# Bizonytalanság azonosítása természetes nyelvű szövegekben

A DOKTORI ÉRTEKEZÉS TÉZISEI

Vincze Veronika

MTA-SZTE Mesterséges Intelligencia Kutatócsoport

és

Szegedi Tudományegyetem

2014. július

Témavezető: Prof. Dr. Csirik János



Szegedi Tudományegyetem

Informatika Doktori Iskola

# 1. Bevezetés

A nyelvi bizonytalanság azonosítása a nyelvfeldolgozás számos területén bír jelentőséggel. A legáltalánosabb értelemben véve a bizonytalanságot információhiányként határozhatjuk meg: az információ befogadója (a hallgató vagy az olvasó) nem lehet biztos valamely információban. A bizonytalan propozíciók tehát azok, melyek igazságértékét vagy megbízhatóságát információ hiányában nem lehetséges eldönteni. A tényszerű (azaz igaz vagy hamis) és a bizonytalan információk megkülönböztetése igen lényeges mind a nyelvészet, mind a természetesnyelv-feldolgozás szempontjából. Vegyük például az információkinyerést, ahol is az alkalmazások célja a tényszerű információ kinyerése a szövegekből, a módosított szövegrészeket tehát elkülönítve kell kezelni. Egy jellemző példa lehet a fehérje-fehérje interakciók kinyerése biológiai szövegekből, ahol a különféle biológiai entitások közti kapcsolatokat kell összegyűjteni. Noha a felhasználó számára a bizonytalan kapcsolatok is érdekesek lehetnek, ezeket mindenképpen érdemes elkülöníteni a tényszerű szövegrészekből kinyert megbízható információktól. A gépi fordításban szintén lényeges a bizonytalanság azonosítása, hiszen a forrás- és a célnyelv időnként eltérő eszközökkel fejezi ki a bizonytalanságot (az angol nyelv például modális segédigéket alkalmaz ott, ahol a magyarban morfémákat találunk). Végül az orvosi dokumentumok osztályozásában a dokumentumok csoportosításának egy lehetséges módja, ha aszerint soroljuk osztályokba őket, hogy a páciens egy adott betegségben szenved-e, nem szenved vagy valószínűleg szenved.

A bizonytalanság vizsgálatára számos tanulmány irányult a közelmúltban, ezek azonban eltérő terminusokat alkalmaznak a némileg eltérő nyelvi jelenségek megnevezésére. A modalitás fogalmát sokszor kapcsolják a bizonytalansághoz (Palmer, 1986), de a tényszerűség (factuality) (Saurí és Pustejovsky, 2012), igazolhatóság (veridicality) (de Marneffe et al., 2012), evidencialitás (evidentiality) (Aikhenvald, 2004) és elköteleződés (commitment) (Diab et al., 2009) terminusok is használatosak. Továbbá, léteznek olyan számítógépes nyelvészeti alkalmazások, melyek adott doménhez tartozó szövegekben azonosítják a bizonytalan szövegrészeket (például biológiai szövegek (Velldal et al., 2012) vagy hírek (Saurí és Pustejovsky, 2012)), azonban a terminológiabeli különbségek miatt ezen megközelítések, illetve eredményességük összevetése is problematikus. A legtöbb megközelítést an-

notált adatbázisokon értékelik ki, mivel az utóbbi években számos bizonytalanságra annotált korpusz született, l. BioScope (Vincze et al., 2008), Genia (Kim et al., 2008), FactBank (Saurí és Pustejovsky, 2009), a CoNLL-2010 verseny korpuszai (Farkas et al., 2010) stb. Egységes annotációs elvek hiányában azonban nem tudjuk összehasonlítani ezeket a korpuszokat egymással, így az ezekre épülő bizonytalanságazonosító rendszerek mindegyike csak az adott korpuszon vagy doménen működik jól. Ez azt is jelenti, hogy a meglévő rendszerek nem (vagy csak nehezen) alkalmazhatók más doménre, azonban minden egyes doménre külön erőforrást és rendszert létrehozni igen költséges és hosszadalmas. Mindezt megoldhatja egy egységesített átfogó megközelítés, amelyet könnyen lehet alkalmazni az adott domén speciális igényeire szabva, további költségek nélkül, mindemellett a modell nyelvfüggetlensége is fontos szempont.

