# Egy sok szálon futó nyelvelemző program moduljainak kialakítása és harmonizációja


Indig Balázs

*A PhD Disszertáció tézisei*

Pázmány Péter Katolikus Egyetem
Információs Technológiai és Bionikai Kar
Roska Tamás Műszaki és Természettudományi Doktori Iskola

Témavezető:
Dr. Prószéky Gábor
az MTA doktora

Budapest, 2017

# 1. Bevezetés

Hagyományosan a nyelvtechnológiai eszközök egy *csővezeték (pipeline)* formájában (hívják még *szerelőszalagnak* is) működnek az architektúra minden előnyével és hátrányával. A csővezeték a nyers szövegtől indul és az elemzés tetszőleges szintjén ér véget. A tradicionális modulok a következők:

- Mondatra bontó, tokenizáló
- Morfológiai elemző, szófaji egyértelműsítő
- Főnévicsoport- és névelemkereső (sekély elemzés)
- Szintaktikai elemző
- Szemantikai elemző
- Információkinyerő, gépi fordító, stb.

A csővezeték egyszerűségéből adódóan az egyes moduloknak elméletben nem kell tudniuk semmit a többi modulról. Így hagyományosan a tesztelésük is a *gold sztenderd korpuszon vagy más néven referenciaadaton (gold standard)* történt, azaz a tökéletes bemenetből tökéletes kimenetet kellett előállítaniuk. Viszont az éles használat során a csővezetéknek a bemenettel nem közvetlenül érintkező moduljai az őket megelőző modulok hibáit felhalmozva korántsem tökéletes bemenettel kell, hogy dolgozzanak. Az irodalomban kevés utalást találunk arra, hogy mennyire robusztusak ezek a rendszerek – egy csővezeték részeként – hibás bemenet esetén.

Ez a tény motiválja azt a kérdést, hogy hogyan működhetnének jobban együtt ezek a modulok, hogy a potenciális hibák ne halmozódjanak fel a feldolgozás során a csővezetékben. Ezért dolgozatomban áttekintem a szabadon elérhető, magyar nyelvű state-of-the-art módszereket, valamint megoldást keresek a harmonizációjukkal kapcsolatos problémákra.

A magyar nyelvben az egyszerű mondatok két jól elkülöníthető komponensre bonthatók. Az egyik a *közvetlen összetevős szerkezetek*, ezekben az elemek sorrendje kötött, és nem mozognak szabadon a mondatban. Ilyenek a *főnévi csoportok*. Míg a másik osztály az *igei szerkezetek*, melynek elemei között megtalálhatjuk az imént említett közvetlen összetevős szerkezeteket mint az igék argumentumait. A dolgozatban ezt a két osztályt tárgyalom bővebben.

# 2. Módszerek

A dolgozatban használt, különböző típusú gépi tanulást végző programok és mintakereső módszerek a szabványos körülmények közötti összehasonlítás és elemzés nélkül nem tudnak tudományosan értékelhető eredményt adni. Ezért a használt programok kiértékeléséhez a tudományterületen megszokott, a *pontosság* és *fedés* harmonikus közepeként előálló *F-mértéket*, bemenetként és elvárt kimenetként pedig a főbb elérhető korpuszokat használtam.

A tudományos vizsgálatokra szánt szövegek speciális szempontok szerint előállított korpuszok formájában érhetőek el. A korpuszok tartalmazzák a szövegekhez tartozó, többségében automatikusan készült annotációt. A felügyelt tanításra épülő módszereknek szüksége van referenciaadatra is, mely egy előre meghatározott formátum és eljárásrend szerint kézzel készül. Az ilyen korpuszok előállítási költsége az annotáció emberierőforrás-igénye miatt igen magas. Magyar nyelvre a *Szeged Korpusz* (Csendes et al. 2003) a jelenleg egyetlen kézzel annotált korpusz, mely 70 000 mondatot és 1 194 348 tokent tartalmaz. A függőségi elemzéssel ellátott változata a *Szeged Treebank* (Vincze et al. 2010). A dolgozatban magyar nyelvre ezt a korpuszt használtam én is tanítóanyagként a maximális főnévi csoportok kereséséhez.

A nyelv modellezéséhez viszont elegendő a lehető legnagyobb mennyiségű szöveg gyűjtése, mivel a feldolgozás emberi erőforrást nem igényel. Ezekkel a szövegekkel kapcsolatos egyedüli kritérium, hogy a megfelelő nyelven legyenek és normalizált formában, egységes egészt alkossanak. Az internetes kommunikáció erősödésével manapság nagyon könnyű különböző minőségű szövegeket szisztematikusan legyűjteni az internetről, így a magyar nyelvre elérhető, géppel elemzett korpuszok száma is egyre nő (Indig 2018).

