DOI: 10.15774/PPKE.BTK.2021.17

The Right Edge of the Hungarian NP

A Computational Approach

Noémi Ligeti-Nagy

A thesis presented for the degree of
Doctor of Philosophy

Pázmány Péter Catholic University
Doctoral School of Linguistics
Director: **Prof. Balázs Surányi DSc**

Language Technology Program

Supervisor: **Prof. Gábor Prószéky DSc**
Head of the Language Technology Program

Budapest
2021

DOI: 10.15774/PPKE.BTK.2021.17

A magyar főnévi csoport végződése

Számítógépes megközelítésből

Ligeti-Nagy Noémi

Doktori (PhD) értekezés

Pázmány Péter Katolikus Egyetem

Nyelvtudományi Doktori Iskola

Vezetője: **Prof. Surányi Balázs**

egyetemi tanár, az MTA doktora

Nyelvtechnológiai Műhely

Témavezető: **Prof. Prószéky Gábor**

egyetemi tanár, az MTA doktora

a Nyelvtechnológiai Műhely vezetője

Budapest

2021

DOI: 10.15774/PPKE.BTK.2021.17

to Balázs

DOI: 10.15774/PPKE.BTK.2021.17

Acknowledgements

Although I did most of the counting, writing, struggling, despairing during the last few years, many people have contributed to the following pages in some way. I am sincerely grateful to them, including but not limited to:

The other 5 Ligeti's - making the biggest sacrifice by letting a wife and mother sometimes put work at the top of the priority list. Especially to Balázs, who often had to hold up with me as a co-worker, or a roommate in an office, or a professional consultant, upon his everyday duties as a husband and a father. I dedicate this thesis to him.

My supervisor, Professor Gábor Prószéky, for always convincing me that there is a point in what I am doing. And for teaching me the correct pronunciation of *John McLaughlin* (among other, probably more useful and important things).

Professor Katalin É. Kiss, who not only introduced me to the magic of linguistics in 2007, at a lecture on Hungarian generative syntax, but also taught me how to do research at all. I have learnt a lot from all the other teachers at the Faculty of Humanities and Social Sciences during my master and PhD studies: Andrea Reményi, András Cser and Csaba Olsvay have all inspired me with their knowledge and enthusiasm. I am also grateful to Kornél Szovák who unwillingly pushed me towards linguistics with the task he assigned to me during my medieval studies in the Scriptorium specialisation. And a big, wholehearted thank you goes to the staff at the Faculty of Information Technology and Bionics for the welcoming and helpful athmosphere and for the courses on programming they provided.

My opponents: Bálint Sass and Ágoston Tóth who reviewed my dissertation as thoroughly as possible and contributed a total of 20 pages of comments to make my work more accurate and understandable.

The members of the π, the members of the "hive" and the "girls": Noémi Vadász, Andrea Dömötör, Ágnes Kalivoda, László Laki, Attila Novák, Borbála Novák and Zijian

DOI: 10.15774/PPKE.BTK.2021.17

Győző Yang - making it worthwhile to go to work; or just staying home; drinking coffee, having very long lunch breaks, travelling around the globe, being friends while being colleagues.

My bff, Magdolna Gilányi for providing me with a connection to my old love, medieval studies.

My parents, for always being there for me.

And a special thanks goes to Júlia Keresztes - being my friend and my English vocabulary in one.

"My help comes from the Lord, who made heaven and earth." (Psalm 121:2)

DOI: 10.15774/PPKE.BTK.2021.17

Contents

DOI: 10.15774/PPKE.BTK.2021.17

DOI: 10.15774/PPKE.BTK.2021.17

DOI: 10.15774/PPKE.BTK.2021.17

List of Figures

DOI: 10.15774/PPKE.BTK.2021.17

List of Tables

DOI: 10.15774/PPKE.BTK.2021.17

DOI: 10.15774/PPKE.BTK.2021.17

Abstract

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 subtitle of this thesis only foreshadows that the approach used here will be "computational", however, "corpus-driven" is also a defining feature – if not more so – of the research described in the following chapters.

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.

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 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.
- More on suffixless nominals; 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').

DOI: 10.15774/PPKE.BTK.2021.17

- 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.

Although the topics listed above seem to be diverse both in the method they require and in the levels of language their analysis affects, studying and understanding all of them is crucial for any parser.

I begin with an introduction, then I discuss what motivated my thesis, the background of my research will be presented by describing the principles of `AnaGramma` and by encountering the corpora that was used (1). First, I focus on suffixless nominals. I discuss the design and implementation of an algorithm disambiguating them in the sentence (2). In this part of my research I realised the importance of extended named entities, and so the third chapter focuses on those (3). Finally, I turn to the morphemes undoubtedly finalising any noun phrase: case suffixes (4) and postpositions(5). A conclusion finalises my dissertation by presenting multiple ways to continue what was started in this thesis (6).

DOI: 10.15774/PPKE.BTK.2021.17

Foreword

NP is NP and appears NP.

This might be considered the "skeleton" of the sentence *The introduction is the first chapter of your thesis or dissertation and appears after the table of contents*, where all the noun phrases were cut off and replaced by "NP". And this is where my research, summarised in the following chapters, started. I always thought of verbs as being vibrant, lively, the ones aiming to be the ruler of a sentence while noun phrases represent something stable, calm and earnest, the true holders of power instead of (or behind) the verb. At the beginning of my PhD studies, when looking at sentences and sentence skeletons I had to choose between the verbs and their arguments as a research topic, this silly picture of monarchs and rulers in my mind's eye led me to choose the latter without thinking. This is how my relationship with noun phrases started.

DOI: 10.15774/PPKE.BTK.2021.17

Chapter 1

Introduction

" 'My way is to begin at the beginning,'
said Lord Byron, who knew his way
around polite society."

The Pendragon Legend
Antal Szerb

1.1 The subject

This thesis focuses on linguistic phenomena related to the right edge of the noun phrases in Hungarian. These phenomena are significant in the parsing process of Hungarian texts, especially during NP chunking. The computational approach highlighted in the title of the thesis originates from the parser called `AnaGramma` (Prószéky and Indig, 2015; Prószéky et al., 2016). The majority of the substudies presented here was conducted with `AnaGramma`'s principles in mind, and after discussing noun phrases, NP chunking, and other basic concepts (1.2) I will turn to the introduction of this parser (1.3) before presenting the corpora I used for my research (1.4).

1.2 Noun phrases and parsers

1.2.1 Nouns and noun phrases

Nouns are one of the major classes of parts of speech. Noun phrases are the constituents of a sentence that contain a noun - often only one noun in their simplest form (1).

DOI: 10.15774/PPKE.BTK.2021.17

(1) a. *Mary*

b. *book*

c. *scissors*

Nouns can be preceeded by a determiner, can be modified by one or more adjectives and can take prepositions or postpositions within the same constituent (2).

(2) a. *The Standard Book of Spells*

b. *one plain pointed hat*

c. *three sets of plain work robes*

The complexity of these constituents resulted in the distinction of several different phrases within the framework of X Bar theory: D(eterminer) P(hrases), P(repositional) P(hrases) and Post(positional) P(hrases). Here I use the term NP for all of these: NP is a constituent consisting of a noun and an optional determiner, one or more optional modifiers and one optional postposition (or preposition). Unless I state otherwise, NP is the maximal projection of the noun, the top-level NP.

By baseNP I refer to NPs that do not contain any other NP (this is the definition of NP chunks by Ramshaw and Marcus, 1995, see 1.2.3). Maximal NPs, on the other hand, may be made up of two or more baseNPs: coordination (3a), participles (3b) or possessive structures (3c) are examples of these.

(3) a. *the boy and the girl*

b. *the boy waiting for the girl*

c. *the friend of the boy*

1.2.2 Noun phrases in linguistic literature

Many segments of (Hungarian) noun phrases are studied in several papers of which I am going to cite the most relevant ones in each chapter and section of my thesis. However, there are some comprehensive works on Hungarian noun phrases mainly, but not exclusively, with a theoretical background. First, I have to mention the paper of Kornai (1985) in which he aims to describe the internal structure of Hungarian noun phrases (first, the

DOI: 10.15774/PPKE.BTK.2021.17

"easy parts", the lower bar-levels, then the possessive constructions) in a context-free rule schemata. This is the first significant step in the computationally motivated discussion of Hungarian NPs. The rules themselves were further specified and then implemented in Recski (2014) in an attempt to create a rule-based NP chunker for Hungarian.

There are numerous studies on Hungarian NPs from a theoretical background: É. Kiss (2000) also studies the internal structure of Hungarian noun phrases, while Éva Dékány's doctoral thesis (2012) is a broader investigation into noun phrases. The most recent (and most exhaustive) summary of Hungarian noun phrases is the first volume of the series *Syntax of Hungarian* (Alberti and Laczkó, 2018). If necessary, I will cite the relevant results or ascertainment of the above papers in each chapter of my thesis.

1.2.3 Chunks and chunkers

Noun phrases are among the subjects of the task *chunking* in natural language processing, which is the separation and segmentation of a sentence into its constituents.[1] In this phase of the parsing we aim to identify the constituents of the sentence. A definition for various types of chunks can be found in the description of the chunking task of CoNLL 2000 (Tjong Kim Sang and Buchholz, 2000). The most widely used definition of NP chunks, however, is the NP definition of Ramshaw & Marcus (1995). In their groundbreaking paper they focus on non-recursive "baseNPs": baseNPs are NPs that do not contain any other NP. As NP chunking is most widely considered an algorithm whose goal is to provide basic information on sentence structure, most existing tools (chunkers) were designed to identify non-recursive noun phrases. However, complex tasks in natural language processing (NLP) such as information extraction, information retrieval, named entity recognition or machine translation could also benefit from the extraction of maximal constituents, top-level NPs among them (Recski, 2010b: 3).

NP chunking is also known as IOB tagging of NPs. The IOB format (short for inside, outside, beginning) was introduced by Ramshaw and Marcus (1995). The B- prefix indicates the beginning, the I- prefix indicates the inside of a token sequence, respectively. The tag O is used to distinguish tokens belonging to no chunk. The B- tag is used only when it is followed by a tag of the same type without O tokens between them. In an extension

[1] It has to be noted that in some early papers chunks are units that do not necessarily coincide with syntactic constituents (e.g. Abney, 1992).

DOI: 10.15774/PPKE.BTK.2021.17

of IOB tagging E- prefix indicates the end of a token sequence, thus each chunk can be formulated as a sequence of the tags: B, I and E. Additionally, chunks only consisting of one token may be tagged with a 1. Example (4) illustrates IOB tagging (without using the labels E- and 1-).

