SZEMANTIKAI ERŐFORRÁSOK ÉS ALKALMAZÁSAIK A MAGYAR TERMÉSZETESNYELV-FELDOLGOZÁSBAN

Ph.D. disszertáció tézisei

Mihálty Márton

Témavezető:
Dr. Prószéky Gábor

Pázmány Péter Katolikus Egyetem
Információs Technológiai Kar
Multidiszciplináris Műszaki Tudományok Doktori Iskola

Budapest, 2010

1. Bevezetés

A természetesnyelv-feldolgozás (vagy nyelvtechnológia) az információs technológiának egy olyan ága, melynek érdeklődésében az emberi (természetes) nyelven megfogalmazott szövegek feldolgozására képes algoritmusok és szoftveres alkalmazások fejlesztése áll.

Ahogy a természetes nyelvekben megkülönböztethetünk különböző szerkezeti szinteket, úgy a nyelvtechnológiában is leírhatunk különböző feldolgozási lépéseket. A szövegfeldolgozásban ezek például a következők lehetnek[1]: a szegmentálás (nyers (feldolgozatlan) szöveg mondatokra és szavakra, írásjelekre bontása), morfológiai elemzés/szófaji egyértelműsítés (a szavakat alkotó morfémák és azok tulajdonságainak azonosítása), szintaktikai elemzés (a mondatokat alkotó szerkezeti elemek feltárása), és szemantikai feldolgozás (a szöveg „jelentésének" kezelése: többjelentésű szavak helyes értelmének azonosítása, dokumentumon belüli hivatkozások feloldása stb.)

Disszertációmban a nyelvtechnológia ez utóbbi, szemantikai oldalára koncentráltam, főképpen magyar nyelven írt (vagy magyarra fordított) szövegek feldolgozásának összefüggésében.

A szemantikai feldolgozás gyakran szemantikai tudástárak, vagy másképpen ontológiák használatára alapul, melyek speciális, világról

[1] A természetesnyelv-feldolgozás az itt feltüntetetthez képest másképpen is értelmezhetjük, illetve sok más feladat létezik még a nyelvtechnológiában, melyekről itt nem teszünk említést.

való tudásunk bizonyos aspektusait modellező adatbázisok. Munkám első részében egy ilyen ontológia-formalizmusra koncentráltam, melynek neve *WordNet*.

A WordNet eredetileg olyan lexikális szemantikai adatbázis, melyet angol nyelvre készítettek a Princeton Egyetemen [28], [29]. Célja a mentális lexikon felépítésével – természetes nyelvi fogalmi egységek (szavak és többszavas lexémák) jelentésének és azok egymás közötti kapcsolatainak modellezésével – kapcsolatos nyelvészeti és pszicholingvisztikai elméletek tesztelése és implementációja volt. A WordNet olyan hálózat, melynek alapvető építőelemei a fogalmak, melyeket szinonimák halmazaival (synsetekkel) definiáltak. Ezeket különböző szemantikai kapcsolatok kötik össze, melyek közül a legfontosabb ilyen, a hipernima-reláció (az öröklődési hálózatok „az-egy” (is-a) relációjával megfeleltethető) egy hierarchikus hálózatot hoz létre.

Létrehozása után a WordNet hasznos eszköznek bizonyult a természetesnyelv-feldolgozás alkalmazásaiban is [28], és megkezdődött a wordnetek fejlesztése az angol után további nyelvekre. A különböző nyelvek fogalmi hálózatait összekötő projektek indultak [30], [31].

Kutatásom első részében **olyan módszerek alkalmazásával és továbbfejlesztésével, valamint újabb ilyenek létrehozásával foglalkoztam, melyek célja egy magyar nyelvű WordNet ontológia létrehozásának támogatása**. Noha megbízható szemantikai adatbázisok végső soron csak emberi szakértők munkájával tökéletesíthetők, több javaslat is született már ennek

automatikus támogatására [32],[33],[34]. **Olyan módszerekkel kísérleteztem, melyek célja szemantikai és szerkezeti viszonyok kinyerése számítógéppel feldolgozható szótárak anyagából az ún. kiterjesztéses modell [31] támogatására**. Ez utóbbi lényege, hogy az angol nyelvű Princeton WordNet (PWN) fogalmi vázából származtatva készüljön egy magyar wordnet a magyar nyelv szemantikai tulajdonságaihoz igazítva.

Kutatásaim második érdeklődési területe a **jelentés-egyértelműsítésre** (word sense disambiguation, WSD) irányult, mely a természetes nyelvi jelentések gépi feldolgozásának egy következő vetülete. A WSD célja szemantikailag többértelmű szavak aktuális jelentésének meghatározása szövegkörnyezetük alapján. A lexikális szemantikai többértelműség maga sokat kutatott terület az elméleti nyelvészeten belül, mely a homonímiától a poliszémiáig tartó jelenségeket felölelő spektrum [35], ahol bizonyos esetekben a finom szemantikai különbségek az emberi megítélést is kihívás elé állítják egy-egy szójelentés meghatározásában. Munkám során gyakorlati megközelítést alkalmaztam, és egy adott szó jelentéseit A nyelven a szó B nyelven adható lehetséges fordításainak halmazával definiáltam. Ez az elv természetes módon vezetett a gépi fordításhoz kapcsolódó kísérletezéshez. **Felügyelt gépi tanulásos módszereket alkalmazva, egy szabályalapú angol-magyar gépi fordító rendszer keretei között végeztem kísérleteket lexikai egységek jelentés-egyértelműsítésével.** Mivel a felügyelt gépi tanulás számára elengedhetetlen megfelelő mennyiségű, ám kézi munkával igen költségesen előállítható tanítópéldák biztosítása, **olyan,**

szinkronizált párhuzamos korpuszokban található információkra támaszkodó módszerek fejlesztése is érdeklődési körömbe esett, melyek segítségével automatizálható ilyen tanítópéldák előállítása.

Vizsgálódásaim harmadik területe, **főnévi csoportok (NP-k) koreferencia-viszonyainak azonosítása magyar szövegekben** többek között szintén a (magyar) WordNet alkalmazására támaszkodott. NP-k koreferencia-feloldása (coreference resolution, CR) egy dokumentumban ugyanarra a való világbeli entitásra hivatkozó kifejezések azonosítását jelenti. Ez a feladat is természetes nyelvi jelenségek egész sorát érinti, melyek közül a következőket kíséreltem meg kezelni: koreferencia kifejezése ismétléssel, tulajdonnév-változatokkal, szinonimákkal és hipernímákkal/hiponímákkal, valamint személyes névmások és zérónévmások.

