Theses of doctoral (PhD) dissertation

NOÉMI LIGETI-NAGY

THE RIGHT EDGE OF THE HUNGARIAN NP
A COMPUTATIONAL APPROACH

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

Doctoral Programme in Language Technology

Supervisor:
Prof. Gábor Prószéky
Professor, Doctor of Sciences

Budapest
2021

1. The aim and the background of the thesis

This thesis focuses on some central linguistic phenomena related to the right edge of the noun phrases in Hungarian which in some ways proved to be significant in the parsing process of Hungarian texts. The results presented here may be used in the automatic parsing of Hungarian, especially in one of the tasks of computational language processing, the so-called „NP-chunking" (automatic detection of noun phrases in a sentence).

The computational approach originates from the linguistic studies that ground and support the creation of a parser called AnaGramma (Prószéky and Indig, 2015; Prószéky et al., 2016). The aim of AnaGramma was to model human sentence processing by parsing the text word-by-word, from left to right. All the substudies presented here were conducted with AnaGramma's principles in mind. The following issues are addressed here:

- When nothing marks the end of the noun phrase – the cases of "suffixlessness" and their role in the parsing process.
- A problem from inside the noun phrase: noun phrases consisting of a proper name and a common noun (like *Angela Merkel kancellár* 'Angela Merkel chancellor').
- Marked endings of a noun phrase:
    - Locative case suffixes: categorisation with respect to adverbial adjuncts in a sentence.
    - Postpositions in Hungarian: literature review and categorisation.

2. Data and methods

Each linguistic phenomenon is examined more or less by following the process described in the steps below:

- What does the literature reveal about this phenomenon? (This covers a literature review of the topic.)
- What does the corpus say? (A corpus-driven data collection is provided in this section of the chapters.)
- What can be learned about this phenomenon based on the corpus? (This may be the most important part of each substudy; here, I analyse the data retrieved from the corpus.)
- How should the phenomenon be handled in the parsing process of AnaGramma? (Finally, if possible, I provide a suggestion for an algorithm to parse noun phrases that are somehow affected by the phenomenon in question.)

The Hungarian Gigaword Corpus (HGC, Oravecz et al., 2014) proved to be the best choice for most of the tasks. The biggest advantage of HGC (besides its respectable size of 1.5 billion tokens) is the query interface which allows complex searches on every layer of the annotation.

As the studies presented here all focus on noun phrases, a syntactically annotated, or at least shallow parsed corpus is required as well. For this purpose, the Szeged Treebank was used (Csendes et al., 2005). The 2.0 version has a deep phrase-structured syntactic analysis, while the Szeged Dependency Treebank contains a dependency annotation for the sentences.

3. The structure and the main theses of the dissertation

The dissertation is divided into 6 chapters: an introduction, four chapters expounding the four phenomena listed above, and a conclusion. Chapter 1, the introduction, describes the principles of AnaGramma and lists the corpora that was used. A section is dedicated to the task of NP-chunking and its challenges in Hungarian, with a brief enumeration of the algorithms designed to tackle this problem so far (Váradi, 2003; Hócza, 2004; Recski – Varga 2012; among others).

Chapter 2 focuses on cases when nothing marks the end of the noun phrase – the cases of "suffixlessness" and their role in the parsing process. I present an algorithm called *nom-or-what* that specifies the role of the suffixless nominals in the sentence based on the information retrieved from a two-token-wide look-ahead parsing window. The design of the algorithm required the collection of the roles a suffixless nominal may bear in the sentence. The algorithm was tested and evaluated on a test corpus comprised of 1 000 manually annotated sentences with a high precision. I also implemented an upgraded version of the algorithm which included some rules written to find nominative predicates in the sentence (by Andrea Dömötör, see Dömötör, 2018). The main results of this chapter are the following:

- the algorithm itself (*nom-or-what*). A rule-based method the task of which is to disambiguate suffixless nominals. It operates with high precision: the algorithm correctly tagged 2

112 instances of suffixless nominals reaching a precision of 92.88% (and a recall of 93.45%, with an F-measure of 93.16%).

– I supported that a two-token-wide look-ahead parsing window is indeed sufficient in short-range parsing tasks (such as this disambiguation; the role of a suffixless token may be specified with great certainty based on the two-token-wide parsing window). I compared the results of the manual tagging that relied on the parsing window to the results of the manual tagging that identified the whole sentence and showed that the window-based annotation reaches a very high precision (98.26%). AnaGramma's goal was to make decisions as precisely as possible so that in any later phase of the parsing process no backtracking is needed. These results show that the use of a two-token-wide parsing window can meet this expectation.

