, hogy a tanítófájlokban nincsenek olyan névmások negatív tanítópéldaként, amelyek alapvetően nem utalnak vissza. A jellemzők és távolságszámítási módszerek tekintetében a teljes tanítófájlnál a cp2_Pron érte el, azaz az összetettebb távolságszámolási módszer, valamint az az alap jellemzőkészlethez hozzáadott jellemző, amely azt mutatja, hogy az antecedens névmás-e. A szűrt tanítófájlok esetében a legjobb fedést szintén a cp2_Pron érte el, azonban összességében a cp2_Def érte el a legjobb eredményt. Az általam hozzáadott jellemzők közül egyedül a szűrt tanítófájl felhasználása során az antecedens pozíciójára vonatkozó jellemzők (Pos) mutatkoznak eredménytelennek, a többi esetben nem mutatkoznak mérvadó eltérések. Az azonban egyértelműen látható, hogy a cp2 számolási módszer javít a párok felismerésén így a továbbiakban ezzel a távolságszámítási módszerrel építettem a tanítófájlok segítségével modelleket. Ahhoz, hogy megvizsgáljam, abban az esetben, ha egyetlen antecedenst kellene azonosítania a modellnek, milyen eredményeket érne

el, minden esetben kiválasztottam ugyanazt az egy darab teszt fájlt, és a cp2 számolási módszerrel együtt alkalmazott jellemzőkészletekkel új modelleket építettem. Minden esetben megvizsgáltam a névmáshoz legnagyobb valószínűséggel hozzárendelt antecedenst (Best) és a legközelebbi hozzárendelt antecedenst is (Closest). A következő két táblázatban először a teljes tanítófájl alapján épített modell eredményei utána pedig a szűrt tanítófájl eredményei láthatóak.

PrsPer3Exp1		P	R	F
Best	Base	38,30	15,93	22,50
Closest		38,30	15,93	22,50
Best	Pron	39,13	15,93	22,64
Closest		39,13	15,93	22,64
Best	Def	37,78	15,04	21,52
Closest		37,78	15,04	21,52
Best	Length	42,50	15,04	22,22
Closest		42,50	15,04	22,22
Best	Pos	39,02	14,16	20,78
Closest		39,02	14,16	20,78

9. táblázat Egyetlen antecedens azonosításának eredményei a teljes tanítófájl segítségével épített modell esetében

PrsPer3Exp2		P	R	F
Best	Base	16,67	30,97	21,67
Closest		19,52	36,28	25,39
Best	Pron	14,88	28,32	19,51
Closest		18,60	35,40	24,39
Best	Def	17,31	31,86	22,43
Closest		18,27	33,63	23,68
Best	Length	15,74	30,09	20,67
Closest		18,06	34,51	23,71
Best	Pos	06,09	10,62	07,74
Closest		06,09	10,62	07,74

10. táblázat Egyetlen antecedens azonosításának eredményei a pozitív és negatív példákra is szűrt tanítófájl segítségével épített modell esetében

PrsPer3Exp3		P	R	F
Best	Base	13,45	28,32	18,23
Closest		19,75	41,59	26,78
Best	Pron	12,50	26,55	17,00
Closest		19,17	40,71	26,06
Best	Def	10,55	22,12	14,29
Closest		19,41	40,71	26,29
Best	Length	13,14	27,43	17,77
Closest		16,53	34,51	22,35
Best	Pos	11,54	23,89	15,56
Closest		17,95	37,17	24,21

11. táblázat Egyetlen antecedens azonosításának eredményei a kizárólag negatív példák szűrésével létrejött tanítófájl segítségével épített modell esetében

PrsPer3Exp4		P	R	F
Best	Base	17,41	38,05	23,89
Closest		10,93	23,89	15,00
Best	Pron	17,41	38,05	23,89
Closest		11,34	24,78	15,56
Best	Def	17,41	38,05	23,89
Closest		10,53	23,01	14,44
Best	Length	17,81	38,94	24,44
Closest		11,74	25,66	16,11
Best	Pos	18,62	40,71	25,56
Closest		10,12	22,12	13,89

12. táblázat Egyetlen antecedens azonosításának eredményei a pozitív és negatív példákat egységes arányban tartalmazó tanítófájl segítségével épített modell esetében

A tesztfájlban összesen 247 harmadik személyű személyes névmás volt, ebből 113 visszautaló névmás, azaz 45,75%-a a névmásoknak visszautal a szövegben. A pontosság tekintetében a teljes tanítófájl (Exp1) alapján épített Length jellemzővel kiegészített modell volt a legeredményesebb. A fedés tekintetében a negatív példákra szűrt tanítófájl (Exp3) segítségével az alap jellemzőkészlet mellett a Closest-first módszer mutatja a legjobb eredményt. A harmadik személyű személyes névmások esetében elmondható, hogy a jellemzőkészletek között nem mutatkozik nagyobb eltérés, a Closest-first és a Best-first módszerek között azonban már

nagyobb különbségek láthatóak. A pozícióra vonatkozó jellemzőkkel kiegészített tanulási kísérleten kívül minden esetben a Closest-first módszer mutatkozik eredményesebbnek.

A személyes névmásokkal kapcsolatban a második lehetőség az, ha a korpuszban megtalálható annotációt egészítem ki. Először a morfológiai elemzés alapján megkerestem azokat a névmás és antecedensjelölt párokat, amelyek egyaránt egyes szám első személyűek vagy egyaránt többes szám első személyűek voltak, majd feltételezve, hogy a szövegek nem tartalmaznak függő beszédet, mivel nem dialógusok, hanem fogalmazások, és ezek a kifejezések a mindenkori beszélőre utalnak, anaforikusra cseréltem a párokat. Ezzel figyelembe vettem az összes személyes névmást, az egyes szám első személyű és többes szám első személyű névmásoknál pedig nem csak a legközelebbi antecedens volt jelölve, tehát a validálás során pontosabb értékeket kaptam. Az alábbi változtatások mellett már jelentősen nagyobb számmal voltak a pozitív példák jelen. A következő két táblázat a teljes tanítófájlon épített modellek, majd a szűrt tanítófájlon épített modellek eredményeit mutatják.

PrsFullExp1v	RandomForest		
	P	R	F
cp1_Base	93,30	48,70	64,00
cp2_Base	92,90	48,80	64,00
cp12_Base	94,30	48,60	64,10
cp1_Pron	93,20	52,60	67,30
cp2_Pron	93,30	52,90	67,50
cp12_Pron	94,30	52,50	67,50
cp1_Def	93,40	48,50	63,80
cp2_Def	93,70	48,90	64,20
cp12_Def	94,60	48,60	64,20
cp1_Length	93,60	49,40	64,70
cp2_Length	93,40	49,80	64,90
cp12_Length	94,90	49,20	64,80
cp1_Pos	93,00	48,90	64,10
cp2_Pos	94,10	48,80	64,20
cp12_Pos	94,90	48,60	64,30

13. táblázat A teljes tanítófájl segítségével épített modell eredményei

PrsFullExp2v	RandomForest		
	P	R	F
cp1_Base	22,62	50,24	31,12
cp2_Base	37,03	51,19	42,90
cp12_Base	36,09	50,45	41,96
cp1_Pron	26,62	50,91	34,63
cp2_Pron	35,69	51,84	42,20
cp12_Pron	35,10	51,93	41,55
cp1_Def	28,27	50,21	35,49
cp2_Def	42,88	51,25	46,56
cp12_Def	41,67	50,57	45,15
cp1_Length	27,99	49,40	35,18
cp2_Length	41,62	50,47	45,58
cp12_Length	42,25	50,35	45,83
cp1_Pos	18,07	15,69	16,64
cp2_Pos	20,45	15,70	17,66
cp12_Pos	17,84	15,66	16,61

14. táblázat A szűrt tanítófájl segítségével épített modell eredményei

Ebben az esetben egyértelműen a teljes tanítófájl alapján épített modell volt a sikeresebb, aminek fő oka az első személyű személyes névmások nagy száma. A morfológiai egyeztetés hiánya miatt az osztályozó sok negatív párt felismert, ezzel növekedett a pontosság, azonban a fedésen rontott, hiszen emiatt sok harmadik személyű személyes névmási visszautalást is negatívnak ítélt. Az első tesztekben szintén a cp2_Pron érte el a legjobb eredményeket, de itt sem tapasztalhatók nagy különbségek. A második kísérletben a szűrt tanító fájlon való modell építés során a pontosság sokat romlik, viszont emellett a fedés nő. A legjobb értéket ebben az esetben a cp2_Def mutatja.

Az összes teszt esetében megfigyelhető a rendkívül alacsony pontosság, aminek két oka van. Az egyik, hogy a tesztelésnek ebben a fázisában az algoritmus nem kizárólag egy antecedenst keresett a személyes névmásokhoz, hanem az összes őt megelőző antecedenset igyekezett azonosítani, így pedig automatikusan több esetet is pozitívnak talált egy névmás esetében. A másik ok a tanuló fájlok jellegéből adódik. Több kutatás is alátámasztja, hogy tanulás során nem csak a negatív, de a pozitív tanítópéldák szűrése is pozitív hatással van a tanulás sikerességére (Ng–Cardie 2002b; Uryupina 2004), azonban esetünkben azt eredményezi, hogy a tesztelés során is sokkal kevesebb pozitív párt azonosít a rendszer. Ezért ezekben az esetekben is megvizsgáltam egy tesztfájlon, hogy egy antecedens azonosítása esetén milyen eredményeket ér el az

osztályozó, ha a legvalószínűbben pozitívnak értékelt főnévi csoportot tekintem antecedensnek, illetve ha a legközelebbit. Ebben az esetben is ugyanazon az egy tesztfájlon végeztem a kísérleteket és a jellemzőkészleteket a cp2 távolságszámolási módszer mellett alkalmaztam, hiszen ezekben a kísérletekben is jobb eredményeket értek el általánosan a cp1-nél. A teljes tesztfájlban 540 névmás volt megtalálható, ebből 320 volt anaforikus. A modellek eredményei a következő két táblázatban láthatóak. Az első táblázatban a teljes tanítófájlon tanuló modell sikeressége, a másodikban pedig a szűrt tanítófájl segítségével épített modell sikeressége látható.

PrsFullExp1		P	R	F
Best	Base	87,92	56,88	69,07
Closest		87,92	56,88	69,07
Best	Pron	87,73	60,31	71,48
Closest		87,73	60,31	71,48
Best	Def	88,29	56,56	68,95
Closest		88,29	56,56	68,95
Best	Length	89,16	56,56	69,22
Closest		89,16	56,56	69,22
Best	Pos	89,55	56,25	69,10
Closest		89,05	55,94	68,71

15. táblázat Egyetlen antecedens azonosításának eredményei a teljes tanítófájl segítségével épített modell esetében

PrsFullExp2		P	R	F
Best	Base	53,26	63,75	58,04
First		53,26	63,75	58,04
Best	Pron	53,45	65,31	58,79
First		54,99	67,19	60,48
Best	Def	54,03	62,81	58,09
First		54,03	62,81	58,09
Best	Length	55,17	65,00	59,68
First		53,85	63,44	58,25
Best	Pos	47,77	50,31	49,01
First		47,18	49,69	48,40

16. táblázat Egyetlen antecedens azonosításának eredményei a negatív és pozitív példákra is szűrt tanítófájl segítségével épített modell esetében

Az összes visszautaló névmást tartalmazó tanulás és tesztelés esetében is megvizsgáltam azt a stratégiát, amelynek során a tanítófájlban kizárólag a negatív példák számát csökkentem (EXP3), hasonlóan a kizárólag harmadik személyű személyes névmásokat tartalmazó tanítófájl esetében. Ezen felül megvizsgáltam azt az esetet is, amelyben a tanítófájlban a pozitív és negatív példák aránya megegyezik (EXP4). A modellek eredményei a következő táblázatokban láthatók.

PrsFullExp3		P	R	F
Best	Base	21,19	34,38	26,22
Closest		15,80	25,63	19,55
Best	Pron	23,02	37,19	28,43
Closest		16,44	26,56	20,31
Best	Def	19,62	31,88	24,29
Closest		15,38	25,00	19,05
Best	Length	26,01	42,19	32,18
Closest		14,64	23,75	18,12
Best	Pos	18,52	29,69	22,81
Closest		14,81	23,75	18,25

17. táblázat Egyetlen antecedens azonosításának eredményei a kizárólag negatív példák szűrésével létrejött tanítófájl segítségével épített modell esetében

PrsFullExp4		P	R	F
Best	Base	44,25	72,19	54,87
Closest		14,18	23,13	17,58
Best	Pron	45,90	75,31	57,04
Closest		14,10	23,13	17,51
Best	Def	44,25	72,19	54,87
Closest		14,18	23,13	17,58
Best	Length	45,68	74,38	56,60
Closest		14,40	23,44	17,84
Best	Pos	45,21	73,75	56,06
Closest		14,56	23,75	18,05

18. táblázat Egyetlen antecedens azonosításának eredményei a pozitív és negatív példákat egységes arányban tartalmazó tanítófájl segítségével épített modell esetében

Az összes kísérlet közül a legjobb fedést a pozitív és negatív példákat azonos arányban tartalmazó (Exp4) tanítófájlok segítségével épített modellek közül érte el az, amelyik a Pron jellemzővel volt kiegészítve és a Best-first módszert alkalmazta. Ez a modell a 320 visszautalásból 241-et azonosított sikeresen. Pontosság tekintetében a teljes tanítófájl segítségével épített modellek értek el jobb eredményt, ezekben az esetekben a hossz és a pozíció jellemzőkkel kiegészített kísérletek értek el magasabb pontosságot. Az F-érték tekintetében összességében a teljes tanítófájl és a Pron jellemző segítségével épített modell érte el a legjobb eredményt. A Best-first és a Closest-first módszerek eredményei között kizárólag az utolsó két kísérletben mutatkoztak nagyobb eltérések, ahol a negatív példák szűrésével jött létre a tanítófájl. Ezekben az esetekben rosszabb eredményeket értek el a modellek, aminek oka valószínűleg az, hogy a pozitív tanítópéldák közé sok távoli visszautalás került be, hiszen a mindenkori beszélőre és hallgatóra nagyobb távolságokból is vissza tudunk utalni, azonban ezeket az eseteket nem egészítették ki a negatív példák a szűrés miatt, így a teszelés során nem tudta helyesen azonosítani az egyes eseteket az osztályozó. Ezt támasztja alá az is, hogy a kizárólag harmadik személyű személyes névmásokat vizsgáló kísérletekben a Closest-first stratégia mutatott jobb eredményeket, itt azonban a Best-first stratégia ért el jobb eredményeket.