A birtokosviszony-feloldás a koreferencia-azonosításhoz hasonló feladat, melynek célja az elvált, akár más szavakkal vagy mondatrészekkel közbeékelődésével egymástól távolra került birtokos és birtok párok azonosítása.

Mindkét feladatban **érdeklődésem olyan szabályalapú rendszer kifejlesztésére irányult, amely képes a különböző forrásokból származó információk és az ezekre épülő módszerek integrációjával magas pontosság és lefedettség elérésére, így alkalmassá válhat gyakorlati alkalmazásokra természetesnyelv-feldolgozási rendszerekben.**

Munkám során eredményeim gyakorlati, a gépi fordítás, információ-kivonatolás, tartalomelemzés témakörébe tartozó projektek keretei között történő alkalmazásával is foglalkoztam, ezekről részletesen a 4. Fejezetben számolok be.

2. Vizsgálati módszerek

Munkám során mind szabályalapú módszerekkel – az adott területhez kapcsolódó tudásra támaszkodó heurisztikák fejlesztésével – mind felügyelt gépi tanuló algoritmusokkal végeztem kísérleteket. Módszereim kifejlesztése és kiértékelése során több különböző nyelvtechnológiai eszközt felhasználtam a különböző természetes nyelvi erőforrások (számítógéppel olvasható szótárak és korpuszok) feldolgozására, ezekről részletesen az alábbiakban, a különböző kutatási irányokról szólva adok számot.

Munkám **első részében** az ún. **kiterjesztéses modellt** követtem, melyet a EuroWordNet projektek több résztvevője alkalmazott [31]. Ez azt jelenti, hogy a Princeton WordNet kiválasztott synseteit implementáljuk magyarul, örököljük az angol szemantikai relációkat, majd az így előállt fogalmi hierarchiát adaptáljuk a magyar nyelv sajátosságaihoz. A módszer választását egyfelől az alternatív, összevonásos modell alkalmazásához szükséges magyar nyelvű strukturált szemantikai erőforrások hiánya, másfelől az automatikus módszerekkel felgyorsítható synset-fordítás lehetősége motiválta. Az eljárás alkalmazásához ezen felül szükség volt arra a feltételezésre, hogy az angol és a magyar fogalmi rendszer között hasonlóság megfelelő mértékű lesz, legalább a főnevek esetében, hiszen azok (a

kulturális különbségeket természetesen leszámítva) egy többé-kevésbé közös valóság fizikai és absztrakt entitásainak felelnek meg.

A célom olyan módszerek létrehozása volt, melyek megkísérlik egy angol-magyar kétnyelvű számítógépes szótár magyar oldalának címszavait angol WordNet synseteknek megfeleltetni. A feladat során kétszintű többértelműséggel kellett szembenézni. Egy tetszőleges *w* magyar szónak átlag *n* jelentése lehet a kétnyelvű szótárban, amely jelentések mindegyike átlag *m* különböző Wordnet synsetbe tartozhat, így az algoritmusnak átlagosan *n*m* darab synsetből kell kiválasztania a megfelelőt (a gyakorlatban ez *n*m*=3,69 értéket jelentett.) Ennek megvalósításához különböző heurisztikák készletét alkalmaztam, melyek az egyértelműsítéshez szükséges információkat a kétnyelvű és egy egynyelvű szótár anyagából kinyert adatokból merítették.

Munkám **második része** a több lehetséges magyar fordítással is rendelkező angol főnevek jelentés-egyértelműsítésével foglalkozott. Ennek megoldására felügyelt gépi tanulásos megközelítéssel tettem kísérletet, ahol is minden egyes többértelmű szóhoz külön, a tanítópéldák környezeti jellemzőin betanított osztályozó tartozik. Felügyelt gépi tanulással értek már el korábban sikereket jelentés-egyértelműsítésben [38], és az angol nyelvhez több különböző kézzel annotált tanítókorpusz is szabadon hozzáférhető. Ezek közül a Senseval verseny English lexical sample task feladat [41], valamint az Open Mind Word Expert projekt [40] példaállományát alkalmaztam, melyek összesen 45 különböző többértelmű angol főnévhez biztosítottak tanítóanyagot.

A tanítópéldákat az annotátorok Princeton WordNet jelentés-azonosítókkal látták el. Annak érdekében, hogy az angol-magyar gépi fordításban működő WSD rendszer számára alkalmas legyen a jelentéscímke-készlet, kézzel megfeleltettem minden angol fogalmat egy-egy magyar fordítással. A kezdeti 45 főnév közül 34 esetben a magyar fordítások száma alacsonyabb volt az angol WordNet jelentések számánál, vagyis a magyar fordítások a jelentések egy durvább minőségű felosztását adták. 7 további angol főnév esetében minden angol jelentés ugyanazzal a magyar fordítással volt visszaadható, így ezek a szavak kiestek a kísérletből, hiszen nem volt szükség jelentés-egyértelműsítésükre az angol-magyar fordításban. A maradék 4 szó esetében az angol jelentések és a magyar fordítások száma megegyezett. A továbbiakban így azzal a 38 szóval kísérleteztem, melyeknél a magyar fordítások száma kisebb vagy egyenlő volt az angol jelentések számával. Az angol szavak eredetileg átlagosan 3,97 különböző jelentésbe tartoztak, a magyar fordításokkal történt megfeleltetés után viszont egy angol szó már csak átlag 2,49 lehetséges fordítással rendelkezett, így az átlagos többértelműség csökkent.

A rendszer az egyszerű és jól ismert Naiv Bayes osztályozó algoritmust használja, amely azt a jelentést választja, amelynek a legmagasabb a feltételes valószínűsége a rendelkezésre álló környezeti jellemzők együttesére nézve. A különböző, Weka szoftverkörnyezetben [58] rendelkezésre álló felügyelt gépi tanulásos algoritmusok közül a legmagasabb pontosságot a Naiv Bayes adta, így ezt választottam. Az együttes feltételes

valószínűségeket a tanítókorpuszban megfigyelt relatív gyakoriságokkal lehet megbecsülni. Noha az algoritmus nevét adó alapfeltételezés (mely szerint a kontextus jellemzői független valószínűségi változók) nem teljesül természetes nyelvi adatokra, a módszerrel mégis sikereket értek elé a jelentés-egyértelműsítés területén [37], [38].