Az értekezésben angol és magyar nyelvű szövegekben azonosítjuk a nyelvi bizonytalanságot. A kutatás tárgya kettős szempontból is vizsgálható, hiszen a számítógépes nyelvészet területébe tartozik, így mind nyelvészeti, mind informatikai vonatkozásai is vannak. A kutatásban elsődlegesen a kérdés informatikai vonatkozásait helyezzük előtérbe, mindemellett nyelvészeti szempontokat is figyelembe veszünk. A korábbi tanulmányokkal ellentétben, melyek pusztán egyes doménekre, illetve elsődlegesen az angol nyelvre koncentráltak, jelen értekezésben egy átfogó, nyelv- és doménfüggetlen megközelítést nyújtunk a bizonytalanság azonosítására, mely könnyen alkalmazható több doménre és nyelvre is. A kutatás első lépéseként felvázoljuk a bizonytalanság egy nyelvi modelljét elméleti és számítógépes nyelvészeti háttérre támaszkodva, melyet a későbbiekben a bizonytalanság automatikus azonosításában alkalmazunk több doménen és nyelven, felügyelt tanulási módszereket felhasználva.

# 2. Célkitűzés

Az értekezés fő céljai az alábbiakban foglalhatók össze, elsőként az informatikai, majd a nyelvészeti célokat felsorolva. Először, azonosítjuk a szemantikai bizonytalanságot angol és magyar nyelvű szövegekben, megvizsgáljuk a bizonytalanságot jelző kulcsszavak eloszlását különböző doménekben, így biológiai szövegekben, Wikipédia-szócikkekben és

hírekben, továbbá megmutatjuk, hogy a szemantikai bizonytalanság doméneken átívelő azonosításában hogyan alkalmazhatók a különféle doménadaptációs eljárások. Másodszor, angol és magyar nyelvű szövegekben is azonosítjuk a diskurzusszintű bizonytalanságot, tesztelve ezáltal a bizonytalansági osztályok nyelvfüggetlenségét. Harmadszor, megmutatjuk, hogy a bizonytalanság azonosítása egy valós életbeli alkalmazásban, nevezetesen orvosi információkinyerésben is hasznosítható, ahol a fő feladat a dokumentumok osztályba sorolása aszerint, hogy a páciens egy adott betegségben szenved, nem szenved vagy feltehetőleg szenved. Negyedszer, mindehhez egy egységes keretet kívánunk adni, melyben a nyelvi bizonytalanság minden fajtáját el lehet helyezni, két fő csoportba sorolva: szemantikai bizonytalanság és diskurzusszintű bizonytalanság. Az osztályozás alapjául nyelvészeti és informatikai szempontok is szolgálnak. Ötödször, bemutatjuk korpuszainkat, melyekben kézzel annotáltuk a bizonytalanságot jelző nyelvi elemeket a fenti egységes elméleti keret alapján. Hatodszor, összehasonlítjuk a rendelkezésre álló korpuszokat és annotációs sémákat, külön hangsúlyt fektetve a hatókör alapú és az eseményalapú annotációkra, és amellett érvelünk, hogy ezek a különbségek a korpuszok gyakorlati alkalmazhatóságára is nagy hatással bírnak.

# 3. Az értekezés felépítése

Az alábbiakban összegezzük az értekezés legfontosabb eredményeit, az értekezés szerkezetét követve. Az értekezés első részében bemutattuk a bizonytalanság azonosításának alapjait és a legfontosabb gépi tanuló eljárásokat. Az értekezés második részében a természetes nyelvekben előforduló bizonytalansággal kapcsolatos nyelvi jelenségeket és az általunk létrehozott, bizonytalanságra annotált korpuszokat tekintettük át. Az értekezés harmadik részében végül megmutattuk, hogy a nyelvi bizonytalanság egyes típusai hogyan azonosíthatók automatikus módszerekkel különféle természetes nyelvű szövegekben.

## 3.1. I. rész: Háttér

Az értekezés első részében bemutattuk a bizonytalanság azonosításának általános alapjait és az értekezés módszertana szempontjából legfontosabb gépi tanuló eljárásokat, melyeket később a bizonytalanság automatikus azonosításában alkalmazunk.

