# Számítógépes anafora- és koreferenciafeloldás magyar nyelvű szövegeken

**Vadász Noémi**

témavezető: Prószéky Gábor

Pázmány Péter Katolikus Egyetem
Bölcsészettudományi Kar
Nyelvtudományi Doktori Iskola

2024

# A disszertáció felépítése és a tézisek

A dolgozat a számítógépes anafora- és koreferenciafeloldás témáját járja körül két szemszögből megközelítve: az erőforrások és az eszközök felől. A dolgozat első felében ismertetek néhány erőforrást, amelyekben ezek a nyelvi jelenségek állnak a középpontban. Ezek az erőforrások hozzáadott emberi munkával készültek. Az erőforrások ismertetésekor kitérek az erőforrás megtervezésével, a munkafolyamat felállításával, vagy az annotátorok munkájának hatékonyabbá tételével és a kézi annotáció minőségének kiértékelésével kapcsolatos kérdésekre. Az anafora és a koreferencia jelenségével kapcsolatos erőforrások bemutatása mellett kitérek egy további, a magyar nyelvtechnológia számára hasznos korpuszra is, amelynek építésekor a morfológiai annotáció elkészítéséért voltam felelős.

Az általam készített erőforrások között elsőként egy kézzel annotált korpuszt ismertetek. A KorKor korpusz a hagyományos nyelvi elemzési rétegek (egyértelműsített morfológiai elemzés, függőségi mondatelemzés) mellett anafora- és koreferenciaannotációt tartalmaz, minden elemzési rétege kézzel ellenőrzött minőségű. A reprodukálhatóság jegyében az erőforrás mellett az építés teljes munkafolyamatát, a felhasznált szkripteket és az annotálási útmutatókat is elérhetővé tettem, ami más korpuszok építéséhez is segítséget nyújthat. A KorKor korpusz alkalmasnak bizonyult arra, hogy bekerüljön a CorefUD gyűjteménybe (Nedoluzhko et al., 2022) is, ami standardizált formátumba rendezi a különböző nyelvekre készült koreferenciakorpuszokat is. A fejezethez az alábbi tézisek tartoznak:

**1. tézis: Többrétegű, kézzel ellenőrzött minőségű korpuszt építettem, amelynek fő annotációs rétege az anafora- és koreferenciaannotáció.** A munka eredményeként létrejött korpusz, a KorKor mellett nemcsak a munkafolyamat pontos dokumentációját és az annotálási útmutatókat publikáltam, hanem elérhetővé tettem a korpuszépítéshez használt eszközöket is,

amelyeket szintén magam fejlesztettem. A dokumentációk és eszközök lehetővé teszik a korpusz bővítését, vagy alapját képezhetik más típusú annotációval ellátott korpuszok elkészítésének is. A KorKor – bár eltérő annotációs sémákat és címkekészleteket használ –, összevontan is használható előzményével, a SzegedKoref korpusszal (Vincze et al., 2015). A dolgozat fejezeteiben több példát is láthatunk a két korpusz összevont alkalmazására. A tézist alátámasztó publikáció: Vadász (2020); Vadász (2022).

**2. tézis: Társszerzővel feltérképeztem és egységes dokumentációval láttam el a legfontosabb magyar morfológiai címkekészleteket, valamint konvertereket készítettem bizonyos morfológiai címkekészletek között.** A kutatást és a dokumentációkat társszerzővel együtt végeztem, a konvertereket magam implementáltam. A konverterek közül az emmorph2ud és az annak alapján készült emmorph2ud2 konverterek az `emtsv` (Indig et al., 2019; Simon et al., 2020) keretrendszerében is használhatók. Az emmorph2ud konverter használata megkerülhetetlen az egyértelműsítő és a függőségi elemző moduljai között. A tézist alátámasztó publikáció: Vadász és Simon (2019).

**3. tézis: A KorKor korpuszt alkalmassá tettem, hogy beilleszkedjen a CorefUD gyűjteménybe, amelynek célja, hogy összegyűjtse és harmonizálja a különböző nyelvekre készült koreferenciakorpuszokat mind az annotációs sémát, mind pedig a formátumot tekintve.** A korpusz így nagyobb láthatóságot kapott, az egységes annotációs sémának és formátumnak hála pedig más nyelvek koreferenciakorpuszaival is összevethető lett. A CorefUD projekt publikációja Novák et al. (2022) támasztja alá ezt a tézist.

