Miskolci Egyetem

Gépészmérnöki és Informatikai Kar

Ontológia-alapú szemantikai annotáció és tudásábrázolás nyelvtantanuló rendszerben

Ph.D. értekezés tézisei

Készítette:

Baksáné Varga Erika

okleveles mérnök-informatikus
okleveles mérnök-közgazdász

aki doktori fokozat elnyerésére pályázik

Hatvany József Informatikai Tudományok Doktori Iskola
Alkalmazott számítástudomány tématerület
Adat- és tudásbázisok, tudásintenzív rendszerek témacsoport

Doktori Iskola vezető:
Prof. Tóth Tibor
a műszaki tudomány doktora

Témavezető:
Dr. habil. Kovács László

Miskolc, 2011.

Baksáné Varga Erika

Ontológia-alapú szemantikai annotáció és tudásábrázolás nyelvtantanuló rendszerben

Ph.D. értekezés tézisei

Miskolc, 2011.

Védési bizottság

Elnök:

Dr. Tóth Tibor, DSc	ME, egyetemi tanár

Titkár:

Dr. Körei Attila, PhD	ME, egyetemi docens

Tagok:

Dr. Baranyi Péter, DSc	MTA SZTAKI
Dr. habil. Radeleczki Sándor, CSc	ME, egyetemi docens
Dr. habil. Szigeti Jenő, CSc	ME, egyetemi tanár
Dr. Tar József, CSc	Óbudai Egyetem, egyetemi docens

Opponensek:

Dr. Cser László, DSc	Corvinus Egyetem, egyetemi tanár
Dr. Dudás László, CSc	ME, egyetemi docens

Tartalomjegyzék

1. Bevezetés

A kutatás fő célja egy általános, szemantikai annotációt alkalmazó statisztikai szabálytanulási módszertan kidolgozása. Miután a szimbolikus nyelvek szabályrendszere (nyelvtana) bír a legnagyobb gyakorlati jelentőséggel, ezért ezt vesszük alapul a módszertan alapjainak meghatározásakor. A statisztikai módszerek alkalmazása a nyelvtantanulásban [Charniak, 1996], [Manning & Schütze, 1999] azt jelenti, hogy a vizsgált nyelv szintaktikailag nem kerül elemzésre, pusztán az egyes szintaktikai elemek gyakorisági adatai alapján következtetünk a szabályokra. Azonban Gold mérföldkövet jelentő publikációja [Gold, 1967] óta tudjuk, hogy a Chomsky-hierarchiában [Chomsky, 1956] szereplő nyelvosztályok egyike sem tanulható csak pozitív mintából. Ezért a tanítómintát ki kell egészíteni negatív példákkal vagy szerkezeti információval (annotációval). Ez alapján, a kutatás abból az előfeltevésből indult ki, hogy a szabálytanulásban alkalmazott statisztikai módszerek szemantikával történő ötvözése pozitívan befolyásolja a tanuló algoritmusok hatékonyságát. A szemantikai információ tárolásához a mesterséges intelligencia egy napjainkban felfutó ágának, az ontológiának a lehetőségeit és módszereit használjuk fel, mivel az ontológiákat egyre szélesebb körben alkalmazzák olyan területeken ahol a szemantikai információ felhasználása további előnyökkel kecsegtet.

A feladat mérete és bonyolultsága miatt a dolgozat nem terjed ki az új módszertant alkalmazó tanuló ágensek teljeskörű működési modelljének leírására és egy ilyen ágens implementálására. Csupán a feldolgozás első, adatelőkészítő fázisára koncentrál, ahol *elsődleges cél egy megfelelő, nyelvtantanulásra optimalizált szemantika alapú tudásábrázolási módszer kidolgozása és kiterjedt célorientált vizsgálata.*

1.1. Irodalmi áttekintés

Az ágenstechnológia, a nyelvtantanulás és az ontológia egyaránt a mesterséges intelligencia (MI) tárgykörébe tartozó fogalmak. Eredetileg a mesterséges intelligencia célja az volt, hogy olyan számítógépes rendszereket hozzon létre, amelyek intelligens módon képesek feladatokat megoldani. Az új szemléletű, viselkedésalapú megközelítés szerint azonban a mesterséges intelligencia célja az, hogy a feladatmegoldást olyan ágensekkel végeztesse el, amelyek az intelligens viselkedés bizonyos vonásaival rendelkeznek. Egy ágens lehet bármi, ami érzékelői segítségével

észleli környezetét, majd – megfelelő döntéseket hozva – tevékenységével visszahat rá [Futó, 1999].

A nyelvtantanulás (Grammar Induction / Grammar Inference, GI) a nyelvtechnológia egyik részterülete. A nyelvtechnológia (Natural Language Processing, NLP) a mesterséges intelligencia azon határterülete, amelynek célja a számítógépekkel természetes nyelven történő kommunikáció megvalósítása [Jurafsky & Martin, 2000]. Sajnos azonban az emberéhez hasonló mélységű gépi megértésről egyenlőre nem beszélhetünk, mert az emberi megértés igen bonyolult és hosszú elsajátítási folyamat eredménye, melyben a nyelvi eszközökön kívül sok más nem nyelvi intelligenciakomponens is részt vesz [Futó, 1999]. Ezért a nyelvtechnológia alkalmazott irányzata nem azt várja el, hogy a számítógép megértse a természetes nyelvű bemenetet, hanem mindössze azt, hogy az elvártnak megfelelő válaszokat adja (racionálisan tudjon következtetni).

1.1.1. A fogalomalkotás folyamata

A nyelvtantanuló rendszer modellezéséhez először az emberi információ feldolgozást kellett tanulmányozni. Az emberek velük született kognitív képességeiknek köszönhetően képesek érzékelni a környezetükből érkező jeleket, majd a fogalomalkotás (*conceptualization*) során kialakul azok belső reprezentációja (új információ + korábbi ismeretek = tudás). Ha ezt a modellt egy kommunikációs közegben helyezzük el, akkor még hozzátesszük, hogy az ember a környezetéről ily módon alkotott ismereteit, megfigyeléseit jelek segítségével adja tovább, közli másokkal. A jelölés az a folyamat, amelynek során egy bonyolultabb jelenséghez egy azzal bizonyos szempontok alapján azonosított, egyszerűbb jelenséget kapcsolunk (szemiózis) [Sowa, 2000]. A jelek osztályozási rendszere Peirce műve (1867), és e szerint az emberi kommunikáció alapvető eszköze, a nyelv, szimbolikus természetű jelrendszer. Akárcsak Arisztotelész, Peirce is egy háromszöggel (*semiotic triangle*) írta le a környezet objektumai valamint az azokat helyettesítő jelek (szimbólumok) viszonyát, a jelek értelmezésének folyamatát [Hartshorne et al., 1958]. Peirce elméletét Ogden és Richards [Ogden & Richards, 1923] a nyelvi szimbólumok jelentésének meghatározására használta. Modelljük szerint a nyelvi szimbólumok értelmezése során minden korábbi tapasztalat és kontextus felidézésre kerül, amely alapján egyértelművé válik a hivatkozott objektum. Mivel azonban mindenki más tapasztalatokkal rendelkezik és

esetleg eltérő kontextusban találkozik ugyanazzal a szimbólummal, így mindenki számára mást jelent(het) ez a jel.