7.2. Mutató névmások

A mutató névmások esetében hasonlóan a személyes névmásokhoz két tesztet végeztem el, amely során megvizsgáltam, hogy a modell építése során a tanítófájlok szűrése javít-e a sikerességen. Ehhez mind a két esetben azt az öt jellemzőcsomagot használtam fel, mint a személyes névmás esetében háromszor ismételve a különböző távolságszámítási módszerek tekintetében. A teszteket a teljes korpusz 20%-án végeztem el és ezt ötször ismételtem. Hasonlóan a személyes névmásnál a tesztelés során minden kézzel annotált antecedensre pozitív példaként tekintettem. A különböző módszerekkel és jellemzőkészletek mellett létrehozott modellek eredményei a következő két táblázatban láthatóak. Az első táblázatban a teljes tanítófájl felhasználásával készült modellek kiértékelései, a másodikban pedig a pozitív és negatív példákra is leszűrt tanítófájl segítségével épített modellek eredményei láthatóak.

DemExp1v	RandomForest		
	P	R	F
cp1_Base	40,27	10,40	16,25
cp2_Base	45,75	13,40	20,48
cp12_Base	45,67	13,24	20,28
cp1_Pron	40,34	10,58	16,45
cp2_Pron	46,71	13,79	20,99
cp12_Pron	45,66	13,40	20,46
cp1_Def	41,09	10,80	16,91
cp2_Def	48,61	13,74	21,26
cp12_Def	47,53	13,63	21,01
cp1_Length	42,53	12,50	18,81
cp2_Length	46,11	15,49	22,95
cp12_Length	46,91	14,88	22,35
cp1_Pos	41,71	9,75	15,53
cp2_Pos	42,04	12,57	19,16
cp12_Pos	40,47	12,55	19,06

19. táblázat A teljes tanítófájl segítségével épített modell eredményei

DemExp2v	RandomForest		
	P	R	F
cp1_Base	06,37	75,71	11,72
cp2_Base	07,42	78,78	13,52
cp12_Base	07,47	80,84	13,64
cp1_Pron	06,48	75,83	11,91
cp2_Pron	07,40	78,90	13,49
cp12_Pron	07,90	80,42	14,34
cp1_Def	07,30	76,37	13,29
cp2_Def	08,23	79,63	14,87
cp12_Def	08,44	80,87	15,24
cp1_Length	07,85	71,75	14,11
cp2_Length	08,74	74,26	15,59
cp12_Length	08,83	76,42	15,80
cp1_Pos	02,71	23,24	04,83
cp2_Pos	02,70	23,51	04,83
cp12_Pos	02,93	23,96	05,20

20. táblázat A szűrt tanítófájl segítségével épített modell eredményei

Hasonlóan, mint a személyes névmási visszautalások esetében, a teljes tanítófájlt felhasználva készített modellek a pontosság tekintetében, a szűrt tanítófájlon alapuló modellek pedig a fedés tekintetében értek el jobb eredményeket. A távolságszámítási módszereket tekintve az általam

definiált cp2 jellemző minden esetben javított az eredményeken, minimális különbségek mutatkoznak a tekintetben, hogy mind a két távolságszámítási módszer vagy csak a cp2 az eredményesebb. A jellemzők közül az antecedens hosszára utalók érték el a legjobb eredményeket.

Ebben az esetben is megvizsgáltam, hogy milyen eredményeket érnek el a modellek, ha kizárólag egy antecedens azonosítása a cél. Ehhez a teljes korpusz 20%-ából készítettem egy tesztfájlt és ezen vizsgáltam meg a két módszer segítségével, legvalószínűbb jelölt (Best) és legközelebbi jelölt (Closest), a modelleket. Mivel a fenti két kísérletben közel azonos módon teljesített a cp12 és a cp2 távolságszámítási módszer, ezért azt vettem figyelembe, hogy a fedés általában a cp12 esetében volt magasabb, így ezek mellett a távolságszámítási módszerek mellett alkalmaztam a jellemzőket. A mutató névmás esetében is megvizsgáltam a kizárólag negatív példák szűrésével létrejött tanítófájl és a pozitív és negatív példákat azonos arányban tartalmazó tanítófájl segítségével épített modellek eredményeit is. A két kísérlet eredményei a következő táblázatokban láthatók. Az első táblázatban a teljes tanítófájl segítségével létrehozott modell (EXP1), a másodikban a szűrt tanítófájl alapján létrehozott modell (EXP2) a harmadikban, hasonlóan a személyes névmási visszautalásoknál a kizárólag negatív példákra szűrt modell (EXP3), a negyedikben pedig a pozitív és negatív példákat azonos arányban tartalmazó modell (EXP4) eredményei láthatók.

DemExp1		**P**	**R**	**F**
Best	Base	67,86	14,50	23,90
Closest		67,86	14,50	23,90
Best	Pron	67,86	14,50	23,90
Closest		67,86	14,50	23,90
Best	Def	70,00	16,03	26,09
Closest		70,00	16,03	26,09
Best	Length	70,97	16,79	27,16
Closest		70,97	16,79	27,16
Best	Pos	60,00	13,74	22,36
Closest		60,00	13,74	22,36

21. táblázat Egyetlen antecedens azonosításának eredményei a teljes tanítófájl segítségével épített modell esetében

DemExp2		P	R	F
Best	Base	14,66	42,75	21,83
Closest		15,97	46,56	23,78
Best	Pron	14,32	43,51	21,55
Closest		15,58	47,33	23,44
Best	Def	15,05	42,75	22,27
Closest		15,59	44,27	23,06
Best	Length	14,45	38,93	21,07
Closest		15,30	41,22	22,31
Best	Pos	04,97	14,50	07,41
Closest		06,28	18,32	09,36

22. táblázat Egyetlen antecedens azonosításának eredményei a negatív és pozitív példákra is szűrt tanítófájl segítségével épített modell esetében

DemExp3		P	R	F
Best	Base	11,05	46,56	17,86
Closest		12,86	54,20	20,79
Best	Pron	10,69	45,04	17,28
Closest		12,68	53,44	20,50
Best	Def	11,20	44,27	17,87
Closest		13,51	53,44	21,57
Best	Length	10,87	41,98	17,27
Closest		12,45	48,09	19,78
Best	Pos	11,41	45,04	18,21
Closest		12,57	49,62	20,06

23. táblázat Egyetlen antecedens azonosításának eredményei a kizárólag negatív példák szűrésével létrejött tanítófájl segítségével épített modell esetében

DemExp4		P	R	F
Best	Base	10,73	58,02	18,12
Closest		9,60	51,91	16,21
Best	Pron	11,02	59,54	18,59
Closest		9,46	51,15	15,97
Best	Def	10,45	56,49	17,64
Closest		9,46	51,15	15,97
Best	Length	10,88	58,78	18,36
Closest		9,46	51,15	15,97
Best	Pos	10,45	56,49	17,64
Closest		9,60	51,91	16,21

24. táblázat Egyetlen antecedens azonosításának eredményei a pozitív és negatív példákat egységes arányban tartalmazó tanítófájl segítségével épített modell esetében

A mutató névmási visszautalások tekintetében a teljes tesztfájlban 708 névmás szerepelt, ebből 131 volt visszautaló névmás, tehát az összes mutató névmás 18,5%-a visszautaló. A pontosság tekintetében a teljes tanítófájlon és a hossz (Length) jellemző segítségével épített modell lett a legsikeresebb, a fedés tekintetében a negatív és pozitív példákat azonos arányban tartalmazó tanítófájl, a Pron jellemzőkészlet segítségével épített és Best-first módszert alkalmazó modell lett a legsikeresebb. Azonban azt is fontos kiemelni, hogy a mutató névmási visszautalások esetében kisebb eltérések mutatkoztak a kísérletek eredményei között, a szűrt tanítófájlokon épített modellek a pozíció jellemző kivételével hasonlóan 50 és 60 közötti, a negatív példákra szűrt tanítófájlon épített modellek pedig egységesen 60-70 közötti számú visszautalást ismertek fel. Az alap jellemzőkészlethez képest még a hossz jellemző rontott az eredményeken a szűrt tanítófájlok esetében, ami abból a szempontból érdekes, hogy a teljes tanítófájlon épített modelleknél viszont ezzel a jellemzővel érte el a modell a legnagyobb pontosságot. Ezek alapján a Length jellemző a falspozitív esetek szűrésében a Pron jellemző pedig az antecedens azonosításának valószínűségén növelt.

7.3. Vonatkozó névmás

A vonatkozó névmási visszautalás általában a megelőző első vagy második főnévi csoportra utal, ezért első ránézésre nem okozhat gondot az anaforafeloldás szempontjából. Jelen esetben

azonban nem kizárólag az antecedens felismerése a feladat, hanem annak felismerése is, hogy az adott névmás egyáltalán visszautal-e. A vonatkozó névmás szerepelhet szerkezetben *A lány, akit tegnap láttam,* de utalhat vissza teljes tagmondatra is *Tegnap összebarátkoztam egy lánnyal, aminek nagyon örülök*. Ezekben az esetekben a modellnek fel kellene ismernie, hogy a névmásnak nincs a felsoroltak között antecedense. A vonatkozó névmási visszautalás esetében is öt részre osztottam fel a korpuszt a koreferencialáncokban szereplő vonatkozó névmások száma alapján, minden esetben négy részből készült el a tanítófájl és egy részből a teszt fájl. Az első kísérletben a tanítófájlban szereplő párokat nem szűrtem le (EXP1v), tehát az összes vonatkozó névmást tartalmazza, azokat is, amelyek nem utalnak vissza. A második kísérletben mind a negatív, mind a pozitív példákat szűrtem (EXP2v). Az eredmények a következő két táblázatban láthatóak.

RelExp1v	**RandomForest**		
	P	**R**	**F**
cp1_Base	63,95	52,46	57,47
cp2_Base	63,21	52,61	57,30
cp12_Base	63,28	51,99	56,95
cp1_Pron	63,59	51,56	56,74
cp2_Pron	63,32	52,16	57,07
cp12_Pron	62,72	52,02	56,71
cp1_Def	66,38	55,71	60,39
cp2_Def	66,12	55,53	60,21
cp12_Def	66,39	55,54	60,27
cp1_Length	65,95	52,70	58,46
cp2_Length	65,42	53,30	58,58
cp12_Length	65,66	53,79	59,03
cp1_Pos	61,33	50,76	55,34
cp2_Pos	60,45	51,69	55,48
cp12_Pos	60,41	52,64	56,03

25. táblázat A teljes tanítófájl segítségével épített modell eredményei

RelExp2v	RandomForest		
	P	R	F
cp1_Base	02,24	97,88	04,38
cp2_Base	02,48	97,72	04,84
cp12_Base	02,11	97,73	04,12
cp1_Pron	02,13	97,72	04,17
cp2_Pron	02,29	97,88	04,48
cp12_Pron	02,03	97,73	03,98
cp1_Def	02,32	97,73	04,53
cp2_Def	02,42	97,89	04,73
cp12_Def	02,03	97,88	03,98
cp1_Length	02,36	97,42	04,61
cp2_Length	02,26	97,42	04,42
cp12_Length	02,32	97,42	04,53
cp1_Pos	02,15	38,55	04,08
cp2_Pos	02,20	38,84	04,17
cp12_Pos	02,15	40,66	04,09

26. táblázat A szűrt tanítófájl segítségével épített modell eredményei

A két kísérlet eredményei atekintetben hasonlóak a többi névmás eredményeihez, hogy a negatív tanítópéldák számának növelése sokat javít a pontosságon, azonban abban eltér tőlük, hogy a szűrt tanítófájlok alapján épített modell szinte többségi címkézést végez. A magas fedést az eredményezi, hogy a modell majd minden párt pozitívnak ítél, ezt pedig az okozza, hogy a tanítófájlokban is nagyon sok a pozitív tanító példa. Mivel a szűrt tanítófájlokban a negatív példák a névmás és az antecedense közötti főnévi csoportokból jönnek létre, de a vonatkozó névmási visszautalások esetében nincs sok ilyen főnévi csoport, nem áll rendelkezésre elegendő negatív példa a modell számára. Ezért a jellemzők sikerességéről vagy a távolságszámítási módszerekről a szűrt tanítófájlok esetében a tesztek alapján nem vonható le konklúzió. A teljes tanítófájl alapján épített modell esetében azonban az antecedens határozottságára utaló Def jeggyel kiegészített modell érte el a legjobb eredményt.