Az osztályozók tanításához, [37] és [39] alapján a többértelmű szavak tanítópéldáinak kontextusaiban definiált jellemzőket alkalmaztam, melyek két csoportba sorolhatók. Az első csoportba eső jellemzőket csak a többértelmű szót tartalmazó mondatból számítjuk, a szósorrend és a relatív távolság figyelembe vételével. Ezek a jellemzők a kontextus szintaktikai tulajdonságait, gyakori kollokációkat, jellemző módosítókat stb. reprezentálnak. Ide tartoznak: a vizsgált szó felszíni alakja, funkciószavak a szó körüli 2+2-es méretű ablakban, valamint tartalmas szavak a szó körül 3+3-as ablakban. A jellemzők második csoportja a teljes rendelkezésre álló szövegkörnyezet (általában a többértelmű szót tartalmazó teljes bekezdés) szemantikai tartományát, témáját modellezi. Ezt az információt bizonyos, a környezetben gyakori tartalmas szavak előfordulását leíró bináris vektor reprezentálja.

Mivel jelentésekkel annotált korpuszok csupán korlátozott mennyiségben állnak rendelkezésre, szükségem volt egy olyan megoldásra, amellyel a rendszer további skálázását lehetett biztosítani. Az egyik megoldás a többértelmű szavak előfordulásainak kézi annotálása lenne, ám ez igen időigényes és ezért költséges eljárás. Egy másik, kedvezőbb megoldást

szinkronizált, **párhuzamos korpuszok** alkalmazása jelenthet, mivel ezekben a megfelelő fordítások kinyerésével automatikusan előállíthatók tanítópéldák [45], [48].

A Hunglish Korpusz [49] a legnagyobb rendelkezésre álló, mondatszinten szinkronizált angol-magyar párhuzamos korpusz, amely 44,6 millió angol és 34,6 millió magyar szót tartalmaz, 5 különböző műfajból. A korpusz angol oldalát egy szófaji egyértelműsítővel [46] dolgoztam fel, és a Humor morfológiai elemző [42] segítségével meghatároztam a szóalakok szótöveit, valamint elvégeztem a magyar oldal tövesítését is.

A *state* többértelmű angol szóval végeztem kísérleteket, hogy felderítsem a párhuzamos korpusz angol-magyar gépi fordításhoz való tanítópéldák automatikus előállítására történő felhasználása közben fellépő lehetséges problémákat.

Először azonosítottam azokat a korpuszbeli előfordulásokat, melyek a *state* szót többszavas kifejezésekben tartalmazták az angol oldalon. A célszó ezekben a kollokációkban a környezettől függetlenül mindig ugyanabban a jelentésben szerepel, így a többszavas kifejezés lexikális transzfer-szabályokkal egyszerűen fordítható. Ehhez összeállítottam egy listát a *state* szóval alkotott többszavas kifejezésekből, amihez több forrást felhasználtam ([29], [43], [47]), valamint alkalmaztam a *Terminology Extractor* (version 3.0c, Copyright (C) 2002 Chamblon Systems Inc.) nevű programot releváns kollokációk azonosításához a korpusz angol oldalán. A végleges kollokáció-lista 348 különböző kifejezést tartalmazott.

A kétnyelvű szótár [47] segítségével összeállítottam egy listát az összes lehetséges egyszavas magyar fordításról is (19 tétel), ennek segítségével kiválasztottam a Hunglish korpusz azon mondatpárjait, melyek angol oldala tartalmazta a *state* főnevet, magyar oldala pedig az egyik lehetséges magyar fordítást. A mondatpárokat ezután 3 csoportba soroltam: a) az angol mondat egy vagy több ismert, *state*-tel alkotott kollokációt tartalmazott (93%), b) az angol mondat ismert kollokációt és egyéb, nem kollokációs előfordulást is tartalmazott (3%), és c) az angol mondat nem tartalmazott egyet sem az ismert kollokációk közül (4%). A b) és c) csoportba tartozó mondatpárokban megvizsgáltam, hogy a magyar mondat hányat tartalmaz az ismer egyszavas fordítások közül.

A magyar oldalon pontosan egy ismert fordítást, az angol oldalon kollokációt nem tartalmazó, vagyis annotált tanítópélda előállítására közvetlenül alkalmas mondatpárból összesen 2,473 darabot lehetett előállítani a *state* főnévhez. Korábbi kísérletek [37] megmutatták, hogy ekkora mennyiség megfelelő egy nagy pontosságú WSD osztályozó betanításához.

Munkám **harmadik része** a magyar szövegekben található entitások (főnévi csoportok) koreferenica- és birtokosviszonyainak azonosítására koncentrált.

A koreferencia-feloldás legutóbbi kutatásában az adatalapú, gépi tanulásra támaszkodó megközelítések kiszorították a hagyományos, szabályalapú rendszereket [52]. Ehhez azonban nagy mennyiségű annotált tanítópélda szükséges, és mivel magyar nyelvre kutatásaim

ideje alatt nem létezett ilyen adatbázis, a szabályalapú megközelítést választottam.

Az általam javasolt rendszer többféle tudásra támaszkodik: a MetaMorpho gépi fordítórendszer magyar mély nyelvi elemzőjének [43], [44] kimenetében található morfológiai, szintaktikai és szemantikai információkra; a Kormányzás- és Kötéselmélet magyar szintaxis-elméletben megfogalmazott változatára [50], továbbá a magyar mondatmegértés pszicholingvisztikájában elért kutatási eredményekre [53], [54] támaszkodó szabályokra; a Magyar WordNetben [6] található szemantikai tudásra; valamint karakteralapú heurisztikákra, hasonlóan [55] által javasolt módszerekhez.

A rendszer a MetaMorpho parser segítségével azonosítja a bekezdés-, mondat- és tokenhatárokat, a tagmondatokat és ezekben a maximális igei és főnévi csoportokat és ezek morfológiai, nyelvtani és szemantikai tulajdonságait. Ezután az anaforikus NP-khez a dokumentumban balról jobbra haladva, azok típusától függő szabályokkal keresi meg a legközelebbi antecedenseket.

A magyarban háromféle olyan **birtokos szerkezetről** beszélhetünk, melyekben a birtokos és a birtok közé egyéb mondatrészek ékelődhetnek. Ezek közül két jelenséget (birtoklásmondat, dativus-os birtokos) a MetaMorpho elemző képes pusztán nyelvtani eszközökkel kezelni, a harmadik típusú szerkezetek esetében (zérónévmás birtokos) azonban csak a zérónévmások koreferencia-feloldásához hasonló módszerek

alkalmazhatók, ehhez egy szabályalapú megközelítést alkalmazó megoldásra tettem javaslatot.

3. Új tudományos eredmények

I. Téziscsoport: Módszerek a Magyar WordNet ontológia automatikus létrehozására

I.1. Megmutattam, hogy a kiterjesztéses modellben sikerrel alkalmazhatók automatikus módszerek a magyar wordnet ontológia létrehozásának támogatására.