(4) Mr. NNP B-NP
Meador NNP I-NP
had VBD B-VP
been VBN I-VP
executive JJ B-NP
vice NN I-NP
president NN I-NP
of IN B-PP
Balcor NNP B-NP[2]
. . O

If using the "extended" version, the same sentence would be tagged as:

(5) Mr. NNP B-NP
Meador NNP E-NP
had VBD B-VP
been VBN E-VP
executive JJ B-NP
vice NN I-NP
president NN E-NP
of IN 1-PP
Balcor NNP 1-NP
. . O

As can be seen, each tagged sequence has a type that corresponds to the name of the parsing unit, e.g. VP, NP, PP, etc. In general, the task in IOB tagging is to assign

[2]The tagging of *of Balcor* may be confusing, but the CoNLL 2000 shared task (Tjong Kim Sang and Buchholz, 2000) stated that most PP chunks consist of just one word (the preposition) with the part-of-speech tag IN. I follow their analysis in these two examples.

DOI: 10.15774/PPKE.BTK.2021.17

these labels to the tokens correctly. Table (1.1) from Hong Shen and Anoop Sarkar (2005) summarises the standard representations in NP chunking tasks. Note that O marked a token not belonging to any chunks in examples (4) and (5); in NP chunking, on the other hand, it marks the tokens that do not belong to NP chunks (are "outside" of them).

Word	**IOB1**	**IOB2**	**IOE1**	**IOE2**	**Start/End**
In	O	O	O	O	O
early	I	B	I	I	B
trading	I	I	I	E	E
in	O	O	O	O	O
Hong	I	B	I	I	B
Kong	I	I	E	E	E
Monday	B	B	I	E	S
,	O	O	O	O	O
gold	I	B	I	E	S
was	O	O	O	O	O
quoted	O	O	O	O	O
at	O	O	O	O	O
$	I	B	I	I	B
366.50	I	I	E	E	E
an	B	B	I	I	B
ounce	I	I	I	E	E
.	O	O	O	O	O

Table 1.1. The noun chunk tag sequences for the sentence *In early trading in Hong Kong Monday, gold was quoted at $366.50 an ounce.* In the so-called Start/End representation S stands for single words within a chunk. Table quoted from Hong Shen and Anoop Sarkar (2005) (with one modification: the header of the last column was changed)

.

The formalism is not only used in (NP-)chunking, but was successfully applied for Named Entity Recognition as well (CoNLL-2003 shared task, Tjong Kim Sang and De Meulder, 2003).

Several approaches are used in performing NP chunking in the domain of natural language processing. They can be categorised as follows :

- Rule-based chunking
- Statistical-based chunking

DOI: 10.15774/PPKE.BTK.2021.17

- Hybrid approach for chunking (for a comprehensive summary and categorisation of NP chunkers – mostly for Asian languages – see Sarma and Barman, 2015)

A rule-based approach is generally used with languages where a large amount of adequate data is not available. The rules used in this approach are either handcrafted or are extracted from some linguistic resources.

Statistical approaches do not need linguistic knowledge, though they highly depend on the available resources on the language. The method is language-independent, thus it can be applied to different languages with common features. This method extracts statistical information from an NP-annotated corpus. The extracted statistical information consists of occurring phrases, the frequency of occurrence of the words, etc. The statistical methods are mainly based on probability measures, including unigrams, bigrams, trigrams and n-grams.

Hybrid methods aim to increase both the precision and the recall of the chunker by combining the benefits of the above approaches.

1.2.4 NP chunking in Hungarian

Traditionally, NP chunking for English takes the POS-tags as input and provides chunks as output. English has properties that make it a suitable candidate for NP chunking: the word order encodes most of the information required to identify the chunks and it has a low percentage of non-projective dependencies. In Hungarian, on the other hand, NPs can stand at any position in the sentence regardless of their role.[3] The dependencies among the words are mainly encoded not in prepositions but rather in suffixes; as the word order in itself is not sufficient to find the correct NP chunks, one has to rely on the morphological annotation of the words. In example (6), the possessive structure within this maximal NP – the object of the verb, *a gallérnak a bélés alól való kitüremkedését* 'the protrusion of the collar from under the padding' – is encoded in the suffixes of the nouns. The members of the possessive structure, the possessor (*a gallér* 'the collar') and the possessee (*kitüremkedés* 'protrusion') are located further away from each other as a present participle with its own postpositional modifier – *a bélés alól való* 'being from under the padding' – is inserted between them, thus modifying the possessee. As can be

[3]This is a simplification; more precisely, the word order in Hungarian is relatively free but only in the part of the sentence after the verb.

DOI: 10.15774/PPKE.BTK.2021.17

seen, all the morphological information of the tokens is required to precisely detect the maximal NP in this sentence.

(6) *A* *varrónő* *észrevet-te* *a* *gallér-nak* *a* *bélés* *alól*
the seamstress notice-PastSg3 the collar-Gen the padding under.from
val-ó *kitüremkedés-é-t*
be-PrsPtcp protrusion-PossSg3-Acc
'The seamstress noticed the collar protruding from under the padding.'

The first paper presenting a solution for Hungarian NP chunking was that of Váradi (2003). The approach uses so-called cascaded regular grammars. It was tested on a morphologically tagged and disambiguated corpus of 928 sentences representing a sample of written style of journalism. The system achieved an F-score of 58.78%. This rule-based methodology was further developed in Váradi and Gábor (2004).

Hócza (2004) offers a statistical method for the noun phrase recognition task. This approach uses noun phrase tree patterns described by regular expressions from an annotated corpus. The tree patterns are then completed with probability values. The noun phrase recognition parser tries to find the best-fitting trees for a sentence using backtracking technique. It achieved an F-score of 83.11% on a test corpus consisting of news texts.

The next significant contribution to Hungarian NP chunking was the design and implementation of HunTag (in some papers referred as HunChunk), which was presented in the Master's thesis of Gábor Recski (2010b), in the PhD thesis of Dániel Varga (2012) and in some papers of Recski and Varga (Recski, 2010a; Recski and Varga, 2012). HunTag uses a combination of Maximum Entropy learning and Hidden Markov Models (HMM) to perform NP-chunking of tokenised and morphologically annotated texts and is a reimplementation and generalisation of a Named Entity Recognizer built by Dániel Varga and Eszter Simon (2006).

Miháltz (2011) provides a comparison of different modules and systems designed for the task of Hungarian NP recognition. Apart from HunTag, the rule-based method of Váradi and Gábor (2004), and the parser used by MetaMorpho machine translation system (Prószéky et al., 2004) were evaluated on a set of sentences from Szeged Treebank 2.0 (Vincze et al., 2010). In this comparison HunTag performed better than the others, with an F-score of 81.71% (the others achieved 57.73% and 45.99% respectively).

DOI: 10.15774/PPKE.BTK.2021.17

HunTag proved to be an inspirational system. Some of its applications are described in Recski et al. (2009) and Recski et al. (2010). Recski (2014) uses the output of a rule-based chunker (made by implementing Kornai's grammar of NPs (Kornai, 1985) to improve its performance).

The most recent and most widely used system for the extraction of NPs in Hungarian was HunTag3 (Endrédy and Indig, 2015), which, as its name implies, is a(n official) successor of the HunTag project. It is also a sequential tagger for NLP combining a linear classifier and Hidden Markov Models. Based on training data, HunTag3 can perform any kind of sequential sentence tagging and has been used for NP chunking and Named Entity Recognition for English and Hungarian. It was tested on the test set of Mihálty (2011) and on the Szeged NER corpus (Szarvas et al., 2006). The best F-score HunTag3 reached was 93.59% (on the former test corpus) thus achieving the best result among Hungarian NP chunkers before the appearance of deep learning models. HunTag3 is now a part of the new version of e-magyar language processing system (Váradi et al., 2018, 2017) where the chunking module derived from HunTag3 is called emChunk.

The last few years of natural language processing were undoubtedly dominated by neural networks and deep learning models that have achieved state-of-the-art results on various NLP tasks, NP chunking among them: for Hungarian, Nemeskey (2021: p. 9) reported an F-score of 97% both for baseNP and for maximal NP chunking. The model presented in his PhD thesis – called huBERT – consists of two preliminary BERT Base models (Nemeskey, 2020). BERT (Bidirectional Encoder Representations from Transformers) is the state-of-the-art language model for NLP published in a paper by researchers at Google AI Language (Devlin et al., 2019). The innovation here is that BERT applies a bidirectional training to language modelling. This is in contrast to previous (neural network-based) efforts which looked at a text from left to right (or from left to right and from right to left separately). The paper's results show that this bidirectionally trained language model can have a deeper sense of language context than single-direction language models. BERT and its "offsprings" (e.g. BioBERT for biomedical text mining, Lee et al., 2019) proved to be the best in many areas of natural language processing (and other scentific areas as well), beating not only other rule-based or statistical methods, but all the other neural network-based models as well.

DOI: 10.15774/PPKE.BTK.2021.17

Although there are some attempts with rule-based NP chunking (Recski, 2014), the majority of Hungarian approaches are nevertheless statistical ones, most recently almost exclusively neural network-based. In the next section I turn to my personal relationship with NP chunking and I briefly present my own naive attempts to extract NPs from Hungarian texts with handcrafted rules.

1.2.5 NP chunking and me

In the early stage of my research I carried out some preliminary investigation into NP extraction. Endrédy (2014) presents a mini-corpus built from the texts of short news sent via e-mail from InfoRádió news portal. These short items of news consist of a title and then 2-3 sentences summarising the news. This corpus was later supplemented with the content of mno.hu, and the text of a book called *Piszkos Fred, a kapitány*.

From this corpus, NPs were extracted with some simple, intuitive rules: an article always starts a noun phrase; a punctuation mark or a verb always finishes the preceding noun phrase, etc. This way we obtained a long list of NP candidates. Endrédy (2014) designed an online interface to search this list. In Ligeti-Nagy (2015) I presented my study on this list of NPs focusing on false hits. I was looking for any gap in the current morphological annotation of these texts where NP chunking might fail. I suggested some new tags that might be useful for an NP chunker.

Example (7) illustrates the process described above. The NP candidates in (7a) and (7b) are completely identical with regard to their morphological annotation (third line of the examples). However, while the string in (7b) is in fact a noun phrase, the string in (7a) is not; it consists of two noun phrases and has a phrase boundary inside (marked by a |). The difference can be captured by tagging the difference in the third noun of the strings: *beszéd* 'speech' and *kancellár* 'chancellor'. The latter is an occupation, or title, and may follow a proper name within the same NP. The former cannot follow a proper name within the same NP. Thus we need to tag the latter so that a rule-based NP chunker will be able to rely on this difference when extracting NPs. Example (8) illustrates the same strings with a more distinguished morphological tagging: proper names are tagged with an N|Prop label, and nouns marking an occupation, or title, are labelled as N|Occup. Therefore, the difference between these two strings becomes more overt, and more sophisticated rules can be written for the NP extraction task.