Ebben az esetben is elvégeztem egy tesztfájlon a három kísérletet, amelyikben csak egy antecedenst azonosítok a névmásokhoz. A vonatkozó névmási visszautalásoknál a távolságszámolási módszerre nézve már nem olyan egyértelműek az eredmények, mint a többi névmás esetében, két esetben a cp1 két esetben pedig a cp2 ért el jobb eredményt, mivel a többi névmást is a cp2 módszerrel vizsgáltam tovább ezért itt is a cp2 távolságszámítási módszert alkalmaztam. A négy kísérlet eredményei a következő táblázatokban láthatók.

RelExp1		P	R	F
Best	Base	57,02	49,24	52,85
Closest		57,89	50,00	53,66
Best	Pron	57,02	49,24	52,85
Closest		57,89	50,00	53,66
Best	Def	59,81	48,48	53,56
Closest		60,75	48,48	54,39
Best	Length	60,36	50,76	55,14
Closest		58,56	49,24	53,50
Best	Pos	55,56	45,45	50,00
Closest		55,56	45,45	50,00

27. táblázat Egyetlen antecedens azonosításának eredményei a teljes tanítófájl segítségével épített modell esetében

RelExp2		P	R	F
Best	Base	28,33	77,27	41,46
Closest		30,28	82,58	44,31
Best	Pron	29,44	80,30	43,09
Closest		30,56	83,33	44,72
Best	Def	28,49	77,27	41,63
Closest		31,01	84,09	45,31
Best	Length	28,13	76,52	41,14
Closest		30,36	82,58	44,40
Best	Pos	09,64	24,24	13,79
Closest		13,25	33,33	18,97

28. táblázat Egyetlen antecedens azonosításának eredményei a negatív és pozitív példákra is szűrt tanítófájl segítségével épített modell esetében

RelExp3		P	R	F
Best	Base	26,94	73,48	39,43
Closest		31,67	86,36	46,34
Best	Pron	25,56	69,70	37,40
Closest		31,67	86,36	46,34
Best	Def	27,50	75,00	40,24
Closest		31,94	87,12	46,75
Best	Length	25,83	70,45	37,80
Closest		31,11	84,85	45,53
Best	Pos	25,83	70,45	37,80
Closest		31,11	84,85	45,53

29. táblázat Egyetlen antecedens azonosításának eredményei a kizárólag negatív példák szűrésével létrejött tanítófájl segítségével épített modell esetében

RelExp4		P	R	F
Best	Base	42,11	84,85	56,28
Closest		43,23	87,12	57,79
Best	Pron	44,70	89,39	59,60
Closest		43,94	87,88	58,59
Best	Def	43,02	86,36	57,43
Closest		43,40	87,12	57,93
Best	Length	42,80	85,61	57,07
Closest		43,18	86,36	57,58
Best	Pos	42,80	85,61	57,07
Closest		43,56	87,12	58,08

30. táblázat Egyetlen antecedens azonosításának eredményei a pozitív és negatív példákat egységes arányban tartalmazó tanítófájl segítségével épített modell esetében

A vonatkozó névmási visszautalások esetében a tesztfájlban 360 darab névmás volt megtalálható, ezek közül 132 volt visszautaló névmás, amely főnévi csoportra utalt vissza. A pontosság tekintetében a teljes tanítófájl alapján és a Def és Length jellemzőkön alapuló modellek érték el a legjobb eredményeket. A fedés tekintetében a negatív és pozitív példákat azonos arányban tartalmazó tanítófájl segítségével épített modell a Pron jellemzővel kiegészítve, a Best-first módszerrel érte el a legjobb eredményt Az eredmények alapján elmondható, hogy a

vonatkozó névmási visszautalások tekintetében a szűrt tanítófájlon épített modellek általánosságban kevéssé eredményesek a negatív példák hiánya miatt, azonan abban az esetben ha pozitív és negatív példák is egyenlő arányban vannak jelen, már jobb eredményeket érhetünk el.

7.4. A kísérletek összegzése

Az előző fejezetben minden egyes névmás esetében bemutattam az elvégzett kisérleteket, jelen fejezetben ezeknek a kísérleteknek az eredményeit összegzem úgy, hogy az egyes névmások eredményei is összehasonlíthatók legyenek. Az első két kísérletben minden névmás esetében az összes antecedens azonosítása volt a cél, a távolabbi és nagy mennyiségű pozitív pár miatt pedig ezekben az esetekben a különböző távolságszámítási módszerek is összehasonlíthatóvá váltak. Ezeknek a kísérleteknek az eredményeit a következő táblázat mutatja.

F	EXP1v	EXP2v	EXP1v	EXP2v	EXP1v	EXP2v	EXP1v	EXP2v
	PrsPer3		PrsFull		Dem		Rel	
cp1_Base	18,30	07,20	64,00	31,12	16,25	11,72	57,47	04,38
cp2_Base	21,30	09,88	64,00	42,90	20,48	13,52	57,30	04,84
cp12_Base	19,80	05,98	64,10	41,96	20,28	13,64	56,95	04,12
cp1_Pron	19,10	07,56	67,30	34,63	16,45	11,91	56,74	04,17
cp2_Pron	**21,50**	09,96	**67,50**	42,20	20,99	13,49	57,07	04,48
cp12_Pron	19,60	06,14	**67,50**	41,55	20,46	14,34	56,71	03,98
cp1_Def	17,90	08,40	63,80	35,49	16,91	13,29	**60,39**	04,53
cp2_Def	20,80	10,54	64,20	46,56	21,26	14,87	60,21	04,73
cp12_Def	19,90	07,22	64,20	45,15	21,01	15,24	60,27	03,98
cp1_Length	15,70	07,18	64,70	35,18	18,81	14,11	58,46	04,61
cp2_Length	19,60	12,14	64,90	45,58	**22,95**	15,59	58,58	04,42
cp12_Length	18,50	06,88	64,80	45,83	22,35	15,80	59,03	04,53
cp1_Pos	19,00	05,30	64,10	16,64	15,53	04,83	55,34	04,08
cp2_Pos	20,10	05,56	64,20	17,66	19,16	04,83	55,48	04,17
cp12_Pos	18,50	04,84	64,30	16,61	19,06	05,20	56,03	04,09

31. táblázat Az EXP1v és EXP2v kísérletek eredményeinek összesítése F-mérték alapján

Az EXP2v azaz a szűrt tanítófájlok segítségével épített modellek minden névmás esetében rosszabb eredményt értek el az EXP1v-nél, hiszen ezekben az esetekben a távolabbi névmás – főnévi csoport kapcsolatokra nem volt példa a tanítófájlokban, a tesztelés során azonban minden egyes ilyen kapcsolat felismerése cél volt. A tesztekben a vonatkozó névmási visszautaláson

kívül minden névmás esetében javított a CP2 távolságszámítási módszer a CP1 módszerhez képest. A kognitív alapon megfogalmazott jellemzők tekintetében az látható, hogy a Pron jellemző a személyes névmás, a Length jellemző a mutatónévmás a Def jellemző pedig a vonatkozó névmás esetében érte el a legjobb eredményt. A pozícióra vonatkozó Pos jellemző átlagosan rosszabb eredményeket ért el, ami szintén abból következhet, hogy a távolabbi visszautalások esetében már nem olyan mérvadó az antecedens pozíciója, mint a megelőző tagmondatra történő visszautalás esetében, ahol fontos mutatója az alanyváltásnak vagy az alany megtartásának. Ezt támasztja alá az is, hogy a szűrt tanítófájlok segítségével épített modellekben kizárólag a legközelebbi annotált antecedens volt pozitív példa, amelynél még mérvadó lehetett ez a szempont, ezért a tesztelés során, ahol az összes pár azonosítása cél volt sokkal rosszabbak lettek az eredmények, ellenben ahol a tanítófájlban távolabbi visszautalások is megjelentek, ahol már nem volt annyira fontos a pozíció a tesztelés során sem lettek sokkal rosszabbak az eredmények a többi jellemzőhöz képest.

Az első két kísérlet után minden névmás esetében további négyet végeztem el, ahol egyetlen antecedens azonosítása volt a célom. A következő táblázatok közül az első a kísérletek eredményeinek F-mértékét, a második pedig a kísérletek során kapott fedéseket mutatja.

F		PrsPer3				PrsFull				Dem				Rel			
		EXP 1	EXP 2	EXP 3	EXP 4	EXP 1	EXP 2	EXP 3	EXP 4	EXP 1	EXP 2	EXP 3	EXP 4	EXP 1	EXP 2	EXP 3	EXP 4
Base	B.	22,50	21,67	18,23	23,89	69,07	58,04	26,22	54,87	23,90	21,83	17,86	18,12	52,85	41,46	39,43	56,28
	C.	22,50	25,39	**26,78**	15,00	69,07	58,04	19,55	17,58	23,90	23,78	20,79	16,21	53,66	44,31	46,34	57,79
Pron	B.	22,64	19,51	17,00	23,89	**71,48**	58,79	28,43	57,04	23,90	21,55	17,28	18,59	52,85	43,09	37,40	**59,60**
	C.	22,64	24,39	26,06	15,56	**71,48**	60,48	20,31	17,51	23,90	23,44	20,50	15,97	53,66	44,72	46,34	58,59
Def	B.	21,52	22,43	14,29	23,89	68,95	58,09	24,29	54,87	26,09	22,27	17,87	17,64	53,56	41,63	40,24	57,43
	C.	21,52	23,68	26,29	14,44	68,95	58,09	19,05	17,58	26,09	23,06	21,57	15,97	54,39	45,31	46,75	57,93
Length	B.	22,22	20,67	17,77	24,44	69,22	59,68	32,18	56,60	**27,16**	21,07	17,27	18,36	55,14	41,14	37,80	57,07
	C.	22,22	23,71	22,35	16,11	69,22	58,25	18,12	17,84	**27,16**	22,31	19,78	15,97	53,50	44,40	45,53	57,58
Pos	B.	20,78	7,74	15,56	25,56	69,10	49,01	22,81	56,06	22,36	7,41	18,21	17,64	50,00	13,79	37,80	57,07
	C.	20,78	7,74	24,21	13,89	68,71	48,40	18,25	18,05	22,36	9,36	20,06	16,21	50,00	18,97	45,53	58,08

32. táblázat Az EXP1,EXP2,EXP3 és EXP4 kísérletek eredményeinek összesítése F-mérték alapján

R		PrsPer3				PrsFull				Dem				Rel			
		EXP 1	EXP 2	EXP 3	EXP 4	EXP 1	EXP 2	EXP 3	EXP 4	EXP 1	EXP 2	EXP 3	EXP 4	EXP 1	EXP 2	EXP 3	EXP 4
Base	B.	15,93	30,97	28,32	38,05	56,88	63,75	34,38	72,19	14,50	42,75	46,56	58,02	49,24	77,27	73,48	84,85
	C.	15,93	36,28	**41,59**	23,89	56,88	63,75	25,63	23,13	14,50	46,56	54,20	51,91	50,00	82,58	86,36	87,12
Pron	B.	15,93	28,32	26,55	38,05	60,31	65,31	37,19	**75,31**	14,50	43,51	45,04	**59,54**	49,24	80,30	69,70	**89,39**
	C.	15,93	35,40	40,71	24,78	60,31	67,19	26,56	23,13	14,50	47,33	53,44	51,15	50,00	83,33	86,36	87,88
Def	B.	15,04	31,86	22,12	38,05	56,56	62,81	31,88	72,19	16,03	42,75	44,27	56,49	48,48	77,27	75,00	86,36
	C.	15,04	33,63	40,71	23,01	56,56	62,81	25,00	23,13	16,03	44,27	53,44	51,15	48,48	84,09	87,12	87,12
Length	B.	15,04	30,09	27,43	38,94	56,56	65,00	42,19	74,38	16,79	38,93	41,98	58,78	50,76	76,52	70,45	85,61
	C.	15,04	34,51	34,51	25,66	56,56	63,44	23,75	23,44	16,79	41,22	48,09	51,15	49,24	82,58	84,85	86,36
Pos	B.	14,16	10,62	23,89	40,71	56,25	50,31	29,69	73,75	13,74	14,50	45,04	56,49	45,45	24,24	70,45	85,61
	C.	14,16	10,62	37,17	22,12	55,94	49,69	23,75	23,75	13,74	18,32	49,62	51,91	45,45	33,33	84,85	87,12

33. táblázat Az EXP1,EXP2,EXP3 és EXP4 kísérletek eredményeinek összesítése fedés (R) alapján

A fedés mutatja, hogy a gold standard korpuszban található anaforikus kapcsolatokból az osztályozó mennyit azonosított sikeresen. Az általam meghatározott hipotézisek alapján ezeken az értékeken vártam javulást a kognitív jellemzők hatására. Az egyes névmásokon végzett kísérletek alapján egyértelműen látszik, hogy az anaforikus kapcsolatok azonosításán javít a tanítófájlban szereplő pozitív példák nagyobb aránya, tehát az EXP3 és EXP4-es tanítófájlok értek el jobb eredményeket, ezekben az esetekben nagyobb eltérések láthatók a Best-firs és Closest-first módszerek között is. A jellemzők tekintetében nem láthatók nagy eltérések, azonban a legjobb eredményt a Pron jellemzővel kiegészített jellemzőkészlet segítségével épített osztályozó érte el az összes névmást vizsgáló, a mutató névmást vizsgáló és a vonatkozó névmást vizsgáló kísérletben is.