Az általam alkalmazott heurisztikák első csoportját [32] és [33] alkalmazta először a spanyol és katalán wordnetek fejlesztésére a kiterjesztéses modellben. A *Variant*, *Mono* és *Intersection* heurisztikák csupán a kétnyelvű szótárban és a PWN-ben található strukturális információkra támaszkodnak. Egy negyedik eljárás, melyet [32] javasolt, egynyelvű (értelmező) szótárból kinyert információkra támaszkodik. A szótári definíciókat elemzésével minden címszóhoz hozzá lehetett rendelni egy genus proximumot, melyet az ún. fogalmi távolság képlet segítségével fel lehet használni a címszó PWN párjának azonosításához. Ehhez a módszerhez felhasználtam a Magyar Értelmező Kéziszótár (EKSz) [36] egy elektronikus változatát, melyben a definíciókat a Humor morfológiai elemzővel [42] és egyszerű kivonatoló mintákkal feldolgozva

sikerült hipernima/hiponima, szinonima és meronima/holonima szavakat azonosítani a címszavakhoz.

A módszerek kiértékeléséhez a magyar wordnet fejlesztése során kétféle eljárást is alkalmaztam. Az első kiértékelési körben 400, a kétnyelvű szótár magyar oldaláról véletlenszerűen kiválasztott magyar főnévhez kézzel egyértelműsítettem a megfelelő PWN synseteket, és ehhez a referenciához képest határoztam meg minden heurisztika pontosságát és fedését. A [32] és [33] által javasolt módszerek az én implementációmban a magyar adatokon 49-92% pontosságot értek el, míg [32] 61-85%-ról számol be 10% véletlen mintán végzett kézi kiértékeléssel. [33] nyomán kísérletet tettem a módszerek kombinálására is. Így létre tudtam hozni egy előzetes, 10.786 magyar synsetből (9.986 magyar szóból) álló halmazt, melyben a magyar szó-PWN synset kapcsolatok átlagos becsült pontossága 75% volt. [33] 6.552 spanyol synsetről (7.922 spanyol szó) számol be 75%-os becsült átlagos kapcsolati pontossággal.

I.2. Bemutattam 4 új heurisztikát magyar synsetek automatikus létrehozására a kiterjesztéses modellben. A módszerek magyar főneveket rendelnek angol synsetekhez, és a magyar nyelv sajátosságaira, valamint a rendelkezésre álló erőforrásokra támaszkodnak.

Az általam javasolt heurisztikák a következők voltak:

- Az értelmező szótárból kinyert, valamint egy tezauruszból származó szinonimákat használtam fel úgy,

hogy a magyar címszót ahhoz a PWN synsethez rendeltem, ami a legtöbbet tartalmazta a szinonimák angol fordításai közül.

- A morfológiai elemző segítségével azonosítottam endocentrikus N+N főnévi összetételek fejét (utótagját), melyek „deriváviós hipernímaként" felhasználhatók a fogalmi távolság képlet alkalmazásával. A módszert alkalmaztam olyan főnévi többszavas lexémákra is, melyekben az utolsó tag főnév volt.
- Az EKSz szócikkek egy részében (állat- és növényfajok, taxonomikus csoportok, betegségek stb.) rendelkezésre álló latin megfelelőket felhasználtam a címszavaik olyan PWN synsetekhez rendeléséhez, melyek tartalmazták a latin kifejezéseket az angol szavak szinonimáiként.
- Azokban az esetekben, ahol az EKSz-ben azonosított genus/szinonima angol fordítása nem állt rendelkezésre, így a fogalmi távolság képletet nem lehetett alkalmazni, a lefedettség további növelésére felhasználtam a hipernima/szinonima relációk tranzitív tulajdonságát. A módszer megpróbálta felhasználni a genus/szinonima derivációs hipernimáját vagy definíciókból azonosított hipernimáját (utóbbi esetben csak akkor, ha a genus/szinonima nem volt többértelmű az EKSz-ben.)

A kiértékelés második részében arra voltam kíváncsi, hogy a módszereim pontossága és lefedettsége hogyan alakul a végleges,

emberi munkával létrehozott, 42.000 magyar synsetből álló Magyar WordNet (HuWN) ontológia ([6], [10], [12]) fényében. A HuWN létrehozása során az általam fejlesztett heurisztikák futtatásának eredményét lexikográfusok vették át, és ahol szükséges volt, szerkesztették, törölték, kiegészítették stb. az automatikusan javasolt magyar szinonimákat, illetve a PWN 2.0-ból átvett szemantikai relációkat.

A pontosságot a heurisztikák által javasolt *és* a lexikográfusok által jóváhagyott <magyar lexikai egység, PWN 2.0 synset> fordítási párok számának, valamint a heurisztikák által javasolt fordítási párok számának arányaként határoztam meg. A fedést a javasolt és jóváhagyott fordítási párok, valamint az összes, heurisztikák által lefordított synsetekben található fordítási párok arányaként definiáltam. Ezeket az értékeket meghatároztam minden, a HuWN-ben az automatikus synset-fordításban érintett szófajra (főnevek, igék, melléknevek). Az eredmények az 1. Táblázatban láthatók.

	Össz.	Főnevek	Igék	Melléknevek
Pontosság	24.61%	31.53%	13.89%	17.36%
Fedés	64.81%	63.77%	64.46%	71.96%

1. Táblázat: az automatikus synset-fordítási módszerek eredményeinek kiértékelése a végleges Magyar WordNetben.

II. Téziscsoport: Felügyelt gépi tanulásos jelentés-egyértelműsítés az angol-magyar gépi fordításban

II.1. Bemutattam egy jelentés-egyértelműsítő rendszert, amely képes lehet szabályalapú angol-magyar gépi fordításban a lexikális fordítási pontosság javítására. (Az egyértelműsítés nélkül a fordítórendszer a forrásnyelvben többértelmű szavakat mindig a leggyakoribb jelentés célnyelvi megfelelőjével fordítja.)

Az egyértelműsítő osztályozók kiértékeléséhez 10-szeres kiegyenlített keresztellenőrzést végeztem a 38 vizsgált öbbértelmű főnév tanítókorpuszán. A pontosságot a helyesen egyértelműsített példák és az összes példa arányaként definiáltam. Összehasonlítási alapként (baseline) meghatároztam minden szó leggyakoribb jelentésű előfordulásainak relatív gyakoriságát.

A kiértékelést mind az eredeti angol jelentések, mind a megfeleltetett magyar fordítások fölötti egyértelműsítésre elvégeztem. Az angol jelentések esetében az átlagos pontosság 77,99% volt, 64,16%-os átlagos baseline-értéke mellett. A magyar fordítások felhasználásával az osztályozók átlagosan 85,00%-os pontosságot értek el, ami 11,52%-kal múlta felül a baseline átlagos értékét. Az utóbbi esetben a 38 osztályozó közül 10 kivételével az összes pontossága meghaladta a baseline értéket, és csupán egy esetben csökkent a pontosság a baseline szintje alá.