In Chapter 3 I investigate a problem from inside of a noun phrase. I highlight a phenomenon not analysed before which looks similar to extraposed modifiers but is nevertheless somewhat different; noun phrases consisting of a proper name and a common noun (such as *Angela Merkel kancellár* 'Angela Merkel chancellor'). The structure is referred to as Extended Named Entity (XNE). I collected similar structures from a syntactically annotated corpus to be able to define some categories among these phrases. Furthermore, I inspect what kind of words may fit in between the two parts of these structures: *Angela Merkel* ***német*** *kancellár* 'Angela Merkel **German**

chancellor'. The main results of this part of my thesis are the following:

- the categorization of XNEs: the six categories into XNEs' second part, the common noun may fit are 1) words like *néven* 'called', *címmel* 'titled' (I referred to them as NÉVEN) 2) geographical common nouns, 3) courtesy formulas, 4) occupations, 5) names of institutions, 6) brand name – type name pairs.
- I show that nothing may appear in between the proper name and the common noun in XNEs of the first two categories (NÉVEN and geographical common nouns). The common noun ending of the other four groups, on the other hand, may be modified.
- I distinguish 7 categories of modifiers that may intersect the proper name and the common noun in an XNE: 1) the ending itself is complex, consisting of more than one word, 2) the modifier further specifies the meaning of the common noun, 3) the modifier defines the place of operation, 4) the modifier defines the origin of the given person, 5) the modifier states something about the time of the operation of the given XNE, 6) the modifier specifies the exact time when the operation took place, 7) the modifier refers to some additional attribute of the given person.

Chapter 4 discusses locative case suffixes; their categorisation with respect to adverbial roles in a sentence. In this section I present an annotation that would be appropriate when designing a training corpus for a Question-Answering system (see Novák et al., 2019). I

focus on those elements of the dependency treebank annotated with Obl edge that bear one of the case suffixes of the directional triad of locative suffixes. I define 28 categories – altogether 50 counting the subcategories – into which the words can be sorted. In some cases, with some case suffixes, particular words may play a role different from the one defined by the default category, the labelling of which was also a task. The categorisation presented here provides appropriate features in a training corpus to create a QA system. The main results of this chapter are the following:

- the 50 categories into which adverbial adjuncts with a locative case suffix fit. These cover 28 adverbial roles in a sentence with a well-definable question they may answer.
- I manually sort 1 097 lemmas into these categories.
- In addition to the default categories of the lemmas, I further specify the behaviour of the lemmas with regard to the nine locative case suffixes: I define what additional adverbial role the lemma may have with a given suffix (in addition or instead of its default role).

In Chapter 5 a detailed, corpus-driven analysis of Hungarian postposition-like elements is presented. I collect, compare, and unify the diverse categorisation of postpositions in the linguistic literature (Kiefer, 1992; Keszler, 2000; É. Kiss 2002; Dékány, 2012). Then I examine how six (binary) features characterise these words when studied in a corpus. The numerous postposition candidates could be arranged based on these features. The main results of this study are the following:

- I systematise the main linguistic sources' opinion on postpositions.
- I define six distributional features that are suitable for describing the behaviour of postposition candidates. The six features have already been mentioned in one or more linguistics papers but have not been applied together as a feature list.
- I prove that *szemből* 'from oppose to', though it is considered a postposition in many papers, is not a postposition at all. Its behaviour simply can not be evaluated based on the six features because it does not appear in postpositional places in the corpus.
- I outline three main groups of postpositions. The groups can be described with their feature vector. The group of typical postpositions is the group with a vector 1 1 1 1 1 1, meaning that they always strictly follow a caseless noun, they follow the wh-word in questions, they can appear with a personal pronoun (in which case the agreement marker appears on the postposition) and they are copied onto the demonstrative when combined with it. Words with a vector 1 * 1 * * * are all postpositions in a sense that they always follow the noun strictly adjacently (regardless of the case marking it has). The vector 1 0 1 * * * represents case assigning postposition (always following the noun, adjacently). The group almost exclusively comprises postpositions with a clear possessive structure taking a noun with the dative suffix. The vector

0 0 0 0 0 0 marks the group of less typical postpositions, or adverbs: words appearing both before and after their complement, which bears a lexical case.

– With the categorisation of postpositions I refine the categories drawn in the literature. Some words that are uniformly categorised as typical postpositions in linguistics papers are not part of the typical postpositions based on their appearance in the corpus. On one hand, they are the postpositions the base form of which is homonymous to the one attached to a third person singular personal pronoun: *elé* 'to in front of sg' and 'to in front of him/her'. The results show that postpositions with an overt possessive structure form a separate group and are closer to typical postpositions (1 1 1 1 1 1) than to other words (in contrast to the literature's view, where they are generally a member of a bigger group with other case assigning postpositions; or are only considered a transitional class, see Keszler, 2000).