Az F mérték tekintetében már nagyobb különbségek láthatók az egyes névmástípusok között, ebből arra következtetek, hogy a nem anaforikus kapcsolatok azonosításában okoznak nagyobb eltéréseket a kognitív alapon megfogalmazott jellemzők. Ez alól kivételt képez az egyes szám harmadik személyű személyes névmás, ahol a fedéssel azonos módon az alap jellemző készlet és az EXP3 módon összeállított tanítófájl a legsikeresebb a vizsgált módszerek közül. A jellemzők alapján az összes névmás és a vonatkozó névmás esetében a Pron, a mutatónévmás esetében a Length jellemzőkkel kiegészített jellemzőkészlet volt a legsikeresebb módszer. Az egy antecedens kiválasztását elősegítő módszerek közül az összes névmás és a mutatónévmás esetében, ahol a tanítófájlban a negatív példák túlnyomó többségben voltak a pozitív példákkal szemben (EXP1), nem mutatkozott eltérés, a kizárólag egyes szám harmadik személyű

személyes névmási visszautalások esetében a Closest-first, a vonatkozó névmások esetében a Best-first mutatkozott eredményesebbnek.

7.5. Tesztelés

A fenti kísérleteket kizárólag a SzegedKoref korpusz segítségével végeztem el, mivel ez a korpusz tartalmazott elegendő visszautalást a kísérletekhez. Ahhoz, hogy megvizsgáljam a hipotéziseim helytállóságát az antecedens felismerését illetően, további két tesztet végeztem el. Ezekben a tesztekben már kizárólag a harmadik személyű személyes névmást, a mutató névmást és a vonatkozó névmást vettem figyelembe. Mivel mind a három névmástípus esetében a negatív példákra szűrt tanítófájl volt a legsikeresebb a fedés tekintetében, ezért ezzel a módszerrel építettem modelleket a teljes SzegedKoref korpuszon. Az első kísérletben kézzel létrehozott saját teszteseteken értékeltem ki a modelleket, a második esetben pedig a KorKorpusz névmási visszautalásain.

7.5.1. A saját teszteseteken végzett kísérlet

A fenti kísérletek során az általam megfogalmazott, különböző kognitív alapú jellemzők alapján épített modellek között mutatkoztak eltérések, azonban az nem látható, hogy pontosan melyik részlet javít vagy ront az eredményeken. A korpuszban megtalálható előfordulások alapján nehéz megmondani, hogy a kognitív alapú jellemzők milyen tulajdonságokkal rendelkező antecedensek azonosításán javított és esetlegesen milyeneken rontott. A végeredményt pedig nagyban befolyásolja, hogy az egyes antecedens típusokból mennyi található a korpuszban. A kísérlet célja, hogy kizárólag az antecedens általam is megfogalmazott kognitív tulajdonságait változtatva megvizsgáljam, hogy az alap jellemzőkészlethez képest az általam is megfogalmazott jellemzők hogyan módosítják az adott antecedens típus felismerésének sikerességét. Ehhez minimálpárokat vizsgálok meg, így ellenőrizhető környezetben kizárólag csak azokat a paramétereket tudom változtatni, amelyeket aktuálisan vizsgálok. Tehát a következő példák esetében a kérdés nem az, hogy a névmáshoz az osztályozó mely antecedensjelöltet választja, hanem az, hogy az adott jellemzők segítségével épített osztályozó döntését negatív vagy pozitív irányba befolyásolja az antecedens nyelvi jellemzőinek módosítása. Ezzel kapunk egyfajta kvalitatív elemzést is az egyes modellekhez.

7.5.1.1. Személyes névmás

A személyes névmási visszautalás alatt ebben az esetben a harmadik személyű névmást értem, hiszen az első és második személyű névmás a mindenkori beszélőre, illetve a mindenkori hallgatóra utal, ezeket a visszautalásokat pedig a morfológiai jegyek egyeztetésén keresztül egyszerű felismerni. A személyes névmás tekintetében kilenc példát készítettem és mindig csak az általam megfogalmazott adott jellemző tekintetében módosítottam az antecedens tulajdonságain. Így a hossz tekintetében három példát, a pozíció, az antecedens névmási volta és a határozottság tekintetében pedig két-két példát hasonlítottam össze. A lentebbi példák közül nyolcat sikeresen azonosított a kilencből az alap (Base) jellemzőkészlet segítségével épített osztályozó is, egyedül azt az esetet nem ismerte fel, amelyben a névmáshoz tartozó antecedens is névmás volt 61).

Hossz

Az általam megfogalmazott az antecedens hosszára vonatkozó jellemzők segítségével épített osztályozó a hosszú határozott kifejezést 56) nem ismerte fel antecedensként. Emellett érdemes megemlíteni, hogy az osztályozó az alap jellemzőkészlet segítségével épített modellel ellenben minden esetben sikeresen azonosította az egymással nem anaforikus kapcsolatban álló kifejezéseket. Tehát a hossz mint jellemző a fals pozitív esetek szűrésében mutatkozik eredményesnek a személyes névmási visszautalás esetében.

56) A piros szoknyás lány a nyulat kergette, de sehogyan sem tudta elkapni azt. Egyszer hirtelen eltűnt a talaj a lábai alól és ő is és a nyúl is egy mély gödörbe zuhantak.

57) A lány a nyulat kergette, de sehogyan sem tudta elkapni azt. Egyszer hirtelen eltűnt a talaj a lábai alól és ő is és a nyúl is egy mély gödörbe zuhantak.

58) Alíz a nyulat kergette, de sehogyan sem tudta elkapni azt. Egyszer hirtelen eltűnt a talaj a lábai alól és ő is és a nyúl is egy mély gödörbe zuhantak.

Pozíció

Az általam az alap jellemzőkészlethez adott, az antecedens pozíciójára utaló jellemző nem változtatott az antecedens azonosításának sikerességén, azonban hasonlóan az antecedens hosszára utaló jellemzőhöz, ezzel a kibővített jellemzőlistával is sikeresen azonosította a modell az összes egymással nem anaforikus kapcsolatban álló kifejezéspárt.

59) A lány a nyulat kergette, de sehogyan sem tudta elkapni azt. Egyszer hirtelen eltűnt a talaj a lábai alól és ő is és a nyúl is egy mély gödörbe zuhantak.

60) A nyulat kergette a lány, de sehogyan sem tudta elkapni azt. Egyszer hirtelen eltűnt a talaj a lábai alól és ő is és a nyúl is egy mély gödörbe zuhantak.

Pron

A Pron jellemzővel kiegészített jellemzőkészlet ugyanazt az eredményt érte el, mint az alap jellemzőkészlet, tehát az első esetet, ahol az antecedens névmás 61), nem azonosította helyesen a hozzáadott jellemző segítségével sem.

61) Ő a nyulat kergette, de sehogyan sem tudta elkapni azt. Egyszer hirtelen eltűnt a talaj a lábai alól és ő is és a nyúl is egy mély gödörbe zuhantak.

62) A lány a nyulat kergette, de sehogyan sem tudta elkapni azt. Egyszer hirtelen eltűnt a talaj a lábai alól és ő is és a nyúl is egy mély gödörbe zuhantak.

Def

Ebben az esetben is megegyeznek az osztályozó által hozott döntések az alap jellemzőkészlet segítségével épített modellével.

63) Egy lány a nyulat kergette, de sehogyan sem tudta elkapni azt. Egyszer hirtelen eltűnt a talaj a lábai alól és ő is és a nyúl is egy mély gödörbe zuhantak.

64) A lány a nyulat kergette, de sehogyan sem tudta elkapni azt. Egyszer hirtelen eltűnt a talaj a lábai alól és ő is és a nyúl is egy mély gödörbe zuhantak.

A személyes névmás esetében a tesztek alapján elmondható, hogy a különböző tulajdonságokkal rendelkező antecedensek felismerésén nem módosítanak a különböző jellemzők segítségével épített osztályozók. A tesztesetek alapján a különbség főként az egymással nem anaforikus kapcsolatban álló párok azonosításából adódik, ezeken javítanak az egyes modellek.

7.5.1.2. Mutató névmás

A mutató névmás esetében a személyes névmáshoz hasonlóan a jellemzők alapján kilenc tesztesetet készítettem, majd megvizsgáltam az egyes osztályozók döntéseit az adott esetekre. Az alap jellemzőkészlet segítségével épített osztályozó a kilenc eset közül kettőt azonosított helyesen, azokat, amelyek a névmás jellemző vizsgálatára készültek 72) 73). Tehát elöljáróban

elmondható, hogy a mutató névmáshoz tartozó antecedens azonosításának esetében az alap jellemző készlet segítségével épített osztályozó nem mutatkozik sikeresnek.

Hossz

Az alap jellemzőkészlet segítségével épített modell az egyik esetet sem azonosította helyesen, a négy, az antecedens hosszára utaló jellemző hozzáadása után a hosszú határozott leírást 65) sikeresen azonosította az osztályozó. A nem anaforikus kapcsolatok felismerésében hasonló eredményeket ért el, mint az alap jellemzőkészlet. Tehát a bővített jellemzőkészlet segítségével épített osztályozó javított a hosszú antecedens felismerésén.

65) A lány a fürge fehér nyulat kergette, de sehogyan sem tudta elkapni azt
66) A lány a nyulat kergette, de sehogyan sem tudta elkapni azt.
67) A lány Hógolyót kergette, de sehogyan sem tudta elkapni azt.

Pozíció

Az alap jellemzőkészlet segítségével épített modell az alábbi két esetet sem azonosította sikeresen, az antecedens pozíciójára utaló két jellemző hozzáadása után a tagmondat első pozíciójában szereplő antecedenst 68) az osztályozó sikeresen azonosította. A nem anaforikus kapcsolatok felismerésében azonban rosszabb eredményeket ért el az alap jellemzőkészlet segítségével épített modell eredményeinél.

68) A nyulat kergette a lány, de sehogyan sem tudta elkapni azt.
69) A lány a nyulat kergette, de sehogyan sem tudta elkapni azt.

Def

Az alap (Base) jellemzőkészlet alapján az osztályozó nem ismerte fel egyik visszautalást sem, a határozottságra utaló (Def) jellemző hozzáadása a tanulás alapjául szolgáló jellemzőkészlethez azonban a határozatlan névmást tartalmazó antecedens 71) felismerését lehetővé tette. A nem anaforikus kapcsolatok felismerésében a döntések megegyeznek az alap jellemzőkészlet segítségével épített modell döntéseivel.

70) A lány a nyulat kergette, de sehogyan sem tudta elkapni azt.
71) A lány egy nyulat kergetett, de sehogyan sem tudta elkapni azt.

Pron

A következő két példában szereplő visszautalást az alap jellemzőkészlettel is azonosította az osztályozó, aminek oka valószínűleg az antecedenshez rendelt Szófaji címke PRON címkéje. Hiszen a Szeged Korpuszban gyakran előfordul, hogy a koreferencialáncban a sokadik névmási visszautalás már névmási antecedensre történik. Az általam megfogalmazott jellemző az osztályozó döntésén ebben a tekintetben nem változtatott, azonban a nem anaforikus kapcsolatok felismerésén rontott.

72) Alíz a nyulat kergette, de sehogyan sem tudta elkapni azt. Egyszer hirtelen eltűnt a talaj a lábai alól és azzal együtt egy mély gödörbe zuhant.

73) Alíz a nyulat kergette, de sehogyan sem tudta elkapni az állatot. Egyszer hirtelen eltűnt a talaj a lábai alól és azzal együtt egy mély gödörbe zuhant.

A mutatónévmás esetében az alap jellemzőkészlet (Base) segítségével épített osztályozó már nem ilyen eredményes, a kilenc eset közül összesen kettőt azonosított helyesen. A bővített jellemzőkészletek közül javított a hossz (Length) segítségével épített osztályozó a hosszú, tehát háromnál több szavas antecedens felismerésén, szintén javított a pozíció (Pos) segítségével épített osztályozó a tagmondat első pozíciójában szereplő antecedens felismerésén, a határozottságra utaló jellemző (Def) hozzáadása szintén javított az alap jellemzőkészlet segítségével épített osztályozóhoz képest a határozatlan névelős antecedens felismerésén. A Pron jellemző segítségével épített osztályozó az antecedens felismerésén nem változtatott, azonban a nem anaforikus kapcsolatok azonosításán rontott.

7.5.1.3. Vonatkozó névmás

A vonatkozó névmási visszautalás tekintetében bonyolultabb a helyzet, hiszen az antecedens közelsége miatt kevés negatív tanítópélda kerül a tanítófájlba, ez pedig szinte többségi címkézést eredményez az anaforikus kapcsolat javára. Az alap jellemzőkészlet minden esetet anaforikusnak ítélt a tesztfájlban, ezen egyik jellemző sem változtatott. Ebből arra következtetek, hogy a korábbi kísérletekben pusztán a Best-First és Closest-First módszerek, tehát az egyetlen antecedens azonosítására való szűkítés, kimenete miatt nem történt többségi címkézés, ami azt jelenti, hogy az általam hozzáadott kognitív alapú jellemzők a vonatkozó névmás esetében nem módosítanak az osztályozó eredményességén, pusztán az antecedens közelsége magyarázhatja a az eredményeket.