A dolgozatban nyelvmodellezésre két korpuszt használtam. Az egyik a *Magyar Nemzeti Szövegtár* első és második (2.0.3) verziója (Oravecz, Váradi és Sass 2014), az első változat 187 millió, a második pedig 785 millió szót (978 millió tokent) tartalmaz változatos forrásokból (beszélt szövegek átiratai, határon túli újságok, jogi szövegek, parlamenti naplók, stb.). A második korpusz pedig a teljes egészében az internetről gyűjtött szövegekből készült *Pázmány Korpusz*, mely 1,2 milliárd szavas (Endrédy 2016).

A dolgozatban szerepel továbbá az *InfoRádió Korpusz*, amely csak szerkesztett rövidhíreket tartalmaz, néha többmondatos megnyilatkozások formájában. 2 millió szavával egy kisebb doménspecifikus korpusz, mely a bemutatott elemzőmodell által megelemzendő szövegek prototípusát alkotja. Az angol nyelvű, *közvetlen összetevők keresését* célzó vizsgálatokat pedig a *CoNLL-2000 korpuszon* (Tjong Kim Sang és Buchholz 2000) (259 104 token) végeztem.

# 3. Új tudományos eredmények összefoglalása

A dolgozatban bemutattam a magyar nyelvre jelenleg is használt nyelvtechnológiai szerelőszalag működését. A szerelőszalag-architektúra számtalan előnnyel és hátránnyal rendelkezik. Napjainkra a régóta ismert előnyök mellett lassan a hátrányok is megmutatkoznak. Az egyes modulok csak a szomszédos modullal érintkeznek, így a bemenetük és kimenetük nagyban eltérhet. Manapság több eszköz is elérhető egy adott feladat megoldására, ezért szükségessé vált azok egységesítése, együttműködésük vizsgálata.

Felvázoltam az ilyen eszközök ökoszisztémáinak működéséhez szükséges feltételeket. Ismertettem az általam fejlesztett, pszicholingvisztikailag motivált nyelvelemző modellel, az AnaGramma elemzővel szemben támasztott elvárásokat, melynek célja egy emberi elemzőhöz hasonlító számítógépes szövegelemzési modell létrehozása. Az AnaGramma célja továbbá az, hogy megszüntesse a soros architektúrából származó hibákat, melyek a szerelőszalag végére felerősödnek és értékelhetetlenné teszik az eredményt. Az általam készített elemzőrendszer eredendően párhuzamos, így minden modul egyszerre, egymást javítva képes futni benne. A dolgozatban a modell architektúrájához szükséges eljárásokat tekintettem át. A magyar nyelvben a főnévi csoportok a bennük lévő elemek kötött sorrendje miatt jó kiindulópontnak bizonyultak, valamint a mondatban szereplő igék argumentumaiként fontos szerepük van. A dolgozat így ezen két csoport működésének harmóniájára épül.

A dolgozatomban bemutatott téziseket négy csoportba lehet osztani. A főnévi csoportok keresésének state-of-the-art megoldásától módszeresen a szekvenciális címkézés különböző közös tulajdonságain át az n-gram modellek vizsgálatával eljutottam az elemzőhöz szükséges korpuszminták újelvű alkalmazásához. Ezt követően a főnévi csoportok igei argumentumként történő azonosításának vizsgálatakor a létező erőforrások összekapcsolásával és nyelvfüggetlen információk átvitelével a finom osztályozások módját vizsgáltam, melynek segítségével pontosítani lehet az eljárásokat. Végül az elemző architektúrájának ismertetésével összefüggésben bemutattam két feltérképezett és kezelt nyelvi jelenséget.

## I. Főnévi csoportok automatikus meghatározása

Az első téziscsoportban a főnévi csoportok keresésére koncentráltam. Magyar és angol nyelven vizsgáltam meg a jelenleg használt state-of-the-art módszereket, hogy megértsem, miként lehetne őket felhasználni az elemző működéséhez. Először az angol nyelvű state-of-the-art megoldást (Shen és Sarkar 2005) akartam adaptálni a magyar nyelvre. Ennek lényegi újítása, hogy a különböző

IOB reprezentációkon tanított és címkézett korpuszt egyszerű többségi szavazással javítja, kihasználva a eltérő reprezentációk erősségeit. Az angol módszer reprodukálása során azonban kiderült, hogy az IOB reprezentációk szavazásával elérhető nyereség csak mérési hiba eredménye és műtermék. Így az angol nyelven legjobb módszer nem volt alkalmazható magyar nyelvre, hiszen a vele elért eredmény nem valós.