A KorKor projekt eredményeit a társszerzővel készített NYTK-NerKor korpusz építésekor is hasznosítani tudtuk: egyrészt alapanyagként használtuk fel a KorKor anyagát, másrészt a KorKor építési folyamatát, az építéskor készült eszközöket és útmutatókat is fel tudtuk használni. Az NYTK-NerKor korpusz rövid ismertetése mellett részletesebben kitérek az egyértelműsített

morfológiai elemzés annotációs rétegének elkészítésére és a munkafolyamat során szerzett tapasztalatokra. A fejezet az alábbi tézishez kapcsolódik:

**4. tézis: Társszerzővel együtt elkészítettük az NYTK-NerKor korpuszt, ami az egymillió tokenes méretével jelenleg a legnagyobb magyar névelemannotált korpusz.** A korpusz elemzési rétegei gold standard minőségűek. A korpusz építése során az egyértelműsített morfológiai elemzési réteg megtervezéséért és elkészítéséért voltam felelős. A morfológiai elemzési réteg elkészítéséhez a KorKor korpusz építésekor kidolgozott munkafolyamat releváns szakaszára és az ahhoz kapcsolódó eszközökre támaszkodtam, valamint a KorKor teljes anyagát beépítettem az NYTK-NerKor korpuszba. A tézist alátámasztó publikáció: Simon és Vadász (2021).

Egy másik erőforrás, amit a doktori disszertáció ismertet, egy társszerzővel készített sémagyűjtemény-készlet. Elkészült a Winograd-sémák (Levesque et al., 2012) magyar fordítása, valamint egy párhuzamos gyűjtemény, ami a Winograd-sémák elérhető fordításait tartalmazza hét nyelven. Emellett elkészült a Definite Pronoun Resolution Dataset (Rahman és Ng, 2012) 1 882 mondatának magyar fordítása is. A munka nem pusztán fordítási feladat volt, minden egyes séma esetében figyelni kellett arra, hogy az eredményül kapott magyar megfelelőben megőrződjön a Winograd-sémákra jellemző szerkezeti többértelműség. A fejezet az alábbi tézishez kapcsolódik:

**5. tézis: Társszerzővel együtt elkészítettem a Winograd-sémák magyar fordításait és néhány további hasonló erőforrást.** A sémagyűjtemények hiánypótlók, hiszen eddig nem létezett hasonló magyar erőforrás. A többértelmű névmási anafora feloldásának feladata jó indikátora lehet annak, hogy egy nyelvmodell mennyire képes megérteni a nyelvet, hiszen a feloldáshoz a grammatikai szerkezetek felismerésén túl szükség van lexikális ismeretekre, világismeretre és következtetési képességre is, éppen ezért volt szükség ezeknek az erőforrásoknak az elkészítésére, amelyek alkalmasak a nagy ne-

urális nyelvmodellek kiértékelésére. A tézist alátámasztó publikáció: Vadász és Ligeti-Nagy (2022).

A dolgozat második felében olyan eszközöket mutatok be, amelyek az anafora- vagy a koreferenciafeloldás feladatát, vagy a hozzájuk köthető alfeladatok egyikét, a zérónévmások beillesztését oldják meg. Az általam készített, a zérónévmások beillesztését végző program egyszerű szabályokon alapul és működésekor az elemzett szöveg többi elemzési rétegére támaszkodik, tehát morfológiai és szintaktikai információkra. A zérónévmások beillesztésére egy másik módszerrel, neurális nyelvmodell, a huBERT (Nemeskey, 2021) finomhangolásával is kísérletet tettem. A fejezet az alábbi tézist fejti ki:

**6. tézis: Zérónévmás-beszúrót készítettem magyar nyelvre szabályalapú és neurális módszerekkel.** A magyarra eddig nem létezett olyan eszköz, amely a zérónévmások beillesztését végezte volna el. A szabályalapú zérónévmás-beszúrót a KorKor építéséhez előannotáció elkészítésére fejlesztettem ki, majd az elkészült korpuszt tanító- és kiértékelőadatként egy neurális megoldás elkészítésére használtam fel. A szabályalapú zérónévmás-beszúró emZero néven az `emtsv` keretrendszerében is használható. A tézist alátámasztó publikáció: Simon et al. (2020).