Egy ágens belső tudásbázisának felépülését Peirce elméletére alapozva Sieber és Kovács [Kovács & Sieber, 2009] többszintű szemantikai adatmodellje írja le. A modell szerint a fogalomalkotás folyamata több lépésben zajlik. A szintek száma, a folyamat bonyolultsága az ágens kognitív képességeitől függően változik. Az értelmezés első szintje mindenképpen a környezet objektumainak és a közöttük fennálló viszonyoknak direkt leképzése a belső reprezentációra, ami egy szemantikus háló. Mivel a gyakorlati ágensek környezete időben változó, a belső tudásbázist is az időben dinamikusan változónak tételezzük fel. Ebből Ogden és Richards elmélete szerint az következik, hogy egy jel jelentése a belső tudásbázis korábbi állapotainak függvénye.

1.1.2. Tudásábrázolás és ontológia

A nyelvtantanuló ágens ismereteinek, tudásának ábrázolásához szükséges a létező tudásábrázolási eszközök számbavétele. Ezek közül is az MI-n belül jelenleg legintenzívebben művelt terület, az ontológia mint tudásreprezentációs modell állt a vizsgálat középpontjában. Az ontológia eredetileg a filozófia egyik ágazata, a lételmélet (a létező dolgok tudománya), amely a létező dolgok szisztematikus számbavételével foglalkozik. Egy szakterület ontológiája az adott területre jellemző kategóriákat (fogalmakat, objektumokat, kifejezéseket), illetve a köztük fennálló kapcsolatokat írja le – jelentésükkel együtt.

Az MI-n belül a jelenleg elfogadott meghatározás szerint egy adott tárgyterület vonatkozásában az ontológia a fogalomalkotás explicit specifikációja: egy tárgyterület fogalmainak és az azok között fennálló kapcsolatoknak formális specifikációja, amelyhez általában természetes nyelvű leírás is társul [Gruber, 1993]. Egy adott tárgyterület ontológiája egy olyan reprezentációs szójegyzék, amely a tárgyterület leírandó fogalmairól és objektumairól, azok tulajdonságairól és kapcsolatairól szól. Tartalmazza azok olvasható formában leírt megnevezését, a nevek jelentését (interpretációját) és jellemzését (pl. az interpretációs korlátozásokat) [Sántáné-Tóth, 2006]. Azaz minden ontológia megad egy olyan kommunikációs szövegkörnyezetet (*domain of discourse*), amelyben az adott terület fogalmai vitathatók, egyértelműen elemezhetők [Szeredi et al., 2005]. Ezáltal az ontológia alkalmas eszköz a számítógéphálózatokon keresztül történő információ- és tudásmegosztás és újrafelhasználás támogatására.

Emellett az ontológia, mivel jelentést hordoz és tartalmi (szemantikai) kérdésekkel foglalkozik, lehetőséget biztosít a szöveges adatok tartalomorientált feldolgozására is.

Az ontológia leíró nyelvekről részletes jellemzést és összefoglalást nyújt [Bechhofer, 2002], [Calí et al., 2005] és [Scriptum, 2005]. Egy ontológia ábrázolható szöveges vagy grafikus formában. Az ontológiát szöveges formában modellező nyelvek nagyobb része a logikai tudásreprezentációs eszközök családjába tartozik, de léteznek más, például keret-alapú megközelítések is. Grafikus ontológia modellező nyelv nem létezik, de miután a fogalmi adatsémák és az ontológiák sok hasonlóságot mutatnak, számos kísérletet tettek már a létező fogalmi modellek (főleg az UML) ontológia-modellezésben történő alkalmazására [Xueming, 2007], [Jarrar et al., 2003], [Wang & Chan, 2001], [Cranefield & Purvis, 1999].

Az ismeretalapú rendszerekben az elsődleges deklaratív tudásábrázolási mód a logika, bár eredeti formájában (predikátumkalkulus + rezolúció) gyakorlatilag nem használják. Ez a nyelv kellően rugalmas a bonyolult állítások formális leírásához, és pontos szintaxissal, jól definiált szemantikával rendelkezik. Továbbá a nyelvhez tartozó bizonyító, következtető eljárás helyes és teljes, azaz minden formalizálható (és megoldható) feladat megoldható vele, bár nem hatékonyan. A nulladrendű predikátumkalkulussal (vagy ítéletkalkulussal) kevés gyakorlati problémát lehet leírni, ezzel szemben az elsőrendű predikátumkalkulus nyelve jóval nagyobb kifejező erővel rendelkezik. Ez utóbbit általában összehasonlítási alapként használják a reprezentációs eszközök kifejező erejének meghatározásánál, de a gyakorlatban a legtöbb probléma megoldásánál valamely nem-standard logikát részesítik előnyben [Futó, 1999].

A logika-alapú ismeretábrázolási nyelvek közül a leíró logikák (Description Logics, DL) osztálya [Baader et al., 2003], [Bognár, 2000] a legjelentősebb. Kutatásuk a korai szemantikus háló kutatásokból indult ki, formális és operációs szemantikát adva azoknak. A kutatók az elsőrendű logika egy olyan szegmensét keresték, amely elég magas kifejező erővel rendelkezik, de (még) adható hozzá eldönthető és hatékony következtető eljárás. A leíró logikák kifejező ereje az elsőrendű logikával összehasonlítva csekély, viszont a következtetési feladatok számítógéppel jól kezelhetők és polinomiális idejű algoritmusokkal a problémák mindig eldönthetők. A leíró logika segítségével le tudjuk írni egy szakterület fogalmi rendszerét, mert alapvető elemei a fogalmak, a szerepek és az egyedek. A fogalmak az egyedek valamely összességének közös sajátosságait írják le, és az egyedek halmazaiként értelmezett unáris predikátumnak

tekinthetők. A szerepek egyedek közötti bináris relációk (tulajdonságok, attribútumok). Minden leíró logika tartalmaz olyan nyelvi szerkezeteket, amelyek segítségével új fogalmakat és szerepeket képezhetünk; és összetett leírások megadása is lehetséges, beleértve a szerepek bináris relációira vonatkozó megszorításokat.

A leíró logikán alapuló modellek matematikai megalapozottsága és végrehajtási hatékonysága miatt ontológia modellező nyelvnek a szabványos OWL (Web Ontology Language) [Bechhofer et al., 2004] nyelvet célszerű választani, ami 2004. február óta hivatalos W3C ajánlás, és amelyet az RDF séma nyelv [Brickely & Guha, 2004] kibővítéseként dolgoztak ki. Egy OWL leírás nem más, mint jól-definiált jelentéssel bíró XML elemek és attribútumok halmaza, amelyek felhasználásával termeket, relációikat és azok kapcsolatait írhatjuk le. Az OWL nyelv fontos tulajdonsága, hogy nyílt világszemléletet alkalmaz és nem él azzal a feltevéssel, hogy a különböző szóalakok különböző fogalmakat, egyedeket jelölnek. Legfőbb hiányossága, hogy változókat nem lehet használni benne, emiatt kisebb a kifejező ereje, mint egy olyan nyelvnek, amely megenged elsőrendű logikai formulákat a definíciókban; továbbá a kettőnél nagyobb aritású relációk kifejezése kissé körülményes. Előnye viszont, hogy számos projektben alkalmazzák és jól alkalmazható szerkesztő eszközöket, ellenőrző programokat fejlesztettek ki hozzá. Ezeknek köszönhetően elég sok az ontológia-építéssel kapcsolatos tapasztalat.