The list of interesting phenomena of NPs in Hungarian, of course, could be further expanded. I mention some possible research questions in each chapter. There are some promising issues at the beginning of Hungarian noun phrases as well. Here I focused on phenomena influencing the algorithmic detection of the ending of NPs. My results can certainly further refine the image of Hungarian noun phrases drawn in the linguistic literature.

References

Csendes, Dóra – Csirik, János – Gyimóthy, Tibor – Kocsor, András (2005). The Szeged Treebank. In V. Matousek et al. (Eds.) *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 [All that Predicates is not VP – The Identification of Nominal Predicate in Automatic Parsing]. In Zs. Ludányi – V. Krepsz – T. E. Gráczi (Eds.) *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2018*, (pp. 3–10). Original document in Hungarian.

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* [Hungarian grammar]. Nemzeti Tankönyvkiadó. Original document in Hungarian.

Kiefer, Ferenc (1992). *Strukturális magyar nyelvtan: Mondattan* [A Structural Grammar of Hungarian: Syntax]. Akadémiai Kiadó. Original document in Hungarian.

Novák Attila – Laki László János – Novák Borbála – Dömötör Andrea – Ligeti-Nagy Noémi – Kalivoda Ágnes: Egy magyar

nyelvű kérdezőrendszer [A Hungarian Questioning System]. In Berend G. – Gosztolya G. – Vincze V. (Eds.): *XV. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, JATEPress, 225–234. Original document in Hungarian.

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 (Eds.) *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 [Psycholingvistically Motivated Analysis of Hungarian Texts with Computer]. *Alkalmazott Nyelvtudomány*, *15*(1–2), 29–44. Original document in Hungarian.

Prószéky, Gábor – Indig, Balázs – Vadász, Noémi (2016). Performanciaalapú elemző magyar szövegek számítógépes megértéséhez [A Performance-based Parser to the Comprehensive Understanding of Hungarian Texts]. In Kas, B. (Ed.) *"Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére*, (pp. 223–232). Budapest: MTA Nyelvtudományi Intézet. Original document in Hungarian.

Recski, Gábor – Varga, Dániel (2012). Magyar főnévi csoportok azonosítása [Detecting Hungarian Noun Phrases]. *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. Relevant publications

Publications:

2019 Ligeti-Nagy Noémi – Novák Attila: Hol ugat a kutya? Örömében [Where does the dog bark? In a hurry.]. In Berend G. – Gosztolya G. – Vincze V. (Ed.): *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 [A Hungarian Questioning System]. In Berend G. – Gosztolya G. – Vincze V. (Eds.): *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. (Eds): *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 [Zero or Nothing? Case Disambiguation in the Window]. In Vincze Veronika (Ed.): *XIV. Magyar*

Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 25–37.

2018 Ligeti-Nagy Noémi: Névutók, előre! Korpuszvezérelt elemzés a névutószerű elemekről [Postpositions, Come Forward! Corpus-driven Study on Postposition-like Elements]. In Vincze V. (Ed.): *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 [Noun Phrases and What They Leave Behind. Automatic Phrase-extraction from Hungarian texts]. In Reményi Andrea Ágnes – Sárdi Csilla – Tóth Zsuzsa (Eds.): *Távlatok a mai magyar alkalmazott nyelvészetben*. Budapest: Tinta Könyvkiadó, 249–260.

Conference presentations:

2019 XV. Magyar Számítógépes Nyelvészeti Konferencia

(Szeged, 2019. január 24–25.)

Talk: *Hol ugat a kutya? Örömében* (with co-author Attila Novák)

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* (with co-authors Andrea Dömötör and Noémi Vadász)

2018 19th International Conference on Computational Linguistics and Intelligent Text Processing

(Hanoi, Vietnám, 2018. március 18–24.)

Poster: *Corpus-driven Study on Hungarian Postpositions*

2018 XIV. Magyar Számítógépes Nyelvészeti Konferencia.

(Szeged, 2018. január 18–19.)

Talk: *Nulla vagy semmi? Esetegyértelműsítés az ablakban* (with co-authors Noémi Vadász, Andrea Dömötör and Balázs Indig)

Talk: *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.)

Talk: *A főnévi csoportok és ami utánuk marad – automatikus szintagmakinyerés magyar szövegekből*

2014 „Többnyelvűség és kommunikáció Közép-Kelet-Európában" XXIV. Magyar Alkalmazott Nyelvészeti Kongresszus

Talk: *Szövegkorpuszok pontosabb annotációja gépi elemzéshez*