## 3.2. II. rész: Bizonytalanság a nyelvben

Az értekezés második részében a nyelvben előforduló bizonytalansági jelenségeket mutattuk be. Pusztán szemantikai vizsgálatok nem elégségesek a nyelvi bizonytalanság vizsgálatához, hiszen bizonyos esetekben szintaktikai eszközök is fejezhetnek ki bizonytalanságot, például az angolban a passzív mondatok többsége nem jelöli meg a cselekvőt, ezzel az ún. weasel (tkp. forrás nélküli mondatok) jelenséget illusztrálják (Ganter és Strube, 2009). Más esetekben a bizonytalanságot jelző nyelvi elemek a mondat (közlés) kontextusában – pragmatikai vagy nyelven kívüli tényezőknek köszönhetően – válnak többértelművé, például ha egy tudományos cikkben egy állítást nem támaszt alá semmilyen hivatkozás vagy bizonyíték, vagy ha a beszélő köztudottan nem szokott igazat mondani, a propozíció igazságtartalmát nem lehet megítélni. Így tehát figyelembe kell venni a bizonytalanság pragmatikai aspektusait is, illetve a diskurzus előrehaladó természetét is (vö. de Marneffe et al. (2012), Saurí és Pustejovsky (2012)). A fentiekből következik, hogy a bizonytalanság mint nyelvi jelenség leírásához mind szemantikai, mind szintaktikai, mind pedig pragmatikai szempontokat figyelembe kell venni. Az értekezésben felvázoltunk egy interdiszciplináris egységesített keretet a bizonytalanság összes típusának osztályozására, mely szintaktikai, szemantikai, pragmatikai és számítógépes nyelvészeti szempontokra egyaránt épít, és e keretben bemutattuk a szemantikai és diskurzusszintű bizonytalanság több fajtáját.

Annak meghatározásához, hogy egy adott propozíció bizonytalan-e vagy sem, szükség lehet egy véges szótárra, mely tartalmazza a bizonytalanságot jelölő nyelvi elemeket (kulcsszavakat). Ugyanakkor a bizonytalanság automatikus azonosítását célzó felügyelt tanulási módszerek alkalmazásához szükségesek kézzel annotált korpuszok is. Az értekezésben bemutattuk a fenti egységes keretre épülő annotált korpuszainkat mind angol, mind magyar nyelvre, továbbá ismertettük a kulcsszavak eloszlásának nyelvekre, korpuszokra,

doménekre és műfajokra jellemző sajátságait.

A bizonytalanság fogalmának értelmezése mellett a bizonytalanként jelölt nyelvi egység is különbözhet korpuszról korpuszra. Továbbá egyes korpuszok a bizonytalanság különböző szintjeit is elkülönítik egymástól, azaz a nagyobb vagy kisebb valószínűséggel rendelkező propozíciók eltérően vannak jelölve. Az értekezésben összehasonlítottuk a Genia Event és a BioScope 1.0 korpuszokban alkalmazott eseményalapú és hatókör alapú annotációkat, és megmutattuk azt is, hogy a szemantikai és a diskurzusszintű bizonytalanság szintjén egyaránt megjelenik a bizonytalanság fokozatosságának kérdése.

## 3.3. III. rész: A bizonytalanság azonosítása

Az értekezés harmadik részében természetes nyelvi szövegekben azonosítottuk a bizonytalan szövegrészeket. A bizonytalanságot jelölő kulcsszójelöltek nem utalnak minden egyes előfordulásukban bizonytalanságra. Például a *valószínű* szó főleg valószínűségszámítási értelemben fordul elő matematikai szövegekben, viszont bölcsészettudományi írásokban ez az értelmezése kevésbé gyakori. Hasonló példa a *megvizsgál*, mely utalhat orvosi értelemben végzett vizsgálatra (ilyenkor nem bizonytalan a jelentése), illetve különféle összefüggések feltárására irányuló kísérletre (ilyenkor bizonytalan a jelentése), vö.

(1) Az orvos alaposan **megvizsgálta** a beteget.

(2) **Megvizsgáltuk** az esszenciális aminosavak szerepét a testépítésben.

A kulcsszójelöltek bizonytalan és nem bizonytalan használatát elkülönítendő kifejlesztettünk egy gépi tanuló algoritmust: ennek során figyelmet fordítottunk a doménspecifikus különbségekre is az angol nyelvű szemantikai bizonytalanság azonosítása esetében.

Az értekezésben bemutattuk magyar nyelvű szövegekben történő bizonytalanságazonosításra kifejlesztett gépi tanuló megoldásainkat is.