Hossz

74) A lány a nyulat kergette, akinek fehér bundája volt.

75) A lány a fürge kicsi nyulat kergette, akinek fehér bundája volt.

76) A lány Hógolyót kergette, akinek fehér bundája volt.

Pozíció

77) Azt a nyulat kergette a lány, akinek fehér bundája volt.

78) A lány a nyulat kergette, akinek fehér bundája volt.

Pron

79) Azt kergette a lány, amelyiknek fehér bundája volt.

80) Azt a nyulat kergette a lány, akinek fehér bundája volt.

Def

81) Egy nyulat kergetett a lány, amelyiknek fehér bundája volt.

82) Azt a nyulat kergette a lány, akinek fehér bundája volt.

7.5.2. A KorKorpuszon végzett teszt

A névmási visszautalások esetében a KorKorpuszban azokat az eseteket, ahol kopulát tartalmazott a mondat, ezért nem volt egységes a szintaktikai elemzése a Szeged Korpuszéval, illetve ahol valamilyen annotációs hiba volt, például a névmás jelölve volt, mint visszautaló névmás, de nem volt hozzá antecedens rendelve, nem vettem figyelembe. Sajnos a kopulás mondatok nagy száma miatt ez a megszorítás nagy mennyiségű visszautalást szűrt ki, így ezek után a mutató névmási visszautalásokat nem tudtam megvizsgálni, mivel hat darab maradt a teszt fájlban. A szűrés után a KorKorpuszban 221 harmadik személyű személyes névmás maradt, amiből 64 volt visszautaló névmás. Vonatkozó névmásból 402 darabot azonosítottam, amiből 214 volt visszautaló névmás. A harmadik személyű személyes névmásokra és a vonatkozó névmásokra is a negatív példákra szűrt tanítófájlok (Exp3) segítségével épített modelleket teszteltem. A KorKorpuszból kinyert névmási visszautalások tekintetében a következő eredményeket érte el a modell.

PrsPer3Exp3		P	R	F
Best	Base	04,95	15,63	07,52
Closest		06,93	21,88	10,53
Best	Pron	03,38	10,94	05,17
Closest		06,76	21,88	10,33
Best	Def	04,95	15,63	07,52
Closest		07,43	23,44	11,28
Best	Length	05,05	15,63	07,63
Closest		07,58	23,44	11,45
Best	Pos	03,98	12,50	06,04
Closest		07,96	25,00	12,08

34. táblázat A harmadik személyű személyes névmási visszautalások tesztelése a KorKorpuszban

RelExp3		P	R	F
Best	Base	30,87	56,54	39,93
Closest		37,24	68,22	48,18
Best	Pron	30,79	56,54	39,87
Closest		37,15	68,22	48,11
Best	Def	29,85	54,67	38,61
Closest		37,24	68,22	48,18
Best	Length	31,55	57,94	40,86
Closest		36,39	66,82	47,12
Best	Pos	32,65	59,81	42,24
Closest		37,24	68,22	48,18

35. táblázat A vonatkozó névmási visszautalások tesztelése a KorKorpuszban

A két kísérlet alapján látszik, hogy általánosan rosszabbak az eredmények, mint a SzegedKoref Korpusz esetében, de ezt a visszautalások számán kívül maguk a szövegek típusa és a mondatok szintaktikai szerkezete is okozhatja, hiszen a két korpusz eltérő típusú szövegeket tartalmaz. A harmadik személyű személyes névmási visszautalás tekintetében a pozíció jellemzőkkel kiegészített jellemzőlista segítségével épített modell érte el a legjobb eredményt a Closest-first szűrési stratégiával kiegészítve. A vonatkozó névmási visszautalások tekintetében

az alap jellemzőkészlet segítségével épített modell érte el a legjobb eredményt a Closest-first startégiával kiegészítve.

7.6. Hibaelemzés

A kísérletek alapján látható, hogy az Exp1, tehát a tesztfájlhoz azonos arányú negatív példát tartalmazó tanítófájl segítségével épített modell eredményein kívül, az épített modellek a pontosság tekintetében gyengébben teljesítenek. A pontosság azt mutatja meg, hogy az anaforikusnak ítélt párok közül mennyi volt ténylegesen anaforikus. Jelen esetben a rossz pontossági értéknek két oka van, vagy egy nem visszautaló névmáshoz azonosít az osztályozó antecedenst, vagy a visszautaló névmáshoz hibásan azonosít az osztályozó antecedenst. Ahhoz, hogy megvizsgáljam a leggyakoribb hibákat, először a visszautaló és nem visszautaló névmások arányát hasonlítottam össze, majd ehhez hasonlítom a fals pozitív esetek számát. Mind a három névmás esetében gyakori a nem visszautaló névmáshoz történő antecedens azonosítása, azaz majd minden a szövegben előforduló névmáshoz azonosítanak az osztályozók legalább egy antecedenst. Tehát a hibák többségét az okozza, hogy az osztályozó nem ismeri fel a nem visszautaló névmásokat.

A mutató és vonatkozó névmások esetében a nem visszautaló és visszautaló névmások aránya helytállónak tűnik. A mutató névmások között gyakori lehet a deixis, illetve a nem főnévi csoportra, hanem teljes propozícióra való utalás, ugyanez igaz a vonatkozó névmásra is, ami gyakran teljes tagmondatra utal vissza. A harmadik személyű személyes névmás esetében azonban kimagaslóan sok a nem visszautaló névmások aránya, ami nem feltétlenül várható a korpuszokat alkotó szövegek típusából. Tehát meg kell vizsgálni a nem visszautaló, harmadik személyű személyes névmásokat a tesztfájlokban.

A tesztfájlok vizsgálata után a következő hibákat fedeztem fel, amelyek befolyásolták a modellek eredményességét: 1 A tesztfájlokban gyakoriak a harmadik személyű, teljes propozícióra utaló személyes névmások, amelyeknek ugyan van antecedense, de nem főnévi csoport, a leggyakrabban ezek az *utána* és *előtte* szavak.

83) (…) ott maradtunk még 10 percet, de *utána* feleslegesnek tartottuk magunkat.

2 Az általam kialakított módszertan során először főnévi csoportokból álló párokat generálunk, és ezután vizsgáljuk meg, hogy az adott főnévi csoport anaforikus-e vagy sem. Abban az esetben, ha az annotáció nem a teljes főnévi csoportot fedte le, hanem csak egy részét,

tehát a konstituens elemzésben az annotált kifejezés nem NP címkéjű, végeredményben a pár negatívként jelent meg a tesztfájlban. Ez egy technikai jellegű hiba, ami a későbbiek során javítható.

3 A kizárólag harmadik személyű személyes névmási visszautalásokat tartalmazó tesztfájlokból kizártam a többi személyes névmást, a morfológiai elemzés alapján azonban egyes kifejezések: *enyém*, *miénk* a fájlokban maradtak a birtokos jel miatt, hozzájuk antecedensként annotálva viszont a mindenkori beszélő volt, így ezek az esetek is negatív példaként kerültek a tesztfájlokba.

4 Előfordultak a tesztfájlokban igekötők is, amelyekhez nem volt antecedens jelölve a szövegben, viszont a morfológiai elemzésben személyes névmásként voltak elemezve: *ekkor döbbentem rá, rá se hederítettem*.

5 Természetesen az is előfordult, hogy egy-egy visszautalás figyelmetlenségből nem volt annotálva a fájlokban, ezek is negatív példaként jelentek meg a tesztfájlokban.

Összességében elmondható, hogy az általam épített modellek a legtöbb esetben azonosítanak legalább egy antecedenst a névmásokhoz, tehát a nem visszautaló névmások felismerésében nem teljesítenek jól. A pontosság értékének javítása három módon lehetséges.

Egyrészt a korpuszban található fals negatív példák csökkentésével, azaz a technikai jellegű hibák javításával. Ez valószínűleg a fedésen is javítana, hiszen a be nem jelölt visszautalások jelölésével növekedne a pozitív példák aránya a szövegben. Az általam alkalmazott módszertan újragondolása is növelhetné ezt az arányt. Ha nem kizárólag teljes főnévi csoportokat keresnék, hanem főneveket is, akkor azok az esetek is pozitívak lennének, amelyekben csak a főnévi csoport egy része volt beannotálva, azonban ezzel a módszerrel automatikusan nőne a negatív példák száma is, hiszen a módszertan alapján minden esetben hozzá kellene rendelni a névmáshoz párként az összes névmást megelőző főnevet is.

A másik módszer, hogy a tanítófájlokba olyan névmások is bekerüljenek, amelyeknek nincs antecedense egyáltalán a szövegben. A következő lépés, hogy meghatározzuk, hány megelőző főnévi csoportot rendelünk hozzá a nem visszautaló névmáshoz, mint lehetséges pár. Ez azonban azt is eredményezi, hogy a tanítófájlokban még több lesz a negatív példa, ami az algoritmus esetében azt jelenti, hogy még nagyobb a valószínűsége, hogy a pozitív párokat nem ismeri fel az osztályozó, tehát a pontosságon ugyan valószínűleg javít, a fedésen viszont ront.

A harmadik módszer, hogy egy előelemző lépéssel kiszűrjük vagy az összes nem főnévi csoportra visszautaló névmást, tehát azokat is, amelyek tagmondatra utalnak vissza, vagy legalább azokat a névmásként elemzett szavakat, amelyek nem utalnak vissza egyáltalán, például

az igekötőket. Ezzel a lépéssel a pontosságon egyedül azok a névmási visszautalások rontanának, amelyeknek van kézzel annotált antecedense a szövegben, de másik főnévi csoportot azonosít hozzá az osztályozó. Ezzel az osztályozó feladatát redukálnánk, hiszen jelenleg fel kell ismernie, ha egy névmás nem visszautaló, valamint a visszautaló névmáshoz azonosítania kell a megfelelő antecedenst, ha a nem visszautaló névmásokat kiszűrnénk, az osztályozónak egyedül az lenne a feladata, hogy a visszautaló névmáshoz azonosítsa a megfelelő antecedenst.

8. Konklúzió

Az elvégzett gépi tanulási kísérletek egy nullhipotézisen és három további hipotézisen alapultak: A nullhipotézisem az volt, hogy lehetséges szemantikai információk nélkül is gépi tanulás segítségével anaforikus párokat azonosítani a szövegekben. Ez a hipotézisem helytállónak bizonyult, hiszen a kísérletek során minden esetben azonosított az osztályozó helyesen anaforikus párokat. A gépi tanulási kísérletek során a tanító és tesztfájlok létrehozásához a korpuszokban megtalálható szintaktikai elemzést, a tanulás alapját adó jellemzőkhöz pedig szintén kizárólag a korpuszokban megtalálható morfológiai és szintaktikai nyelvi jellemzőket, valamint általam megfogalmazott, kognitív alapú jellemzőket vettem figyelembe. Tekintettel arra, hogy a névmás referenséről, önmagában a névmás alapján nem tudhatunk semmit, a szemantikai elemzés sem garancia a jobb eredményekre. Ugyan a névmás [±élő] jegye adhat támpontot a referensre nézve, ezek sem kizárólagos szabályok, előfordulhat az élőlényre történő visszautalás [-élő] és [+élő] jeggyel is, erről bővebben (Kocsány 2016) A teljes automatikus névmási anaforafeloldás következtetést és a szövegkörnyezet, a kontextus ismeretét igényli. Ennek fényében az általam kidolgozott módszer a lehetséges antecedensjelöltek előszűrésére tűnik alkalmasnak.