Az OWL három résznyelvet foglal magába, amelyek kifejező erő szempontjából eltérnek egymásól. Az OWL Full a teljes OWL nyelv. Itt minden, az RDF által megengedett konstrukció használható (pl. egy osztály példánya lehet egy másik osztálynak), ami esetenként komoly problémákat vet fel a következtetésnél. Az OWL DL bizonyos megkötésekkel megszorított OWL nyelv, ami a leíró logikákon alapszik. Ez a nyelv biztosít egyidejűleg elég magas kifejező erőt, valamint eldönthető és hatékony következtető eljárást. Az OWL Lite bizonyos OWL konstrukciókat nem enged meg, kifejező ereje nem sokkal haladja meg az RDF sémáét.

A klasszikus grafikus tudásreprezentációs modelleket [Kremer, 1998] tárgyalja részletesen. Közülük a szemantikai adatmodellek [Kovács, 2004], a szemantikus hálók [Quillian, 1968], azon belül is [Klyne & Carroll, 2004] az RDF modell és a fogalmi gráfok (Conceptual Graph, CG) [Sowa, 1976], [Sowa, 1991], valamint a keret-alapú modell [Minsky, 1975] feladat-specifikus vizsgálata valósult meg. A szemantikai adatmodelleknél az egyedtípusok és az egyedelőfordulások éles elválasztása, eltérő kezelése, továbbá

a kapcsolatok nem egyértelmű ábrázolása kifogásolható. Ezek a modellek nem predikátum-központúak és a fogalomalkotás különböző szintjeit nem lehet velük modellezni. A keret-alapú modell előnye, hogy természetes módon tudja kezelni a megkötéseket, viszont a fogalmak közötti kapcsolatok ábrázolása itt sem egyértelmű. Ráadásul a logikán alapuló OWL szöveges leíráshoz a szemantikus háló grafikus reprezentáció megfelelőbb választás lenne. Azonban az RDF szemantikai gráfban nincs különbség az állítások predikátum és nem-predikátum fogalmainak ábrázolása között. A vizsgálat szempontjából a fogalmi gráfok legnagyobb hátránya az erős nyelvi kötődés. Igaz ugyan, hogy a CG modell predikátum-központú, de a predikátum nyelvi megfogalmazásától függően azonos szemantikai tartalmú állításokhoz eltérő fogalmi gráf ábrázolás tartozhat. A részletes elemzés és összehasonlítás a [3], [4], [5] publikációkban olvasható.

Újszerű megközelítés Ilieva univerzális grafikus jelölésrendszere, amely egységes keretben képes ábrázolni a természetes nyelvű állításokat és az azokban megfogalmazott szakterület-specifikus tudást [Ilieva, 2007]. Az ábrázolás előkészítő lépéseként a mondatokat mély szintaktikai elemzésnek vetik alá, majd a kinyert szintaktikai és szemantikai információkat táblázatos formában tárolják. A grafikus nyelv fő építőelemei a fogalmak (a mondat főnevei), amelyeket ellipszissel ábrázolnak, valamint a közöttük fennálló kapcsolatok (predikatív, prepozíciós, ok-okozati, feltételes stb.), amelyeket irányított, címkézett élek reprezentálnak. Az így felépülő gráf egy speciális szemantikus háló. A gyakorlatban a természetes nyelven megadott felhasználói követelmények UML-re (vagy más, a szoftverfejlesztésben alkalmazott diagrammra) történő automatikus átfordítására használják köztes nyelvként. A disszertációban tárgyalt nyelvtantanuló ágens tudásbázisának grafikus ábrázolásánál nem alkalmazható, mert a természetes nyelv szintaktikai elemzésére épül. További hátránya, hogy a predikátumot nem fogalomként kezeli, hanem kapcsolat-típusként.

1.1.3. Annotálási technikák a nyelvtantanulásban

A nyelvtan definíció szerint az a szabályrendszer, amely leírja, hogy hogyan jönnek létre a nagyobb nyelvi egységek az alacsonyabb szintű formális elemekből. A nyelvtantanulás tehát egyfajta szabálytanulás, az induktív gépi tanulás egy speciális esete. Egy nyelvtantanuló ágens a

környezetét képező adatokból képes megtanulni az adatok nyelvét előállító formális nyelvtant produkciós szabályok formájában [Bach, 2004]. A feladat nehézségét bizonyítja, hogy Gold [Gold, 1967] szerint a Chomsky-hierarchiában [Chomsky, 1956] szereplő nyelvtanok közül egyik sem tanulható pusztán pozitív minták alapján. Az egyik megközelítés a probléma megoldására a tanítóminta bővítése negatív példákkal, illetve szerkezeti információkkal. Ez utóbbi a mintaadatok annotálását jelenti (kézi vagy automatikus technikával), és az ilyen (címkézett) adatokból tanuló módszereket felügyelt tanulási módszereknek nevezzük, amelyekről [McEnery et al., 2005] nyújt áttekintést. Ezek a módszerek hatékonyabbak és pontosabb eredményt szolgáltatnak, mint a nem-felügyelt tanulási módszerek, amelyek annotáció nélküli adatokból tanulnak. Ennek ellenére a nem-felügyelt tanulási módszereket is intenzíven kutatják [Clark, 2001], [Roberts & Atwell, 2002], mert az annotált adatok előállítása idő- és erőforrásigényes, és ennek következtében hozzáférhetőségük korlátozott. A nem-felügyelt tanulási módszerek összefoglalása az [1], [2] publikációkban olvasható, egy konkrét megvalósítást pedig [11] dokumentál.

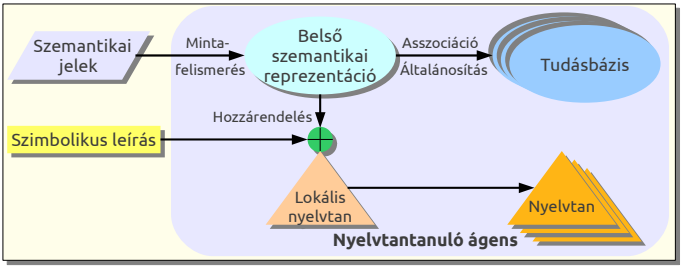
A gyakorlatban szintaktikai és szemantikai annotációs sémákat különböztetünk meg. A szintaktikai (nyelvtani) annotálás kétféleképpen valósulhat meg [Atwell et al., 2000]: vagy megadjuk minden szóhoz, hogy milyen mondatrész szerepét tölti be (*Part-Of-Speech tagging*); vagy minden szó esetén meghatározzuk a főigétől való függését (*dependency-based tagging*). A szemantikai kódolás megvalósítására a szakirodalom szintén kétféle módszert említ [Reeve & Han, 2005]. Egyrészt minden szóhoz hozzárendelhető a mondatban betöltött szemantikai szerepe, másrészt a szavakhoz megadhatjuk azt az útvonalat, amely leírja, hogy egy rögzített (rendszerint szakterület-specifikus) ontológiában hol helyezkedik el. Ez utóbbi, ontológia-alapú szemantikai annotálás csak néhány éve került a kutatók érdeklődésének középpontjába, a szemantikus web koncepciójának [Berners-Lee et al., 2001] megszületésével párhuzamosan. A kitűzött cél a weben elérhető szövegek és multimédiás adatok szó-alapú fogalmi annotációjának automatizálása.