II.2. Az angol wordnet-beli jelentés-címkék magyar fordításokra leképezésével a jelentések átlagos száma csökkenthető, az egyértelműsítés átlagos pontossága pedig növelhető.

A WordNetben található finom jelentés-megkülönböztetések nehézzé teszik nagy pontosságú jelentés-egyértelműsítő módszerek kifejlesztését, amennyiben azok WordNet synsetekből álló jelentés-címkékre építenek. Mivel az általam adott magyar jelentés-fordítások úgy készültek, hogy mindegyikük tartalmaz bizonyos mértékű többértelműséget, a fordítás során csökkent az megkülönböztetendő osztályok száma. Az angol synsetek magyar szavakra leképezésével az egyértelműsítés átlagos pontossága 7,01%-kal javult. A 38 esetből 27-ben a pontosság magasabb volt a magyar azonosítókkal, míg 11 esetben nem változott az angolhoz képest.

II.3. Megmutattam, hogy nagyméretű, szinkronizált párhuzamos korpusz segítségével, a kézi korpuszannotációhoz képest kevesebb munkával lehet előállítani annotált tanítópéldákat az angol-magyar gépi fordításban működő jelentés-egyértelműsítés számára. A bemutatott megközelítésben alapvető fontosságú a célszóval alkotott idiomatikus többszavas kifejezések felismerése a korpuszban.

A Hunglish korpusszal végzett kísérleteim megmutatták, hogy tanítópéldák előállításához szükséges 1) a lehetséges egyszavas fordítási párok listája, pl. egy kétnyelvű szótárból, 2) a célszóval alkotott többszavas kifejezések listája, különböző lexikai forrásokból, vagy korpuszalapú kollokáció-feltáró módszerekkel. A többértelmű esetek kiszűrése után a Hunglish korpusz tekintélyes méretei miatt a jelentésegyértelműsítő betanításához még mindig elegendő példamondat áll rendelkezésre (2.473 példa a *state* esetében, valamint 1.334 további példa, ami kollokációt is tartalmaz).

III. Téziscsoport: Szabályalapú koreferencia- és birtokosviszony-feloldás a magyarban

III.1. Bemutattam egy, különböző tudásforrásokra, valamint a nyelvi elemző hibáit felismerő heurisztikákra támaszkodó algoritmust főnévi csoportok koreferencia-viszonyainak feloldására magyar szövegekben.

A dokumentumokban szereplő NP-k koreferencia-feloldása a megszorítások és preferenciák módszeren alapul [51]. Az antecedens-jelöltek listájának előállítására, a lista szűrésére, valamint a jelölt kiválasztására használt algoritmus az anaforikus NP típusától függ.

Tulajdonneveknél az antecedens-jelölteket az anaforát megelőző teljes dokumentumban szereplő összes tulajdonnév adja. Az antecedens az anaforával legkisebb Levenshtein-távolságot mutató

jelölt lesz, normalizálás (determinánsok elhagyása, fej szótövesítése) után, valamint egy küszöbérték alkalmazásával, mely biztosítja, hogy a rendszer számára nem kötelező kiválasztani egyet a jelöltek közül.

Határozott névelős közneveknél az algoritmus először megpróbálja kizárni a közös világismeretből levezethető, unikus entitásokra referáló kifejezéseket egy lista segítségével. Az antecedens-jelöltek a tulajdonnevek és köznevek (determináns típusától függetlenül) az anafora teljes megelőző bekezdésében, az anafora VP-jéig (a Kötéselmélet kizárja az anafora VP-jében a főige által dominált jelölteket.) Az antecedens kiválasztása az anaforához legközelebb eső, vele azonos fejű jelölt azonosítását, vagy a legközelebbi, a Leacock-Chodorow hasonlósági képlettel [37] szemantikailag leghasonlóbb, Magyar WordNet segítségével azonosított szinonima vagy hipernima/hiponima megtalálását jelenti.

A rendszer képes **személyes névmások** kezelésére is, kiegészülve az „az” mutató névmással, amely alanyi helyzetben, nem utalószói szerepben áll. Antecedens-jelölteknek az anafora mondata előtti második mondattól kezdve (amennyiben az létezik a bekezdésben) választjuk ki az összes NP-t, az anaforát tartalmazó tagmondat határáig. A rendszer ezután ezeket megszűri az anafora és az antecedens-jelölt nyelvtani számának, személyének és két szemantikai jegyének (+/-ÉLŐ, +/-EMBER) egyezését vizsgálva, valamint azoknak az antecedens-jelölteknek a kizárásával, amelyekre már koreferenciát állapított meg a vizsgált anaforával egy tagmondatban szereplő valamelyik másik névmási vagy zérónévmási anaforára nézve (ld. Kötéselmélet). Ha egy mondatban több

személyes névmás van, akkor a rendszer azokat az oblikuszi hierarchiának megfelelő sorrendben oldja fel, a már kötött jelöltek kizárása érdekében. A köznevek és tulajdonnevek feloldása egy adott mondatban mindig megelőzi a névmások feldolgozását, hogy a lehetséges jelöltek számának csökkentésével tovább segítse a névmások feloldását.

A tagmondatában alanyi szerepű névmási vagy zérónévmási anafora antecedensének meghatározásában a magyar mondatmegértés pszicholingvisztikájában végzett kutatások eredményeire támaszkodtam [53], [54]. Az algoritmus a szerkezeti párhuzamosság feltételezéséből indul ki, mely szerint az alanyi helyzetű anafora az előzménymondat alanyára utal vissza. Ezt felülbírálhatja az alanyi szerepben álló „az" mutatónévmás, ami alanyváltást jelöl. Ha a megelőző tagmondatban több nem-alanyi szerepű NP is található, a rendszer ezek közül az oblikuszi hierarchia és az anaforától való távolság alapján választ (az algoritmus a mondat végéhez közelebb álló NP-ket preferálja). Nem alanyi pozícióban álló névmások, zérónévmások esetén több, az alannyal nem koreferens antecedens-jelölt közül szintén az előbbi szabály alkalmazásával választ a rendszer.

A koreferencia-feloldó algoritmus **kiértékeléséhez** készítettem egy kisméretű kézzel annotált referencia-korpuszt (10 szövegrészlet, összesen 99 mondattal, 81 annotált NP-vel.) A koreferencia-feloldás pontossága ezen a korpuszon mérve 68,92% volt, a fedés 62.96%-ot ért el. A leggyakoribb anafora-típusok esetén a pontosság 71% és 80% között mozgott, 61% és 83% közötti fedéssel. A WordNet-

alapú módszerek alacsony teljesítményt mutattak (0-33% F-mérték), de mivel csak 6 ilyen példa volt a korpuszban, a mért értékek nem biztos, hogy realisztikusak.