**1. Tézis.** *Méréssel kimutattam, hogy nem helytálló az a szakirodalomból ismert állítás, amely szerint a különböző IOB-reprezentációk közötti szavazás szignifikáns javulást hoz az angol nyelvű főnévi csoportok meghatározásának minőségén.*

A tézist alátámasztó közlemények: [3]

Ezek után a magyar nyelvű state-of-the-art módszer (Recski és Varga 2012) vizsgálata során a maximális NP-k keresésének feladatában felismertem, hogy a state-of-the-art módszer csak bigram címkeátmeneti modellt használ, mert a névelem-felismerésből jövő módszerből származik. Méréssel igazoltam, hogy a modell eredménye javítható trigram címkeátmeneti modell használatával.

**2. Tézis.** *Az általam kifejlesztett HunTag3 program segítségével méréssel igazoltam (társszerzővel közösen), hogy a trigrammok használatával javulás érhető el a bigrammokhoz képest a magyar nyelvű maximális főnévi csoportok meghatározásában.*

A tézist alátámasztó közlemények: [8]

## II. Lexikalizációs eljárások

A második téziscsoportban a főnévi csoportok keresésének feladatán elindulva felismertem, hogy a szekvenciális címkézési feladatok sok tulajdonságukban különböznek ugyan, de sokban hasonlítanak is. Ez a megfigyelés segítségemre volt az elemző architektúrájának tervezésében. Ezért a szekvenciális címkézés feladatain – melyek közös tulajdonsága, hogy az emberi elemzőhöz hasonlóan balról jobbra dolgozzák fel a szöveget – általánosan alkalmazható módszereken kezdtem dolgozni, melyet a már meglévő eredményeim javítására használtam. Ehhez kapcsolódóan a dolgozatban bemutattam az általam vizsgált lexikalizációs ejárások működését és hatását.

**3. Tézis.** *Létrehoztam egy új, általános, szekvenciális címkézésre alkalmazható lexikalizációs eljárást, melynek első konkrét alkalmazása tetszőleges részszerkezetek hatékony azonosítását szolgálja.*

A tézist alátámasztó közlemények: [2, 3]

Az általam feltalált lexikalizációs eljárással és az optimális küszöbérték meghatározásával és alkalmazásával meghaladtam az angol nyelvű közvetlen összetevős keresés feladatán a state-of-the-art módszer teljesítményét.

**4. Tézis.** *Az általam kidolgozott eljárás angol nyelvű főnévi csoportokra méréssel igazolhatóan felülmúlja a jelenleg ismert módszerek F-mértékét.*

A tézist alátámasztó közlemények: [2, 3]

Következésképpen kimondható, hogy a finomabb osztályozással nagyobb pontosságot lehet elérni, és ahol a pontosság a cél, ott a módszeremmel javíthatóak az eredmények. Emiatt fontosnak tartottam megvizsgálni, hogy az `IOB` reprezentációk konverziójánál a konverter, valamint a tesztelésnél a címkéző program fenn tudja-e tartani a jólformáltságot a kimeneti címkesorozatok zárójelezésében. Ennek mérésére kidolgoztam egy metrikát, amit gyakorlatban alkalmaztam az angol nyelvű közvetlen összetevők keresésének feladatán.

**5. Tézis.** *Kidolgoztam egy zárójelezési módszert, mely egyfajta metrikaként a címkézési feladatra készített módszereket minőség szerint rendezni tudja.*

A tézist alátámasztó közlemények: [2, 3]

## III. Erőforrások összekapcsolása

Mivel a maximális főnévi csoportok az igék argumentumaiként funkcionálnak a mondatban, megvizsgáltam a rendelkezésre álló magyar nyelvű igei erőforrásokat (Indig, Vadász és Kalivoda 2017; Kalivoda 2016; Kornai, Nemeskey és Recski 2016; Sass 2015; Sass et al. 2010). A vizsgálat során arra jutottam, hogy egyikben sincs szemantikai információ a szintaktikai mellett, aminek segítségével tovább lehetne finomítani a főnévi csoportok osztályozását. Bemutattam a *Linked Data*[1] fogalmát, és a módszer az erőforrásokra vonatkoztatott változatának ismertetése után bemutattam néhány angol nyelvű példát az összekapcsolt erőforrásokra (Prószéky, Miháltz és Kuti 2013; Vossen et al. 1998). Majd ezen a vonalon elindulva a kétnyelvű, magyar-angol MetaMorpho adatbázis (Prószéky, Tihanyi és Ugray 2004) és az angol VerbIndex (Loper, Yi és Palmer 2007) összekapcsolását tűztem ki célul azért, hogy nyelvfüggetlen szemantikai annotációt tudjak automatikusan átvinni a információkban jóval gazdagabb VerbIndexből a MetaMorpho adatbázisba.