Egy másik, a dolgozatban bemutatott szabályalapú algoritmus a névmási anafora antecedenskeresését végzi. Szintén morfológiai és szintaktikai információkra támaszkodik, a mögöttes szabályrendszer alapja pedig az ún. Pléh-Radics algoritmus (Pléh és Radics, 1976). Az algoritmus eredetileg az AnaGramma (Prószéky és Indig, 2015) elemzőrendszer keretei között készült, később ezt az algoritmust fejlesztettem tovább, és az ez alapján készült szkriptet használtam a KorKor korpusz építése során az előannotáció elkészítésére. A fejezethez tartozó tézis:

**7. tézis: A Pléh-Radics algoritmus alapján elkészítettem egy anaforafeloldó algoritmust, amely az AnaGramma elemzőrendszer működési alapelveinek is megfelel, majd az algoritmust implementáltam is.** Az AnaGramma kereteibe illeszthető algoritmus a humán szövegfeldolgozás főbb jellemzőit veszi figyelembe. Az algoritmus alapján készített szabályalapú program bizonyos névmások antecedensét azonosítja az előzményben. A program a KorKor korpusz releváns annotációs rétegének előannotálására készült, hogy meggyorsítsa és megkönnyítse a humán annotátorok munkáját. Az anaforafeloldó elkészítését az indokolta, hogy ugyan több magyar megoldás is fellelhető volt a szakirodalomban, sajnos egyik sem volt hozzáférhető. A tézist alátámasztó publikációk: Vadász (2017, 2020); Vadász (2022).

A koreferenciafeloldás feladatára egy neurális megoldást is ismertetek, amelyet társszerzőmmel a huBERT finomhangolásával készítettünk. A finomhangoláshoz a két elérhető magyar nyelvű koreferenciakorpuszt, a SzegedKoref korpuszt (Vincze et al., 2018) és a disszertációban bemutatott KorKor korpuszt egységesítve használtuk. A fejezet az alábbi tézishez tartozik:

**8. tézis: Társszerzővel együtt neurális anaforafeloldót fejlesztettünk.** Az anaforafeloldó elkészítéséhez rendelkezésünkre állt megfelelő mennyiségű és minőségű tanítóanyag (a KorKor és a SzegedKoref korpusz), valamint finomhangolható mély neurális nyelvmodell (huBERT). Ennek ellenére mégsem létezett korábban neurális koreferenciafeloldó magyar nyelvre. A megoldásunk ezt a hiányt pótolja. A tézist alátámasztó publikáció: Vadász és Nyéki (2023).

Végül ismertetek egy kísérletet, amiben azt vizsgáltam, hogy mennyire alkalmas a ChatGPT a magyar nyelvű szövegek esetében az anaforafeloldás feladatára. A kísérletben a fent említett, a doktori kutatás keretében elkészített sémafordításokat használtam. A fejezethez tartozó tézis:

**9. tézis: Kísérleteket végeztem azzal kapcsolatban, hogy a ChatGPT mennyire teljesít jól a többértelmű anaforák előzményének megtalálásában.** A többértelmű anaforák sikeres feloldása megmutathatja, hogy a nyelvmodell milyen nyelvértési képességekkel rendelkezik. A feladatot a Turing-teszt alternatívájaként is szokták emlegetni, hiszen a sikeres anaforafeloldáshoz világismeretre és következtetési képességre is szükség van. A kísérleteimben a ChatGPT válaszadási stratégiáit is vizsgáltam amellett, hogy mennyire képes sikeresen feloldani a többértelmű névmási anaforákat. A fentiek mellett arra is kerestem a választ, hogy milyen következetesen kapunk helyes választ a feltett kérdéseinkre. A tézist alátámasztó publikáció: Vadász (2023).

A dolgozatban ismertetett erőforrásokat és eszközöket a reprodukálhatóság jegyében készítettem. Igyekeztem hangsúlyt fektetni a dokumentációra, hogy az általam készített korpuszok és programok más projektek számára is hasznosíthatók legyenek. Az annotációs útmutatók amellett, hogy lehetővé teszik a korpuszok bővítését a későbbiekben, a korpuszok annotációjának megértését is segítik a korpusz felhasználói számára. Szintén az elérhetőséget és az újrafelhasználhatóságot segíti, hogy az erőforrások szabadon hozzáférhető szövegek felhasználásával készültek, az eszközök pedig szintén nyílt licenc alatt érhetők el.

A doktori munka során az eszközök és erőforrások elkészítésekor több szemszögből jártam körül a számítógépes anafora- és koreferenciafeloldás témáját, valamint az ezekhez kapcsolódó egyéb alfeladatokat.

# A témában végzett publikációs tevékenység

Novák Michal, Popel Martin, Žabokrtský Zdeněk, Zeman Daniel, Nedoluzhko Anna, Acar Kutay, Bourgonje Peter, Cinková Silvie, Cebiroğlu Eryiğit Gülşen, Hajič Jan, Hardmeier Christian, Haug Dag, Jørgensen Tollef, Kåsen Andre, Krielke Pauline, Landragin Frédéric, Lapshinova-Koltunski Ekaterina, Mæhlum Petter, Martí M.Antònia, Mikulová Marie, Nøklestad Anders, Ogrodniczuk Maciej, Øvrelid Lilja, Pamay Arslan Tuğba, Recasens Marta, Solberg Per Erik, Stede Manfred, Straka Milan, Toldova Svetlana, Vadász Noémi, Velldal Erik, Vincze Veronika, Zeldes Amir és Žitkus Voldemaras. Coreference in Universal Dependencies 1.1 (CorefUD 1.1), 2022. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Simon Eszter, Indig Balázs, Kalivoda Ágnes, Mittelholcz Iván, Sass Bálint és Vadász Noémi. Újabb fejlemények az e-magyar háza táján. In: Berend Gábor, Gosztolya Gábor és Vincze Veronika szerk. *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020*, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2020, 29–42.