Elvégeztem az algoritmus által vétett hibák elemzését is, ami megmutatta, hogy névmások esetében (ami a leggyakoribb anafora-típus a korpuszban) a hibák közel fele a nyelvtani elemző hibás kimenetéből adódott. Hibátlanul elő-elemzett input esetén az átlagos feloldási pontosság 75%-ra, a névmások/zérónévmások feloldásának pontossága 91%-ra emelkedne.

III.2. Bemutattam egy, a névmási anafora-feloldáshoz hasonló módszert az elvált birtokos-birtok NP-k azonosítására magyar szövegekben.

Az algoritmus a következő feltételezéseken alapul: 1) a birtokot domináló VP alanyi vonzata az alapértelmezett birtokos, 2) a birtokos számban és személyben egyezik a birtokon azonosított birtokos jel számával és személyével. A második feltételezés felülbírálhatja az elsőt, vagyis, ha a birtokos VP-jének alanya nem egyezik számban és személyben, akkor az előző tagmondat alanya töltheti be a birtokos szerepét, feltéve, hogy ugyanabban a diskurzus-szegmentumban (bekezdésben) található.

Ennek fényében a birtokosviszony-feloldás algoritmusa először azonosítja a birtok NP-t tartalmazó mondat előtti mondatokban található, alanyi szerepű NP-ket, legfeljebb két megelőző mondatnyi távolságban, de nem lépve messzebb az aktuális bekezdés legelső

mondatánál, majd ezek közül a birtokon azonosított birtokos jel számával és személyével egyező számú és személyű, a birtok NP-hez legközelebb eső NP-t választja. Amennyiben nem áll rendelkezésre teljes szintaktikai elemzés, vagyis információ az NP-k nyelvtani szerepéről a szintaktikai elemző kimenetében, akkor az algoritmus a birtok előtt álló, számban és személyben egyező, alanyesetű NP-k közül választja ki a legjobboldalibbat.

A javasolt módszer kiértékelését a koreferencia-feloldás kiértékeléséhez használt korpuszon végeztem el, melyben 38 elvált birtokos szerkezetet annotáltam kézzel. Ezen a halmazon a birtokos-birtok azonosítás 76,47%-os pontosságot és 68,42%-os fedést ért el (F mérték = 72,22%).

4. Az eredmények gyakorlati alkalmazása

A disszertációban tárgyalt összes munka kapcsolódott olyan projektekhez, melyekben az eredményeim gyakorlati alkalmazásra kerültek.

A **magyar wordnet** ontológia létrehozására javasolt módszereket a Magyar WordNet Projekt keretei között valósítottam meg, mely az Európai Unió ECOP programja támogatásában (GVOP-AKF-2004-3.1.1.), 2005-2008 között több vezető magyar akadémiai és ipari partner részvételével (Magyar Tudományos Akadémia Nyelvtudományi Intézet, Szegedi Tudományegyetem, MorphoLogic Kft.) valósult meg. A projektben a Princeton WordNet 2.0-s verziója szolgált a kiterjesztéses modell alapjául, és az én heurisztikáim felhasználásával készült el a főnévi, igei és melléknévi

synsetek automatikus fordítása, amit lexikográfus szakértők kézi munkával ellenőriztek és javítottak ki. A projekt eredményeképpen így előállított Magyar WordNet több, mint 40,000 synsetet tartalmaz.

Az előállított ontológia alkalmazásra került egy **információ-kivonatoló** projektben is [6], melynek során részt vettem egy rövid szöveges üzleti hírekből, eseménykeretekre épülő, tényeket kivonatoló rendszer fejlesztésében. A kivonatoláshoz 124 db, igei vonzatkeretekre, morfológiai és szemantikai megszorításokra épülő eseménykeret készült el. A szemantikai megszorításokat a főnévi Magyar WordNet ontológia csomópontjainak megfeleltetett szemantikai osztályok biztosították.

A disszertációban bemutatott **jelentés-egyértelműsítő** rendszer kifejezetten a MorphoLogic Kft. *MetaMorpho* angol-magyar fordítóprogramja számára készült. A fordítóprogram magját alkotó, manuálisan létrehozott, kontextusfüggetlen nyelvtani elemző és generáló szabályok csupán korlátozott mértékben képesek szemantikai információt kódolni, így segítségül kellett hívni egy külső „orákulumut", amely képes a rendelkezésre álló kontextus alapján dönteni a megfelelő jelentésekről. Egy, a disszertációban bemutatott módszerekre épülő WSD modul beépítésre került a MetaMorpho motorjába. A modul feladata a forrásnyelvi bemeneti bekezdés előfeldolgozása (szegmentálás, morfológiai elemzés, tövesítés) után egy speciális nyelvtani jegy értékének beállítása, amely azonosítja a felismert szavak aktuális jelentését. A forrásnyelvi szintaktikai elemzés további lépései építhetnek erre a jegy-értékre. A célnyelvi generálás fázisában egy elágazás választja

ki a megfelelő célnyelvi fordítást a jelentésazonosító jegy értékének megfelelően. Az angol jelentések és a magyar fordítások közötti megfeleltetés a fordítási szabályokban van rögzítve, ami könnyű karbantarthatóságot biztosít.

A disszertációban ismertetett **koreferencia- és birtokosviszony-feloldó módszerek** a „*Nemzeti és etnikai identitás vizsgálata történelmi eseményekre vonatkozó szövegek narratív alakzatainak számítógépes tartalomelemzése révén”* című projekt [57] keretei között kerültek megvalósításra. A projekt több különböző magyar intézmény részvételével (Pécsi Tudományegyetem, MTA Pszichológiai Kutatóintézet, Szegedi Tudományegyetem, MorphoLogic Kft, MTA Nyelvtudományi Intézet) 2006-2008 között a Nemzeti Kutatási és Fejlesztési Hivatal támogatásával (NKFP6 00074/2005, Jedlik Ányos Program) valósult meg. A projektben iskolai történelemkönyvi szövegekből álló, szintaktikai, morfológiai és szemantikai információkkal (összetevők, nyelvtani szerepek, tematikus szerepek és szemantikai kategóriák) annotált korpuszon futtattak speciális lekérdezéseket, melyek a projekt által vizsgált pszichológiai jelenségeket vizsgálták. Az általam kifejlesztett koreferencia- és birtokosviszony-feloldó módszereket sikerrel alkalmazták a vizsgálatok lefedettségének kiterjesztésére a lekérdezésekben érintett entitásokra referáló kifejezések azonosítása segítségével.