A bizonytalanság azonosítása valós életbeli alkalmazásokban is fontos szereppel bír. Ennek igazolására ismertettük a klinikai szövegekből történő információkinyerés példáját, mely során a beteg elhízottsági státuszát és egyéb kapcsolódó betegségeket kellett megállapítani.

# 4. Az értekezés eredményei

Az értekezésben elért főbb eredmények az alábbiakban foglalhatók össze, az informatikai szempontól lényeges eredmények kiemelésével. Felsoroljuk továbbá a kapcsolódó publikációkat is, hangsúlyozva az értekezés szerzőjének főbb hozzájárulásait az eredményekhez, melyeket a társszerzők jóváhagytak.

## 4.1. A szemantikai bizonytalanság azonosítása

Kísérleteket végeztünk a szemantikai bizonytalanság azonosítására magyar és angol nyelvű szövegekben, tudomásunk szerint a magyar nyelvre ezek az első publikált eredmények a témában. Rendszerünk a szemantikai bizonytalanság négy osztályának (episztemikus, doxasztikus, vizsgálat és feltételes) azonosítására képes. Eredményeink azt igazolják, hogy már egyszerű jellemzők használatával is jó eredményeket lehetséges elérni a szemantikai bizonytalanság azonosításában mind angol, mind magyar nyelvre. Doménadaptációs technikák használatával szintén jó eredményeket kaptunk angol nyelvre a doméneken és műfajokon átívelő bizonytalanságazonosításban. Jellemzőkészletünket pedig a magyar nyelv esetében szemantikai és pragmatikai jellemzőkkel is bővítettük (**1. tézispont**).

A főbb eredmények:

- a szemantikai bizonytalanság négy osztályának (episztemikus, doxasztikus, vizsgálat és feltételes) elkülönítésére képes rendszer;
- az első eredmények magyar nyelvű szemantikai bizonytalanság azonosításában;
- annak igazolása, hogy már egyszerű jellemzők használatával is jó eredményeket lehet elérni a szemantikai bizonytalanság azonosításában angol és magyar nyelven is;
- új (szemantikai és pragmatikai) jellemzők bevezetése a szemantikai bizonytalanság azonosítását célzó magyar nyelvű gépi tanuló rendszerbe;
- doménadaptáció segítségével jó eredményeket értünk el angol nyelven a doméneken és műfajokon átívelő bizonytalanságazonosításban doménen kívül adatok felhasználásával, minimalizálva így az annotációs költségeket;

- annak igazolása, hogy a domének sajátságai jelentősen befolyásolják a gépi tanuló rendszerek hatékonyságát a magyar nyelvű szemantikai bizonytalanság azonosításában.

Szarvas et al. (2012) ismerteti a doménadaptáción alapuló, szemantikai bizonytalanság azonosítására kifejlesztett rendszert. A szerző részt vett az adatok előkészítésében és a korpusz annotálásában, meghatározta az annotálandó kategóriákat, a gépi tanuló algoritmusba beépített jellemzők közül néhányat ő adott meg, ő hasonlította össze a domének és műfajok sajátos jellemzőit a bizonytalanságazonosításra való tekintettel, továbbá ő végezte el a kísérletek hibaelemzését. A társszerzők implementálták a gépi tanuló algoritmust, és végezték el az angol nyelvű méréseket, azonban a kísérleti eredmények közös és oszthatatlan hozzájárulásnak tekintendők minden szerző részéről. Vincze (2014) bemutatja a magyar nyelvű szemantikai bizonytalanság azonosítására kifejlesztett gépi tanuló módszert, mely szemantikai és pragmatikai jellemzőket is tartalmazó gazdag jellemzőtéren alapul. A magyar nyelvű kísérletek kizárólag a szerző nevéhez köthetők.

## 4.2. A diskurzusszintű bizonytalanság azonosítása

A diskurzusszintű bizonytalanság három típusának (weasel, hedge és peacock) automatikus azonosítására szintén kidolgoztunk egy rendszert. Alapmegoldást nyújtottunk az angol Wikipedia-szövegekben rejlő diskurzusszintű bizonytalanság automatikus azonosítására, illetve a magyar nyelvű szövegek esetében szekvenciajelölésre épülő, gazdag jellemzőtérrel rendelkező felügyelt tanulási módszert alkalmaztunk. Megoldásaink jó eredményt nyújtottak mindkét nyelv esetében (**2. tézispont**).

