**Doktori (PhD) értekezés tézisei**

LIGETI-NAGY NOÉMI

# A MAGYAR FŐNÉVI CSOPORT VÉGZŐDÉSE SZÁMÍTÓGÉPES MEGKÖZELÍTÉSBŐL

Pázmány Péter Katolikus Egyetem
Bölcsészet- és Társadalomtudományi Kar
Nyelvtudományi Doktori Iskola

**Nyelvtechnológia Műhely**

**Témavezető:**
**Prof. Dr. Prószéky Gábor**
egyetemi tanár, akadémikus

**Budapest**
**2021**

## 1. Célkitűzés, a kutatás háttere

A dolgozat a magyar főnévi csoport (NP) szerkezetével, azon belül az NP végződéséhez kapcsolódó négy jelenséggel foglalkozik. A korpuszvizsgálatokon alapuló új elméleti megállapítások és gyakorlati eredmények a magyar nyelv automatikus szintaktikai elemzésében, illetve a számítógépes nyelvfeldolgozás egyik feladatában, az ú.n. „NP-chunking" (névszóicsoport-azonosítás) során hasznosíthatóak.

A számítógépes megközelítés kiindulópontja egy elemző, az AnaGramma (Prószéky – Indig, 2015; Prószéky et al., 2016), és az ezt megelőző, ennek létrejöttét támogató nyelvészeti kutatások köre. Az AnaGramma célja, hogy az emberi szövegfeldolgozást modellálja, a szöveget balról jobbra, szóról szóra haladva dolgozva fel. A disszertációmban bemutatott kutatások mind az AnaGramma elveit szem előtt tartva készültek.

A következő jelenségeket vizsgálom:

- Amikor nincs, ami jelezze egy NP végét: az esetragnélküliség és az esetrag nélküli névszók szerepe az elemzésben.
- Az esetrag nélküli névszók egy csoportjával külön foglalkoztam. Ez egy, a főnévi csoportok belsejét érintő probléma: az egy tulajdonnévből és egy köznévből álló főnévi csoportokat vizsgáltam (*Angela Merkel kancellár*).
- A főnévi csoportok végét jelző elemek:
  - Helyhatározói esetragok: a mondatban betöltött szabad határozói szerepük alapján kategorizálom őket.

– A magyar névutók: a szakirodalom olykor egymásnak ellentmondó szempontrendszerének bemutatása után hat disztribúciós tulajdonság alapján csoportosítom a névutókat.

## 2. Adatok és módszerek

Mind a négy nyelvi jelenséget nagyjából a következő lépésekben vizsgáltam:

– Mit mond a szakirodalom a jelenségről? (Ez egy hosszabb szakirodalmi áttekintést takar.)
– Mit mond a korpusz? (Ezekben a fejezetrészekben általában egy korpuszvezérelt adatgyűjtés eredménye került bemutatásra.)
– Mit tudunk meg a jelenségről a korpuszadatok alapján? (A fejezetek egyik legfontosabb egysége: ebben a részben elemeztem a korpuszból kinyert adatokat.)
– Hogyan lehetne az adott jelenséget az AnaGramma elemzési folyamatában kezelni? (Végül, ahol lehetséges volt, javaslatot tettem arra, hogy azokat a főnévi csoportokat, amelyek érintettek a kérdéses jelenségben, hogyan dolgozza fel az elemző.)

A vizsgálat során használt korpusz egyike a Magyar nemzeti szövegtár (MNSz2.0, Oravecz et al., 2014), melynek legnagyobb előnye tekintélyes méretén (1.5 milliárd token) túl a keresőfelület,

amely lehetőséget biztosít arra, hogy az annotáció bármely rétegében konplex lekérdezéseket hajthassunk végre.

Mivel az itt bemutatott kutatáns minden szegmense a főnévi csoportokra koncentrál, egy szintaktikailag elemzett, vagy legalábbis az összetevők szintjén elemzett korpuszra is szükség volt. Erre a célra a Szeged Treebanket használtam (Csendes et al., 2005). A Szeged Treebank 1.0 verziójában a főnévi csoportok és a tagmondatok vannak azonosítva. A Szeged Treebank 2.0-ban a frázisstruktúra-címkék mellett az igék és vonzataik közötti nyelvtani viszonyok is jelölve vannak. A Szeged Treebank 2.0-ból előállított Szeged Dependency Treebank pedig a mondatok függőségi elemzett változatát tartalmazza.

