Computational anaphora and coreference resolution in Hungarian texts

Noémi Vadász

supervisor: Gábor Prószéky

Pázmány Péter Catholic University
Faculty of Humanities and Social Sciences
Doctoral School of Linguistics

2024

Structure of the dissertation, theses

The thesis coveres the topic of computational anaphora and coreference resolution, approached from two perspectives: from the perspective of resources and tools. In the first half of the thesis, I described some resources that focus on these linguistic phenomena. These resources were created using human annotation. In describing the resources, I cover the issues related to the design of the resource, setting up the workflow or making the annotators' work more efficient and evaluating the quality of the manual annotation. In addition to presenting resources on the phenomena of anaphora and coreference, I also discuss another corpus useful for Hungarian language technology, which I was responsible for building in order to produce morphological annotation.

One of the resources I have built is a manually annotated corpus. KorKor corpus contains an anaphora and coreference annotation in addition to the traditional linguistic analysis layers (disambiguated morphological analysis, syntactic annotation), and all its analysis layers are of hand-checked quality. In order to ensure reproducibility, I have made available the complete workflow, the scripts used and the annotation guidelines alongside the resource, which may also be helpful for the construction of other corpora. The KorKor corpus has also proven suitable for inclusion in the CorefUD collection [Nedoluzhko et al., 2022], which organizes coreference corpora for different languages in a standardized format. The following theses are associated with this chapter:

Thesis 1: I built a multi-layered, hand-checked quality corpus with anaphora and coreference annotation as the main annotation layer. In addition to the resulting corpus, KorKor, I have not only published precise documentation of the workflow and annotation guidelines, but also made available the tools used for corpus construction, which I also developed my-

self. The documentation and tools can be used to extend the corpus or as a basis for other types of annotated corpus. KorKor – although it uses different annotation schemes and tagsets – can be used in combination with its predecessor, the SzegedKoref corpus [Vincze et al., 2015]. In the chapters of this thesis, several examples of the combined use of the two corpora are presented. The publication supporting this thesis is: Vadász [2020], Vadász [2022].

Thesis 2: With co-author I mapped and provided uniform documentation of the most important Hungarian morphological tagsets, and created converters between certain morphological tagsets. The research and documentation was done together with co-author, the converters were implemented by myself. Among the converters, emmorph2ud and the emmorph2ud2 converters based on it can be used in the `emtsv` [Indig et al., 2019, Simon et al., 2020] pipeline. The use of the emmorph2ud converter is inescapable between the disambiguation and dependency parser modules. The publication supporting this thesis is: Vadász and Simon [2019].

Thesis 3: I have made the KorKor corpus suitable for inclusion in the CorefUD collection, which aims to collect and harmonize coreference corpora in different languages, both in terms of annotation scheme and format. The corpus has thus been given greater visibility and, thanks to the uniform annotation scheme and format, it has been made comparable with coreference corpora of other languages. The CorefUD project publication [Novák et al.] supports this thesis.

The results of the KorKor project were also used in the construction of the NYTK-NerKor corpus with co-author: on the one hand, we used the KorKor corpus as a raw material, and on the other hand, we were able to use the KorKor construction workflow, tools and guidelines. In addition to a brief description of the NYTK-NerKor corpus, I go into more detail on the anno-

tation layer of the disambiguated morphological analysis and the experience gained during the work process. The chapter is linked to the following thesis:

Thesis 4: Together with co-author we have created the NYTK-NerKor corpus, which is currently the largest Hungarian named entity corpus with its one million tokens. The analysis layers of the corpus are of gold standard quality. During the construction of the corpus, I was responsible for the design and preparation of the disambiguated morphological analysis layer. For the construction of the morphological analysis layer, I relied on the relevant stage of the workflow and associated tools developed during the construction of the KorKor corpus, and incorporated the complete KorKor material into the NYTK-NerKor corpus. The publication supporting this thesis is: Simon and Vadász [2021].

Another resource, described in the doctoral thesis, is a collection of schemas created with a co-author. A Hungarian translation of the Winograd schemas [Levesque et al., 2012] has been produced, as well as a parallel collection of available translations of Winograd schemas in seven languages. In addition, the Hungarian translation of 1 882 sentences of the Definite Pronoun Resolution Dataset [Rahman and Ng, 2012] has been completed. The work was not just a translation task, for each schema care had to be taken to ensure that the resulting Hungarian equivalent preserved the structural ambiguity characteristic of Winograd schemas. The chapter is linked to the following thesis:

Thesis 5: Together with co-author, I have prepared the Hungarian translations of the Winograd schemas and some other similar resources. The schema collections are a niche, as no similar Hungarian resource has existed so far. The task of resolving the ambiguous pronominal anaphora can be a good indicator of the ability of a language model to understand the language,

since in addition to the recognition of grammatical structures, the resolution requires lexical knowledge, world knowledge and inference skills, which is why it was necessary to prepare these resources, which are suitable for evaluating large neural language models. The publication supporting this thesis: Vadász and Ligeti-Nagy [2022].

In the second half of the paper, I present tools that solve the task of anaphora or coreference resolution, or one of their related subtasks, the insertion of zero pronouns. The algorithm I have created for inserting zero nouns is based on simple rules and relies on the other analysis layers of the text being analysed, i.e. morphological and syntactic information. I have also experimented with another method of inserting zero pronouns by fine-tuning a neural language model, huBERT [Nemeskey, 2021]. The chapter develops the following thesis:

Thesis 6: I created tools for inserting zero pronouns using rule-based and neural methods. No tool for inserting zero pronouns has existed for Hungarian so far. I developed a rule-based zero pronoun insertion tool for building a pre-annotation for KorKor, and then used the corpus for fine-tuning a neural model and evaluating it. The rule-based zero pronoun insertion can also be used in the `emtsv` framework as emZero. The publication supporting the thesis: Simon et al. [2020].