A főbb eredmények:

- az első eredmények magyar nyelvű diskurzusszintű bizonytalanság azonosításában;
- alapmegoldás az angol nyelvű Wikipedia-szövegekben található diskurzusszintű bizonytalanság azonosítására;
- az első gépi tanuló eredmények magyar nyelvű diskurzusszintű bizonytalanság azonosításában;

- új (szemantikai és pragmatikai) jellemzők bevezetése a diskurzusszintű bizonytalanság azonosítását célzó magyar nyelvű gépi tanuló rendszerbe;

- annak igazolása, hogy a domének sajátságai jelentősen befolyásolják a gépi tanuló rendszerek hatékonyságát a magyar nyelvű diskurzusszintű bizonytalanság azonosításában.

Vincze (2013) az angol nyelvű diskurzusszintű bizonytalanság azonosítására mutat be néhány alapmegoldást, majd összeveti az eredményeket a korábbi tanulmányokban leírtakkal. Vincze (2014) egy gépi tanuló módszert ismertet a magyar nyelvű diskurzusszintű bizonytalanság azonosítására, mely szemantikai és pragmatikai jellemzőket is magában foglaló gazdag jellemzőtéren alapszik. Ezek a cikkek a szerző önállóan elért eredményeit írják le.

## 4.3. Bizonytalanság azonosítása orvosi szövegekben

Bemutattuk a bizonytalanság azonosításának egy gyakorlati példáját is: klinikai zárójelentések szövege alapján következtettünk arra, hogy a beteg elhízott-e, illetve szenved-e 15 másik betegség valamelyikében. A nem egyértelmű esetek felcímkézésében jelentős szerep jutott a bizonytalanságazonosító rendszerünknek, ami igazolja, hogy egy valós életbeli információkinyerési feladatban is sikeresen alkalmazható a bizonytalanságazonosító rendszerünk (**3. tézispont**).

A főbb eredmények:

- klinikai zárójelentések szövege alapján betegségek azonosítására szolgáló automatikus rendszer;

- annak megmutatása, hogy a bizonytalanság azonosítása jól használható információkinyerési feladatokban;

- az információkinyerő rendszerbe beépített bizonytalanságazonosító;

- az eredmények igazolják, hogy egy szótáralapú megközelítésen és a bizonytalanság

/ tagadás azonosításán alapuló egyszerű módszer is hatékonyan alkalmazható a feladatra.

Farkas et al. (2009) megmutatta, hogy egy valós életbeli feladatban – betegségek azonosítása klinikai szövegekből – hogyan lehet alkalmazni a bizonytalanság azonosítását. A szerző nyelvi szabályokat fogalmazott meg a bizonytalanság és tagadás azonosítására, az orvosi doménre jellemző bizonytalansági kulcsszavakat gyűjtött, meghatározta ezen kulcsszavak nyelvi hatókörét és összeállított egy szótárat a szükséges orvosi szakkifejezésekből és betegségnevekből. Utóbbi feladatot az egyik társszerzővel közösen végezte, míg más társszerzők a terminusok statisztikai alapú azonosításáért, a kontextus megállapításáért és a biomarkerek rendszerbe való beépítéséért feleltek. A kísérleti eredmények közös és oszthatatlan hozzájárulásnak tekintendők minden szerző részéről.

## 4.4. A bizonytalanság típusainak kategorizálása

A nyelvi bizonytalanság különféle típusainak kategorizálására létrehoztunk egy nyelvészeti és számítógépes alapokon nyugvó egységes, nyelvfüggetlen osztályozást. Mind a szemantikai, mind a diskurzusszintű bizonytalanság típusait besoroltuk a rendszerbe, összehasonlítottuk a korábban létrehozott, bizonytalanságra annotált korpuszok irányelveit, majd beillesztettük a korábbi (számítógépes) nyelvészeti tanulmányokban vizsgált nyelvi jelenségeket az általunk definiált keretrendszerbe. E keretrendszer képezi az általunk létrehozott, bizonytalanságra annotált korpuszok elméleti hátterét (**4. tézispont**).

A főbb eredmények:

- a szemantikai bizonytalanság típusainak nyelvfüggetlen osztályozása;
- a diskurzusszintű bizonytalanság típusainak nyelvfüggetlen osztályozása;
- a bizonytalanságra annotált korpuszok annotációs elveinek összehasonítása;
- egységes osztályozás, amelyben a korábbi tanulmányokban tárgyalt bizonytalansági jelenségek mindegyike elhelyezhető.