## 3. A dolgozat szerkezete és főbb tézisei

A dolgozat 6 fejezetből áll: egy bevezetésből, négy fejezetből, melyek a fent említett négy jelenséget járják körbe, és egy összefoglalásból. Az első fejezet bemutatja az AnaGramma működési elveit és ismerteti a kutatás során használt korpuszokat. Egy alfejezet az NP-chunking feladattal és annak magyar nyelvvel kapcsolatos kihívásaival foglalkozik, számba véve az eddig a magyar nyelvre írt NP-azonosító algoritmusokat (többek között Váradi, 2003; Hócza, 2004; Recski – Varga, 2012).

A második fejezet azokra az esetekre fókuszál, amikor a főnévi csoport végét nem jelzi semmi. Bemutatom a *nom-or-what* nevű, általam tervezett és implementált algoritmust, amely a testes

esetragot magukon nem viselő névszók mondatbeli szerepének egyértelműsítését végzi el egy kételemű, előretekintő elemzési ablak információi alapján. Az algoritmus tervezéséhez szükség volt annak tisztázására, milyen szerepeket tölthet be a mondatban egy esetrag nélküli névszó. Az implementált algoritmust egy 1 000 mondatból álló, kézzel annotált korpuszon teszteltem, melyen magas pontosságot ért el. Ennek az algoritmusnak egy továbbfejlesztett változatát is bemutatom, amelyben az eredeti szabályok kiegészültek a predikatív névszó detektálására írt szabályokkal (ezeket Dömötör Andrea írta, ld. Dömötör, 2018). A fejezet főbb eredményei a következők:

– Létrehoztam egy szabályalapú algoritmust (*nom-or-what*), melynek célja az esetrag nélküli névszók egyértelműsítése. Magas pontossággal teljesít: 2 112 esetrag nélküli névszót címkézett meg helyesen, ezzel 92.88% pontosságot (és 93.45%-os fedést, így 93.16%-os F-mértéket) ért el.
– Eredményeim alátámasztják, hogy a kétszavas elemzési ablak megfelelő az itt tárgyalthoz hasonló, lokális elemzési feladatokra. Összehasonlítottam a kézi annotálást, amely csak a kétszavas elemzési ablak alapján döntött azzal a kézi annotálással, amely az egész mondatot figyelembe véve döntött, és azt találtam, hogy az ablakon alapuló kézi annotáció is magas pontosságot ért el (98.26%). Az AnaGramma egyik célja az volt, hogy minden lépésben a lehető legpontosabban döntsön, hogy az elemzés későbbi fázisaiban minél kevesebb visszalépésre és módosításra

legyen szükség. Eredményeim azt mutatják, hogy a kétszavas elemzési ablak használata megfelel ennek az elvárásnak.

A harmadik fejezetben vizsgált probléma tulajdonképpen a főnévi csoportok belsejével kapcsolatos. Egy olyan jelenséget járok körül, amelyet korábban még nem elemeztek, és bár hasonlít a hátravetett jelzőre, mégis különbözik attól: ezek azok az NP-k, amelyek egy tulajdonnévből és egy köznévből állnak (pl. *Angela Merkel kancellár*). A szerkezetet *Extended Named Entity*-nek (XNE) hívom. Szintaktikailag elemzett korpusz segítségével gyűjtöttam hasonló szerkezeteket. Megvizsgáltam, hogy milyen szavak ékelődhetnek a névelemek tulajdonnévi és köznévi része közé: *Angela Merkel* ***német*** *kancellár*). Disszertációm ezen fejezetének főbb eredményei a következők:

- A korpuszból kinyert XNE-k köznévi eleméről megállapítottam, hogy 6 kategóriába sorolhatóak: 1) a *néven*, *címmel* típusúak, 2) a földrajzi köznevek, 3) udvariassági formulák, megszólítások, 4) foglalkozások, 5) intézménynevek, 6) márkanév – típusnév párok.
- Megmutattam, hogy a tulajdonnévi és a köznévi elem közé az első két kategóriába eső XNE-k esetében nem ékelődhet semmi.
- A többi kategóriánál azonban előfordulhat módosító a köznévi elem előtt. Ez a következő 7 típus egyike lehet: 1) maga a köznévi végződés összetett, és több szóból áll, 2) a módosító a köznévi elem jelentését tovább specifikálja

(*domonkos szerzetes*), 3) a módosító a működés helyéről mond valamit, 4) a módosító a származás helyét specifikálja, 5) a módosító a működés relatív idejéről mond valamit, 6) a működés pontos idejét határozza meg, 7) a köznévi elem valamely egyéb attribútumát jelzi.