5. Köszönetnyilvánítás

Szeretnék köszönetet mondani témavezetőnek, Prószéky Gábornak. Köszönettel tartozom minden kollégámnak, akik ötleteikkel és hasznos hozzászólásaikkal hozzájárultak a munkámhoz: Hatvani Csaba, Kuti Judit, Merényi Csaba, Naszódi Mátyás, Pohl Gábor, Schönhofen Péter, Szarvas György, Tihanyi László, Vajda Péter, Varasdi Károly és még sokan mások. Hálával tartozom a Pázmány Péter Katolikus Egyetem Információs Technológiai Kar Doktori Iskolájának azért, hogy lehetőséget biztosított kutatásaim folytatására. Köszönet Vásárhelyi Gábornak, Bankó Évának és a Doktori Iskola többi hallgatójának segítségükért, hogy fontos információkkal segítették a disszertációm elkészülését. És végül, de nem utolsó sorban különösen hálás vagyok minden barátnak és családtagnak támogatásukért.

A disszertációban ismertetett munkát részben a GVOP-AKF-2004-3.1.1. és az NKFP6 00074/2005 (Jedlik Ányos Program) projektek támogatták.

6. Publikációk

Folyóiratokban

[1] **Miháltz Márton**: Tudásalapú koreferencia- és birtokosviszony-feloldás magyar szövegekben. To appear in: *Általános Nyelvészeti Tanulmányok*

[2] Prószéky, Gábor, **Miháltz Márton**: Magyar WordNet: az első magyar lexikális szemantikai adatbázis. In: *Magyar Terminológia* 1 (2008) 1, pp. 43-57.

[3] Németh, Dezső, Ivády Eszter Rozália, **Miháltz Márton**, Krajcsi Attila, Pléh Csaba: A verbális munkamemória és morfológiai komplexitás. In *Magyar Pszichológiai Szemle*. 61. évf., 2. szám, pp. 265-298.

Konferenciákon

[4] **Miháltz Márton**: Információ-kivonatolás szabad szövegekből szabályalapú és gépi tanulásos módszerekkel. In: *VI. Magyar Számítógépes Nyelvészeti Konferencia* kiadványa, Szeged, pp.49-58, 2009.

[5] **Miháltz, Márton**: Knowledge-based Coreference Resolution for Hungarian. In: *Proceedings of The Sixth International Conference on Language Resources and Evaluation* (LREC 2008), Marrakesh, Morocco, 2008.

[6] **Miháltz, Márton**, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, Tamás Váradi: Methods and Results of the Hungarian WordNet Project. In: *Proceedings of The Fourth Global WordNet Conference*, Szeged, Hungary (2008), pp. 311–321.

[7] **Miháltz Márton**, Naszódi Mátyás, Vajda Péter, Varasdi Károly: NP-koreferenciák feloldása magyar szövegekben a Magyar WordNet ontológia segítségével. In: *V. Magyar Számítógépes Nyelvészeti Konferencia kiadványa*, Szeged (2007), pp. 138–146.

[8] Hatvani Csaba, Kocsor András, **Miháltz Márton**, Szarvas György, Szécsi Katalin: Főnevek a Magyar WordNetben. *IV. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, pp. 109-116.

[9] **Miháltz, Márton**, Gábor Pohl: Exploiting Parallel Corpora for Supervised Word-Sense Disambiguation in English-Hungarian Machine Translation. *Proceedings of the 5th Conference on Language Resources and Evaluation*, 1294–1297. Genoa, Italy (2006)

[10] Alexin, Zoltán, János Csirik, György Szarvas, András Kocsor, **Márton Miháltz**: Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. In *Proceedings of the Third International WordNet Conference* (GWC 2006), Seogwipo, Jeju Island, Korea, January 22-26, 2006, pp. 291-292.

[11] **Miháltz Márton**, Pohl Gábor: Javaslat szemantikailag annotált többnyelvű tanítókorpuszok automatikus előállítására jelentés-egyértelműsítéshez párhuzamos korpuszokból. *III. Magyar számítógépes nyelvészeti konferencia*, Szeged, 2005. december 8-9, pp. 418-419.

[12] **Miháltz Márton**, 2005: Magyar EuroWordNet projekt: bemutatás és helyzetjelentés. *III. Magyar számítógépes nyelvészeti konferencia*, Szeged, 2005. december 8-9, pp.68-78.

[13] **Miháltz, Márton**, 2005: Towards A Hybrid Approach To Word-Sense Disambiguation In *Machine Translation. Workshop „Modern Approaches in Translation Technologies" at Recent Advances in*

Natural Language Processing (RANLP-2005) Conference, Borovets, Bulgaria.

[14] Németh, Dezső, Ivády Eszter Rozália, **Miháltz Márton**, Pléh Csaba: "Phonological loop and morphological complexity" XIVth ESCOP - *Conference of European Society for Cognitive Psychology*, August 31 - September 3, 2005, Leiden

[15] **Miháltz Márton**, 2004: Angol-magyar gépi fordítórendszer támogatása jelentés-egyértelműsítő modullal. *Második Magyar Számítógépes Nyelvészeti Konferencia* (MSzNy-2004), Szeged, pp. 92-99.

[16] **Miháltz, Márton**, 2004: Word Sense Disambiguation Using Random Indexing. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

[17] **Miháltz, Márton**, Gábor Prószéky, 2004: Results and Evaluation of Hungarian Nominal WordNet v1.0. In *Proceedings of the Second International WordNet Conference* (GWC 2004), Brno, Czech Republic, pp. 175-180.

[18] **Miháltz, Márton**, 2003: Magyar főnévi WordNet létrehozása automatikus módszerekkel (Constructing a Hungarian WordNet Ontology with Automatic Methods). *Első Magyar Számítógépes Nyelvészeti Konferencia* (MSzNy-2003), Szeged, pp. 153-160.

[19] **Miháltz, Márton**, 2003: Constructing a Hungarian ontology using automatically acquired semantic information. In *Proceedings of the 5th International Workshop on Computational Semantics* (IWCS-5), Tilburg, The Netherlands, pp. 475-478.

[20] Prószéky, Gábor and **Márton Miháltz**, 2002: Automatism and User Interaction: Building a Hungarian WordNet. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, Vol 3, pp. 957-961.

[21] Prószéky, Gábor and **Márton Miháltz**, 2002: Semi-Automatic Development of the Hungarian WordNet. In *Proceedings of the LREC 2002 Workshop on WordNet Structures And Standardization, And How These Affect WordNet Applications And Evaluation*, Las Palmas de Gran Canaria, Spain, pp. 42-46.