Szarvas et al. (2012) és Vincze (2013) részletesen tárgyalja a szemantikai és diskurzusszintű bizonytalanság osztályozását, mely a szerző kizárólagos eredményének tekinthető.

## 4.5. Bizonytalanságra annotált korpuszok létrehozása

Számos korpuszt hoztunk létre (BioScope, FactBank, WikiWeasel, hUnCertainty), melyekben kézzel megjelöltük a bizonytalanságot jelző kulcsszavakat, a fenti egységes osztályozásra alapozva. Ezeket a korpuszokat használtuk a későbbiekben a bizonytalanság automatikus azonosítására irányuló gépi tanulási kísérleteinkben. A kulcsszavak eloszlását statisztikai módszerekkel is megvizsgáltuk a korpuszok alapján, ami során kiderült, hogy az egyes domének és műfajok jellemző sajátosságokat mutatnak a bizonytalanságot jelölő kulcsszavak eloszlása terén (**5. tézispont**).

A főbb eredmények:

- a BioScope, FactBank és WikiWeasel angol nyelvű korpuszokban megjelöltük a szemantikai bizonytalanság kulcsszavait;
- a WikiWeasel korpuszban a diskurzusszintű bizonytalanság kulcsszavait is jelöltük;
- a hUnCertainty magyar nyelvű korpuszban megjelöltük a szemantikai és diskurzusszintű bizonytalanság kulcsszavait;
- a hUnCertainty és a WikiWeasel 3.0 korpuszok azonos annotálási elvek alapján készültek, és a kulcsszóeloszlásuk hasonlóságokat mutat, ami alátámasztja a bizonytalansági jelenségek osztályozásának nyelvfüggetlenségét;
- statisztikai adatok a kulcsszavak eloszlási gyakoriságáról;
- a szemantikai bizonytalanság kulcsszavainak eloszlásának összehasonlítása korpuszbeli adatok alapján domének és műfajok között, ami rávilágít a bizonytalanság azonosításának domén- és műfajfüggő aspektusaira.

A korpuszok részletes leírása a Vincze et al. (2008), Vincze (2010), Farkas et al. (2010), Szarvas et al. (2012), Vincze (2013) és a Vincze (2014) cikkekben található. A szerző

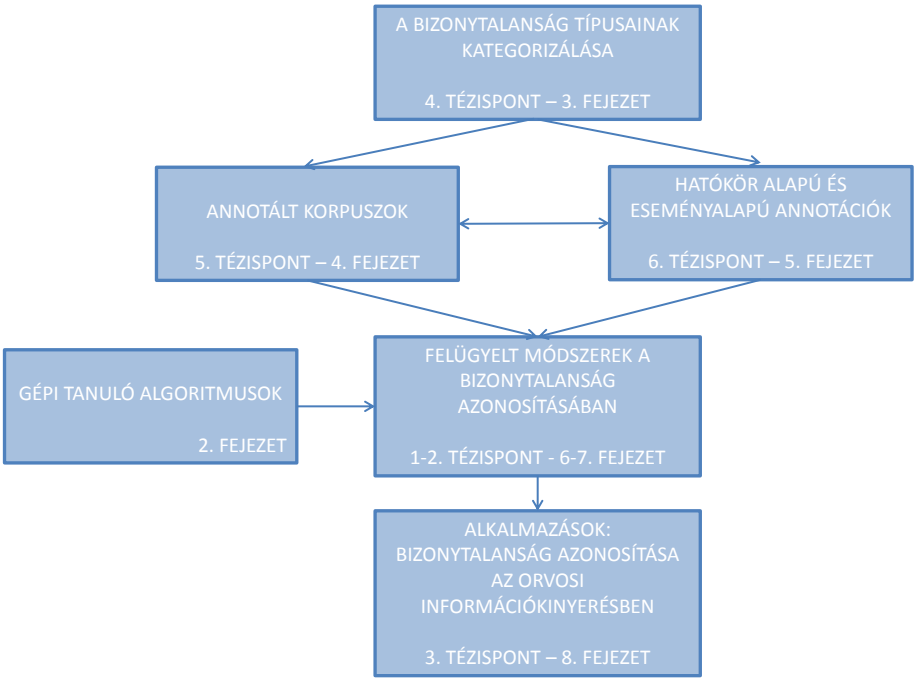
feladata volt a korpuszépítés módszertanának megtervezése, az annotációs útmutatók elkészítése, az annotációs munkálatok felügyelete, továbbá ő is részt vett az annotációban és az adatok ellenőrzésében. A szerző végezte a kulcsszóeloszlások statisztikai vizsgálatát minden korpuszban, igazolva ezáltal a bizonytalansági jelenségek doménfüggő voltát. A fenti cikkek társszerzői a fenti eredményekhez csak kismértékben járultak hozzá: ők végezték a BioScope 1.0 statisztikai adatelemzését, illetve a korpusz általános annotálási elvei közül néhány az ő nevükhöz fűződik.