A negyedik fejezet a helyhatározói esetragokkal foglalkozik. Bemutatok egy olyan annotációt, amely alkalmas lehet arra, hogy egy kérdés-válasz rendszer számára (ld. Novák et al., 2019) megfelelő tanítóanyagot állítsunk össze vele. A függőségi elemzett korpusznak azon elemire fókuszálok, amelyek szabadhatározók, és a 9 helyhatározói esetrag egyikét hordozzák magukon. 28 kategóriát állapítok meg – alkategóriákkal együtt összesen 50-et, – melyekbe az ezen kritériumnak megfelelő szavak besorolhatóak. Bizonyos esetekben, bizonyos esetraggal egyes szavak más szerepet töltenek be, mint amit az alapértelmezett kategóriájuk meghatározna, így ezeket az alapértelmezettől eltérő szerepeket is címkéznem kellett. Az itt bemutatott kategorizáció megfelelő jegyeket tartalmaz egy kérdés-válasz rendszer tanítóanyagának címkézéséhez. A fejezet főbb eredményei a következők:

– Meghatároztam azt az 50 kategóriát, amelybe a helyhatározói esetragot magukon viselő szabadhatározói szerepű névszók besorolhatók. Ez lényegében 28 olyan határozói szerepet jelöl, melyek jól meghatározható kérdésekre válaszoló egységeket jelentenek.
– 1 097 lemmát kézzel besoroltam a fenti kategóriákba.

– A lemmák alapértelmezett kategóriáján túl tovább specifikáltam azok viselkedését aszerint, hogy az egyes esetragokkal ellátva milyen szerepet töltenek be, külön-külön: mindegyik lemmánál megadom, ha valamilyen, az alapértelmezetten túli szerepkört tölt be egy adott esetraggal.

Az 5. fejezetben a névutószerű elemek részletes, korpuszvezérelt elemzését mutatom be. Összegyűjtöttem, összehasonlítottam és egységesítettem azt az egyébként igen változatos kategorizációs rendszert, amely a nyelvészeti szakirodalomban található a névutókkal kapcsolatban (Kiefer, 1992; Keszler, 2000; É. Kiss, 2002; Dékány, 2012). Ezt követően megvizsgáltam, hogy az ezekben a forrásokban említett névutók korpuszbeli viselkedését mennyiben jellemezhetjük hat, bináris jeggyel, amelyek hat disztribúciós tulajdonságot fednek le. A fejezet főbb eredményei a következők:

– Rendszerezem a főbb nyelvészeti irodalmi tételek névutókról megfogalmazott álláspontját.
– Meghatároztam hat olyan disztribúciós tulajdonságot, amelyek megfelelőek a névutójelöltek viselkedésének leírására. A hat tulajdonság mindegyike említve lett már egyik vagy másik irodalmi tételben, de együtt még nem alkalmazták őket ilyen célokra.
– Megmutattam, hogy a *szemből*, bár több forrás is névutóként hivatkozik rá, nem névutó semmilyen tekintetben. Viselkedése egyszerűen nem is értékelhető ki a hat

tulajdonság alapján, mivel nem fordul elő a névutókra jellemző pozíciókban, szerkezetekben a korpuszban.