Another rule-based algorithm presented in the thesis performs antecedent search for the pronominal anaphora. It also relies on morphological and syntactic information and the underlying rule system is based on the so-called Pléh-Radics algorithm [Pléh and Radics, 1976]. The algorithm presented here was originally developed in the framework of the AnaGramma [Prószéky and Indig, 2015] text processing system, and later on I improved this algorithm and used the script based on it to perform the pre-

annotation during the construction of the KorKor corpus. Thesis for this chapter:

Thesis 7: Based on the Pléh-Radics algorithm, I have constructed an anaphora resolution algorithm that also satisfies the principles of the AnaGramma analysis system, and implemented is al well. The algorithm, which fits into the AnaGramma framework, takes into account the main features of human text processing. A rule-based program based on the algorithm identifies the antecedent of certain pronouns. The program is designed to pre-annotate the relevant annotation layer of the KorKor corpus to speed up and facilitate the work of human annotators. The rationale for this work was that, although several Hungarian solutions were available in the literature, unfortunately none were accessible. Publications supporting this thesis: Vadász [2017, 2020], Vadász [2022].

A neural solution to the problem of coreference resolution is introduced as well, which my co-author and I created by fine-tuning huBERT. For the fine-tuning, we used the two available Hungarian coreference corpora, the SzegedKoref corpus [Vincze et al., 2018] and the KorKor corpus presented in the dissertation, in a unified way. The chapter is related to the following thesis:

Thesis 8: Co-author and I fine-tuned a neural language model for the task of coreference resolution. To develop our solution, we had at our disposal a sufficient amount and quality of training material (the KorKor and SzegedKoref corpora) and a fine-tunable deep neural language model (huBERT). Nevertheless, no neural coreference resolver for Hungarian has been available before. Our solution fills this gap. The publication supporting this thesis: Vadász and Nyéki [2023].

Finally, I presented an experiment to investigate the suitability of ChatGPT for the task of anaphora resolution for Hungarian texts. In the exper-

iment, I used the above-mentioned schema translations prepared as part of the doctoral research. Thesis for this chapter:

Thesis 9: I have conducted experiments on how well ChatGPT performs in finding the antecedents of ambiguous pronouns. Successfully resolving ambiguous anaphors can show the language understanding capabilities of the language model. The task is also referred to as an alternative to the Turing test, since successful anaphora resolution requires both world knowledge and inference skills. In my experiments, I also tested the response strategies of ChatGPT in addition to its ability to successfully resolve ambiguous pronominal anaphors. In addition to the above, I was also looking to see how consistently we get correct answers to the questions we ask. The publication supporting this thesis: Vadász [2023].

The resources and tools described in this thesis are designed with reproducibility in mind. I have tried to focus on documentation so that the corpora and tools I have created can be used for other projects. The annotation guidelines, in addition to allowing the corpus to be extended in the future, also help the users of the corpus to understand the annotation of the corpus. Accessibility and reusability are also facilitated by the fact that the resources are created using open access texts and the tools are also available under an open licence.

In the course of my doctoral work, I have explored the topic of computer anaphora and coreference resolution, as well as other related sub-tasks, from several angles in the development of the tools and resources.

Relevant publications

Noémi Vadász. Korkorpusz: kézzel annotált, többrétegű pilotkorpusz építése. In Gábor Berend, Gábor Gosztolya, and Veronika Vincze, editors, *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020*, pages 141–154, Szeged, 2020. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Noémi Vadász. 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, October 2022. Association for Computational Linguistics.

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

Eszter Simon and Noémi Vadász. Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus. In Kamil Ekstein, Frantisek Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings*, volume 12848 of *Lecture Notes in Computer Science*, pages 222–234. Springer, 2021. doi: 10.1007/978-3-030-83527-9_19.

Noémi Vadász and Noémi Ligeti-Nagy. Winograd schemata and other datasets for anaphora resolution in hungarian. *Acta Linguistica Academica*, 2022. doi: https://doi.org/10.1556/2062.2022.00575.

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

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

Noémi Vadász and Bence Nyéki. Koreferenciafeloldás magyar szövegeken bert-tel. In Gábor Berend, Gábor Gosztolya, and Veronika Vincze, editors, *XIX. Magyar Számítógépes Nyelvészeti*

Konferencia (MSZNY 2022, pages 119–131, Szeged, 2023. Szegedi Tudományegyetem Informatikai Intézet.

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

References

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

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

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

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

Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman, Anna Nedoluzhko, Kutay Acar, Peter Bourgonje, Silvie Cinková, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, Petter Mæhlum, M.Antònia Martí, Marie Mikulová, Anders Nøklestad, Maciej Ogrodniczuk, Lilja Øvrelid, Tuğba Pamay Arslan, Marta Recasens, Per Erik Solberg, Manfred Stede, Milan Straka, Svetlana Toldova, Noémi Vadász, Erik Velldal, Veronika Vincze, Amir Zeldes, and Voldemaras Žitkus. Coreference in universal dependencies 1.1 (CorefUD 1.1).

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 552–561. AAAI Press, 2012. ISBN 9781577355601.

Altaf Rahman and Vincent Ng. 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*, pages 777–789, 2012.

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

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

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

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