DOI: 10.15774/PPKE.BTK.2021.17

(7) a. *Angela Merkel / beszéd-et [mond]*
Angela Merkel | speech-Acc [say]
N N | N-Acc

'Angela Merkel [gave a] speech'

b. *Angela Merkel kancellár-t [meghív-ták]*
Angela Merkel chancellor-Acc [invite-PastPl3]
N N N-Acc

'[they invited] chancellor Angela Merkel'

(8) a. *Angela Merkel / beszéd-et [mond]*
Angela Merkel | speech-Acc [say]
N.Prop N.Prop | N-Acc

'Angela Merkel [gave a] speech'

b. *Angela Merkel kancellár-t [meghív-ták]*
Angela Merkel chancellor-Acc [invite-PastPl3]
N.Prop N.Prop **N.Occup**-Acc

'[they invited] chancellor Angela Merkel'

As a next step, I applied my tags on the texts of the InfoRádió corpus mentioned above. Based on the morphological annotation supplied with this novel tagset, I wrote a rule-based NP-extraction method. It is presented in Ligeti-Nagy (2016). On a randomly selected and manually evaluated mini-corpus it reached a relatively high accuracy (90%). However, it still needed a lot of fine-tuning and a gold standard corpus with my tags to evaluate it on.

The point of this short insight into my early research was to illustrate how I came closer to the problems around Hungarian NP chunking and Hungarian NPs themselves.

1.3 AnaGramma

The idea of sentence skeletons – mentioned in the foreword – being a significant research topic came up during the research project of `AnaGramma`. The goal of this project was

DOI: 10.15774/PPKE.BTK.2021.17

to create a psycholinguistically motivated parser (called `AnaGramma`) for Hungarian. The project included informaticians and linguists as well; during the design of the parser many linguistic questions were raised, providing a fruitful base for many papers and some dissertations as well (e.g. Indig et al., 2016a,b; Vadász, 2017; Vadász et al., 2017, etc.). In this section I briefly introduce this parser[4] and its motivation and background, as it also proved to be the inspiration and framework of my own research.

The theoretical background, the impetus, and the basic principles of `AnaGramma` are summarised in Prószéky et al. (2014), Prószéky and Indig (2015), and Prószéky et al. (2016). In each chapter of this dissertation I will highlight the features of this parser that were the starting point or the motivation of the given research. Here I present the basic principles and the working mechanism of `AnaGramma` based on the three papers mentioned above.

`AnaGramma` was presented as a new paradigm and framework for syntactic and semantic analysis. The main principles of its working mechanism are the following:

1. psycholinguistically motivated
2. performance-based
3. left-to-right processing
4. parallel architecture
5. the processing units are utterances, not sentences
6. the representation is a connected graph with different types of coloured edges
7. supply and demand threads are parallel to each other

Psycholinguistically motivated means that the model aims to hold on to the algorithms of human language processing as much as possible. This results – among others – in the third feature: the parser processes the texts strictly left-to-right incrementally without using or referencing any part of the sentence succeeding the current token.

The system's performance-based nature results in two main characteristics: instead of sentences the parser processes 2-3 sentences-long utterances (the fifth feature); more importantly, the goal is not to be able to analyse theoretically existing yet almost never

[4]The parser is available at `https://github.com/ppke-nlpg/AnaGramma-Parser`.

DOI: 10.15774/PPKE.BTK.2021.17

before seen phenomena, but rather to be able to interpret any text in Hungarian that actually appears in corpora disregarding its grammaticality.

The fourth feature of the parser is related to its architecture and design. In a traditional approach an analysis of a sentence is generated at the end of a pipeline of different modules. `AnaGramma` processes the actual word using parallel threads (a morphological analyser thread, a corpus statistics thread, etc.). These threads analyse each word in parallel and communicate with each other to correct each others' errors and to make a final decision in the analysis, thus the architecture is parallel.

As mentioned above, when discussing the performance-based nature of the parser, the framework's processing and representational units are not individual sentences, but rather utterances consisting of one or more sentences. Thus it is possible to handle intra- and intersentential anaphoric relations in a unified way.

The parser also needs some kind of grammar which enumerates the possible roles for every linguistic unit. This kind of description of the phenomena of the language are handled by parallel threads in `AnaGramma`. Two basic thread types seemed necessary: a supply-type thread provides information on the current element (e.g. this element is in nominative case), and a demand-type thread is looking for a required element with a specific property (e.g. a possessed noun looks for its possessor, a determiner seeks the NP head, a transitive verb needs its object, etc.). Every word may have demands: for example, verbs demand their arguments. And every word may have some features to supply: the nouns have a grammatical case. The two have to meet; a demand must be satisfied with a supply to form an edge between the two elements.

The principles described so far result in the representation of the sentences / utterances as a connected graph – a forest, instead of a single tree – where different types of connections are marked with different colours of edges.

To turn to the parsing process itself, `AnaGramma` uses a two-token-wide look-ahead window that provides information of the right context of the word to capture influence of the context on a word, while the information of the previously processed elements is always available in the so-called *pool*. The process is based on a sentence processing model, the *Sausage Machine*, where the parsing process consists of two main phases. The first phase is – as Frazier and Fodor (1978) put it – the *Preliminary Phrase Packager* where lexical and phrasal nodes are assigned to groups of words within the string input. The look-ahead

DOI: 10.15774/PPKE.BTK.2021.17

window of `AnaGramma` implements this first phase. In this phase the components of the sentence are prepared, e.g. the disambiguation of case-ambiguous nominals (see Chapter 2). In the second phase, these packaged phrases acquire their roles in the sentence by adding non-terminal nodes. The second phase is called the *Sentence Structure Supervisor*, as the packages – the pieces of the sausage – receive their role in the sentence. This is where a verb is connected to its arguments, for example (see Vadász et al., 2017).

With the above described features, `AnaGramma` is meant to be as fast as human processing; it is also meant to make the same mistakes as humans do with backtracking occurring in the parsing process only when really necessary; it uses every resource while parsing, mixing the statistics of frequent n-grams with rules provided by the grammar of supplies and demands.

Figure 1.1 illustrates the supply-demand architecture of `AnaGramma`. The numbers in the circles mark the places of clock signal, which is the basic processing unit of the parser. At every clock signal (or word boundary) several processing threads are launched. The first of these is the morphological analysis resulting in features that facilitate the higher levels of the analysis. The morphosyntactic features of a given token may be of the type *supply* or the type *demand*. As mentioned before, the goal of the parsing is to correctly combine these features so that every demand is satisfied by a supply once the utterance is over.

The first token is *be.üzemel-ték*: in.install-PAST.PL.3. Being a finite verb form it has a supply (FIN) that may or may not be required by an other node in the sentence. The verb argument lexicon provides the information for the parser that the stem of the word, *beüzemel* 'install', may *demand* a NOM or an ACC case ending, thus two demand threads are launched here with this information (last two lines under the word on the figure). They are further specified based on the morphosyntactic information on the token as NOM+PL+3, the possible subject of the verb is a third form plural as the ending of the verb token implies. The object of the installing, if overt, is definite: ACC+DEF. As both the subject and the object of the verb are optionally overt in a sentence, the demands may remain unfulfilled in the analysis of this sentence: NOM?+PL+3, ACC?+DEF?.

The determiner *a* is a supply demanded by the ending of a noun phrase later. *Balaton* is a noun with no overt case suffix on it: thus may be a subject or an unmarked possessor as well (more on the possible roles of nouns with no overt case suffix in 2). Here the

DOI: 10.15774/PPKE.BTK.2021.17

subject must be a third person plural, therefore, by simplifying the method a little, it can be stated that *Balaton* is a possessor here: it launches a demand thread (Pers?).

Vihar-előrejelző 'storm signalling' is an adjective, only supplying itself as a modifier for a noun. *Rendszer-é-t* is a noun with an accusative suffix: system-Pers.3Sg-Acc. The stem of the word is a noun, demanding one or more optional modifiers searching strictly in the pool: it becomes connected to the adjective *viharelőrejelző*'s supply. The possessive suffix launches a demand thread satisfied by the supply of *Balaton*: Pers?. The case suffix launches a supply thread immediately satisfying the demand of the verb for an argument (Acc?+Def?). It also launches a demand for a determiner that can be fulfilled by the supply of the determiner.

The sentence is stopped by a punctuation mark, which means that no subject came to fulfil the demand of the verb (so that thread remains unsatisfied meaning that here we have the generic subject).

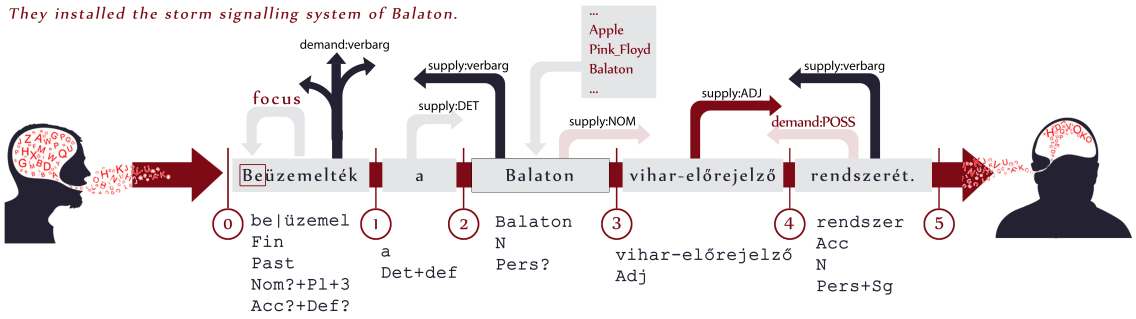

Figure 1.1. An illustration of the parsing process of `AnaGramma` on the sentence *Beüzemelték a Balaton vihar-előrejelző rendszerét.* 'They installed the storm signalling system of Balaton.'

Here I intended to present, in a very general way, `AnaGramma`, the performance-based, psycholinguistically motivated parser that provides the background to many of the research topics discussed in the following chapters. Some key aspects and features of the parser, such as the supply-demand framework, the two-token-wide look-ahead parsing window, among others, will appear later, forming a strict framework around my work. However, the parsing process of `AnaGramma` as a whole needs to be narrowed down, as my research topic is the structure of noun phrases: in the next section I give a brief overview of the relationship between parsers and noun phrases.

DOI: 10.15774/PPKE.BTK.2021.17

1.4 Corpora

In this section I briefly describe the corpora I used for the studies presented in the following chapters.

Some papers of which I am a co-author, and are the result of the work of the MTA-PPKE Hungarian Language Technology Research Group, present their results on the Pázmány Korpusz (Endrédy, 2016; Endrédy and Prószéky, 2016). Pázmány Korpusz was meant to be – at least when published – the largest Hungarian annotated corpus with 1.2 billion tokens. It has been created mainly as a background and text material for the different studies supporting `AnaGramma`. At the beginning of the psycholinguistically motivated research on Hungarian language ("the `AnaGramma` project") no large (greater than a billion tokens) text corpus with several different annotations had been available yet to support the project. To overcome this limitation, a crawler has been designed (Endrédy and Novák, 2013) whose task was to download the Hungarian content of the internet with high quality and speed. After several years of running, the crawler collected 1.2 billion tokens. The corpus is stored in an XML-like format used by the Bonito corpus manager and its graphical interface tool (Rychlý, 2007). However, no matter how large and well-annotated the Pázmány Korpusz is, it was unfortunately never made publicly available, only the members of the research group could query it.