– A névutók három fő csoportját határozom meg a jegyvektorok alapján: a tipikus névutók csoportja az 1 1 1 1 1 1 értékeket kapóké (azaz mindegyik tulajdonságukban a névutókra jellemző viselkedést mutatták a korpuszban): mindig közvetlenül követik az esetragot magán nem viselő névszót, kérdőszavas kifejezésekben követik a kérdőszót, megjelenhetnek személyes névmással, mely esetben az egyeztetés a névutón jelenik meg, és ha mutató névmással kombinálódnak, arra átmásolódnak. Az 1 * 1 * * * vektorú szavak mind névutók abban az értelemben, hogy mindig közvetlenül a névszó után állnak (annak esetragosságától függetlenül). Az 1 0 1 * * * vektor az esetadó névutók vektora, amik mindig közvetlenül a névszó után jönnek. Ebbe a csoportba szinte kivétel nélkül olyan szavak tartoznak, melyekben még látható birtokos szerkezetre utaló jel, és datívuszi esetragos névszót vonzanak. A 0 0 0 0 0 0 vektor a kevésbé tipikus névutók csoportját jelzi, amelyik inkább határozószók: az esetragos névszó előtt és után is, tőle távol is megjelenhetnek.
– A névutók fent bemutatott kategorizációjával tovább finomítottam az irodalomban felvázolt csoportosítást. Néhány szó, mely egységesen névutóként szerepel a forrásokban, korpuszbeli viselkedése alapján nem része a tipikus névutók csoportjának. Ezek egy része olyan névutó, ahol a névutó

alapalakja egybeesik annak egyesszám harmadik személyű személyes névmáshoz illesztett alakjával, pl. *elé*. Az eredmények azt is mutatják, hogy a láthatóan birtokos szerkezetű névutók egy külön csoportot alkotnak, és közelebb állnak a tipikus névutókhoz, mint a többi szóhoz – szemben az irodalomban látható vélekedéssel, mely szerint ezek vagy egy nagyobb csoportnak a részei, vagy valamilyen átmeneti csoport, pl. Keszler, 2000).

A magyar névutókkal kapcsolatos érdekes jelenségek sora természetesen tovább bővíthető. Mindegyik fejezetben megemlítek néhány vonatkozó kutatási kérdést, illetve értelemszerűen a főnévi csoportok kezdetének a vizsgálata is várat még magára. Disszertációmban a főnévi csoportok végére, és annak algoritmikus feldolgozására fókuszáltam. Eredményeim tovább finomítják a nyelvészeti szakirodalomban a magyar főnévi csoportokról alkotott képet.

**Hivatkozások**

Csendes, Dóra – Csirik, János – Gyimóthy, Tibor – Kocsor, András (2005). The Szeged Treebank. In V. Matousek et al. (szerk.) *Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005, LNAI 3658*, (pp. 123–131). Springer Verlag

Dékány, Éva Katalin (2012). *A profile of the Hungarian DP: the interaction of lexicalization, agreement and linearization with the functional sequence*. PhD thesis, University of Tromsø.

Dömötör Andrea (2018). Nem mind VP, ami állít – A névszói állítmány azonosítása számítógépes elemzőben. In Zs. Ludányi – V. Krepsz – T. E. Gráczi (szerk.) *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2018*, (pp. 3–10).

É. Kiss, Katalin (2002). *The Syntax of Hungarian*. Cambridge Syntax Guides. Cambridge University Press.

Hócza, András (2004). Noun Phrase Recognition with Tree Patterns. *Acta Cybernetica*, *16*, 611–623.

Keszler, Borbála (2000). *Magyar grammatika.* Nemzeti Tankönyvkiadó.

Kiefer, Ferenc (1992). *Strukturális magyar nyelvtan: Mondattan*. Akadémiai Kiadó.

Novák Attila – Laki László János – Novák Borbála – Dömötör Andrea – Ligeti-Nagy Noémi – Kalivoda Ágnes: Egy magyar nyelvű kérdezőrendszer. In Berend G. – Gosztolya G. – Vincze V. (szerk.): *XV. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, JATEPress, 225–234.

Oravecz, Csaba – Váradi, Tamás – Sass, Bálint (2014). The Hungarian Gigaword Corpus. In N. Calzoari (Conference Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (szerk.) *Proceedings of the Ninth International Conference on Language Resources and*

*Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Prószéky Gábor – Indig Balázs (2015). Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. *Alkalmazott Nyelvtudomány*, *15*(1–2), 29–44.

Prószéky Gábor – Indig Balázs – Vadász Noémi (2016). Performanciaalapú elemző magyar szövegek számítógépes megértéséhez. In Kas Bence (szerk.) *"Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére*, (pp. 223–232). Budapest: MTA Nyelvtudományi Intézet.