## 4.6. Hatókör alapú és eseményalapú annotációk

Feltérképeztük és osztályokba soroltuk a nyelvi hatókörökön, illetve az eseményeken alapuló, tagadásra és bizonytalanságra irányuló annotációk közti jellegzetes különbségeket a BioScope 1.0 és a Genia Event korpuszok közös halmazának összevetésével. Eredményeink szerint a hatókör alapú annotáció hatékonyabbnak bizonyul a biológiától eltérő területekre fejlesztett alkalmazások esetében, mivel az eseményalapú annotációs rendszer nagymértékben épít a biológiai domén sajátságaira. Megmutattuk azt is, hogy a szemantikai és a diskurzusszintű bizonytalanság szintjén egyaránt megjelenik a bizonytalanság fokozatosságának kérdése (**6. tézispont**).

A főbb eredmények:

- a tagadás és bizonytalanság nyelvi alapú és eseményalapú annotációi közti különbségek osztályba sorolása a BioScope 1.0 és Genia Event korpuszok közös halmaza alapján;
- az eltéréskategóriák gyakoriságának megbecslése;
- stratégiák az eltérések feloldására: a szintaktikai típusú eltérések függőségi elemzésen alapuló módszerekkel feloldhatók, míg az egységesített annotációs séma egyes szemantikai jellegű különbségek felszámolásában segít;
- a hatókörökre épülő annotációs séma a nem biológiai területeken eredményesebben alkalmazható, mivel az eseményalapú annotáció nagymértékben doménfüggő;




1. ábra. Az értekezés témakörei, fejezetei és a tézispontok.

- annak megmutatása, hogy a szemantikai és diskurzusszintű bizonytalanság szintjén is releváns a fokozatosság kérdése.

Vincze et al. (2011) összeveti a hatókör alapú és eseményalapú bizonytalansági annotáció elveit. A szerző osztályozta és elemezte a korpuszok közti eltéréseket, meghatározta a hatókör alapú annotáció elvi hátterét, feloldási stratégiákat javasolt az eltérésekre, valamint rávilágított az annotációs módszertannak a gyakorlati alkalmazásokban játszott szerepének egyes aspektusaira. A cikk társszerzői határozták meg az eseményalapú annotáció elveit, valamint statisztikailag elemezték az eltéréseket.

A fenti tézispontok és a publikációk kapcsolatát az 1. táblázat szemlélteti, míg az értekezés témaköreinek, fejezeteinek és a tézispontoknak a kapcsolatát az 1. ábra mutatja.

# 5. Összegzés és jövőbeli tervek

Az értekezés fő célja a nyelvi bizonytalanság azonosítása volt természetes nyelvi szövegekben. A legfontosabb eredményeink a következőkben összegezhetők:

| | Thesis | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| BMC 2008 (Vincze et al., 2008) | | | | | • | |
| JAMIA 2009 (Farkas et al., 2009) | | | • | | | |
| NESP 2010 (Vincze, 2010) | | | | | • | |
| CoNLL 2010 (Farkas et al., 2010) | | | | | • | |
| JBMS 2011 (Vincze et al., 2011) | | | | | | • |
| CL 2012 (Szarvas et al., 2012) | • | | | • | • | |
| IJCNLP 2013 (Vincze, 2013) | | • | | • | • | |
| COLING 2014 (Vincze, 2014) | • | • | | | • | |

1. táblázat. A szerző legfontosabb publikációi és a tézispontok kapcsolata.