Ontológiával annotált pozitív mintából megszorítás-alapú nyelvtant tanul Muresan rendszere [Muresan, 2006], ahol a szakterület-specifikus ontológia a szavakat és jelentésüket keret-alapú rendszerben tárolja.

A disszertációban tárgyalt megközelítésben a szemantikai annotálás ontológia-alapú, de állítás-szintű, azaz minden állításhoz külön ontológia (fogalmi háló) tartozik.

1.2. A kutatás célja

Az értekezés az ontológia egy újszerű alkalmazási lehetőségét tárgyalja. Az 1.1. ábrán vázolt nyelvtantanuló ágens tudásbázisának ábrázolására, valamint a tanítóminták mondatszintű szemantikai annotálására szolgál.



1.1. ábra: A nyelvtantanuló rendszer modellje

Az ágens az alábbi előre rögzített képességekkel rendelkezik:

- **mintafelismerés**, azaz az ágens képes érzékelni és felismerni a környezetében lévő objektumokat és azok viszonyát;
- **asszociáció**, azaz az ágens be tudja építeni az új információkat a tudásbázisába;
- **általánosítás**, azaz az ágens a megszerzett és eltárolt ismeretei alapján képes absztrakt – vagyis új, összetett – fogalmakat alkotni.

Ezen feladatok megvalósítása érdekében az ágens olyan szemantika alapú tudásábrázolási modellt igényel, amire az alábbiak jellemzők:

- fő építőelemei a fogalmak és a közöttük fennálló kapcsolatok,
- predikátum-központú, ahol a predikátum egy fogalomtípus,
- szűk, rögzített elemkészlettel rendelkezik,
- különbséget tesz az adott és a tanult (általánosított) fogalmak között,
- képes ábrázolni a fogalomalkotás többszintű folyamatát,
- rugalmas és bővíthető.

A vizsgált létező grafikus tudásábrázolási technikák egyike sem teljesíti maradéktalanul a fenti követelményeket. Ezért a disszertáció elsődleges feladata a deklarált követelményeket kielégítő új szemantikai

modell kidolgozása és kifejező erejének széleskörű vizsgálata. Második feladata egy megfelelő nyelvtani formalizmus kialakítása, amely egységes módon ábrázolja a szimbolikus nyelvi mondatokat és a hozzájuk tartozó szemantikai leírást (annotációt). Harmadik feladata a fogalomalkotás folyamatának modellezése a megalkotott új szemantikai modell segítségével. Végezetül implementálni kellett egy mintarendszert, amelyen bemutatható az elméleti eredmények gyakorlati alkalmazhatósága.

2. Új tudományos eredmények

2.1. Az ECG szemantikai modell

Kidolgoztam a kétszintű fogalomháló (Extended Conceptual Graph, ECG) szemantikai modellt [8], amely rendelkezik egy alkalmasan kiterjesztett magasabb-rendű predikátum logikai leírásmóddal (ECG-HOPL) és egy ezzel ekvivalens grafikus leírásmóddal (ECG Diagram). Igazoltam, hogy a modell teljesíti a vizsgált nyelvtantanuló ágens megvalósításához szükséges tudásábrázolási módszerrel szemben támasztott követelményeket, azaz a modell

- predikátum-központú;
- fő építőelemei a fogalmak, a közöttük fennálló kapcsolatok, és a modell strukturálását lehetővé tevő konténerelemek;
- a modell eszközkészlete rögzített: hét fogalomtípusból és négy kapcsolattípusból építkezik;
- két szintet különböztet meg: az objektum szinten történik a környezet objektumainak közvetlen statikus leképzése, míg az absztrakt szinten az objektum-szintű fogalmak és kapcsolatok általánosítása valósul meg;
- eltérően ábrázolja az objektum- és az absztrakt-szintű fogalmakat és kapcsolatokat;
- a modell modulárisan építkező rendszer, ezért végtelen sok állítás konstruálható a rögzített, szűk elemkészletből.

Mivel az ECG modell fő építőelemei a fogalmak és a közöttük lévő kapcsolatok, ezért ontológia leíró nyelvnek tekinthető. Ebből következik, hogy a modell grafikus eszközkészlete alkalmas ontológiák grafikus megjelenítésére. Ennek igazolására kidolgoztam egy $O(n^2)$ műveletigényű

algoritmust, amely elvégzi az ECG Diagram gráf előállítását OWL szöveges ontológia leírásból (ahol n a megjelenítendő OWL elemek száma).

Elvégeztem a modell természetes nyelvi kifejező erejének vizsgálatát [10]. Mivel a vizsgált nyelvtantanuló ágens nyelvi kifejezőképessége a megfigyeléseire korlátozódik, ezért csak olyan nyelvi jelenségek kerültek megvizsgálásra, amelyekkel igaz logikai értékű, egyértelműen értelmezhető, tényszerű kijelentéseket lehet megfogalmazni. A vizsgálat eredménye alapján kijelenthető, hogy a kompozíció-őrzés kritériumának figyelembe vételével minden ECG-HOPL állítás egyértelműen leképezhető egy vizsgált természetes nyelvi mondatra, ahol a leképzést szemantikai ekvivalencia-osztályokra értelmezzük. Szintén teljesül, hogy amennyiben a nyelv pragmatikai szintjét figyelmen kívül hagyjuk, minden vizsgált természetes nyelvi mondathoz konstruálható vele ekvivalens szemantikai tartalmú ECG-HOPL állítás. A vizsgált ágens tekintetében ez a leképzés is egyértelmű. Ezért az ECG modell alkalmazható mondatszintű szemantikai annotációs nyelvként.

Sikerült belátni, hogy az ECG-HOPL megadható környezetfüggetlen nyelvtannal (Context Free Grammar, CFG) [9]. Ezáltal igazolást nyert, hogy az ECG nyelv szintaktikája elég egyszerű, így készíthető hozzá hatékony tanuló algoritmus, és következésképpen az ECG-vel annotált természetes nyelvi mintákból történő nyelvtantanuláshoz is.

1. tézis:
Megalkottam a fogalomalkotás többszintű folyamatát tükröző, nyelvtantanulásra optimalizált ECG szemantikai modellt, amely alkalmas nyelvtantanuló ágensek tudásának ábrázolására, valamint az ilyen ágensek tanítómintáinak állítás-szintű szemantikai annotálására [8, 9, 10].