Recski, Gábor – Varga, Dániel (2012). Magyar főnévi csoportok azonosítása. *Általános Nyelvészeti Tanulmányok*, *24*. Original document in Hungarian.

Váradi, Tamás (2003). Shallow Parsing of Hungarian Business News. In *Proceedings of the Corpus Linguistics 2003 Lancaster*, (pp. 845–851).

## 4. A témában végzett publikációs tevékenység

**Publikációk az értekezés témakörében:**

2019 Ligeti-Nagy Noémi – Novák Attila: Hol ugat a kutya? Örömében. In Berend G. – Gosztolya G. – Vincze V. (szerk.): *XV. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, JATEPress, 83–96.

2019 Novák Attila – Laki László János – Novák Borbála – Dömötör Andrea – Ligeti-Nagy Noémi – Kalivoda Ágnes: Egy magyar nyelvű kérdezőrendszer. In Berend G. – Gosztolya G. – Vincze V. (szerk.): *XV. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, JATEPress, 225–234.

2019 Attila Novák – László Laki – Borbála Novák – Andrea Dömötör – Noémi Ligeti-Nagy – Ágnes Kalivoda: Creation of a corpus with semantic role labels for Hungarian. In *Proceedings of the 13th Linguistic Annotation Workshop*. Association for Computational Linguistics, 220– 229.

2019 Noémi Ligeti-Nagy – Andrea Dömötör – Noémi Vadász: What does the Nom say? An algorithm for case disambiguation in Hungarian. In Vainumäe, A. – Kaalep, H. (szerk.): *IWCLUL 2019. The fifth International Workshop on Computational Linguistics for Uralic Languages: Proceedings of the Workshop*, 2–41.

2018 Ligeti-Nagy Noémi – Vadász Noémi – Dömötör Andrea – Indig Balázs: Nulla vagy semmi? Esetegyértelműsítés az ablakban. In Vincze Veronika (szerk.): *XIV. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 25–37.

2018 Ligeti-Nagy Noémi: Névutók, előre! Korpuszvezérelt elemzés a névutószerű elemekről. In Vincze V. (szerk.): *XIV. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 52–63.

2016 Ligeti-Nagy Noémi: A főnévi csoportok és ami utánuk marad. Automatikus szintagmakinyerés magyar szövegekből. In Reményi Andrea Ágnes – Sárdi Csilla – Tóth Zsuzsa (szerk.): *Távlatok a mai magyar alkalmazott nyelvészetben*. Budapest: Tinta Könyvkiadó, 249–260.

**Konferencia-előadások az értekezés témakörében:**

2019 XV. Magyar Számítógépes Nyelvészeti Konferencia
(Szeged, 2019. január 24–25.)
Előadás: *Hol ugat a kutya? Örömében* (Novák Attila társszerzővel)

2019 5th International Workshop on Computational Linguistics for Uralic Languages
(Tartu, Észtország, 2019. január 7–8.)

Poster: *What does the Nom say? An algorithm for case disambiguation in Hungarian* (Dömötör Andrea és Vadász Noémi társszerzőkkel)

2018 19th International Conference on Computational Linguistics and Intelligent Text Processing
(Hanoi, Vietnám, 2018. március 18–24.)
Poszter: *Corpus-driven Study on Hungarian Postpositions*

2018 XIV. Magyar Számítógépes Nyelvészeti Konferencia.
(Szeged, 2018. január 18–19.)
Előadás: *Nulla vagy semmi? Esetegyértelműsítés az ablakban* (Vadász Noémi, Dömötör Andrea és Indig Balázs társszerzőkkel)
Előadás: *Névutók, előre! Korpuszvezérelt elemzés a névutószerű elemekről*

2016 „Nyelv – Nyelvtechnológia – Nyelvpedagógia: 21. századi távlatok" XXV. Magyar Alkalmazott Nyelvészeti Kongresszus.
(Budapest, 2015. március 30 – április 1.)
Előadás: *A főnévi csoportok és ami utánuk marad – automatikus szintagmakinyerés magyar szövegekből*

2014 „Többnyelvűség és kommunikáció Közép-Kelet-Európában" XXIV. Magyar Alkalmazott Nyelvészeti Kongresszus
Előadás: *Szövegkorpuszok pontosabb annotációja gépi elemzéshez*