- a nyelvi bizonytalanság modellezhető nyelv- és doménfüggetlen módon;
- a bizonytalanságot jelölő kulcsszavak doménfüggő eloszlást mutatnak;
- felügyelt tanulási módszerek jól alkalmazhatók a bizonytalanság azonosítására;
- a gépi tanuláson alapuló bizonytalanságazonosító módszerek angolban és magyarban is jól működnek;
- doménadaptációs technikák segítségével csökkenthető a domének közti távolság a bizonytalanság azonosításában;
- az annotációs elvek meghatározhatják a korpuszok hasznosíthatóságát, például az eseményalapú annotációt tartalmazó korpuszokat leginkább biológiai információkinyerésben alkalmazzák;
- a bizonytalanságazonosító rendszerek képesek javítani az információkinyerő rendszerek hatékonyságán.

A fentieken kívül az értekezés eredményeit a számítógépes nyelvészet más területein is (például információkinyerés és -visszakeresés, dokumentumosztályozás és gépi fordítás), továbbá más tudományterületeken is (például elméleti és kontrasztív nyelvészet) hasznosítani lehet.

A jövőben más nyelvű, illetve más típusú szövegekben is szeretnénk azonosítani a bizonytalanságot jelző kulcsszavakat. E célból más nyelvű és más doménbe tartozó szövegekből is szeretnénk annotált korpuszt építeni (lásd Vincze et al. (2014) a magyar webes szövegekben fellelhető bizonytalanság azonosításáról), továbbá automatikus módszereinket is szeretnénk az új korpuszokra kiterjeszteni. Tervezzük, hogy a későbbiekben a bizonytalanságazonosító rendszerünket beépítjük különféle információkinyerő, illetve információvisszakereső alkalmazásokba. Véleményünk szerint az értekezésben ismertetett módszerek jól hasznosíthatók számos számítógépes nyelvészeti feladat megoldásában, valamint újfajta megközelítések kidolgozásához és eddig még feltáratlan alkalmazási területek felfedezéséhez is hozzájárulhatnak.

# Hivatkozások

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford University Press, Oxford.

de Marneffe, Marie-Catherine; Manning, Christopher D.; Potts, Christopher. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38:301–333, June.

Diab, Mona; Levin, Lori; Mitamura, Teruko; Rambow, Owen; Prabhakaran, Vinodkumar; Guo, Weiwei. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pp. 68–73, Suntec, Singapore, August. Association for Computational Linguistics.

Farkas, Richárd; Szarvas, György; Hegedűs, István; Almási, Attila; Vincze, Veronika; Ormándi, Róbert; Busa-Fekete, Róbert. 2009. Semi-automated construction of decision rules to predict morbidities from clinical texts. *Journal of the American Medical Informatics Association*, 16:601–605.

Farkas, Richárd; Vincze, Veronika; Móra, György; Csirik, János; Szarvas, György. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pp. 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

Ganter, Viola; Strube, Michael. 2009. Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 173–176, Suntec, Singapore, August. Association for Computational Linguistics.

Kim, Jin-Dong; Ohta, Tomoko; Tsujii, Jun'ichi. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(Suppl 10).

Palmer, Frank Robert. 1986. *Mood and Modality*. Cambridge University Press, Cambridge.

Saurí, Roser; Pustejovsky, James. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Saurí, Roser; Pustejovsky, James. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38:261–299, June.

Szarvas, György; Vincze, Veronika; Farkas, Richárd; Móra, György; Gurevych, Iryna. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38:335–367, June.

Velldal, Erik; Øvrelid, Lilja; Read, Jonathon; Oepen, Stephan. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational Linguistics*, 38:369–410, June.

Vincze, Veronika; Szarvas, György; Farkas, Richárd; Móra, György; Csirik, János. 2008. The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Vincze, Veronika; Szarvas, György; Móra, György; Ohta, Tomoko; Farkas, Richárd. 2011. Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora. *Journal of Biomedical Semantics*, 2(Suppl 5):S8.

Vincze, Veronika; Simkó, Katalin Ilona; Varga, Viktor. 2014. Annotating Uncertainty in Hungarian Webtext. In *Proceedings of LAW VIII*.

Vincze, Veronika. 2010. Speculation and negation annotation in natural language texts: what the case of BioScope might (not) reveal. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 28–31, Uppsala, Sweden, July. University of Antwerp.

Vincze, Veronika. 2013. Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 383–391, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Vincze, Veronika. 2014. Uncertainty detection in Hungarian texts. In *Proceedings of Coling 2014*.