Az első hipotézisem a tanítófájlokban megtalálható pozitív és negatív példák egymáshoz viszonyított arányára vonatkozik. Az algoritmus jellemzőiből következik, hogy a tanítófájlokban megtalálható példák mennyisége nagyban befolyásolja az osztályozó sikerességét. Mivel a Random Forest algoritmus véletlenszerűen választ mintát a tanítóadathalmazból, így a tanítóadatbázisban egyébként is többségben levő osztály kiválasztására nagyobb a valószínűség, így annak felismerésére is. Ez azonban együtt jár a kisebbségi osztály, azaz az anaforikus párok tanítóadathalmazba való kiválasztásának kisebb valószínűségével, amely maga után vonja a párok felismerésének kisebb valószínűségét is. Ezért az általam megfogalmazott jellemzők mellett a tanítóadathalmazban megjelenő pozitív és negatív példák arányaival is kísérleteket végeztem. Négy módszer segítségével építettem fel a tanítóadathalmazokat a gépi tanulási kísérlethez: Az elsőben a párok generálásához figyelembe vettem minden a szövegben megtalálható névmást és hozzárendeltem párként a névmást megelőző összes főnévi csoportot, tehát sem a pozitív tanítópéldák, sem a negatív tanítópéldák számát nem csökkentettem (Exp1). A második esetben kizárólag a visszautaló névmásokból generáltam tanítópéldákat, ezekhez pedig párként hozzárendeltem az őket megelőző főnévi csoportokat az első kézzel is annotált

főnévi csoporttal bezárólag (Exp2). Ezzel a módszerrel mind a negatív, mind a pozitív tanítópéldák számát csökkentettem. A harmadik esetben a kisebbségi csoport, azaz a pozitív példák számát növeltem úgy, hogy a második esetet kiegészítve pozitív párként a névmástól számított távolabbi kézzel is annotált főnévi csoportokat is a tanítóadatbázishoz rendeltem (Exp3). A negyedik esetben a pozitív és negatív példák arányait kiegyenlítettem egymással, úgy, hogy a negatív példák közül annyit adtam a tanítóadatbázishoz véletlenszerűen, amennyi pozitív példát tartalmazott (Exp4). Az alap elképzelés az volt, hogy az osztályozó akkor fog a legjobban teljesíteni, ha a példák mennyisége szempontjából hasonló összetételű fájlon fog tanulni, mint amin a modell végül tesztelésre kerül. A kísérlet eredményeként elmondható, hogy két esetben ez az elképzelés teljesült. Az összes személyes névmás és a mutatónévmás esetében ez a módszer mutatkozott a legeredményesebbnek. Itt azonban ki kell térnem arra, hogy a mutató névmás esetében a névmásoknak mindössze 18,5%-a volt visszautaló, tehát a negatív párok magasabb aránya a tanítófájlban ebből kifolyólag a tesztelés során is javított az eredményen. Az egyes szám harmadik személyű és vonatkozónévmási visszautalás esetében már nem ez mutatkozik a legsikeresebb módszernek. Ez részben az algoritmus működéséből is adódik, hiszen a Random Forest algoritmus a tanítópéldákból is véletlenszerűen választ, így annak az osztálynak a tagjai, amelyből több van a tanítófájlban, valószínűbben vesznek részt a modellépítésben és ezáltal hatékonyabban ismeri fel az osztályozó a tesztelés során őket. Ennek következtében a tanítófájlok létrehozása során törekedni kell arra, hogy a pozitív példák minél nagyobb arányban legyenek jelen, közel azonosban, mint a negatív példák, hogy egyenlő eséllyel kerüljenek be a modellépítés alapjául szolgáló példák közé, a vonatkozó névmási visszautalás esetében például a legközelebbi kézzel is annotált antecedens mindig igen közel található, tehát a negatív példák szűrése nem feltétlenül javít összességében az osztályozó eredményességén. Ezt a gondolatot támasztja alá az is, hogy a kizárólag az egyes szám harmadik személyű személyes névmásokat vizsgáló teszt esetében az EXP3, a vonatkozó névmások esetében pedig az EXP4 mutatkozott jobb stratégiának. Más osztályozó esetében természetesen lehetséges, hogy eltérő eredményeket kaphatunk, így a végső konklúzió levonása további kutatást igényel.

A második hipotézisem a névmáshoz antecedensként azonosított főnévi csoportok szűrésére vonatkozott. Mivel az általam alkalmazott Mention-pair technika több főnévi csoportot is megenged, mint a névmás antecedense, a névmási anaforafeloldás során azonban egy antecedens azonosítása a cél, két szűrési módszert hasonlítottam össze. A hipotézisem az volt, hogy a Best-first módszer alkalmasabb lesz a helyes antecedens azonosítására, hiszen a névmáshoz tartozó legközelebbi antecedens is gyakran több tagmondatnyi távolságra található a névmástól. A

kísérletek alapján ez a hipotézis részben volt helytálló. A két módszer közötti különbségeket nagyban befolyásolja a tanítófájlban szereplő példák aránya. Abban az esetben, ha a pozitív példák nagyobb arányban vannak jelen a tanítófájlokban, tehát távolabbi pozitív esetek is bele kerülnek, akkor a tesztelés során is előfordul, hogy távolabbi antecedenst azonosít az osztályozó és a döntés során a Best-first alkalmasabbnak bizonyul mint pusztán a közelség (Closest-first) alapján való döntés. Ha a legjobb eredményeket elért kísérleteket vesszük figyelembe, akkor az összes személyes névmás és a mutatónévmás esetében, ahol a tanítófájlban a példák aránya megegyezett a tesztfájlban található példák arányával nem mutatkozott különbség a két módszer között. Az egyes szám harmadik személyű személyes névmás esetében a Closest-first, míg a vonatkozó névmás esetében a Best-first módszer ért el jobb eredményt.

A harmadik és egyben utolsó hipotézisem az volt, hogy a morfológiai és szintaktikai jellemzőket kognitív alapon megfogalmazott jellemzőkkel kiegészítve jobb eredményeket fog elérni az osztályozó, mint önmagában a korpuszokban megtalálható nyelvi jellemzők segítségével. Ez a hipotézisem összetettebb volt a többinél, hiszen öt dolgot vizsgáltam ezen a témakörön belül. Egyrészt meghatároztam kognitív nyelvészeti alapon egy távolságszámítási módszert, amelynek során figyelembe vettem a tagmondatok egymással való viszonyát is, majd ezt hozzáadtam a gépi tanulási kísérletekhez és összehasonlítottam azzal a távolságszámítási módszerrel, amelynek során minden a névmás és az antecedense közötti tagmondati határátlépés azonos módon növeli a távolság értékét. Az általam meghatározott távolságszámítás növelte az osztályozó eredményességét a vonatkozó névmási visszautaláson kívül minden esetben, így a hipotézisemnek ez a része helytálló volt. Ezen kívül négy jellemzőcsoportot határoztam meg, az első az antecedensjelölt határozottságát (Def), a második a hosszát (Length), a harmadik a pozícióját (Pos) vizsgálta a negyedik pedig azt vizsgálta, hogy az antecedensjelölt névmás-e (Pron). Ugyan az összes személyes névmás és a vonatkozó névmás esetében a Pron, a mutató névmás esetében pedig a Length jellemző hozzáadásával épített modell érte el a legjobb eredményeket, ezek a jellemzőcsoportok összességében nem, vagy nagyon keveset javítottak az osztályozó eredményességén az egy antecedensjelöltre leszűrt változatokban. Így a hipotézisemnek ez a része nem mondható helytállónak. Ennek oka lehet, hogy a személyes névmás esetében az osztályozó nagyobb mértékben tud a morfológiai tulajdonságok egyeztetésére támaszkodni, a vonatkozó névmás esetében pedig, ahol az antecedens mindig közel található a távolság lesz az egyik legfontosabb jellemző.

Azokban a tesztekben, ahol minden pozitív pár azonosítása cél volt, már nagyobb eltérések láthatók, az egyes névmások alapján az általam megfogalmazott jellemzők között: a Pron

jellemző a személyes névmás, a Length jellemző a mutatónévmás a Def jellemző pedig a vonatkozó névmás esetében érte el a legjobb eredményt. Az általam definiált kognitív alapú jellemzők ezek alapján nagyobb hatást gyakorolhatnak a koreferenciafeloldás során.

Az általam végzett kísérletek alapján összességében elmondható, hogy a névmási anaforafeloldás során mindenképpen érdemes figyelembe venni a két kifejezés közötti tagmondatok egymáshoz való viszonyát, hiszen a Cp2 távolságszámítási módszer majd minden esetben javított az eredményeken a szimpla tagmondati határátlépések egyesével való számolásához viszonyítva. Érdemes gépi tanulási kísérleteket végezni a magyar nyelven elérhető koreferencia és anafora szempontjából kézi annotációt tartalmazó magyar nyelvű korpuszokon, hiszen ilyen mennyiségű visszautalás és szemantikai információ nélkül is azonosít az osztályozó anaforikus párokat. A tanítófájlokban megtalálható párok aránya, a jellemzők valamint az egy jelölt azonosítására vonatkozó módszerek tovább tesztelhetőek más algoritmusokkal, így ezek az eredmények összehasonlíthatóvá válnak a már más nyelvekre elvégzett hasonló kísérletekkel. Szintén fontos tapasztalat az adott feladat szempontjából a szövegtípus kérdése, hiszen a SzegedKoref korpuszon végzett kísérletek eredményei nagyban eltérnek a KorKorpuszon végzett kísérletek eredményeitől. Ezek a tapasztalatok megerősítik, hogy a névmási anaforafeloldás erőteljesen szövegfüggő feladat. A gépi tanulási kísérletek során magának az annotációnak a jellege is fontos, hiszen ez befolyásolja a pozitív és negatív példák mennyiségét és minőségét. A szöveg és annotáció típusán felül maga a névmástípus is fontos szempont, hiszen az általam végzett kísérletek eredményein jól látszik, hogy az egyes névmástípusok eltérően viselkednek a vizsgált szempontok alapján egymástól. Kísérleteim további eredménye az a megállapítás is, hogy az általam vizsgált jellemzők hatása eltérhet abban az esetben, ha nem kizárólag egy antecedens azonosítása a feladat célja.

Felhasznált irodalom

Aaronson, Doris – Steven Ferres 1983. Lexical categories and reading tasks. *Journal of Experimental Psychology: Human Perception and Performance* **9**/**5**:675–699. doi:10.1037/0096-1523.9.5.675.

Aires, Ana Margarida – Jorge Cesar B. Coelho – Sandra Collovini – Paulo Quaresma – Renata Vieira 2004. Avaliação de Centering em Resolução Pronominal da Língua Portuguesa. *Taller de Herramientas y Recursos Lingüísticos para el Español y el Portugués*. Tornantzintla, Mexico.

Aone, C. – S. Benett 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-1995)*. Cambridge. 122–129.

Ariel, Mira 1990. *Accessing Noun-Phrase Antecedents*. Beijing: World Publishing Corporation.

Ariel, Mira 2001. Accessibility theory. An overview. In Ted Sanders – Joost Schilperoord – Wilbert Spooren (szerk.) *Text Representation: Linguistic and Psycholinguistic Aspects*. Amsterdam: John Benjamins Publishing Company. 29–87.

Ariel, Mira 2014. *Accessing noun-phrase antecedents.* (Routledge Library Editions: Linguistics) London: Routledge.

Arnold, Jennifer E. 2001. The Effect of Thematic Roles on Pronoun Use and Frequency of Reference Continuation. *Discourse Processes* **31**/**2**:137–162. doi:10.1207/S15326950DP3102_02.

Arnold, Jennifer E. 2010. How Speakers Refer: The Role of Accessibility: How Speakers Refer. *Language and Linguistics Compass* **4**/**4**:187–203. doi:10.1111/j.1749-818X.2010.00193.x.

Arnold, Jennifer E. – Janet G. Eisenband – Sarah Brown-Schmidt – John C. Trueswell 2000. The rapid use of gender information: evidence of the time course of pronoun resolution from eyetracking. *Cognition* **76**:B13–B26.

Aroonmanakun, Wirote 2000. Zero Pronoun Resolution in Thai : A Centering Approach. In Denis Burnham – Sudaporn Luksaneeyanawin – Chris Davis – Mathieu Lafourcade (szerk.) *Interdisciplinary Approaches to Language Processing*. Bangkok: Chulalongkorn University Printing House. 127–147.

Bagga, Amit – Breck Baldwin 1998. Algorithms for scoring coreference chains. *The first international conference on language resources and evaluation workshop on linguistics coreference*. 563--566.

Bagha, Karim Nazari 2009. Generative Grammar (GG). *Management and Labour Studies* **34**/**2**:291–304. doi:10.1177/0258042X0903400208.

Barlow, Michael – Suzanne Kemmer (szerk.) 2000. *Usage-based models of language*. Stanford, Calif: CSLI Publications, Center for the Study of Language and Information.

Bengtson, Eric – Dan Roth 2008. Understanding the value of feaures for coreference resolution. *Proceedings of the International Conference on Empirical Methods Conference in Natural Language Processing.*. Waikiki: Association for Computational Linguistics. 294–303.

Berger, Adam L. – Stephen A. Della Pietra – Vincent J. Della Pietra 1996. Maximum entropy approach to natural language processing. *Computational Linguistics* **22**/**1**:39–71.

Björkelund, Anders – Richárd Farkas 2012. Data-driven Multilingual Coreference Resolution using Resolver Stacking. In Sameer Pradhan – Alessandro Moschitti – Nianwen Xue (szerk.) *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics. 49–55.

Björkelund, Anders – Jonas Kuhn 2014. Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics. 47–57.

Bock, Kathryn J. – David E. Irwin 1980. Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior* **19**/**4**:467–484. doi:10.1016/S0022-5371(80)90321-7.

Boronkai Dóra 2010. A deixis szerepe a nézőpont jelölésében. *Magyar Nyelv* **134**/**4**:436–452.

Breiman, Leo 2001. Random Forest. *Machine Learning* **45**/**1**:5–32.

Brennan, Susan E. – Marilyn W. Friedman – Carl J. Pollard 1987. A centering approach to pronouns. *Proceedings of the 25th annual meeting on Association for Computational Linguistics* -. Stanford, California: Association for Computational Linguistics. 155–162.

Cairns, Helen S. – Joan Kamerman 1975. Lexical information processing during sentence comprehension. *Journal of Verbal Learning and Verbal Behavior* **14**/**2**:170–179. doi:10.1016/S0022-5371(75)80063-6.

Caramazza, Alfonso – Ellen Grober – Catherine Garvey – Jack Yates 1977. Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behavior* **16**/**5**:601–609. doi:10.1016/S0022-5371(77)80022-4.

Caramazza, Alfonso – Shalini Gupta 1979. The roles of topicalization, parallel function and verb semantics in the interpretation of pronouns. *Linguistics* **17**/**5–6**. doi:10.1515/ling.1979.17.5-6.497.