2.2. Az ECG beágyazása nyelvtan formalizmusba

Második feladat a szimbolikus nyelvi állítások és a szemantikájukat leíró ECG ontológiák (annotációk) összerendelési szabályainak kifejezésére alkalmas nyelvtani formalizmus megalkotása. Több évtizede vitatott kérdés, hogy a természetes nyelvek milyen nyelvtani formalizmussal írhatók le. Napjainkban az az elfogadott elmélet, hogy a természetes nyelvek valószínűleg olyan nyelvosztályba tartoznak, ami a környezetfüggetlen (*context-free*) és környezetfüggő (*context-sensitive*) nyelvosztályok

'között' helyezkedik el. A probléma megoldását a [6] publikáció függőségalapú algoritmussal oldja meg. Az értekezésben a szakirodalomban fellelhető számos javasolt formalizmus közül a TAG (Tree Adjoining Grammar) [Joshi & Schabes, 1997] faegyesítő nyelvtant vettem alapul, mert számítási időkomplexitását tekintve a gyakorlatban alkalmazható, polinomiális időben feldolgozható algoritmuson alapszik; a nyelvi jelenségek széles körét lefedi; és az ECG aciklikus gráfok átalakíthatók ilyen fastruktúrává. A TAG kiterjesztéseként kialakított ECG-TAG formalizmus definíciója:

$$ECG\text{-}TAG(\mathcal{G}) = \langle V, E, R^+, T(S), T(I), T(A) \rangle, \tag{2.1}$$

ahol V a csomópontok véges halmaza úgy hogy $V = C \cup \{\mathbf{S}\}$, ahol C az ECG fogalmak véges halmaza és $\mathbf{S}$ a start szimbólum. E az élek véges halmaza úgy hogy $E = RS \cup \bar{E}$, ahol RS az ECG kapcsolatok véges halmaza és $\bar{E}$ a predikátum fogalmakhoz tartozó élek véges halmaza. Az élek címkézettek, ahol az élcímkék véges halmaza $R^+ = R \cup \{predicate\}$, ahol R a szemantikai szerepek véges halmaza. $T(S)$ az egyelemű start-fa halmaz, $T(I)$ az alapfák (*initial trees*) véges halmaza, és $T(A)$ a bővítményfák (*auxiliary trees*) véges halmaza. A fák egyesítése a TAG formalizmusban alkalmazott behelyettesítés (*substitution*) és kiterjesztés (*adjunction*) műveletekkel valósul meg.

2. tézis:
A TAG kiterjesztéseként megalkottam az élcímkézett lexikális fákból felépülő ECG-TAG formalizmust, ahol az élcímkék szemantikai függőségi viszonyt fejeznek ki. Beláttam, hogy az ECG Diagram gráfok leképzése ECG-TAG formalizmusra veszteségmentes átalakítás, és ennek végrehajtására kidolgoztam egy $O(n^2)$ műveletigényű algoritmust, ahol n az ECG gráf elemeinek (csomópontjainak és éleinek) a száma [7].

Az ECG-TAG formalizmus az állítások szemantikai szintjét ábrázolja, nem foglalja magába a szimbolikus nyelvi szint megjelenítését. Ehhez az ECG-TAG formalizmust ki kellett bővíteni egy szintaktikai szinttel. Az ilymódon kiterjesztett formalizmus az S-ECG-TAG elnevezést kapta. Ezen a szinten valósul meg a szimbolikus nyelvi egységek (összefüggő szószerkezetek) hozzárendelése a szemantikai-szintű fogalmakhoz (csomópontokhoz). Ez a hozzárendelés egy nem kölcsönösen egyértelmű függvény, azaz minden szimbolikus nyelvi egységnek van pontosan egy megfelelője a szemantikai szinten, de nem minden fogalom jelenik

meg a szimbolikus szinten, illetve egy fogalomhoz több szimbolikus nyelvi egység is tartozhat (nem-összefüggő szószerkezetek). Az S-ECG-TAG formalizmus definíciója:

$$S\text{-}ECG\text{-}TAG(\mathcal{G}) = \langle V, E, R^{+n}, T(D) \rangle, \tag{2.2}$$

ahol V a csomópontok véges halmaza úgy hogy $V = C \cup \{\mathbf{S}\} \cup SN$, ahol C az ECG fogalmak véges halmaza, **S** a start szimbólum, és SN a szimbolikus-szintű csomópontok véges halmaza. E az élek véges halmaza úgy hogy $E = RS \cup \bar{E} \cup \tilde{E}$, ahol RS az ECG kapcsolatok véges halmaza, $\bar{E}$ a predikátum fogalmakhoz tartozó élek véges halmaza, és $\tilde{E}$ a szimbolikus-szintű csomópontokhoz tartozó élek véges halmaza. Az élek címkézettek, ahol az élcímkék véges halmaza $R^{+n} = R \cup \{predicate\} \cup \{n_1 \dots n_k\}$, ahol R a szemantikai szerepek véges halmaza, és a szimbolikus-szintű csomópontokat szemantikai-szintű csomópontokhoz kötő élek a szimbolikus nyelvi egységek sorrendiségét leíró megelőzési relációt fejeznek ki. $T(D)$ pedig az egyelemű leszármaztatási-fa (*derivation tree*) halmaz.

3. tézis:
Az ECG-TAG szimbolikus szinttel való kiterjesztésével megalkottam az S-ECG-TAG formalizmust, amely alkalmas a szimbolikus nyelvi állítások és a szemantikájukat leíró ECG ontológiák együttes ábrázolására, valamint a hozzárendelési szabályok tanulásának támogatására. A formalizmusban az összefüggő szószerkezetek sorrendje lokálisan van tárolva a szimbolikus-szintű csomópontokhoz tartozó élek címkéjében, és a nem-összefüggő szószerkezetek ábrázolása a szimbolikus szinten testvér-csomópontokkal valósul meg. Kidolgoztam a szimbolikus nyelvi egységek szemantikai-szintű csomópontokhoz történő hozzárendelésének statisztika-alapú tanuló algoritmusát, amelynek műveletigénye a tanítóminta halmaz rendelkezésre állását és kiválasztását követően a szimbolikus nyelvi mondat hosszának lineáris függvénye.

2.3. A fogalomalkotás folyamatának modellezése ECG gráfokon

A fogalomalkotás a gépi tanulás esetén az a folyamat, mely során az ágens a megfigyelései közötti szabályszerűségek feltárása révén megtanulja besorolni azokat általános kategóriákba (osztályokba). A folyamat számítógéppel történő kezelhetősége érdekében az absztrakció és

az általánosítás műveleteinek alkalmazása elengedhetetlenül szükséges. Peirce [Hartshorne et al., 1958] megközelítését alapulvéve, a disszertációban tárgyalt értelmezésben a fogalomalkotás során a vizsgált nyelvtantanuló ágens a tudásbázisába beépíti (*asszociáció*) és általánosítja (*általánosítás*) megfigyeléseit.

Miután az ágens megfigyeléseit ontológiák írják le és tudásbázisát ECG gráfokkal ábrázoljuk, az asszociáció az ECG gráfok illesztését (*graph matching*) foglalja magába. Az ECG gráfok illesztése pedig az elemek illesztését, összehasonlítását jelenti az elemek kategória-típusa alapján felépített fogalomháló felhasználásával.