[22] Prószéky, Gábor, **Márton Miháltz** and Dániel Nagy, 2001: Toward a Hungarian WordNet. In Proceedings of the NAACL 2001. *Proc. Workshop on WordNet and Other Lexical Resources*, Pittsburgh, USA, pp.174-176.

Egyéb publikációk

[23] **Mihaltz, Márton**: Development of the Hungarian WordNet Ontology and its Application to Information Extraction. Presentation at the *10th International Protégé Conference*, Budapest, Hungary (2007)

[24] **Mihaltz Márton**, Prószéky Gábor: Egy magyar WordNet felé. Előadás a *W3C Szemantikus Web Műhelykonferencián*, MTA SZTAKI W3C Magyar Iroda, Budapest, 2006. április 13.

[25] Németh, Dezső, Rozália Eszter Ivády, **Márton Mihaltz**, Attila Krajcsi, Csaba Pléh, 2005: Verbal Working Memory And Morphology. Poster at the *9th European Congress of Psychology*, Granada, Spain.

[26] Ivády Rozália Eszter, Németh Dezső, **Mihaltz Márton**, Pléh Csaba, 2004: Fonológiai hurok és morfológia komplexitás. Magyar Pszichológiai Társaság Biennális Nagygyűlése, Debrecen, 2004.

[27] Ivády R. E., **Mihaltz M.**, Németh D., Pléh Cs. (2004). A rövidtávú emlékezet és morfológiai komplexitás. In Németh D. (szerk.). *Szegedi Pszichológiai Tanulmányok*, JGYTF Kiadó, Szeged, pp. 21-32.

7. A témához kapcsolódó irodalom

[28] Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller: Introduction to WordNet: an on-line lexical database. Int. J. of Lexicography 3 (1990) 235–244.

[29] Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998)

[30] Tufiş, D., Cristea, D., Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal of Information Science and Technology Special Issue, vol. 7, no. 1-2 (2004)

[31] Vossen, P. (ed.): EuroWordNet General Document, Version 3. University of Amsterdam (1999)

[32] Atserias, J., S., Climent, X., Farreres, G., Rigau, H., Rodríguez: Combining multiple methods for the automatic construction of multilingual WordNets. Proc. of Int. Conf. on Recent Advances in Natural Language Processing, Tzigov Chark (1997)

[33] Farreres, X., G., Rigau, H., Rodriguez: Using WordNet for building Wordnets. Proc. of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal (1998)

[34] Eduard Barbu, Verginica Barbu Mititelu, Automatic Building of Wordnets. In N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov

(Eds.), Recent Advances in Natural Language Processing IV (RANLP-05), 2005.

[35] Kiefer Ferenc (2001). Jelentéstan. Corvina, Budapest.

[36] Juhász, J., I., Szőke, G. O. Nagy, M. Kovalovszky (eds.): Magyar Értelmező Kéziszótár. Akadémiai Kiadó, Budapest (1972)

[37] Leacock, C., Miller, G. A., Chodorow, M.: Using Corpus Statistics and WordNet Relations for Sense Identification. Computational Linguistics, Special Issue on Word Sense Disambiguation. (1998)

[38] Manning, C. D., Schütze, H: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)

[39] Mihalcea, Rada Word sense disambiguation with pattern learning and automatic featureselection. Journal of Natural Language Engineering (special issue on evaluating word sense disambiguation systems, 8 (4) 279-291 (2002)

[40] Mihalcea, R., Chklovski, T.: Building a Sense Tagged Corpus with Open Mind Word Expert. Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions (2002)

[41] Mihalcea, R., Chklovski, T. and Kilgarriff, A.: The Senseval-3 English Lexical Sample Task. In Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Barcelona, Spain (2004)

[42] Prószéky, G.: Humor: a Morphological System for Corpus Analysis. Language Resources and Language Technology, Tihany (1996)

[43] Prószéky, G., Tihanyi, L.: MetaMorpho: A Pattern-based Machine Translation Project. Proceedings of the 24th 'Translating and the Computer' Conference. London, UK, 19–24 (2002)

[44] Prószéky, Gábor; László Tihanyi; Gábor Ugray: Moose: a robust high-performance parser and generator. Proceedings of the 9th Workshop of the European Association for Machine Translation, Foundation for International Studies, La Valletta, Malta, pp. 138–142 (2004)

[45] Diab, M. (2004): Relieving the data acquisition bottleneck for Word Sense Disambiguation. In Proceedings of ACL 2004.

[46] Giménez, J., L. Márquez: SVMTool (2004): A general POS tagger generator based on Support Vector Machines. In Proceedings of the 4th International Conference on Language Re-sources and Evaluation (LREC'04). Lisbon, Portugal.

[47] Országh, L., Magay, T. (2004): Angol-magyar nagyszótár. Budapest: Akadémiai Kiadó.

[48] Specia, L., M. G. Volpe Nunes, M. Stevenson (2005): Exploiting Parallel Texts to Produce a Multi-lingual Sense Tagged Corpus for Word Sense Disambiguation. In Proceedings of Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria

[49] Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón (2005): Parallel corpora for medium density languages. In Proceedings of Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria.

[50] Kenesei István: Az alárendelt mondatok szerkezete. In: Kiefer Ferenc (szerk.): Strukturális Magyar Nyelvtan, I. kötet, Mondattan. Akadémiai Kiadó, Budapest (1992)

[51] Mitkov, Ruslan: Anaphora Resolution: The State of The Art. Working Paper, University of Wolverhampton, 1999.

[52] Ng, Vincent: Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. Proceeding of the 43rd Annual Meeting of the Associaction for Computational Linguistics (2005)

[53] Pléh Csaba, Radics Katalin: „Hiányos mondat", pronominalizáció és a szöveg. In Általános Nyelvészeti Tanulmányok, XI, 261-277 (1976).

[54] Pléh Csaba: Mondatközi viszonyok feldolgozása: az anafora megértése a magyarban. In: Pléh Csaba: Mondatmegértés a magyar nyelvben. Osiris Kiadó, Budapest (1998)

[55] Uryupina, Olga: Evaluating Name-Matching for Coreference Resolution. In Proceedings of the 4th International Conference on Language Resources and Evaluation (2004)

[56] Varasdi Károly: Koreferenciák feloldása. MTA Nyelvtudományi Intézet (2005)

[57] Szalai Katalin, Ferenczhalmy Réka, Fülöp Éva, Vincze Orsolya, László János: Történelmi szövegek narratív pszichológiai vizsgálata a nemzeti identitás tükrében. In VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2009, pp. 259–271.

[58] Witten, I. H., E. Frank, *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2005.