Chambers, Craig G. – Ron Smyth 1998. Structural Parallelism and Discourse Coherence: A Test of Centering Theory. *Journal of Memory and Language* **39**/**4**:593–608. doi:10.1006/jmla.1998.2575.

Chastain, Charles 1975. Reference and Context. *Language, mind, and knowledge*/**7**:194–269.

Chen, Chen – Vincent Ng 2012. Combining the Best of Two Worlds: A Hybrid Approach to Multilingual Coreference Resolution. In Sameer Pradhan – Alessandro Moschitti – Nianwen Xue (szerk.) *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics. 56–63.

Chomsky, Noam 1981. *Lectures on Government and Binding*. Dordrecht: Foris Publications.

Clark, Herbert H. – Eve V. Clark 1977. *Psychology and language: an introduction to psycholinguistics*. New York: Harcourt Brace Jovanovich.

Clark, Kevin – Christopher D. Manning 2015. Entity-Centric Coreference Resolution with Model Stacking. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics. 1405–1415.

Clark, Kevin – Christopher D. Manning 2016. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. [arXiv: 1606.01323]. *arXiv:1606.01323 [cs]*.

Clark, Herbert H. – C. J. Sengul 1979. In search of referents for nouns and pronouns. *Memory & Cognition* **7**/**1**:35–41. doi:10.3758/BF03196932.

Clifton, Charles .Jr. – Fernanda Ferreira 1987. Discourse structure and anaphora: some experimental results. In M. Coltheart (szerk.) *Attention and Performance XII: The Psychology of Reading*. Hove: Lawrence Erlbaum. 635–654.

Coelho, Thiago Thomes – Ariadne Maria Brito Rizzoni Carvalho 2005. Lappin and Leass' Algorithm for Pronoun Resolution in Portuguese. In Carlos Bento – Amílcar Cardoso – Gaël Dias (szerk.) *EPIA'05: Proceedings of the 12th Portuguese conference on Progress in Artificial Intelligence*. (3808) Heidelberg: Springer. 680–692.

Cohen, William W. 1995. Fast effective rule induction. *Proceedings of the 12th International Conference on Machine Learning (ICML-1995)*. Tahoe. 115–123.

Converse, Susan P. 2005. Resolving Pronominal References in Chinese with the Hobbs Algorithm. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Crawley, Rosalind A. – Rosemary J. Stevenson 1990. Reference in single sentences and in texts. *Journal of Psycholinguistic Research* **19**/**3**:191–210. doi:10.1007/BF01077416.

Crawley, Rosalind A. – Rosemary J. Stevenson – David Kleinman 1990. The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research* **19**/**4**:245–264. doi:10.1007/BF01077259.

Csendes, Dóra – János Csirik – Tibor Gyimóthy 2004. The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC 2004) at The 20th International Conference on Computational Linguistics (COLING 2004)*. Geneva. 19–23.

Csendes, Dóra – János Csirik – Tibor Gyimóthy – András Kocsor 2005. The Szeged Treebank. In Václav Matoušek – Pavel Mautner – Tomáš Pavelka (szerk.) *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)*. Karlovy Vary: Springer. 123–131.

Daelemans, Walter – Antal van den Bosch 2005. *Memory-Based Language Processing*. Cambridge: Cambridge University Press,.

Dagan, Ido – Alon Itai 1990. Automatic processing of large corpora for the resolution of anaphora references. *Proceedings of the 13th conference on Computational linguistics* -. Helsinki, Finland: Association for Computational Linguistics. 330–332. doi:10.3115/991146.991209.

Daumé, Hal – Daniel Marcu 2005. Learning as search optimization: approximate large margin methods for structured prediction. *Proceedings of the 22nd international conference on Machine learning - ICML '05*. Bonn, Germany: ACM Press. 169–176.

Deemter, Kees van – Rodger Kibble 2000. On Coreferring: Coreference in MUC and Related Annotation Schemes. **26**/**4**:629–637.

Denis, Pascal – Baldridge 2009. Global joint models for coreference resolution and named entity classication. *Procesamiento del Lenguaje Natural*/**42**:87–96.

Desislava, Zhekova 2013. *Towards multilingual coreference resolution*. University of Bremen.

Dixon, Robert M. W 1972. *The Dyirbal language of North Queensland*.

Donnellan, Keith S. 1966. Reference and definite description. *The Philosophical Review* **75**/**3**:281–304.

Durrett, Greg – Dan Klein 2013. Easy Victories and Uphill Battles in Coreference Resolution. In David Yarowsky – Timothy Baldwin – Anna Korhonen – Karen Livescu – Steven Bethard (szerk.) *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics. 1971–1982.

Durrett, Greg – Dan Klein 2014. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *Transactions of the Association for Computational Linguistics* **2**:477–490.

É. Kiss, Katalin 1987. *Configurationality in Hungarian*. Budapest: Akadémiai Kiadó.

Ehlich, Konrad 1982. Anaphora and deixis: same, similar, or different? In Robert J. Jarvella – Wolfgang Klein (szerk.) *Speech, place and action: studies of deixis and related topics.*. Chichester ; New York: John Wiley & Sons. 315–338.

Ehrlich, Kate – Keith Rayner 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning & Verbal Behavior* **22**/**1**:75–87.

Eibe, Frank – Mark A. Hall – Ian H. Witten 2016. *The WEKA Workbench. Online Appendix for „Data Mining: Practical Machine Learning Tools and Techniques”*. Fourth Edition. Morgan Kaufmann.

Fernandes, Eraldo Rezende – Cícero Nogueira dos Santos – Ruy Luiz Milidiú 2012. Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. In Sameer Pradhan – Alessandro Moschitti – Nianwen Xue (szerk.) *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics. 41–48.

Freund, Yoav – Robert E. Shapire 1999. Large margin classification using the perceptron algorithm. *Machine Learning* **37**/**3**:277–296.

Garnham, Alan 2001. *Mental Models and the Interpretation of Anaphora*. Hove: Psychology Press.

Garnham, Alan – Jane V. Oakhill – Marie France Ehrilch – Manuel Carreiras 1995. Representation and process in the interpretation of pronouns. *Journal of Memory and Language* **34**/**1**:41–62.

Garvey, Catherine – Alfonso Caramazza – Jack Yates 1974. Factors influencing assignment of pronoun antecedents. *Cognition* **3**/**3**:227–243. doi:10.1016/0010-0277(74)90010-9.

Ge, Niyu – John Hale – Eugene Charniak 1998. A Statistical Approach to Anaphora Resolution. In Eugene Charniak (szerk.) *Proceedings of Sixth WorkShop on Very Large Corpora*. Montreal, Quebec, Canada. 161–170.

Gernsbacher, Morton Ann 1985. Surface information loss in comprehension. *Cognitive Psychology* **17**/**3**:324–363.

Gernsbacher, Morton Ann – David J Hargreaves 1988. Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language* **27**/**6**:699–717. doi:10.1016/0749-596X(88)90016-2.

Gernsbacher, Morton Ann – David J. Hargreaves – Mark Beeman 1989. Building and accessing clausal representations: The advantage of first mention versus the advantage of clause recency. *Journal of Memory and Language* **28**/**6**:735–755. doi:10.1016/0749-596X(89)90006-5.

Gibson, Edward 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium.*. 95–126.

Givón, Talmy 1983. Topic continuity in discourse: The functional domain of switch-reference. In John Haiman – Pamela Munro (szerk.) *Switch Reference and Universal Grammar*. (Typological Studies in Language 2) Amsterdam: John Benjamins Publishing Company. 51–82. doi:10.1075/tsl.2.06giv.

Givón, Talmy 1993. *English Grammar: A function-based introduction. Volume I*. Amsterdam: John Benjamins Publishing Company.

Givón, Talmy 2001. *Syntax: an introduction*. Amsterdam ; Philadelphia: John Benjamins.

Gordon, Peter C. – Randall Hendrick – Kerry Ledoux – Chin Lung Yang 1999. Processing of reference and the structure of language: an analysis of complex noun phrases. *Language and Cognitive Processes* **14**/**4**:353–379.

Grimes, Joseph E 2015. *The Thread of Discourse*.

Grober, Ellen H. – William Beardsley – Alfonso Caramazza 1978. Parallel function strategy in pronoun assignment. *Cognition* **6**/**2**:117–133. doi:10.1016/0010-0277(78)90018-5.

Grosz, Barbara J. – Aravind K. Joshi – Scott Weinstein 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics* **21**/**2**:203–225.

Gundel, Jeanette K. – Nancy Hedberg – Ron Zacharski 1993. Cognitive status and the form of referring expressions in discourse. *Language* **69**/**2**:274–307.

Halliday, Mark A. K 1994. *An introduction to functional grammar*. London/Beijing: Arnold: Foreign Language Teaching and Research Press.

Halliday, M. A. K. – Ruqaiya Hasan 1976. *Cohesion in English*. (English language series ; no. 9) London: Longman.

Halliday, M. A. K. – Christian M. I. M. Matthiessen 2004. *An introduction to functional grammar*. 3rd ed. London : New York: Arnold ; Distributed in the United States of America by Oxford University Press.

Harabagiu, Sanda – Răzvan C. Bunescu – Steven J. Maiorano 2001. Text and knowledge mining for coreference resolution. *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL-2001),*. Pittsburgh,. 55–62.

Hendrickx, Iris – Veronique Hoste – Walter Daelemans 2007. Evaluating hybrid versus data-driven coreference resolution. *Anaphora: Analysis, Algorithms and Application. Lecture Notes in Artificial Intelligence*. (4410) Berlin/New York: Springer. 137–150.

Hicks, Glyn 2009. *The derivation of anaphoric relations*. (Linguistik aktuell = Linguistics today v. 139) Amsterdam ; Philadelphia: John Benjamins Pub. Co.

Hinrichs, E. – S. Kübler – K. Naumann – H. Telljohann – J. Trushkina 2004. Recent developments in linguistic annotations of the TüBa-D/Z treebank. In S. Kübler – J. Nivre – E. Hinrichs – Holger Wunsch (szerk.) *Proceedings of the Third Workshop on Treebanks and Linguistic Theories.*. Tübingen, Germany.

Hirst, William – Gary A. Brill 1980. Contextual aspects of pronoun assignment. *Journal of Verbal Learning and Verbal Behavior* **19**/**2**:168–175. doi:10.1016/S0022-5371(80)90152-8.

Hobbs, Jerry R. 1976a. Pronoun Resolution. *Research Report. Department of Computer Sciences*/**76a**.

Hobbs, Jerry R. 1976b. A Computational Approach to Discourse Analysis. *Research Report 76b*.

Hobbs, Jerry R. 1979. Coherence and Coreference. *Cognitive Science* **3**/**1**:67–90.

Hoek, Karen van 1995. Conceptual reference points. A Cognitive Grammar account of pronominal anaphora constraints. *Language* **71**/**2**:310–340. doi:10.2307/416165.

Hoste, Veronique 2005. *Optimization Issues in Machine Learning of Coreference Resolution*. Antwerpen University.

Hou, Yufang – Katja Markert – Michael Strube 2013. Global inference for bridging anaphora resolution. *Proceedings of the 2013 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia. 907–917.

Huang, Yan 2000. *Anaphora: a cross-linguistic approach*. (Oxford studies in typology and linguistic theory) Oxford ; New York: Oxford University Press.

Karttunen, Lauri 1969. Discourse referents. *Proceedings of the 1969 conference on Computational linguistics*. Association for Computational Linguistics. 1–38. doi:10.3115/990403.990487.

Kehler, Andrew – Douglas Appelt – Lara Taylor – Aleksandr Simma 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. *Proceedings of 2004 North American Chapter of the Association for Computational Linguistics Annual Meeting*. Boston. 289–296.

Kennedy, Christopher – Branimir Boguraev 1996. Anaphora for everyone: pronominal anaphora resoluation without a parser. *Proceedings of the 16th conference on Computational linguistics*. Copenhagen, Denmark: Association for Computational Linguistics. 113. doi:10.3115/992628.992651.

Kiefer Ferenc 2007. *Jelentéselmélet*. Budapest: Corvina.

Kieras, David E. 1980. Initial mention as a signal to thematic content in technical passages. *Memory & Cognition* **8**/**4**:345–353. doi:10.3758/BF03198274.

Kocsány Piroska 1996. „Fog ő még gondolni ránk.” A személyes névmás egy különös használatáról. In Terts István (szerk.) *Nyelv, nyelvész, társadalom. Emlékkönyv Szépe György 65. születésnapjára barátaitól, kollégáitól, tanítványaitól*. Pécs: Janus Pannonius Tudományegyetem. 159–161.

Kocsány, Piroska 2011. A komplex anafora kettős arca. In Ágnes Bánki – Gábor Tillinger (szerk.) *Survivance du latin et grammaire textuelle. Mélanges offerts à Sándor Kiss à l'occasion de son 70e anniversaire. (Studia Romanica)*. Debrecen: Debreceni Egyetemi Kiadó. 305–314.

Kocsány Piroska 2016. A mondatközi anafora és az ő névmás szerepei. *Jelentés és Nyelvhasználat* **3**:117–150. doi:10.14232/JENY.2016.1.6.