[1] `http://linkeddata.org/`

**6. Tézis.** *Létrehoztam egy automatikus módszert az 1-, 2- és 3-vonzatú igék magyar–angol vonzatkeretpárjainak összekapcsolására, melynek eredményeképpen sikerült angolról magyarra átvinni a megfelelő tematikus szerepeket.*

A tézist alátámasztó közlemények: [11, 12, 4, 22]

Az összekapcsolás részeként harmonizálni kellett a két erőforrás között az elemek megszorításait leíró ontológiákat, melyek között egy áthidaló fogalmakat tartalmazó ontológiával teremtettem meg az átjárhatóságot.

**7. Tézis.** *Kialakítottam egy ontológiát, amely összekapcsolja a magyar nyelvű MetaMorpho igéinek leírását az angol VerbIndex szintaktikai és szemantikus kategóriáival.*

A tézist alátámasztó közlemények: [11, 12, 4]

A meglévő információk alapján össze lehetett kapcsolni a magyar és az angol nyelvű WordNeteket is. Ezeket a kapcsolatokat is latba vetettem, hogy javítsam a minőséget, de azok nem bizonyultak megfelelőnek a feladat szempontjából.

**8. Tézis.** *Méréssel kimutattam, hogy a magyar és angol nyelvű WordNetek bevonásával nem lehet a fenti ontológia minőségét tovább javítani.*

A tézist alátámasztó közlemények: [11, 12, 4]

Végül az igei vonzatkeretek egy viszonylag jó fedésű alosztályára sikerült jó minőségben szemantikai információt átvinni automatikus úton, mely további osztályozási lehetőségeket nyitott meg.

## IV. A pszicholingvisztikailag motivált elemző architektúrája

A különböző nyelvi jelenségekből levont tanulságok nyomán bemutattam az AnaGramma elemző működését, az elmélet után a megvalósítás szempontjából is. Definiáltam a nyelvi jelenségek kezeléshez használt ablakot, melynek ötlete a két fázisban működő *Sausage Machine*-ből (Frazier és Fodor 1978) származik. Az általam bemutatott ablak – mely a *Sausage Machine* első, PPP fázisának felel meg – és a rajta definiált keresőeljárások megoldást adnak a hatékony, emberi elemzőhöz hasonló, balról jobbra elemzés számos problémájára.

**9. Tézis.** *Létrehoztam egy új megközelítésű (az ún. ablakra épülő) elemzési modell alapjait, amelynek elméletét társszerzővel közösen dolgoztam ki, és melynek segítségével a magyar nyelvű bemenet hatékonyan és az emberi feldolgozáshoz hasonlóan, szigorúan balról jobbra feldolgozható.*

A tézist alátámasztó közlemények: [14, 5, 21, 34, 25, 26]

A definiált ablakon működő eljárások közül ismertettem a jelöletlen (nominatívuszos) és a jelölt (*-nAk* ragos) birtokos szerkezetek (Bánréti et al. 1992) kezelését. A módszerben a kétfázisú mondatelemzés első fázisában, az előretekintő elemzési ablak segítségével megtörténik a testes esetrag nélküli elemek eset-egyértelműsítése, melynek eredményeképpen tisztázódik a mondatbeli szerepük. Birtokos esetén, a bemutatott kereslet-kínálat keretrendszerben a birtok lesz az, amely birtokos-kereslettel él, amely kereslet kielégülésekor birtokos él jön létre a birtok és birtokosa között.

**10. Tézis.** *Létrehoztam egy az ablak segítségével a jelöletlen szerkezetek egyértelműsítését (például a magyar birtokos szerkezet és az alanyeset hatékony, valós idejű elkülönítését) elvégző algoritmust, melynek elméletét társszerzővel közösen közösen dolgoztam ki.*

A tézist alátámasztó közlemények: [5, 21, 25]

Korpuszmérések alapján bizonyítható, hogy az AnaGramma elemzőrendszer keretein belül a finit ige–igekötő kapcsolat létrehozása mellett az infinitívusz–igekötő és a finit ige–infinitívuszi vonzat kapcsolatok létrehozásához is elegendő a feltételezett két token méretű elemzési ablak használata. A tározó és az ablak segítségével a VFrame keresőeljárás a mondatban szereplő igei elemeket (finit és infinit igéket), valamint az igekötőket a megfelelő módon kapcsolja össze.