Az értekezésben az ECG fogalmak általánosításán azt a folyamatot értjük, mely során ismert fogalmak közös elemeinek kiemelésével új, összetett fogalmak jönnek létre, melyek ábrázolásához az ECG modell külön elemeket definiál. Egy adott tématerület új (tanult) fogalmainak előállítását és fogalomhálóba szervezését pedig absztrakciónak nevezzük.

Az általánosítás algoritmusa az asszociáció műveletén belül valósul meg. Ennek során hasonló részgráfokat kell keresni, amelyek csak egy, kategória-típusuk alapján szemantikailag összehasonlítható csomópontban térnek el egymástól. Ehhez be kellett vezetni az ECG gráfok metszetének ($\cap$) és a metszet kiterjesztésének ($\cap^*$) műveletét. Az eltérő csomópontok helyett bevezetésre kerül egy új fogalom, ami az absztrakció során az elemek egyed-típusa alapján felépített fogalomhálóban az eltérő elemek legkisebb közös általánosítása. Ez alá összevonhatók a hasonló részgráfok közös elemei, az eltérő csomópontok pedig hozzákötheők specializációs kapcsolattal.

4. tézis:
A vizsgált tanuló ágens tudásbázisának felépülését leíró fogalomalkotási folyamat modellezésére kidolgoztam egy módszert, amely az asszociáció és az általánosítás algoritmusain alapszik. Ennek során az ágens megfigyeléseit kifejező ECG gráfok az asszociáció algoritmusa szerint inkrementálisan beillesztésre kerülnek egy kezdetben üres ECG gráf halmazba. Az eljárás alapja egy hibrid, kontextus-függő ECG gráf illesztési algoritmus. A beillesztés során az általánosítás algoritmusát alkalmazva a feltárt hasonló részgráfok eltérő csomópontjai helyett új, összetett fogalmak (csomópontok) kerülnek bevezetésre. A folyamat végén kialakuló ECG gráf írja le a vizsgált tanuló ágens megfigyeléseiből kinyert általánosított 'tudását' [13].

A 4. tézis következményei:

1. Az elsődleges-szintű ECG gráfokból valamint az asszociáció és általánosítás végrehajtási lépései után kialakuló összevont gráfokból háló építhető. Az ágens 'tudását' a háló legfelső eleme reprezentálja.

2. Úgyszintén háló építhető az elsődleges-szintű ECG gráfokból és a rajtuk értelmezett metszet műveletének rekurzív végrehajtása során kapott részgráfokból, ahol a háló elemei között $\subseteq$ reláció áll fenn. A háló alsó szintjén elhelyezkedő elemek az egyedi ECG gráfok, míg a felső szintjén lévő elemek a gyakori (általános) részgráfok.

3. Az elméleti eredmények alkalmazása

Az elméleti eredmények alkalmazhatóságának bemutatására elkészült egy Java-ban implementált mintarendszer [12], ami az alábbi funkciókat valósítja meg:

- grafikus felületet biztosít egy előre rögzített elemekből álló mikrovilág létrehozásához (a mikrovilág egyedei síkidomok, amelyeket alakjuk, méretük és színük jellemez),
- a mikrovilágra vonatkozó állításokhoz (amik a mikrovilág egyedei között értelmezhető geometriai és méretviszony relációkra vonatkoznak) megadható azok szimbolikus nyelvi megfogalmazása,
- a program OWL leírást generál minden állításhoz, amely tartalmazza a szituáció szemantikai és szintaktikai leírását,
- az OWL leírásból előállítja annak ECG modell szerinti logikai és grafikus megjelenítését.

Az ilymódon létrejövő ECG gráfokkal szemantikailag annotált mikrovilágra vonatkozó állítások alaphalmazán kerül modellezésre a fogalomalkotás (asszociáció és általánosítás) folyamata. Ehhez elő kellett állítani a mikrovilágra jellemző, az elemek egyed-típusa alapján felépülő fogalomhálót. Az ECG modellben az általánosítás több szinten értelmezhető:

- az első szinten feltárhatók a fogalomsémák a közös jellemzők alapján;

- a második szinten megtanulható az objektumok helyettesíthetősége a predikátumhoz kötődő szerepkörök alapján;
- a harmadik szinten feltárhatók a predikátumsémák.

Jelen kutatás keretein belül azonban csak az 5. tézisben megfogalmazott értelmezés és eljárás kerül bemutatásra, mert a létrehozott mikrovilágban az általánosítás csak az első szinten domináns. A rögzített elemkészlet a másik két szint szemléltetésére nem alkalmas.

4. További kutatási feladatok

A kidolgozott elméletet célszerű olyan példahalmazon is kipróbálni, ahol az általánosítás mindhárom szintje szimulálható. Úgyszintén fontos feladat az általánosítás inverzének, a specializáció műveletének a modellezése.

Mivel a kutatás távlati célja igazolni, hogy a nyelvtantanulás hatékonyan megvalósítható ontológiával annotált pozitív mintából. Ehhez implementálni kell az ábrán látható statisztikai módszereket alkalmazó nyelvtantanuló ágenst. Ezt követően kísérletezésre, összehasonlításra számtalan lehetőség nyílik a szimbolikus nyelv, valamint a formális nyelvtan megválasztásának függvényében.

A javasolt módszertan a gépi fordítás támogatására is alkalmas. Ehhez implementálni kell egy a nyelvtantanuló ágenssel kommunikáló mondatgeneráló ágenst, amely képes szimbolikus nyelvi leíró mondatot társítani egy ontológia modellhez. Amennyiben a célnyelv nyelvtana már rendelkezésre áll, a forrásnyelven megfogalmazott és ontológiával annotált állításokhoz a rendszer elő tudja állítani a célnyelvi leírást a nyelvtan felhasználásával abból kiidulva, hogy az azonos szemantikai tartalmú (különböző szimbolikus nyelvű) állítások ontológia ábrázolása megegyezik.

A mondatgeneráló ágenssel kibővített nyelvtantanuló rendszer az ECG szemantikai modell természetes nyelvű interfészének tekinthető. Amennyiben kiegészül további konvertáló modulokkal, tetszőleges szemantikai modell természetes nyelvű interfészeként alkalmazható.

Érdekes és fontos felhasználási terület lehet, ha a kibővített rendszert képfelismerő ágenshez illesztjük, hogy annak természetes nyelvű interfészeként szolgáljon.

5. Summary

Ontology-based semantic annotation and knowledge representation in a grammar induction system

The main motivation for the research is to develop a new general rule learning methodology that alloys statistics with semantics. With that, our aim is to improve the performance of statistical grammar induction by utilizing semantic information in the learning process. The dissertation covers the first phase in the development of this system, that is the specification and deep examination of an appropriate semantic representation optimized for grammar induction. A learning agent needs abstraction and generalization to make learning feasible and tractable in complex domains. Therefore the process of conceptualization (involving the operations of association and generalization) should also be modeled within the grammar induction system examined by means of the semantic model developed. The new scientific results can be summarized as follows.