As the results of a study carried out on a "phantom" corpus are neither reliable nor reproducible, I generally sought the best corpus possible which has the advantageous features of the Pázmány Korpusz needed for that given research. Most of the time the Hungarian Gigaword Corpus (HGC, Oravecz et al., 2014) proved to be the best choice for a given task. It is an extended version of the Hungarian National Corpus (Váradi, 2002) with an upgraded and redesigned linguistic annotation. It consists of texts from different registers such as journalism, literature, science, personal, official and it has transcribed spoken texts from radio programs as well. By 2014, it reportedly contained 1.5 billion tokens. The biggest advantage of HGC (besides its respectable size) is the query interface which allows complex searches on every layer of the annotation. The full text version is not available, but the web search interface completely satisfies the needs of linguists.

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 Szeged Treebank is the largest fully manually annotated

DOI: 10.15774/PPKE.BTK.2021.17

(thus gold standard) treebank of Hungarian. It was preceded by the creation of the Szeged Corpus (Csendes et al., 2004). The Treebank contains 82 000 sentences, 1.45 million tokens (1.2 million words and 250 000 punctuation marks). Texts were selected from six different domains: fiction, compositions of pupils between 14-16 years of age, newspaper articles, texts in IT, legal texts, business and financial news. The 1.0 version of the Szeged Treebank is annotated for noun phrases and clauses. The 2.0 version has a deep phrase-structured syntactic analysis. The Szeged Dependency Treebank contains a dependency annotation for the sentences.

1.5 Rule-based vs. statistical methodology

There are two basic approaches to the tasks in NLP (as mentioned above in 1.2.3): rule-based systems and machine learning algorithms. A rule-based system requires skilled developers and linguists, but does not require a massive training corpus; it usually produces high precision results, and is good at capturing and describing a specific language phenomenon. Machine learning algorithms, on the other hand, require a large amount of labelled training data, but do not need many experts. They generally achieve a higher recall and are easy to scale. As the results of NP chunking also show (1.2.4) machine learning algorithms, and especially neural network-based methods perform much better on downstream NLP tasks. In spite of all these, the following chapters use an almost exclusively rule-based approach; rule-based systems can be easily mapped to linguistic observations, and are generally useful for tasks where there is no available labelled training data or the available data is not enough for a statistical-based system. I will turn to the help of neural network-based models myself when they are available (4), but as linguistic curiosity is always a motivating factor in my research, the rule-based approach remains my preferred tool in my inquiries.

`AnaGramma` (1.3) and the task of NP chunking were the driving force of my research; the above described corpora (1.4) provided the material for it. The four topics discussed here in more detail are the following: nominals bearing no overt case suffix (Chapter 2), as I found them challenging in NP chunking in my earlier studies (Ligeti-Nagy, 2015); extended named entities, as they are strings of multiple suffixless nominals following each other, making it complicated to find the right edge of the NPs (Chapter 3); locative case

DOI: 10.15774/PPKE.BTK.2021.17

suffixes, as they are somewhat unexplored and play a significant role not only in syntactic parsing, but also in other semantic tasks as well (Chapter 4); and postpositions, as they play a crucial role in marking the right edge of a noun phrase (Chapter 5). The following chapters will explore these points in greater detail.

DOI: 10.15774/PPKE.BTK.2021.17

Chapter 2

The disambiguation of suffixless nominals

"Nothing will come of nothing"

King Lear Act 1, Scene 1

In this chapter I focus on nominals bearing no overt case suffix at all. First, I discuss the roles in a sentence that may be expressed by a suffixless nominal (2.1). Then I introduce an algorithm designed to disambiguate the role of nominal tokens bearing no overt case suffix (2.2). After presenting the motivation for the algorithm I demonstrate its working mechanism (2.2.1) and evaluate its performance (2.2.3). I show that the algorithm is able to define the function of a suffixless nominal with high precision based only on a two-token wide forward-looking window. Finally, I briefly present a second algorithm that was meant to be an extension of the first one, aiming to define predicative nominals as well (2.3).

The first algorithm (`Nom-or-What`) presented in this chapter is the result of a collaboration between me and three colleagues of mine: Andrea Dömötör, Noémi Vadász, and Balázs Indig. The second algorithm (`Nom-or-Not`) is the result of a collaboration between me, Andrea Dömötör and Noémi Vadász. In the following sections, I will always indicate if a part of the process under discussion is not exclusively my achievement, but rather the result of a joint effort.

DOI: 10.15774/PPKE.BTK.2021.17

2.1 The meaning of the lack of case suffix

To know and to be able to define the possible finalising elements is crucial for an NP-chunking task and for a sentence parser as well; these are the points where one can start linking verbs and arguments.

There are two visible and easily detectable markers of the end of a (base) NP: case suffixes and postpositions. These two will be discussed in chapter 4 and chapter 5. However, there are several cases when there is *nothing* at the end of an NP. With *nothing* I refer to the lack of any overt case suffix or postposition. This section concentrates on this special instance, showing no visible sign at the end of a nominal that helps to identify whether this is the end of a given NP or not.

Throughout this section, for the sake of simplicity, I use the term nominal to refer not only to nouns and adjectives, but to participles and pronouns as well. While discussing the difficulties of NP recognition caused by the large number of bare nominals (nominals without a definite article, a quantifier, or a case suffix) in a sentence, pronouns substituting nouns and adjectives can be examined identically as the former ones, thus it is simpler and more concise referring to them with the same term. In (9) *szomszéd* 'neighbour' is a noun, here the possessor of the dog; *ő* 'he/she' is a pronoun, also the possessor of a dog. Both tokens appear without any overt case suffix, and both are unmarked possessors in the sentence.[1]. Additionally, participles also form a group with nouns and adjectives in this study as tokens with a POS-tag "participle" may bear case suffixes just as nouns and adjectives. In (10) (retrieved from MNSZ2.0) the POS-tag of *történtté* is IGE._MIB. (adjectival participle) , bearing the suffix *-té* which is an allomorph of *-vé*, the translative case marker.

(9) a. *a* ***szomszéd*** *kutyája*
the neighbor dog-Poss.3Sg
'the neighbour's dog'

[1]However, it is obvious that in (9) the determiner before the personal pronoun narrows down the possible roles this personal pronoun can take in the sentence, while it does not do that with the noun *szomszéd* 'neighbour'. Therefore it seems to be a false generalisation to say that nouns/adjectives and pronouns can be studied identically here. On the other hand, the algorithm presented here has some strict rules to adjust to and is constructed to define the role of the given nominal based on a two-word wide forward-looking window and based on that window only. This means that when analysing the word *ő* or the word *szomszéd* the algorithm is "blind" to all the preceding tokens, thus it cannot narrow down the roles the pronoun may bear.

DOI: 10.15774/PPKE.BTK.2021.17

b. *az* ***ő*** *kutyája*
the he/she dog-Poss.3Sg
'his/her dog'

(10) *Meg nem* ***történ-t-té*** *azonban már nem lehet te-nni a*
Perf not happen-Ptcp-Fac however already not may do-Inf the
dolg-ok-at.
thing-Pl-Acc
'Things cannot be undone.'

The problem with nominals bearing no overt case suffix thus automatically annotated as `Nom` is illustrated in (11). All the tokens in bold are nominals – without any overt case suffix – tagged by the morphological analyser as `Nom`. However, only one of them is actually the subject of the sentence with a zero nominative case suffix on it (*édesapja* 'father of sy').

(11) *Az* ***ön*** ***édesap-ja*** *is* ***iskolás*** ***gyerek*** *volt az* ***ábrázol-t***
the you father-Poss.3Sg too school kid is.Pst3Sg the depicted
időszak-ban.
period-Ine
'Your father was also a school kid in the depict-ed period.'

The goal here is to clarify the role of each token with a Nom tag: *ön* is an unmarked possessor, *édesapja* is the subject of the sentence, *iskolás* is a modifier of *gyerek*, *gyerek* is the nominal part of the nominative predicate, *ábrázolt* is the modifier of the adverbial. In the next section I summarise the structure we presume is behind suffixless nominals.

2.1.1 The possible roles of a suffixless nominal

A nominal with no overt case suffix may have the following functions[2]:

[2]This list is the result of a cooperation between me, Noémi Vadász and Andrea Dömötör, published in Ligeti-Nagy et al. (2018) and Ligeti-Nagy et al. (2019)

DOI: 10.15774/PPKE.BTK.2021.17

- **subject** of the sentence: in this case, we assume the zero nominative case suffix at the end of the nominal (12a)[3]
- unmarked **possessor**: (12b); about unmarked possessors see 2.1.1.1
- element in **vocative** role: in (12c), *szomszéd* 'neighbour' is addressed directly. About vocative case in Hungarian see 2.1.1.2
- **nominal followed by a postposition** (12d): here I assume that the nominal part of the postpositional phrase does not bear any (zero) case suffix, for details see 2.1.1.3
- a nominal **modifier** of another nominal (12e).
- the member of an **extended named entity**. In (12f) *Máris szomszéd* 'neighbour Máris' is a named entity where the common noun, *szomszéd* 'neighbour' bears the case suffix of the whole unit (zero nominative case suffix, in this case), and *Máris* acts like a modifier bearing no case suffix at all. For a detailed explanation see 2.1.1.4.
- **predicative nominal**. In Hungarian, non-elliptic sentences may be formed without a finite verb. The explanation for this is the so-called zero copula phenomenon. A detailed description and theoretical background of this can be found in É. Kiss (2002). The computational analysis and handling of predicate nominals is studied and processed by Andrea Dömötör (see for example Dömötör, 2018, 2017), therefore I will only partially mention some relevant facts about nominal predicates in this section. In (12g), *szomszédom* 'my neighbour' is the subject of the sentence, and bears a zero nominative case suffix, while *ügyvéd* 'lawyer' is the nominal predicate of the sentence.

(12) a. *A* ***szomszéd*** *tegnap* *érkez-ett.*
the neighbour yesterday arrive-PST.3SG
'The neighbour arrived yesterday.'

[3]The examples in (12) are from Ligeti-Nagy et al. (2019).

DOI: 10.15774/PPKE.BTK.2021.17

b. *A* ***szomszéd*** *kutyá-ja* *ugat.*
the neighbour dog-POSS.3SG bark.3SG
'The neighbour's dog barks.'

c. *Jó* *reggel-t,* ***szomszéd!***
good morning-ACC neighbour
'Good morning, neighbour!'

d. *A* ***szomszéd*** *után* *érkez-t-ünk.*
the neighbour after arrive-PST-1PL
'We arrived after the neighbour.'

e. *A* ***szomszéd*** *gyerek* *kedves* *volt.*
the neighbour kid kind be.PST.3SG
'The neighbour kid was kind.'

f. ***Máris*** *szomszéd* *tegnap* *érkez-ett.*
Máris neighbour yesterday arrive-PST.3SG
'Neighbour Máris arrived yesterday.'

g. *A* *szomszéd-om* ***ügyvéd.***
the neighbour-POSS.1SG lawyer
'My neighbour is a lawyer.'