**11. Tézis.** *Létrehoztam a VFrame eljárást, amelynek elméletét társzerzővel közösen dolgoztam ki, amivel a magyar nyelvben a helyes igekötő megtalálása az igekötők eloszlási mintájának ismeretében a lehetséges igei vonzatkeretek halmazának leszűkítésével történik.*

A tézist alátámasztó közlemények: [14, 26, 27]

A bemutatott módszerek mind pontosság, mind pedig fedés tekintetében jól teljesítettek.

# 4. Az eredmények alkalmazási területei

A bemutatott eredmények frissességük miatt még nem kerültek széleskörben alkalmazásra, viszont a főnévi csoportokkal és a szekvenciális címkézéssel kapcsolatos eredmények nagy érdeklődést vonzottak a nemzetközi konferenciákon. Úgy látom, hogy a jelenleg csak angol nyelvre megvizsgált eredmények némi változtatással átültethetőek magyar nyelvre, valamint más agglutináló nyelvekre is.

Ezek egyike lehet az, hogy a meglévő szófaji egyértelműsítő módszerekkel egybeépítve a közvetlen összetevők és az NP-k határát jelölő annotációt is párhuzamosan el lehessen végezni. Az általam bemutatott enyhe lexikalizáció a dolgozatban ismertetett feladatokon túl számos feladatra általánosítható. A zárójelezés jólformáltságát ellenőrző metrika alkalmazása sok kellemetlenségtől kímélheti meg a jövő kutatóit minden szekvenciális címkézési feladat során.

Az összekapcsolt erőforrásokkal kapcsolatos elméleti eredményeim jól használhatóak később azok számára, akik hasonló erőforrás-összekapcsoláson gondolkoznak. Látható, hogy a rendszerben a szabályalapú összetevők túlsúlya miatt az ember által elkövethető hibák száma is nagyobb, ezért a tapasztalataimat érdemes figyelembe venni egy másik hasonló projekt előtt. A méréseimből az is látszik, hogy a jelenlegi szabályalapú és statisztikai erőforrások együttműködése egy jól használható rendszerként még nem megvalósított, így jobban járunk, ha csak a szabályalapú rendszereket használjuk.

Az ismertetett munka alkalmazható például jó minőségű szemantikai információkat tartalmazó igei adatbázisok előállítására, melyek pontos szemantikai elemzést tesznek lehetővé, és a jövőben számos elméleti nyelvészeti kutatás alapjául szolgálhatnak. Mindamellett ezt már az igék igekötőinek keresésekor az általam kidolgozott elemzőmodellben fel is használtam. Ez példaként szolgál az eredményeim alkalmazásához a számítógépes nyelvészet területén tevékenykedők számára. Távlati cél lehet az angol nyelvű erőforrásokból elérhető nyelvfüggetlen információ megbízható, automatikus átemelése magyar nyelvre a létrehozott ontológiák segítségével, de a neurális hálók előretörésével várhatóan a WordNet és a kézzel készített erőforrások háttérbe szorulnak a statisztikailag megalapozottabb erőforrásokkal szemben, így ebből a szempontból a hosszútávú haszna kétséges.

Az AnaGramma elemző architektúrájának tervezésekor felhasználtam a dolgozatban közölt többi eredményemet, melyek elméleti jelentősége nagyban hozzájárult a további új eredmények létrejöttéhez. Az utolsó fejezetben bemutatott eredményeim az elméleti nyelvészet szempontjából fontosak. Várható, hogy pszicholingvisztikai alkalmazásaik is lesznek, és további kutatások épülnek rájuk.

# A szerző közleményei

## Nemzetközi folyóiratcikkek és könyvfejezetek


[1] Garay, Barnabás Miklós és **Balázs Indig** (2015.). „Chaos in Vallis' asymmetric Lorenz model for El Niño". *Chaos, Solitons & Fractals* 75.1., 253–262. old. issn: 0960-0779.

[2] **Indig**, **Balázs** (2017.a). „Less is More, More or Less... – Finding the Optimal Threshold for Lexicalization in Chunking". *Computación y Sistemas* 21.4.

[3] **Indig**, **Balázs** és István Endrédy (2018.). „Gut, Besser, Chunker – Selecting the best models for text chunking with voting". *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part I (Lecture Notes in Artificial Intelligence)*. Szerk. Alexander Gelbukh. Cham: Springer International Publishing. Fej. 29, 409–423. old. isbn: 978-3-319-75476-5. doi: 10.1007/978-3-319-75477-2_29.

[4] **Indig**, **Balázs**, András Simonyi és Márton Miháltz (2018.). „Exploiting Linked Linguistic Resources for Semantic Role Labeling". *Human Language Technology. Challenges for Computer Science and Linguistics. 7th Language and Technology Conference, LTC 2015, Poznań, Poland, November 27-29, 2015. Revised Selected Papers (Lecture Notes in Artificial Intelligence 10930)*. Szerk. Zygmunt Vetulani, Joseph Mariani és Marek Kubis. Cham: Springer International Publishing. isbn: 978-3-319-93781-6. doi: 10.1007/978-3-319-93782-3.