Kocsány, Piroska 2018. A közelre mutató *ez* névmás anaforikus kifejtő szerkezetekben és a diszkurzív kontinuitás fokozatai. *Jelentés és Nyelvhasználat* **5**/**1**:117–158. doi:10.14232/jeny.2018.1.5.

Kovács, Viktória 2017. Koreferenciaviszonyok vizsgálata enyhe kognitív zavarban szenvedők beszédátirataiban. In Zsófia Ludányi (szerk.) *Doktoranduszok tanulmányai az*

alkalmazott nyelvészet köréből 2017 : XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia. Budapest: MTA Nyelvtudományi Intézet. 120–130.

Kovács, Viktória 2019. Az elérhetőségi elmélet névmási anaforafeloldásra gyakorolt hatása. In Zsófia Ludányi – Tekla Etelka Gráczi (szerk.) *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2019 : XIII. Alkalmazott Nyelvészeti Doktoranduszkonferencia*. Budapest: MTA Nyelvtudományi Intézet. 113–121.

Kovács, Viktória 2020. A tagmondati távolságszámítás módjainak hatása a névmási anaforafeloldásra. In Gábor Berend – Gábor Gosztolya – Veronika Vincze (szerk.) *XVI. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, Informatikai Intézet. 129–139.

Kozminsky, Ely 1977. Altering comprehension: The effect of biasing titles on text comprehension. *Memory & Cognition* **5**/**4**:482–490. doi:10.3758/BF03197390.

Laczkó, Krisztina 2004. A névmási rendszer funkcionális keretben. 1. *Magyar nyelvőr* **128**/**4**:469–479.

Laczkó, Krisztina 2005. A névmási rendszer funkcionális keretben. 2. **129**/**1**:78–88.

Laczkó Krisztina 2008. A mutató névmási deixisről. *Általános Nyelvészeti Tanulmányok XXII*. 309–347.

Laczkó Krisztina – Tátrai Szilárd 2012. Személyek és/vagy dolgok. A harmadik személyű és a mutató névmási deixis a magyarban. In Tolcsvai Nagy Gábor – Tátrai Szilárd (szerk.) *Konstrukció és jelentés. Tanulmányok a magyar nyelv funkcionális kognitív leírására*. Budapest: ELTE. 231–257.

Langacker, Ronald W. 1986. An Introduction to Cognitive Grammar. *Cognitive Science* **10**/**1**:1–40. doi:10.1207/s15516709cog1001_1.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Vol. 1. Theoretical Prerequisites*. Stanford: Stanford University Press.

Lappin, Shalom – Herbert J. Leass 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics* **20**/**4**:535–561.

Lappin, Shalom – Michael McCord 1990a. A syntactic filter on pronominal anaphora in slot grammar. *Proceedings, 28th Annual Meeting of the Association for Computational Linguistics*. 135–142.

Lappin, Shalom – Michael McCord 1990b. Anaphora resolution in slot grammar. *Computational Linguistics* **16**:197-212.

Lasnik, H. 1976. Remarks of co-reference. *Linguistic Analysis* **2**/**1**:1–22.

Le, Zhang 2004. *Maximum Entropy Modeling Toolkit for Python and C++ (version 20041229)*. China: Northeastern University.

Lee, Kenton – Luheng He – Mike Lewis – Luke Zettlemoyer 2017. End-to-end Neural Coreference Resolution. In Martha Palmer – Rebecca Hwa – Sebastian Riedel (szerk.) *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics. 188–197.

Lejtovicz, Katalin, Eszter – Zsolt Kardkovács Tivadar 2006. Anaforafeloldás magyar nyelvű szövegekben. In Zoltán Alexin – Dóra Csendes (szerk.) *IV. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2006*. Szeged: Szegedi Tudományegyetem. 362–363.

Linguistic Data Consortium 2001. Message Understanding Conference (MUC) 7. Philadelphia, PA.: Linguistic Data Consortium.

Linguistic Data Consortium 2003. *Message Understanding Conference (MUC) 6*.

Luo, Xiaoqiang 2005. On coreference resolution performance metrics. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*. Vancouver, British Columbia, Canada: Association for Computational Linguistics. 25–32.

Lyons, John 1977. *Semantics*. Vol. 2 Cambridge: Cambridge University Press.

Manabu, Okumura – Tamura Kouji 1996. Zero pronoun resolution in Japanese discourse based on centering theory. *Proceedings of the 16th conference on Computational linguistics -*. Copenhagen, Denmark: Association for Computational Linguistics. 871. doi:10.3115/993268.993319.

Martschat, Sebastian – Jie Cai – Samuel Broscheit – Éva Mújdricza-Maydt – Michael Strube 2012. A Multigraph Model for Coreference Resolution. In Sameer Pradhan – Alessandro Moschitti – Nianwen Xue (szerk.) *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics. 100–106.

McCarthy, J. 1996. *A trainable approach to coreference resolution for information extraction.* Amherst: University of Massachusetts.

Miháltz, Márton 2012. Tudásalapú koreferencia- és birtokosviszony-feloldás magyar szövegekben. Szerk. István Kenesei – Gábor Prószéky – Tamás Váradi. *Általános Nyelvészeti Tanulmányok* **24**:151–166.

Miháltz, Márton – Károly Varasdi – Péter Vajda – Mátyás Naszódi 2007. NP-koreferenciák feloldása magyar szövegekben a Magyar WordNet ontológia segítségével. In Attila

Tanács – Dóra Csendes (szerk.) *V. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2007*. Szeged: Szegedi Tudományegyetem. 138–146.

Mitkov, Ruslan 2001. Outstanding issues in anaphora resolution. *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. 110–125.

Mitkov, Ruslan (szerk.) 2009. *The Oxford handbook of computational linguistics*. Reprinted. Oxford: Oxford Univ. Press.

Munkácsy, Gergely – Richárd Farkas 2016. Statisztikai koreferenciafeloldó rendszer magyar nyelvre – első eredmények. In Attila Tanács – Viktor Varga – Veronika Vincze (szerk.) *XII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, Informatikai Intézet. 295–297.

Newmeyer, Frederick J. 2000. *Language form and language function*. 2. kiad. Cambridge, Mass.: MIT Press.

Ng, Vincent – Claire Cardie 2002a. Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*. Philadalphia. 104–111.

Ng, Vincent – Claire Cardie 2002b. Combining sample selection and error-driven pruning for machine learning of coreference rules. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*. Philadelphia. 55–62.

Ng, Vincent – Claire Cardie 2002c. Identifying anaphoric and non- anaphoric noun phrases to improve coreference resolution. *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*. Taipei.

NIST 2003. Proceedings of ACE'03 Booklet, Alexandria, VA, September.

Palomar, Manuel – A. Ferrandez – Lidia Morenoy – Patricio Martínez-Barco – Jesús Peral – Maximiliano Saiz-Noeda – Rafael Muñoz 2001. An Algorithm for Anaphora Resolution in Spanish Texts. *Computational Linguistics* **27**/**24**:545–567.

Perfetti, Charles A. – Susan R. Goldman 1974. Thematization and sentence retrieval. *Journal of Verbal Learning and Verbal Behavior* **13**/**1**:70–79. doi:10.1016/S0022-5371(74)80032-0.

Pléh, Csaba 1994. Mondatközi viszonyok feldolgozása: Az anafora megértése a magyarban. *Magyar pszichológiai szemle* **50**/**5–6**:287–320.

Pléh Csaba 1998. *A mondatmegértés a magyar nyelvben. Pszicholingvisztikai kísérletek és modellek*. Budapest: Osiris Kiadó.

Pléh, Csaba – Katalin Radics 1976. „Hiányos mondat", pronominalizáció és a szöveg. *Általános Nyelvészeti Tanulmányok* **11**/**1**:261–277.

Pradhan, Sameer – Alessandro Moschitti – Nianwen Xue – Olga Uryupina – Yuchen Zang 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In Sameer Pradhan – Alessandro Moschitti – Nianwen Xue (szerk.) *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics. 1–40.

Quinlan, Ross J. 1993. *C4.5: Programs for Machine Learning.* San Mateo: Morgan Kaufmann.

Recasens, Marta – Eduard Hovy 2009. A deeper look into features for coreference resolution. *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium*. Goa. 29–42.

Reinhart, Tanya 1976. *The syntactic domain of anaphora*. Cambridge: MIT.

Sanford, Anthony J. – Simon C. Garrod 1981. *Understanding written language: explorations of comprehension beyond the sentence*. Chichester ; New York: Wiley.

Sanford, A. J. – K. Moar – S. C. Garrod 1988. Proper Names as Controllers of Discourse Focus. *Language and Speech* **31**/**1**:43–56. doi:10.1177/002383098803100102.

Santos, Denis Neves de Arruda – Ariadne Maria Brito Rizzoni Carvalho 2007. Hobbs' Algorithm for Pronoun Resolution in Portuguese. (Lecture Notes in Computer Science 4827). 966–974.

Schwarz, Monika 2000. *Indirekte Anaphern in Texten, Studien zur domänengebundenen Referenz und Kohärenz im Deutschen*. Tübingen: Max Niemeyer Verlag.

Sidner, Candace Lee 1979. *Towards a computational theory of definite anaphora comprehension in English discourse.* MIT.

Soon, Wee Meng – Hwee Tou Ng – Daniel Chung Yong Lim 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* **27**:521–544.

Strube, Michael – Stefan Rapp – Christoph Müller 2002. The influence of minimum edit distance on reference resolution. *Proceedings of the 2002 Conference on EmpiricalMethods in Natural Language Processing (EMNLP-2002),*. Philadelphia. 312–319.

Sukthanker, Rhea – Soujanya Poria – Erik Cambria – Ramkumar Thirunavukarasud 2020. Anaphora and coreference resolution: A review. *Information Fusion*/**59**:139–162.

Szécsényi, Tibor – Viktória Kovács 2020. A topikalizálhatóságot befolyásoló tényezők statisztikai vizsgálata. *Általános nyelvészeti tanulmányok 32.: Újabb eredmények a*

grammatikaelmélet, nyelvtörténet és uralisztika köréből. Budapest: Akadémiai Kiadó. 237–247.

Tátrai Szilárd 2010. Áttekintés a deixisről (funkcionális kognitív kiindulópontból). *Magyar Nyelvőr* **134**/**2**:211–233.

Telljohann, Heike – Erhard W. Hinrichs – Sandra Kübler 2004. The TüBa-D/Z Treebank – Annotating German with a Context-Free Backbone. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lissabon, Portugal. 2229–2232.

Tolcsvai Nagy, Gábor 2000. Kérdések a koreferenciáról. *Koreferáló elemek - koreferenciarelációk*. (Officina textologica 4) Debrecen: Kossuth Egyetemi Kiadó. 11–34.

Tolcsvai Nagy Gábor 2000. Vázlat az *ő – az* anaforikus megoszlásról. *Magyar Nyelv* **96**/**3**:282–296.

Tolcsvai Nagy Gábor 2005. Kognitív jelentéstani vázlat az igekötős igéről. *Magyar Nyelv* **2005**/**1**:27–43.

Uryupina, Olga 2004. Linguistically motivated sample selection for coreference resolution. *Proceedings of DAARC-2004*. Azores.

Uryupina, Olga 2006. Coreference resolution with and without linguistic knowledge. *Proceedings of the äth International Conference on Language Resources and Evaluation*. Genoa: European Language Resource Association. 893–898.

Vadász, Noémi 2020. KorKorpusz: kézzel annotált, többrétegű pilotkorpusz építése. In Gábor Berend – Gábor Gosztolya – Veronika Vincze (szerk.) *XVI. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged. 141–154.

Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Versley, Yannick – Alessandro Moschitti – Massimo Poesio – Xiaofeng Yang 2008. Coreference systems based on kernel methods. *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester. 961–968.

Vilain, Marc – John Burger – John Aberdeen – Dennis Connolly – Lynette Hirschman 1995. A model-theoretic coreference scoring scheme. *Proceedings of the 6th conference on Message understanding - MUC6 '95*. Columbia, Maryland: Association for Computational Linguistics. 45.

Vincze, Veronika – Klára Hegedűs – Alex Sliz-Nagy – Richárd Farkas 2018. SzegedKoref: A Hungarian Coreference Corpus. *11th edition of the Language Resources and Evaluation Conference.*. Miyazaki, Japan: European Language Resources Association.

Vincze, Veronika – Dóra Szauter – Attila Almási – György Móra – Zoltán Alexin – János Csirik 2010. Hungarian Dependency Treebank. *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. Valletta. 1855–1862.

Walker, Marilyn A. 1989. Evaluating discourse processing algorithms. *Proceedings of the 27th annual meeting on Association for Computational Linguistics* -. Vancouver, British Columbia, Canada: Association for Computational Linguistics. 251–261. doi:10.3115/981623.981654.

Wasow, Thomas 1979. *Anaphora in generative grammar*. (Sigla 2) Ghent: E. Story-Scientia.

Wiseman, Sam – Alexander M. Rush – Stuart M. Shieber 2016. Learning Global Features for Coreference Resolution. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics. 994–1004. doi:10.18653/v1/N16-1114.

Wunsch, Holger 2006. Anaphora Resolution – What Helps in German.

Yadolah, Dodge 2008. Gini Index. *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York. 231–233. doi:10.1007/978-0-387-32833-1_169.

Yang, Xiaofeng – Guodong Zhou – Jian Su – Chew Lim Tan 2003. Coreference resolution using competitive learning approach. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo. 176–183.