Based on the above described roles and the detailed explanations in 2.1.1.1-2.1.1.4 it can be stated that suffixless nominals either bear a phonologically zero case suffix (when functioning as a subject as in (12a)) or bear nothing (as an unmarked possessor (12b) or when being in vocative role (12c), being followed by a postposition (12d), when modifying another nominal (12e), being part of a complex proper name (12f), or functioning as a predicative nominal (12g)). In the following sections I use the following annotation to distinguish these functions:

- NOM is the zero nominative case suffix
- GEN marks the unmarked possessor

DOI: 10.15774/PPKE.BTK.2021.17

- VOC marks a vocative role
- NONE marks the lack of case suffix before postpositions, inside extended named entities and on modifiers
- tag marking the predicative nominal: PRED
- SUFF stands for the default phonologically zero case suffix of nouns, marking either a NOM or a GEN

In 2.1.1.1-2.1.1.4 I give a brief overview on the roles in question.

2.1.1.1 Genitive case

In the Hungarian possessive construction, the possessed noun bears a suffix indicating the possessedness (while also bearing an agreement marker that matches the person and number feature of the possessor). The possessor, on the other hand, may bear a *-nAk* suffix (13a) or may be suffixless (13b).

(13) a. *A* ***szomszéd-nak*** *a* *kutyá-ja* *ugat.*
the neighbour-DAT the dog-POSS.3SG bark.3SG
'The neighbour's dog barks.'

b. *A* ***szomszéd*** *kutyá-ja* *ugat.*
the neighbour dog-POSS.3SG bark.3SG
'The neighbour's dog barks.'

The *-nAk* suffix in (13a) is a dative suffix (Szabolcsi, 1981; É. Kiss, 2002), although some papers argue that it is a genitive case suffix (most importantly Kiefer, 1992). As the main focus here is on the suffixless possessive structure (13b), I do not wish to investigate this question in detail. However, as the algorithm presented here is an improved and expanded version of the `Nom-or-Gen` procedure (Vadász and Indig, 2017) of ANAGRAMMA, we decided to keep the naming conventions applied there. Vadász and Indig (2017) declared that they follow the terminology of Kiefer (1992) thus using the term "genitive" when discussing possessive structures (and naming their procedure).

DOI: 10.15774/PPKE.BTK.2021.17

The nominal in (13b) is called "nominative possessor" by Kiefer (1992) and others, but Bartos (2001) and É. Kiss (2002) argue that these unmarked possessors are caseless rather than Nominative (Dékány, 2012, also provides further evidence for this).

As our goal is to maintain no more than one nominative nominal in a sentence, which then must be the subject, the unmarked possessor of a sentence in this study will be marked by a GEN tag referring to its specific role in the sentence.[4]

2.1.1.2 The vocative case

László Antal provides a concise summary on what Hungarian linguists had thought about nominal cases in Hungarian before the mid-20th century (Antal, 1961). This summary includes the historical overview of the status of vocative case in Hungarian, therefore I briefly outline the view of the major linguistic works on vocative case based on Antal's chapter titled *Grammatical heritage* (Antal, 1961: p. 389-435).

Vocative case was considered one of the six or seven cases of Hungarian nominals by the authors of the first grammatical studies on Hungarian – who took the Latin nominal paradigms as a basis of their grammatical examinations and descriptions – including the very first Hungarian grammar, Sylvester's Grammatica from 1539 (Sylvester, 1539), Albert Szenczi Molnár's grammar from 1610 (Szenczi Molnár, 2004), Pereszlényi's grammar from 1682 (Pereszlényi, 2006) and Kövesdi's short book from 1686 (Kövesdi, 2010). Unlike these authors from the 16th-17th century, who all based their grammar (written in Latin) on Latin declination, György Komáromi Csipkés is the only one from this period excluding *vocativus* from the list of the cases (Komáromi Csipkés, 2008). He rejected Latin as an example and turned to Hebrew when writing his grammar. He argues against treating *vocativus* as a case: it has no special ending different from that of *nominativus*, therefore it is not a case (Komáromi Csipkés, 2008: p. 25).

The late 18th century was the period of the first grammars written in Hungarian. Földi's grammar (Földi, 1912), however, still operates with the Latin cases, *vocativus* among them, called "hívó eset" 'calling case'; while the argument of Gyarmathy (1794) against *vocativus* being a case in Hungarian is similar to that of Komáromi (2008): the former simply states that all his predecessors wrote Hungarian grammar to the pattern of

[4] Although this naming convention is motivated by practical reasons, it is worth mentioning that the suffixless possessor and its position in a noun phrase is compared to that of the subject within a sentence (Kiefer, 1992). And as the subject bears a non-overt nominative case suffix, the possessor may bear a non-overt case suffix marking the possessive structure there.

DOI: 10.15774/PPKE.BTK.2021.17

Latin declination; however, we have no reason to consider *vocativus* a case, as *Domine!*, a vocative form in Latin, is the same word in Hungarian (*ó, Úr* 'Oh, Lord') as the nominative (*az Úr* 'the Lord'). The so-called "grammar from Debrecen" (Debreceni Grammatika, 1795) begins with a declaration that the number of cases cannot be determined by the number of case endings - if this were true, the Latin *cornu* 'horn', which is a member of the 4th declination, would have only one case; but the meaning of the word in sentences is numerous and cannot be narrowed down to just one. This statement predicts that *vocativus* will be a case in this grammar, and so it is.

The two significant grammars written in Latin from the first third of the 19th century, Révai's *Elaboratior* (1806) and Verseghy's *Analyticae* (1816) still enlist *vocativus* as a case in Hungarian. Although grammars written in Hungarian in the second half of the 19th century (Fogarasi, 1843; Galgóczi, 1848; Riedl, 1866; Szvorényi, 1866) do differ in the number of cases they propose, *vocativus* is consistently omitted from their lists. And from then on almost all the grammars from the 20th century, regardless of the language they were written in, excluded the case *vocativus* from their discussion about Hungarian noun cases. This is also true for the volume on morphology of the Structural Grammar of Hungarian (Kiefer, 2000b) where the 18 Hungarian nominal cases are discussed without even mentioning *vocativus*. The definition of a case suffix discussed in the chapter *Inflection* (Kiefer, 2000b: p. 569-618) excludes the vocative case from the group of Hungarian case suffixes:

(14) A suffix is a case suffix if and only if it binds a complement of the verb (in any form of the verb).

However, there is a specific vocative role in the sentence fulfilled by a noun without a case suffix that needs to be distinguished from the nominative case. Thus we mark the nominal in a vocative role with a VOC tag.

2.1.1.3 The noun and the postposition

In chapter 5 postpositions are discussed in more detail. Here I focus on the nominal before the postposition and the lack of any overt case suffix on it.

The phrase of a nominal and a postposition may be compared to a nominal bearing an overt case suffix (15a). In this case, postpositions are considered semi-bound case endings.

DOI: 10.15774/PPKE.BTK.2021.17

As no nominative case suffix is present on the lemma (e.g. *Alíz*) bearing the case suffix (e.g. *-zAl* INS), no nominative case suffix should be marked on the nominal (e.g. *Alíz*) preceding a postposition (e.g. *mellett* 'near.at').

However, one could argue that a parallel can be drawn between phrases consisting of a nominal without any case suffix and a postposition, and phrases of nominals with a lexical case and a postposition (15b). In this case, the suffixless nominal *Alíz* is compared to the one with a case suffix *Alíz-zal* 'Alíz-INS' as both are followed by a postpositon (*mellett* 'near.at' or *együtt* 'together'). Thus a non-overt (nominative) case suffix should be present on the former nominal.

(15) a. *Alíz* ***mellett*** / *Alíz-**zal***
Alíz near.at | Alíz-INS
'next to Alíz' | 'with Alíz'

b. *Alíz* ***mellett*** / *Alíz-zal* ***együtt***
Alíz near.at | Alíz-INS together
'next to Alíz' | 'together with Alíz'

In this study I rely on the former comparison. The reasons behind that are the following:

- as mentioned before, the goal is to keep one and no more nominative case tags in a sentence to assist a parser in identifying the true subject of the sentence
- the category of postpositions is rather problematic. While most studies agree that postpositions such as *mellett* in (15a) taking a nominal without any case suffix are undoubtedly postpositions, the ones such as *együtt* taking a nominal with a case suffix are categorized as adverbs by many papers. Therefore there is no need to compare the former to the latter (as in (15b)).
- the pair in (15a) has one more thing in common: they have a strict linear order. Neither the lemma and the case suffix, nor the suffixless nominal and the postposition can be separated. This is not true for the structures in the right-hand side of (15b). See (16). (The examples are from É. Kiss, 2002).

DOI: 10.15774/PPKE.BTK.2021.17

- finally, most generative studies argue that the (bare) complemenets of postpositions are caseless rather than nominative (see Asbury, 2008b; Dékány, 2012)

(16) a. *Alíz* **pontosan** ***mellett***
Alíz exactly near.to
'right next to Alíz'

b. *Alíz-zal* *teljesen* ***együtt***
Alíz-INS completely together
'completely together with Alíz'

To sum up, here I focus on the properties of postpositions making them similar to case endings, thus I assume that the suffixless nominal preceding the postposition bears no case suffix at all.

2.1.1.4 Extended named entities and their cases

In chapter 3 I study extended named entities (XNE) and their structure. Here I only wish to emphasise that we do not assume any case suffix being present on the members of these proper names, but rather only on the common noun at the end (note that proper names in XNEs may consist of more than one proper nouns). This is where the difference between XNEs and interpretative structures becomes apparent: in the latter (17a) every member bears the suffix (accusative case suffix in example (17a)), while in XNEs (17b-17c) only the common noun (*barátom* 'my friend') does, *Peti* remains a bare noun.

Here I handle extended named entities as follows:

- the bare nouns inside XNEs are suffixless (they are tagged with a NONE as modifiers are)
- the common noun in the named entity is treated like any other nominal in the sentence by the algorithm, its role is decided based on the two tokens following it (thus may bear a zero nominative case suffix, etc.)

DOI: 10.15774/PPKE.BTK.2021.17

(17) a. *vet-t-em* ***csizmá-t, piros-at***
buy-PST-SG1 boot-ACC,

'I bought boots, red ones'

b. ***Peti barát-om*** *hív-ott*
Peti friend-POSS.SG1

'my friend, Peti called'

c. ***Peti barát-om-at*** *hív-t-am*
Peti friend-POSS-ACC

'I called my friend, Peti'

2.2 The `Nom-or-What` algorithm

The `Nom-or-What` algorithm described in this subsection is an improved and expanded version of the `Nom-or-Gen` procedure functioning in `AnaGramma`, whose objective is to decide, based on a narrow context of the given word, whether a suffixless nominal is an unmarked possessor in a sentence or not (Vadász and Indig, 2017). In `Nom-or-Gen`, every category that is able to modify a noun received the tag `NPMod` in addition to its original features. The procedure presented here is called `Nom-or-What`, since the case tag of nominals without any overt case suffix in the corpora is `NOM`, regardless of their function in the sentence and the task of the case disambiguation is to clarify the exact role of these nominals.