[5] **Indig**, **Balázs**, Noémi Vadász és Ágnes Kalivoda (2016.). „Decreasing Entropy: How Wide to Open the Window?": *Theory and Practice of Natural Computing (Lecture Notes in Computer Science volume 10071)*. Szerk. Carlos Martín-Vide, Takaaki Mizuki és Miguel A. Vega-Rodríguez. Cham: Springer International Publishing, 137–148. old. isbn: 978-3-319-49001-4. doi: 0.1007/978-3-319-49001-4_11.


## Hazai folyóiratcikkek és könyvfejezetek


[6] Prószéky, Gábor és **Balázs Indig** (2015.a). „Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel". *Alkalmazott nyelvtudomány* 15.1-2., 29–44. old.

[7] Prószéky, Gábor, **Balázs Indig** és Noémi Vadász (2016.). „Performanciaalapú elemző magyar szövegek számítógépes megértéséhez”. *“Szavad ne feledd!”: Tanulmányok Bánréti Zoltán tiszteletére*. Szerk. Bence Kas. Budapest: MTA Nyelvtudományi Intézet, 223–232. old.

## Nemzetközi konferenciacikkek

[8] Endrédy, István és **Balázs Indig** (2015.). „HunTag3: a general-purpose, modular sequential tagger – chunking phrases in English and maximal NPs and NER for Hungarian”. *7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. (Poznań, Poland, nov. 27.–2015). Poznań, Poland: Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu, 213–218. old. isbn: 978-83-932640-8-7.

[9] **Indig**, **Balázs** (2017.b). „Mosaic n-grams: Avoiding combinatorial explosion in corpus pattern mining for agglutinative languages”. *8th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. (Poznań, Poland, nov. 17.–2017). Poznań, Poland: Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu, 147–151. old. isbn: 978-83-64864-94-0.

[10] **Indig**, **Balázs** (2018.b). „The stability of the parameter transformation with Zipfian distributions across corpora”. *Computational Linguistics and Intelligent Text Processing: 19th International Conference, CICLing 2018, Hanoi, Vietnam, April 18–24, 2018, Revised Selected Papers, Part I (Lecture Notes in Artificial Intelligence)*. Szerk. Alexander Gelbukh. (Accepted, in press). Cham: Springer International Publishing.

[11] **Indig**, **Balázs**, Márton Miháltz és András Simonyi (2015.). „Exploiting Linked Linguistic Resources for Semantic Role Labeling”. *7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. (Poznań, Poland, nov. 27.–2015). Poznań, Poland: Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu, 140–144. old. isbn: 978-83-932640-8-7.

[12] **Indig**, **Balázs**, Márton Miháltz és András Simonyi (2016.). „Mapping Ontologies Using Ontologies: Cross-lingual Semantic Role Information Transfer”. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Szerk. Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk és Stelios Piperidis. Portorož, Slovenia:

European Language Resources Association (ELRA), 2425–2430. old. isbn: 978-2-9517408-9-1.

[13] **Indig**, **Balázs**, András Simonyi és Noémi Ligeti-Nagy (2018.). „What's Wrong, Python? – A Visual Differ and Graph Library for NLP in Python". *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Szerk. Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis és Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA). isbn: 979-10-95546-00-9.

[14] **Indig**, **Balázs** és Noémi Vadász (2016.b). „Windows in Human Parsing – How Far can a Preverb Go?": *Proceedings of the Tenth International Conference on Natural Language Processing (HrTAL2016) 2016, Dubrovnik, Croatia, September 29-October 1., 2016*. Szerk. Marko Tadić és Božo Bekavac. (Accepted, in press).

[15] Miháltz, Márton, Bálint Sass és **Balázs Indig** (2013.). „What Do We Drink? Automatically Extending Hungarian WordNet With Selectional Preference Relations". *Proceedings of the Joint Symposium on Semantic Processing: Textual Inference and Structures in Corpora*. ( nov. 20.–2013). Szerk. Octavian Popescu és Alberto Lavelli. Trento, Italy: Association for Computational Linguistics (ACL), 105–109. old. isbn: 978-1-6299353-9-3.

[16] Váradi, Tamás, Eszter Simon, Bálint Sass, Iván Mittelholcz, Attila Novák, **Balázs Indig**, Richárd Farkas és Veronika Vincze (2018.). „E-magyar – A Digital Language Processing System". *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Szerk. Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis és Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA). isbn: 979-10-95546-00-9.