Thesis 1: [8], [9], [10]
A novel semantic model is developed, called ECG, which has a logic-based ECG-HOPL and a semantically equivalent graphical ECG diagram representation. The model satisfies the requirements of the knowledge representation format in the investigated grammar induction system, and can be used as an ontology modeling language because its main building blocks are concepts and their relationships. It is predicate-centered and it defines two levels and distinct elements for describing the different phases of conceptualization. It provides high levels of functionality, flexibility and extendibility. It is computationally tractable while highly expressive, that is it covers a wide range of linguistic phenomena.

Consequences of Thesis 1:

1. Since ECG can be considered as an ontology modeling language, ECG diagram can be used for visual ontology representation. The generation of ECG diagram graphs can be accomplished by an $O(n^2)$ algorithm, where n is the number of OWL elements to be displayed.

2. ECG can also be applied as a sentence-level semantic annotation language, because every ECG-HOPL statement can be semantically unambiguously rendered into an NL sentence examined and

every NL sentence under examination can be approximated by an ECG-HOPL statement.

3. ECG-HOPL can be defined with CFG, which proves that the syntax of ECG is simple enough so that a computationally effective learning algorithm can be constructed for inducing a set of grammar rules from ECG, and consequently from the sentences annotated by ECG.

Thesis 2: [7]
ECG fragment diagrams are acyclic graphs, therefore they can be converted to a tree structure the root of which is the kernel predicate. The mapping is proved to be lossless and is accomplished by an $O(n^2)$ algorithm, where n is the number of ECG diagram elements. The new ECG-TAG grammar formalism consists of edge-labeled lexicalized tree structures, the nodes of which correspond to ECG concepts, while the edges represent ECG relationships. The formalism is TAG-based, because it uses the same tree set (with different interpretation) and the same operations for tree construction as the original TAG formalism. At the same time, it is also dependency-based in the sense that edge labels represent semantic dependency relations.

Thesis 3:
The next task is to represent the semantic models and their symbolic language descriptions in a common framework. The algorithm that performs the assignment of symbolic sentence units to ECG concepts results in a new grammar formalism, called S-ECG-TAG, which combines the levels of semantics and syntax. The formalism extends the ECG-TAG formalism with a symbolic level, where the nodes include word sequences, while the edges are labeled by precedence relations representing the order of word sequences in the corresponding symbolic sentence. Hence, the symbolic level encodes word order locally and discontinuous constructions are represented by sibling nodes.

Consequences of Thesis 3:

1. The S-ECG-TAG formalism can be applied as a common framework for representing ECG diagrams and the corresponding symbolic sentences.

2. The S-ECG-TAG formalism can be applied as a formal grammar to be learnt in the grammar induction process.

Thesis 4: [13]
A method is developed for the execution of the conceptualization process within the learning agent examined, which involves the operations of association and generalization. According to the association algorithm, primary-level ECG diagram graphs are matched to and incorporated in an initially empty knowledge base, which is itself another (accumulated) ECG diagram graph. The matching of ECG diagram graphs is based on a hybrid context-dependent ECG diagram graph matching algorithm, and is traced back to the matching of element instances, for the examination of which an element category type lattice is defined.
The generalization algorithm is implemented as part of the association process and proceeds by introducing new (not observed) higher-level concepts into the knowledge base. First, the algorithm searches for maximal similar subgraphs which differ in only one ECG diagram graph node. For their exploration the intersection operation of two ECG diagram graphs and its extension are defined. If the differing nodes are semantically comparable on the basis of the element category type lattice, a new concept is inserted from the element instance type lattice determined as the least common generalization of the differing concepts. Finally, the relationships are updated in the knowledge base.

Consequences of Thesis 4:

1. The two operations of association and generalization together accomplish the process of conceptualization. At the end of the process, the generalized knowledge of the agent can be obtained as the top element of the lattice constructed from the set of primary-level ECG diagram graphs and the set of accumulated ECG diagram graphs resulting from the association and generalization steps executed.

2. Recursively performing the operation of graph intersection on the set of ECG diagram graphs and on the resulting sets of common subgraphs, a lattice can be built. The lower-level nodes of the lattice include individual (infrequent specialized) ECG diagram graphs, while at the top levels of the lattice frequent general subgraphs are located.

Saját publikációk az értekezés témakörében

[1] Varga, E. & Kovács, L. (2005). Review of Unsupervised Grammar Induction Systems. In: 5th International Conference of PhD Students, Miskolc, Hungary, pp. 201–206.

[2] Varga, E. & Kovács, L. (2005). Quality Measures of Language Learning Systems. In: 5th International Conference of PhD Students, Miskolc, Hungary, pp. 207–212.

[3] Baksa-Varga, E. & Kovács, L. (2008). A Semantic Model for Knowledge Base Representation in a Grammar Induction System. In: 1st Workshop on Computational Intelligence in Measurement, Control and Instrumentation (CIMCI 2008), Timisoara, Romania, 3, pp. 27–32.

[4] Kovács, L. & Baksa-Varga, E. (2008). Logical Representation and Assessment of Semantic Models for Knowledge Base Representation in a Grammar Induction System. In: 7th International Conference on Renewable Sources and Environmental Electrotechnologies (RSEE 2008), Oradea, Romania, pp. 48–53.

[5] Kovács, L. & Baksa-Varga, E. (2008). Logical Representation and Assessment of Semantic Models for Knowledge Base Representation in a Grammar Induction System. Journal of Computer Science and Control Systems, University of Oradea, Romania, pp. 48–53.

[6] Kovács, L. & Baksa-Varga, E. (2008). Dependency-Based Mapping between Symbolic Language and Extended Conceptual Graph. In: 6th International Symposium on Intelligent Systems and Informatics (SISY 2008), Subotica, Serbia, pn. 13.

[7] Baksáné Varga, E. & Kovács, L. (2008). Ontológia-alapú nyelvtantanuló rendszer nyelvtan-modellje. A Dunaújvárosi Főiskola Közleményei, A Magyar Tudomány Hete 2008 konferenciasorozat, Informatikai konferencia (DFTH 2008), XXX/1, pp. 219–226.

[8] Baksa-Varga, E. & Kovács, L. (2008). Knowledge Base Representation in a Grammar Induction System with Extended Conceptual Graph. Transactions on Automatic Control and Computer Science, Scientific Bulletin of "Politehnica" University of Timisoara, Romania, 53(67), pp. 107–114.

[9] Baksáné Varga, E. (2009). Magasabb rendű logika a természetes nyelvek szemantikájának reprezentálásánál. A Gépipari Tudományos Egyesület Műszaki Folyóirata (GÉP), LX. évfolyam, 2009/6, pp. 49–55.

[10] Baksa-Varga, E. & Kovács, L. (2009). Semantic Representation of Natural Language with Extended Conceptual Graph. Journal of Production Systems and Information Engineering, Vol. 5, pp. 19–39.

[11] Kovács, L. & Baksa-Varga, E. (2010). Induction of Probabilistic Context-Free Grammar Using Frequent Sequences. Journal of Advanced Computational Technologies, *in press*.