Hence, the `Nom-or-Gen` procedure has been extended so that it can now make decisions about the subject (nominative case), the modifiers, and the nouns preceding a postposition. It would only be possible to clarify the function of a predicative nominal if the narrow context of the given word contained information that facilitates such a decision. However, this is rarely the case, because in the recognition of a predicate, the previously processed information is generally more helpful. Based on the window, a nominal predicate can only be clearly identified if a copula is present in the window, and even this form has to be narrowed down to the 1st or 2nd person singular or plural (otherwise, we cannot be sure that the verb is really a copula). While in example (18a) the window

DOI: 10.15774/PPKE.BTK.2021.17

can be used to clarify the predicative nominal *gyerek* 'child', it is essential to know also the antecedent tokens of the sentence in order to distinguish the case of the noun *gyerek* 'child' in (18b) and (18c) (it is a predicative nominal in (18b), but it is a subject in (18c)). The situation is even more difficult if there is no overt copula in the sentence, in which case we can only rely on two main criteria: 1) there is no finite verb in the sentence at all, 2) the sentence already contains a (disambiguated) nominative. Both features can only be determined based on a larger context, typically on the part of the sentence preceding the predicative nominal. For further details on this topic, see Dömötör (2018) and Dömötör (2017).[5]

(18) a. *Negyedik* ***gyerek*** *vol-t-am* *a* *család-ban.*
fourth child is-PST-1SG the family-INE

'I was the fourth child in the family.'

b. *Erdélyi Dániel maga is iskolás* ***gyerek*** *volt* *a film ábrázolta*
Erdélyi Dániel himself too school kid is.PST3SG the movie depicted
korszak-ban.
period-INE

'Dániel Erdélyi himself was a shoolkid in the period depicted by the movie.'

c. *Kevés zsidó* ***gyerek*** *volt* *a* *falu-m-ban.*
Few jewish child is.PST3SG the village-POSS.1SG-INE

'There were only a few jewish children in my village.'

In conclusion, the assumption behind the current algorithm is that the subject (and its nominative case ending), the case of the unmarked possessor, the role of a nominal before a postposition and the role of a modifier is clarified within the first phase of the two-stage parsing model, when the elements are prepared to take their role in the sentence. The predicative nominal is defined in the second phase of the two-stage sentence analysis, as it requires a broader context. In this section and with this algorithm, we only deal with the first stage of the analysis.

[5]The above discussion of nominal predicates and the two-token-wide parsing window is discussed in Dömötör (2018) and Dömötör (2017) by a colleague of mine, Andrea Dömötör. Here I only summarised her thoughts on the topic.

DOI: 10.15774/PPKE.BTK.2021.17

2.2.1 Design

As a first step of the design of the algorithm I investigated the behaviour of suffixless nominal (and participial) tokens within the (manually annotated) noun phrases of Szeged Treebank 2.0 (Csendes et al., 2005). In Szeged Corpus (Csendes et al. (2004)), the predecessor and the base of the treebank, the MSD morphological coding system was used (Erjavec, 2004)). MSD is a positional coding system developed for several languages. Here lemmas contain derivational suffixes and only inflectional morphemes are distinguished.

It became clear at an early stage of the research that a separate set of rules had to be established for nouns, adjectives, numerals and participles, as they typically behave differently as caseless tokens. I manually created two lists for each part-of-speech category: one for instances where the token of the given category labelled `Nom` is inside a noun phrase, and one for the cases where the token tagged as `Nom` is the last token in a noun phrase. Based on these lists, I have identified the tokens following a `Nom` which help to identify the exact function of this `Nom`. My observations and the working mechanism of the algorithm are illustrated by decision trees (figures (2.1), (2.2) and (2.3)).

I decided to use only the two-token-wide parsing window when disambiguating the role of suffixless nominals and to disregard the pool (the collection of the previously seen information, see section 1.3). The reasons behind this restriction are the following:

1. There were three studies preceeding the design of this algorithm that aimed to examine the advantages and limitations of a two token wide, forward looking parsing window for Hungarian. Frazier and Fodor (1978) stated that the size of the window used in the first phase of the Sausage Machine (see section 1.3 for more details) is approximately six words (in English). However, the results of Indig et al. (2016b) and Vadász et al. (2017) showed that in Hungarian, where most of the information is stored in suffixes, in contrast to the large number of function words and the fixed word order in English, a narrower window is enough: in 99% of the cases the right detached preverb appeared after the verb in a distance smaller than three positions (Vadász et al., 2017: p. 7). Moreover, a two-token-wide window proved to be sufficient to establish the link between a verb and its infinitive argument as well (p. 8). Vadász and Indig (2017) measured the distance between the unmarked possessor and the possessee and found that in 80% of the cases the possessee appears at a maximum of two word distance after the possessor (p. 90). All of these studies

DOI: 10.15774/PPKE.BTK.2021.17

focused on the right context of the word; the algorithm `nom-or-gen` described in Vadász and Indig (2017) which is inserted into `AnaGramma` – and which can be considered as the predecessor of `nom-or-what` – was designed to make a decision based only on the two-token-wide parsing window disregarding the left context.

2. As the discussion of the results in 2.2.3 shows, an erroneous morphological annotation may cause serious damage to the algorithm's performance. Using the information from the pool means using not only the morphological annotation of the preceeding tokens, but also the tags assigned to them by `nom-or-what`. Thus by using the pool we import even more, potentially flawed information to rely on.

Since the algorithm is thus forced to make a decision based on the two tokens in the window, there will obviously be cases where the suffixlessness cannot be disambiguated. Therefore, default values had to be defined that are applied when no disambiguation can take place: in the case of nouns (common nouns and proper nouns as well) the default tag is SUFF marking the zero nominative case suffix and the zero case suffix of the unmarked possessor. In the case of adjectives, numerals and participles the default tag we used in the first set of annotations and testing was a *default*_NONE. We intended to highlight the defaultness of this label: it is not a disambiguation, it does not state that the token with this label is a modifier in the sentence, it can also get a NOM or GEN tag later, after processing the whole sentence. However, the first evaluation of the algorithm's performance (see section 2.2.3) made it clear that this default_NONE is useless and can simply be replaced by a NONE, as it turned out that almost all the cases of *default*_NONEs are in fact NONEs. Anyway, as the first study was carried out with this *default*_NONE label I will continue using this, but bear in mind that it should be seen as a NONE.

In example (19a) the precise role of the token *patak* 'brook' cannot be determined based on the tokens in the two-token-wide parsing window (*tőlük keletre*), thus it receives a SUFF label. It is clear after the whole sentence is processed that *patak* is the subject (NOM), but when only seeing *patak tőlük keletre* a possible continuation could also be similar to (19b); here *patak* is an unmarked possessor (GEN). In example (20a) the suffixlessness of *egyik* cannot be disambiguated based on the information in the parsing window (*egyik Veszprémben ,*), therefore it is labelled with a *default*_NONE. The whole sentence makes it clear that it is a subject (of the second clause), but a hypothetical continuation (20b) could mean that it is a modifier (NONE).

DOI: 10.15774/PPKE.BTK.2021.17

(19) a. *A* ***patak*** *tőlük* *kelet-re* *húzód-ott.*
the brook they.ABL east-SUB stretch-PASTSG3

'The brook stretched east of them.'

b. *A* ***patak*** *tőlük* *kelet-re* *húzód-ó* *ág-a.*
the brook they.ABL east-SUB stretch-PRSPTCP branch-POSSSG3

'The branch of the brook to the east of them'

(20) a. *A* *Nagy_Szent_Bazil-rend-nek* *két* *kolostor-a* *is*
the St.Basil-order-DAT two monastery-POSS.SG3 too
működ-ött, *az* ***egyik*** *Veszprém-ben,* *a* *másik* *Dunapentelé-n.*
operate-PST.SG3, the pne Veszprém-INE, the other Dunapentele-SUP

'The Order of St. Basil had two monasteries, one in Veszprém, one in Dunapentele.'

b. *A* *Nagy_Szent_Bazil-rend-nek* *két* *kolostor-a* *is*
the St.Basil-order-DAT two monastery-POSS.SG3 too
működ-ött, *az* ***egyik*** *Veszprém-ben,* *illetve* *Dunapentelé-n* *is*
operate-PST.SG3, the one Veszprém-INE, and Dunapentele-SUP too
tanít-ó *szerzetes* *meséli* *ez-t.*
teach-PRSPTCP monk tells this-ACC

'The Order of St. Basil had two monasteries, a monk teaching in Veszprém and in Dunapentele tells this.'

The decision tree in Figure 2.1 illustrates the rules applied to nouns (common nouns and proper nouns), plural adjectives, numerals and participles.[6] The root of the tree is the part-of-speech tag of the token in question. The labels of the edges of the first level of the tree represent the information retrieved from the first token in the parsing window. For example, if the first token in the window is a postposition (it has a label `NU`), the algorithm disambiguates the `Nom` tag on the given token as a NONE. The edges on the second level of the tree represent the information read from the second token in the parsing

[6]Note that these tokens are not plural adjectives or participles but only tokens labelled as plural adjectives or participles, etc. The algorithm takes the morphological analysis as it is and does not question it.

DOI: 10.15774/PPKE.BTK.2021.17

window. These edges are activated only if the algorithm could not make a final decision based on the first token and only the default SUFF label has been assigned to the token in question. In such cases, there is still a chance to disambiguate the original `Nom` tag with the information gathered from the second token in the parsing window.

It is an important condition that a possessive case suffix on the second token in the parsing window (`PS`) results in replacing the `Nom` label of the given token with a GEN tag if and only if the given token does not bear a possessive case suffix. This excludes the occurrence of structures such as the one in (21a). Phrases like *Magyarország kormánya nyilatkozatából* 'from the statement of the government of Hungary' (21b) are easy to process for the algorithm.

(21) a.

Magyarország	***kormány-a***	*mostani*	*nyilatkozat-á-ból*
Hungary	government-POSS.3SG	current	statement-POSS.3SG-out.of
N.NOM	**N.Poss.3sg.Nom**	ADJ.NOM	N.POSS.3SG.ELA

'from the current statement of the government of Hungary'

b.

Magyarország	***kormány-a***		*nyilatkozat-á-ból*
Hungary	government-POSS.3SG		statement-POSS.3SG-out.of
N.NOM	**N.Poss.3sg.Nom**		N.POSS.3SG.ELA

'from the statement of the government of Hungary'

NU is a macro used for postpositions, and the words *című* 'titled' and *nevű* 'named', and words of the category `NU_MN`. This latter is not an existing tag in MNSZ2.0, however, it proved to be useful to distinguish the words *alatti* 'the.one.under', *általi* 'the.one.happened.by', *mögötti* 'the.one.behind' etc. These words, similarly to postpositions, disambiguate the `Nom` on the token preceeding them and assign a NONE. *Című* 'titled', *nevű* 'named' etc. tokens have a similar role, therefore they are part of this branch of the decision tree. These words undoubtedly signal their preceeding nominal as being caseless (and not being an unmarked possessor, either). They all instantly disambiguate the NONE tag.