## Hazai konferenciacikkek

[17] **Indig**, **Balázs** (2013.b). „PureToken: egy új tokenizáló eszköz". *IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013)*. Szerk. Attila Tanács és Veronika Vincze. Szegedi Tudományegyetem Informatikai

Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 305–309. old.

[18] **Indig**, **Balázs** (2018.a). „Közös crawlnak is egy korpusz a vége – Korpuszépítés a CommonCrawl .hu domainjából”. *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szerk. Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 125–135. old.

[19] **Indig**, **Balázs**, László János Laki és Gábor Prószéky (2016.). „Mozaik nyelvmodell az AnaGramma elemzőhöz”. *XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016)*. Szerk. Attila Tanács, Viktor Varga és Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 260–270. old.

[20] **Indig**, **Balázs** és Gábor Prószéky (2013.). „Ismeretlen szavak helyes kezelése kötegelt helyesírás-ellenőrző programmal”. *IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013)*. Szerk. Attila Tanács és Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 310–317. old.

[21] Ligeti-Nagy, Noémi, Noémi Vadász, Andrea Dömötör és **Balázs Indig** (2018.). „Nulla vagy semmi? Esetegyértelműsítés az ablakban”. *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szerk. Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 25–37. old.

[22] Miháltz, Márton, **Balázs Indig** és Gábor Prószéky (2015.). „Igei vonzatkeretek és tematikus szerepek felismerése nyelvi erőforrások összekapcsolásával egy kereslet-kínálat elvű mondatelemzőben”. *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*. Szerk. Attila Tanács, Viktor Varga és Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 298–302. old.

[23] Novák, Attila, György Orosz és **Balázs Indig** (2011.). „Javában taggelünk”. *VIII.. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2011)*. Szerk. Attila Tanács és Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 310–317. old.

[24] Prószéky, Gábor, **Balázs Indig**, Márton Miháltz és Bálint Sass (2014.). „Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé”. *X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY*

*2014)*. Szerk. Attila Tanács, Viktor Varga és Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 79–87. old.

[25] Vadász, Noémi és **Balázs Indig** (2018.). „A birtokos esete az ablakkal”. *LingDok: nyelvész-doktoranduszok dolgozatai*. Szerk. György Scheibl. Szegedi Tudományegyetem. Nyelvtudományi Doktori Iskola, old. 85–99.

[26] Vadász, Noémi, Ágnes Kalivoda és **Balázs Indig** (2017.). „Ablak által világosan – Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével”. *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017)*. Szerk. Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 3–12. old.

[27] Vadász, Noémi, Ágnes Kalivoda és **Balázs Indig** (2018.). „Egy egységesített magyar igei vonzatkerettár építése és felhasználása”. *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szerk. Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 3–15. old.

[28] Váradi, Tamás, Eszter Simon, Bálint Sass, Mátyás Gerőcs, Iván Mittelholcz, Attila Novák, **Balázs Indig**, Gábor Prószéky, Richárd Farkas és Veronika Vincze (2017.). „Az `e-magyar` digitális nyelvfeldolgozó rendszer”. *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017)*. Szerk. Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 49–60. old.

## Egyéb közlemények

[29] **Indig**, **Balázs** (2013.a). „An extended spell checker for unknown words”. *Pázmány Péter Catholic University PhD Proceedings* 8., 29–32. old.

[30] **Indig**, **Balázs** (2014.a). „Towards a Psycholinguistically Motivated Performance-Based Parsing Model”. *PhD Proceedings Annual Issues of the Doctoral School Faculty of Information Technology and Bionics* 2014., 133–136. old.

[31] **Indig**, **Balázs** (2014.b). „Towards recognizing thematic roles for verbal frames by linking two independent language resources for a parser based on the supply and demand paradigm”. *PhD Proceedings Annual Issues of the Doctoral School Faculty of Information Technology and Bionics* 2015., 159–161. old.

[32] **Indig**, **Balázs** és Noémi Vadász (2016.a). *POS Comes with Parsing: a Refined Word Categorisation Method*. Konfereciaabsztrakt (konferenciakötetbe nem került), 4th International Conference on Statistical Language and Speech Processing (SLSP 2016), Csehország, Plzeň, 2016. október 11-12. Pilseň, Czech Republic. url: `http://grammars.grlmc.com/SLSP2016/Download/slides/pos-comes-with-parsing-abstract.pdf`.

[33] **Indig**, **Balázs**, Noémi Vadász és Ágnes Kalivoda (2017.). *Manócska – integrált igeivonzatkeret-adatbázis*. url: `https://github.com/ppke-nlpg/manocska`.