[12] Baksáné Varga, E. (2010). Ontológia-alapú szemantikai annotálást végző ágens dokumentációja. Projektjelentés. ME Általános Informatikai Tanszék, Tanszéki Közlemények.
http://www.iit.uni-miskolc.hu/iitweb/opencms/research/TechReports/.

[13] Baksa-Varga, E. & Kovács, L. (2011). Generalization and Specialization Using Extended Conceptual Graphs. In: 11th International Scientific Conference on Informatics (INFORMATICS'2011), Rožňava, Slovakia, *in press*.

Hivatkozások

[Atwell et al., 2000] Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C., & Wilcock, S. (2000). A comparative evaluation of modern English corpus grammatical annotation schemes. ICAME Journal, 24, pp. 7–23.

[Baader et al., 2003] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., & Patel-Schneider, P. (2003). The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press.

[Bach, 2004] Bach, I. (2004). Formális nyelvek. Budapest: Neumann Kht.

[Bechhofer, 2002] Bechhofer, S. (2002). Ontology Language Standardization Efforts. Technical Report IST Project IST-2000-29243, Information Management Group, Department of Computer Science, University of Manchester, UK.

[Bechhofer et al., 2004] Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P., & Stein, L. (2004). OWL Web Ontology Language Reference, W3C Recommendation.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American.

[Bognár, 2000] Bognár, K. (2000). Leíró logikák az ismeretábrázolásban. Alkalmazott Matematikai Lapok, 20(2), pp. 183–193.

[Brickely & Guha, 2004] Brickely, D. & Guha, R. (2004). Resource Description Framework (RDF) Schema Specification. W3C Recommendation.

[Calí et al., 2005] Calí, A., Calvanese, D., Grau, B. C., Giacomo, G. D., Lembo, D., Lenzerini, M., Lutz, C., Milano, D., Möller, R., Poggi, A., & Sattler, U. (2005). State of the art survey. Technical Report WP1 – Assessment of Fundamental Ontology Based Tasks, FP6-7603 Thinking ONtologiES (TONES) project.

[Charniak, 1996] Charniak, E. (1996). Statistical Language Learning. Cambridge, MA: MIT Press.

[Chomsky, 1956] Chomsky, A. (1956). Three models for the description of language. IRE Transactions on Information Theory, 2(2), pp. 113–123.

[Clark, 2001] Clark, A. (2001). Unsupervised Language Acquisition: Theory and Practice. PhD thesis, COGS, University of Sussex.

[Cranefield & Purvis, 1999] Cranefield, S. & Purvis, M. (1999). UML as an ontology modeling language. In In Proceedings of the Workshop on Intelligent Information Integration, 16th International Joint Conference on Artificial Intelligence (IJCAI-99): pp. 46–53.

[Futó, 1999] Futó, I., Ed. (1999). Mesterséges Intelligencia. Aula Kiadó.

[Gold, 1967] Gold, E. (1967). Language identification in the limit. Information Control, 10, pp. 447–474.

[Gruber, 1993] Gruber, T. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), pp. 199–220.

[Hartshorne et al., 1958] Hartshorne, C., Weiss, P., & Burks, A., Eds. (1931–1958). Collected Papers of C. S. Peirce. Cambridge, MA: Harvard University Press.

[Ilieva, 2007] Ilieva, M. (2007). Graphical notation for natural language and knowledge representation. In 19th SEKE.

[Jarrar et al., 2003] Jarrar, M., Demey, J., & Meersman, R. (2003). On using conceptual data modeling for ontology engineering. Journal on Data Semantics, pp. 185–207.

[Joshi & Schabes, 1997] Joshi, A. & Schabes, Y. (1997). Handbook of Formal Languages, chapter Tree-Adjoining Grammars, pp. 69–123. Springer: Berlin.

[Jurafsky & Martin, 2000] Jurafsky, D. & Martin, J. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall.

[Klyne & Carroll, 2004] Klyne, G. & Carroll, J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation.

[Kovács, 2004] Kovács, L. (2004). Adatbázisok tervezésének és kezelésének módszertana. Budapest: ComputerBooks.

[Kovács & Sieber, 2009] Kovács, L. & Sieber, T. (2009). Multi-layered semantic data models. In Encyclopedia of Artificial Intelligence pp. 1130–1135. Hersey (USA): IGI Global Publisher.

[Kremer, 1998] Kremer, R. (1998). Visual languages for knowledge representation. In 11th Workshop on Knowledge Acquisition, Modeling and Management (KAW'98) Banff, Alberta, Canada.

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

[McEnery et al., 2005] McEnery, A., Xiao, R., & Tono, Y. (2005). Corpus-Based Language Studies: An Advanced Resource Book. Routledge Applied Linguistics. Routledge.

[Minsky, 1975] Minsky, M. (1975). A Framework for Representing Knowledge. In P. Winston (Ed.), The Psychology of Computer Vision. New York: McGraw-Hill.

[Muresan, 2006] Muresan, S. (2006). Learning Constraint-based Grammars from Representative Examples: Theory and Applications. PhD thesis, Columbia University, NY.

[Ogden & Richards, 1923] Ogden, C. & Richards, I. (1923). The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism. London: Routledge & Kegan Paul.

[Quillian, 1968] Quillian, M. (1968). Semantic Information Processing, chapter Semantic Memory, pp. 216–270. MIT Press: Cambridge, MA.

[Reeve & Han, 2005] Reeve, L. & Han, H. (2005). Survey of semantic annotation platforms. In 2005 ACM Symposium on Applied Computing Santa Fe, New Mexico: pp. 1634–1638.

[Roberts & Atwell, 2002] Roberts, A. & Atwell, E. (2002). Unsupervised Grammar Inference Systems for Natural Language. Technical Report 2002.20, University of Leeds, School of Computing.

[Sántáné-Tóth, 2006] Sántáné-Tóth, E. (2006). Ontológia – Oktatási segédlet.

[Scriptum, 2005] Scriptum (2005). Ontológia-építő nyelvek értékelése, elemző összehasonlítása. Technical Report MEO projekt, Scriptum Rt.

[Sowa, 1976] Sowa, J. (1976). Conceptual graphs for a database interface. IBM Journal of Research and Development, 20(4), pp. 336–357.

[Sowa, 1991] Sowa, J., Ed. (1991). Principles of Semantic Networks: Explorations in the Representation of Knowledge. San Mateo, CA: Morgan Kaufmann Publishers.

[Sowa, 2000] Sowa, J. (2000). Ontology, Metadata, and Semiotics. In Conceptual Structures: Logical, Linguistic, and Computational Issues, number 1867 in Lecture Notes in AI pp. 55–81. Berlin: Springer-Verlag.

[Szeredi et al., 2005] Szeredi, P., Lukácsy, G., & Benkő, T. (2005). A szemantikus világháló elmélete és gyakorlata. Budapest: Typotex.

[Wang & Chan, 2001] Wang, X. & Chan, C. (2001). Ontology modeling using UML. In 7th International Conference on Object Oriented Information Systems Conference (OOIS'2001: pp. 59–68.

[Xueming, 2007] Xueming, L. (2007). Using UML For Conceptual Modeling: Towards An Ontological Core. PhD thesis, Memorial University of Newfoundland.