There are some words that are never parts of an NP; in this case they are the "outsiders". On the one hand, the macro OUT marks them: verbs (`IGE` tag), punctuation marks (`SPUNCT`), *aki* 'who', *ami* 'what' and their inflexed forms, *hogy* 'that' and *de* 'but'. Seeing them as the subsequent token by all means indicates that the given suffixless token

DOI: 10.15774/PPKE.BTK.2021.17

is the end of an NP and thus receives a NOM tag. On the other hand, definite articles (*a*, *az* 'the', with the tag `ART`), and the function words *is* 'too', *sem* 'neither', *nem* 'no', *pedig* 'in.turn' (these four marked with the macro PART) are also mostly "outsiders" and, as can be seen in Figures 2.1-2.3, often disambiguate the NOM tag.

The decision tree in Figure 2.2 illustrates the rules applied to (singular) adjectives and participles. Here it is also an important condition that a possessive case suffix on the second token in the parsing window (`PS`) results in replacing the `Nom` label of the given token with a GEN tag only if the given token does not bear a possessive case suffix. The macros described above are also part of this figure.

Finally, the decision tree in Figure 2.3 is the summary of rules applied to numerals. Here I also use the macros of the other two figures.

Table 2.1 sums up the common parts of the decision trees: postpositions (and the other words of the macro NU) always indicate that the suffixless nominal before them must be assigned with the tag NONE, and this is valid for nouns, adjectives, and numerals as well. The nominative case suffix is also generally assigned to the suffixless token before verbs, punctuation marks, definite articles and *is* 'too', *sem* 'neither', *nem* 'no', *pedig* 'in.turn', whether the token in question is a noun, an adjective, or a numeral. It is important to note that the unmarked possessor's role can not be determined with the same rules for every group of nominals.

Table 2.1. Table summarizing the common parts of the decision trees: rules that are applied to all the suffixless nominals regardless of their POS-tag.

first word in the window	role of the suffixless nominal
`NU`, `NU_MN`, *című*, *nevű*	NONE
`IGE`, `SPUNCT`, `ART`, *is* 'too', *sem* 'neither', *nem* 'no', *pedig* 'in.turn'	NOM

DOI: 10.15774/PPKE.BTK.2021.17

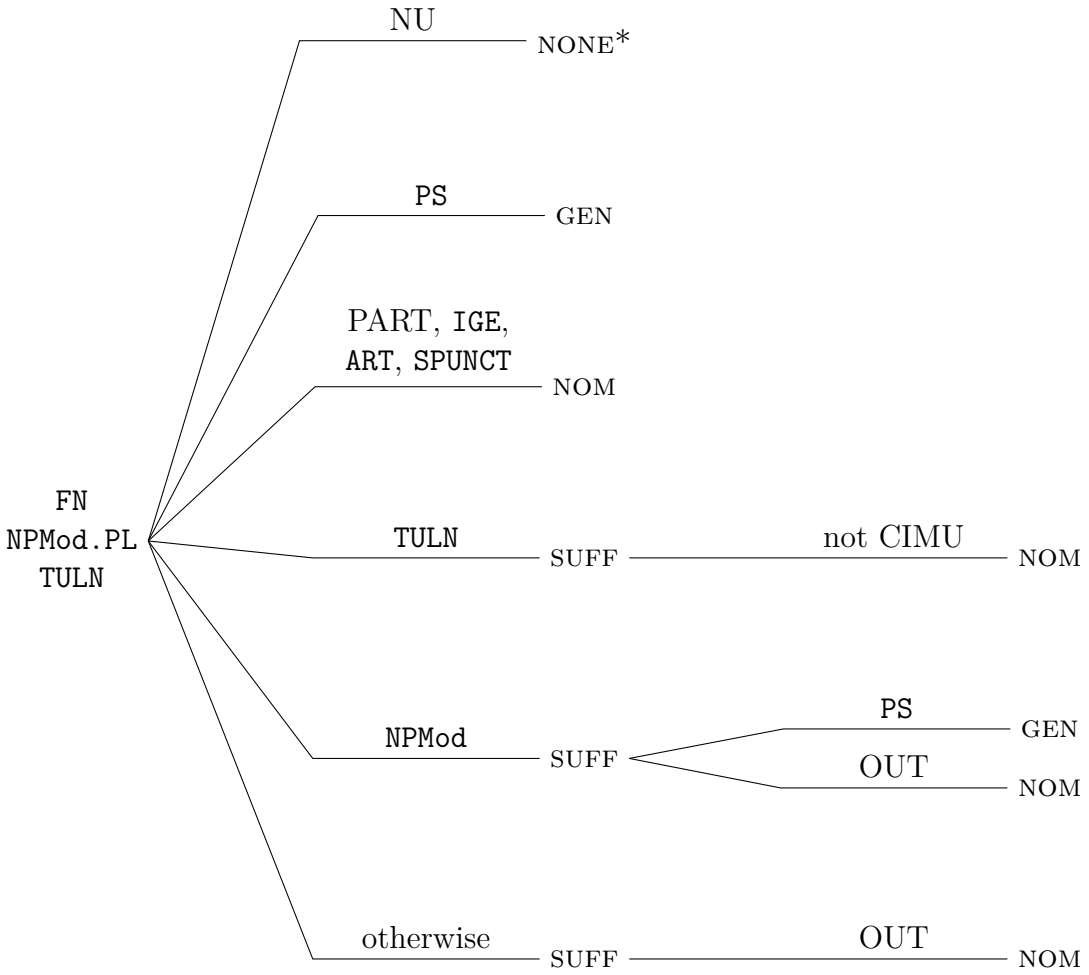

Figure 2.1. Decision tree summarising the rules applied to nouns (common nouns and proper nouns), and plural adjectives, numerals or participles. The root of the tree represents the part-of-speech tag of the current token. The labels on the edges of the first level of the tree illustrate the information retrieved from the first token in the parsing window. The labels on the edges of the second level of the tree represent the information gathered from the second token in the parsing window. The tags on the tree, on the one hand, are the tags used in MNSZ2.0 (the source of the texts we evaluated the algorithm on): `FN` stands for noun, `TULN` means proper name, `IGE` is the tag of verbs, `NU` is the tag of postpositions. `NPMod` is a label assigned to categories that can be modifiers of a noun (adjectives, numerals, participles). Some tags, on the other hand, are macros used only in these figures: NU stands for the tags and tokens [`NU`, `NU_MN`, *című* 'titled', *nevű* 'named']; PART is the macro of [*is* 'too', *sem* 'neither', *nem* 'no', *pedig* 'in.turn'], OUT stands for [`IGE`, `SPUNCT`, *aki* 'who', *ami* 'what', *hogy* 'that', *de* 'but'], CIMU stands for [*című, címen, címmel* 'titled', *nevű, néven, névvel* 'named']

* If the given token with the label `Nom` cannot be the unmarked possessor in a possessive structure in any way, then the algorithm assigns the case NOM to the token and the disambiguation is finished; hence no other branch of the tree is executed. Words of this category are *az*, 'that, *ez* 'this', *mindaz* 'all.of.that', *mindez* 'all.of.this', *aki* 'who', *ami* 'that'.

DOI: 10.15774/PPKE.BTK.2021.17

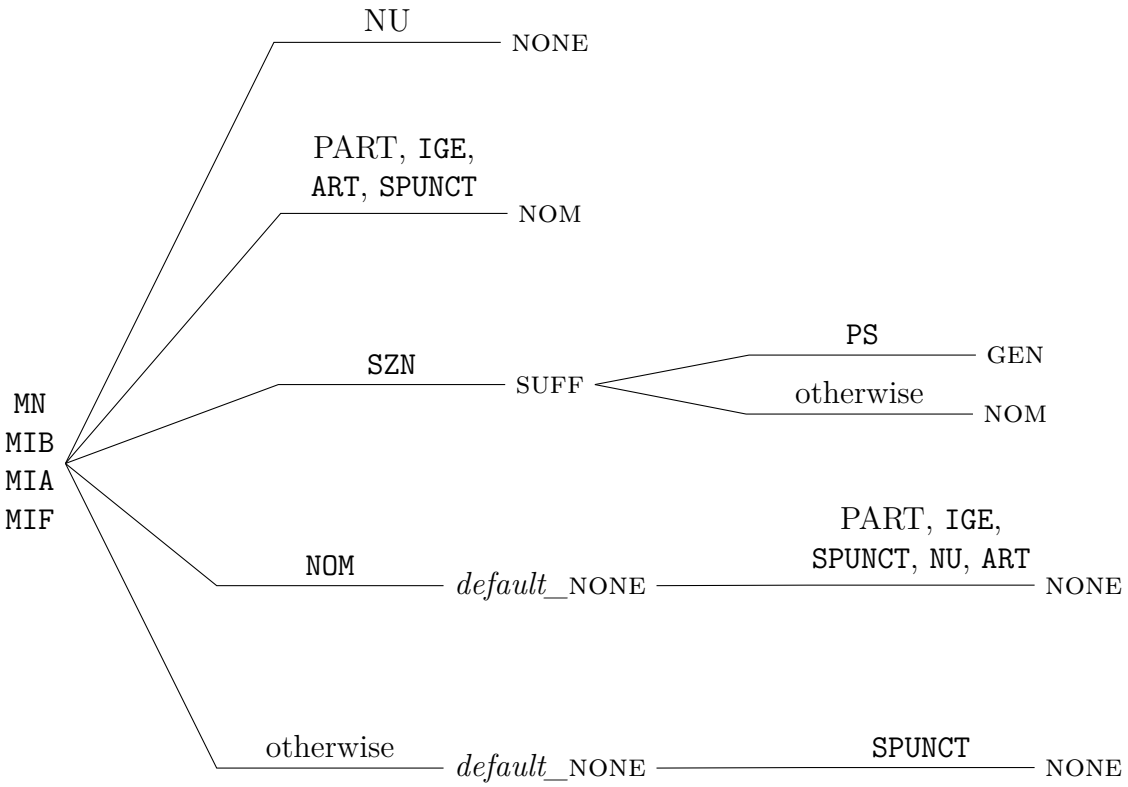

Figure 2.2. Decision tree summarising the rules applied to singular adjectives and participles. The detailed description of the tree can be found in the caption of Figure 2.1. The tags on the tree are the tags used in MNSZ2.0: MN is the tag of adjectives, MIB, MIA and MIF are the labels of participles, SZN means numerals, the other tags are explained in the caption of Figure 2.1.