Simon Eszter és Vadász Noémi. Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus. In: Ekstein Kamil, Pártl Frantisek és Konopík Miloslav szerk. *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, ProceedingsLecture Notes in Computer Science*, 12848. kötet. Springer, 2021, 222–234.

Vadász Noémi. Building a Manually Annotated Hungarian Coreference Corpus: Workflow and Tools. In: *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, Gyeongju, Republic of Korea. Association for Computational Linguistics, October, 2022, 38–47.

Vadász Noémi. Resolving Hungarian Anaphora with ChatGPT. In: Ekštein Kamil, Pártl František és Konopík Miloslav szerk. *Text, Speech, and Dialogue*, Cham. Springer Nature Switzerland, 2023, 45–57.

Vadász Noémi. Anaforafeloldás menet közben – névmások egy pszicholingvisztikailag motivált elemzőben. In: Ludányi Zsófia szerk. *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2017: XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia*. MTA Nyelvtudományi Intézet, Budapest, 2017, 192–205.

Vadász Noémi. KorKorpusz: kézzel annotált, többrétegű pilotkorpusz építése. In: Berend Gábor, Gosztolya Gábor és Vincze Veronika szerk. *XVI. Magyar Számítógépes Nyelvészeti Konfe-*

*rencia (MSZNY 2020*, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2020, 141–154.

Vadász Noémi és Ligeti-Nagy Noémi. Winograd schemata and other datasets for anaphora resolution in Hungarian. *Acta Linguistica Academica*, 2022.

Vadász Noémi és Nyéki Bence. Koreferenciafeloldás magyar szövegeken BERT-tel. In: Berend Gábor, Gosztolya Gábor és Vincze Veronika szerk. *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022*, Szeged. Szegedi Tudományegyetem Informatikai Intézet, 2023, 119–131.

Vadász Noémi és Simon Eszter. Konverterek magyar morfológiai címkekészletek között. In: Berend Gábor, Gosztolya Gábor és Vincze Veronika szerk. *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019)*, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2019, 99–112.

# Hivatkozások

Indig Balázs, Sass Bálint, Simon Eszter, Mittelholcz Iván, Vadász Noémi és Makrai Márton. One format to rule them all – The `emtsv` pipeline for Hungarian. In: *Proceedings of the 13th Linguistic Annotation Workshop*, Florence, Italy. Association for Computational Linguistics, 2019, 155–165.

Levesque Hector J., Davis Ernest és Morgenstern Leora. The Winograd Schema Challenge. In: *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. AAAI Press, 2012, 552–561.

Nedoluzhko Anna, Novák Michal, Popel Martin, Žabokrtský Zdeněk, Zeldes Amir és Zeman Daniel. CorefUD 1.0: Coreference Meets Universal Dependencies. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, June, 2022, 4859–4872.

Nemeskey Dávid Márk. Introducing `huBERT`. In: Berend Gábor szerk. *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, Szeged. 2021, 3–14.

Pléh Csaba és Radics Katalin. „Hiányos mondat", pronominalizáció és a szöveg. *Általános Nyelvészeti Tanulmányok*, 1976, 11(1):261–277.

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

Rahman Altaf és Ng Vincent. Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, 777–789.

Simon Eszter, Indig Balázs, Kalivoda Ágnes, Mittelholcz Iván, Sass Bálint és Vadász Noémi. Újabb fejlemények az e-magyar háza táján. In: Berend Gábor, Gosztolya Gábor és Vincze Veronika szerk. *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020*, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2020, 29–42.

Vincze Veronika, Hegedűs Klára és Farkas Richárd. SzegedKoref: kézzel annotált magyar nyelvű koreferenciakorpusz. In: Tanács Attila, Varga Viktor és Vincze Veronika szerk. *XI. Magyar Számítógépes Nyelvészeti Konferencia*. SZTE TTIK Informatikai Tanszékcsoport, 2015, 312–322.

Vincze Veronika, Hegedűs Klára, Sliz-Nagy Alex és Farkas Richárd. SzegedKoref: A Hungarian Coreference Corpus. In: *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association, May, 2018.