[34] Prószéky, Gábor és **Balázs Indig** (2015.b). *Natural parsing: a psycholinguistically motivated computational language processing model*. Konfereciaabsztrakt (konferenciakötetbe nem került), 4th International Conference on the Theory and Practice of Natural Computing (TPNC 2015), Spanyolország, Astruias, Mieres, 2015. december 15-16. Mieres, Astruias, Spain. url: `http://grammars.grlmc.com/TPNC2015/Slides/d1s503natural_parsing_abstract.pdf`.

# Hivatkozások


Bánréti, Zoltán, István Kenesei, András Komlósy, Tibor Laczkó és Anna Szabolcsi (1992.). *Strukturális magyar nyelvtan I: Mondattan*. Szerk. Ferenc Kiefer és Zsófia Róbert. Akadémiai Kiadó. isbn: 963-05-6468-8.

Csendes, Dóra, Csaba Hatvani, Zoltán Alexin, János Csirik, Tibor Gyimóthy, Gábor Prószéky és Tamás Váradi (2003.). „Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz". *I. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003)*. Szerk. Zoltán Alexin és Dóra Csendes. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 238–245. old.

Endrédy, István (2016.). „Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz". PhD dissz. Budapest: PPKE-ITK.

Frazier, Lyn és Janet Dean Fodor (1978.). „The Sausage Machine: A New Two-Stage Parsing Model". *Cognition* 6.4., 291–325. old.

Indig, Balázs (2018.). „Közös crawlnak is egy korpusz a vége – Korpuszépítés a CommonCrawl .hu domainjából". *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szerk. Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 125–135. old.

Indig, Balázs, Noémi Vadász és Ágnes Kalivoda (2017.). *Manócska – integrált igeivonzatkeret-adatbázis*. url: `https://github.com/ppke-nlpg/manocska`.

Kalivoda, Ágnes (2016.). „A magyar igei komplexumok vizsgálata". Mesterszakos szakdolgozat. PPKE-BTK. url: `https://github.com/kagnes/hungarian_verbal_complex`.

Kornai, András, Dávid Márk Nemeskey és Gábor Recski (2016.). „Detecting Optional Arguments of Verbs". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Szerk. Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk és Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA). isbn: 978-2-9517408-9-1.

Loper, Edward, Szu-Ting Yi és Martha Palmer (2007.). „Combining lexical resources: mapping between PropBank and VerbNet". *Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg*, 118–128. old.

Oravecz, Csaba, Tamás Váradi és Bálint Sass (2014.). „The Hungarian Gigaword Corpus”. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Szerk. Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk és Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA). isbn: 978-2-9517408-8-4.

Prószéky, Gábor, Márton Miháltz és Judit Kuti (2013.). „Lexikális szemantika: a számítógépes nyelvészet és a pszicholingvisztika határán”. *Általános Nyelvészeti Tanulmányok* XXV., 143–172. old.

Prószéky, Gábor, László Tihanyi és Gábor Ugray (2004.). „Moose: A robust high-performance parser and generator”. *Proceedings of the 9th Workshop of the European Association for Machine Translation*. (La Valletta, Malta), 138–142. old.

Recski, Gábor és Dániel Varga (2012.). „Magyar főnévi csoportok azonosítása”. *Általános Nyelvészeti Tanulmányok* XXIV. Szerk. Gábor Prószéky, Tamás Váradi és István Kenesei.

Sass, Bálint (2015.). „28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet”. *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*. Szerk. Attila Tanács, Viktor Varga és Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 399–403. old.

Sass, Bálint, Tamás Váradi, Júlia Pajzs és Margit Kiss (2010.). *Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára.* Budapest: Tinta Könyvkiadó.

Shen, Hong és Anoop Sarkar (2005.). „Voting Between Multiple Data Representations for Text Chunking”. *Proceedings of the Advances in Artificial Intelligence, 18th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2005, Victoria, Canada, May 9-11, 2005*. Szerk. Balázs Kégl és Guy Lapalme. 3501. Lecture Notes in Computer Science. Springer, 389–400. old.

Tjong Kim Sang, Erik F. és Sabine Buchholz (2000.). „Introduction to the CoNLL-2000 Shared Task: Chunking”. *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*. ConLL ’00. Lisbon, Portugal: Association for Computational Linguistics, 127–132. old.

Vincze, Veronika, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin és János Csirik (2010.). „Hungarian Dependency Treebank”. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*. Szerk. Nicoletta Calzolari (Conference Chair), Khalid Cho-

ukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner és Daniel Tapias. Valletta, Malta: European Language Resources Association (ELRA), 1855–1862. old. isbn: 2-9517408-6-7.

Vossen, Piek, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge és Wim Peters (1998.). *The EuroWordNet base Concepts and Top Ontology*. Tech. rep.