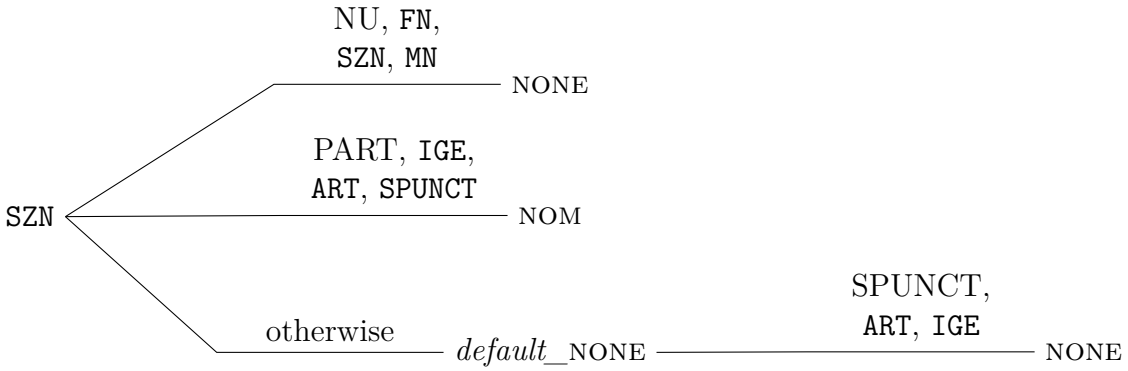

Figure 2.3. Decision tree summarising the rules applied to numerals. The detailed description of the tree can be found in the caption of Figure 2.1.

DOI: 10.15774/PPKE.BTK.2021.17

I demonstrate the working mechanism of the algorithm on the sentence chunk *szoba nagyobbnak tűnt* 'room seemed bigger' (22). The third line in example (22) shows the morphological annotation of the given tokens in the treebank. *Szoba* 'room' is the token with a `Nom` tag to be disambiguated. First, the corresponding tree (Figure 2.1, in this case) is chosen based on the POS-tag of the word (`FN`). Now, we turn to the first token in the parsing window, which is *nagyobbnak* 'bigger'. It is annotated as an adjective (`MN`), and adjectives are part of our category `NPMod`. The algorithm proceeds with the branch of the tree with a label `NPMod` on it. This means that, at this point, a SUFF label is assigned to *room* 'szoba'; based on the first token in the parsing window no certain decision can be made, thus the default tag is chosen. So the algorithm continues; the second token in the parsing window is *tűnt* 'seemed', and its POS-tag says it is a verb (`IGE`). The algorithm can now go on with the branch `IGE` on it: the suffixless *szoba* 'room' token receives a NOM tag; it is disambiguated, and is tagged as a Nominative, the subject of the sentence.

(22) ***szoba*** *nagy-obb-nak* *tűn-t*
room big-CMPR-DAT seem-PASTSG3
`FN.NOM` `MN._FOK.DAT` `IGE.Me3`

'room seemed bigger'

2.2.2 Implementation

The algorithm was implemented in Pyhton 3.8 (see A.1). The code presented here is solely my contribution to this project. The algorithm was planned to be inserted into `AnaGramma`, and Balázs Indig's future work would have been to adjust this code to make it fit into the framework of the parser.

2.2.3 Evaluation

The performance of the algorithm was evaluated on 1 000 sentences. The sentences were retrieved from MNSZ2.0[7] with the constraint that each sentence must contain at least one finite verb. Three modifications were made on the retrieved sentences:

[7]The retrieval was done by my colleague, Ágnes Kalivoda.

DOI: 10.15774/PPKE.BTK.2021.17

- As proper names have a significant role in disambiguating the suffixless nominal preceding them, we decided to manually annotate every proper name in the test sentences. We manually linked the members of extended named entities with an underscore together and replaced their `FN` tag with `TULN`.
- We removed every clause not containing a finite verb or only containing a copula. If the whole sentence had to be removed because of the copula, we retrieved another one to replace it.
- We carried out a cleaning step by removing items that are not part of the sentence, e.g. numbers at the beginning or at the end of the sentence etc.

Each token of the 1 000 sentences (annotated and cleaned as described above) with an original `NOM` tag, received three analyses: one from the algorithm, one during the manual annotation based on the tokens in the two-word-wide parsing window, and one during the manual annotation as a final analysis (based on the whole sentence). An extract of the annotated test corpus can be found in Appendix A.3.

The goal of these two manual annotations is to facilitate a more precise and complex evaluation of the algorithm's performance and of its theoretical background as well. The point of the manual decision that was made based solely on the window (SUFF in example 23) is to judge the extent to which the algorithm accomplishes the behaviour it is expected to carry out. A good performance means that the algorithm is capable of defining as accurately as possible what role the given suffixless nominal has in the sentence based on the tokens in the parsing window.

The labels manually assigned to the suffixless nominals based on the processing of the whole sentence (NOM in example (23)) are to declare the definitive role of the given token (in example (23) *patak* 'brook' is the subject of the sentence). With this annotation we intended to judge the theoretical background of the algorithm: if the two manual annotations match, then the process we are using is correct and accurate; the role of a suffixless nominal can be specified with great certainty based on the two-token-wide parsing window, without looking further (in other words: these roles can be assigned in the first phase of the two-stage parsing model).

(23) An example with the suffixless nominal in bold.

A ***patak*** *tőlük keletre húzódott.* 'The **brook** stretched east of them.'

DOI: 10.15774/PPKE.BTK.2021.17

the window: ***patak*** *tőlük keletre* '**brook** of.them east'

the label assigned by the algorithm: *patak FN.*NOM

manual annotation, based on the window: *patak FN.*SUFF

manual annotation, based on the whole sentence: *patak FN.*NOM

During manual annotation,[8] as a first step, considering only the information on the two tokens in the parsing window, we can assign suffixless nominals with the following tags:[9]

- NOM: nominative case
- GEN: unmarked possessor
- NONE: no case suffix at all (nominal preceding a postposition, or a modifier of another nominal)
- VOC: vocative role
- XNE: a tag marking the inner elements of extended named entities
- SUFF: still unclear; the nominal acquires the default value of nouns (which is to be disambiguated into NOM or GEN later)
- *default*_NONE: still unclear; the nominal acquires the default value of the elements of the `NPMod` category
- *false_pos*: one of the tokens in question has a false POS tag therefore the analysis of the algorithm deviates (e.g. because of a verb analysed as a participle)

In the second phase of manual annotation, when making the decisions based on the whole sentence, nominals waiting to be disambiguated may be assigned with the following labels:

- NOM
- GEN
- NONE

[8] The 1000 sentences were manually annotated by Andrea Dömötör, Noémi Vadász and me.

[9] PRED is not needed, as we excluded sentences with a predicative nominal from our corpus.

DOI: 10.15774/PPKE.BTK.2021.17

- VOC
- XNE
- SUFF or *default*_NONE, in case the whole sentence is ambiguous

There were 128 tokens in the test corpus where the annotation of the suffixless nominal itself was false (e.g a verb was labelled as a participle with `Nom` case), or a token in the parsing window was tagged erroneously (the participle following the suffixless nominal was tagged as verb etc.). These instances were not corrected manually and were not counted in the evaluation. There were 34 nominals in a vocative role. However, as the recognition of a *vocativus* heavily depends on orthography (it has to be separated from the other part of the clause with a comma), `Nom-or-What` is not prepared to identify it. Therefore, when evaluating the performance of the algorithm by comparing its output to the manual annotation made by looking only at the parsing window, we did not encounter the tokens labelled as VOC.

45 tokens received the tag XNE meaning that they are the inner members of an extended named entity (see example (24); the surname *Ottlik* and the first name *Géza* were linked by us in the test sentences; the proper name as one unit received the tags originally assigned to *Géza*; this way *Ottlik_Géza* popped up as a nominal with the tag `Nom`; it should be labelled with a NONE, as it is the member of an extended named entity: *Ottlik_Géza író* 'writer Géza Ottlik'). The identification of a token like this largely depends on world knowledge, thus it cannot be processed by rules relying on a two-word-wide parsing window only. Therefore we decided not to count these tokens when evaluating the performance of the algorithm (comparing its labels to that of the first manual annotation).

(24) ***Ottlik_Géza*** *író*
Ottlik_Géza writer

'writer Géza Ottlik'

Table 2.2 summarises the rules of the evaluation. True positive (TP), false positive (FP) and false negative (FN) hits were defined. The columns of the table are to be read as follows: if the value in the column "label to evaluate" matches the value in the column "gold" then it is a TP, FP or FN hit depending on the row this pair of values can be

DOI: 10.15774/PPKE.BTK.2021.17

found in. This comparison is valid if it is between the label the algorithm assigned to a given token and the manual annotation relying only on the information in the parsing window, or between two manual annotations. TP hits are complete matches. Note that even a default tag can be a TP result: if the algorithm assigns a default tag to a given token and the manual annotation based on the parsing window shows that the role of the given token can not be specified based on the information in the window, then the algorithm's decision is correct, the default tag is the appropriate tag in that case. FP reflects overspecification: if, for example, the algorithm decides a suffixless nominal to be the subject of the sentence (assigning a NOM label to it), but the manual annotation based on the window states that the role of the token cannot be specified exactly yet and assigns a default SUFF to the token, then it is an FP hit. If, on the other hand, the algorithm underspecifies (it labels the token with a default value while its role can be specified based on the window), it is an FN hit.

Table 2.2. Rules of evaluation

category	label to evaluate	gold
TP	NOM	NOM
	GEN	GEN
	NONE	NONE
	SUFF	SUFF
	*default*_NONE	*default*_NONE
FP	NOM	SUFF
	GEN	SUFF
	NONE	*default*_NONE
	all other non-matching cases	
FN	SUFF	NOM
	SUFF	GEN
	*default*_NONE	NONE

2.2.3.1 Evaluating the parsing window

In Table 2.3 the precision and recall of the theoretical background, the idea of disambiguating the "suffixlessness" based on the two-word-wide parsing window can be seen when comparing it to the disambiguation of these tokens knowing and considering the whole sentences (in other words, this table shows the evaluation of the first manual anno-

DOI: 10.15774/PPKE.BTK.2021.17

tation). As can be seen, we reached a high precision (98.26%). It was a fundamental goal of `AnaGramma`, and thus of our study as well, to make decisions as precisely as possible so that in any later phase of the parsing process there would be no need to correct earlier verdicts that proved to be false. We managed to meet this expectation with the precision seen here.

Table 2.3. Evaluating the manual annotation that considered only the two-token-wide parsing window by comparing its labels to the result of the manual annotation considering the whole sentence.

TP	**FP**	**FN**	**precision**	**recall**	**F-measure**
1 753	31	638	98.26%	73.32%	83.98%

Recall, however, is only 73.32%: this low value is mainly caused by the relatively high number of false negative hits which require further analysis. As defined above, FN hit means that a token was labelled with a default tag when only relying on the parsing window, but after processing the whole sentence, the token's role is further specified. In other words, underspecification is seen as a false negative hit. Table 2.4 summarises the false negative results in this comparison.

Table 2.4. False negative results in the comparison of the two different manual labels. The rows show how many times the given underspecification took place: for example, the window-based manual annotation assigned a SUFF to a given suffixless nominal, while the second manual annotation specified the role of the given token as a subject (NOM) 187 times.

error type	#
FN	638
SUFF instead of NOM	187
SUFF instead of GEN	54
*default*_NONE instead of NONE	397

It must be seen, however, that these false negative results are not necessarily inaccurate; the default role of nouns, SUFF, is deliberately constructed to represent the two case suffixes, NOM and GEN: using this label means that based on the parsing window it cannot be clearly decided whether this nominal with the tag SUFF is the subject of the

DOI: 10.15774/PPKE.BTK.2021.17

sentence or the unmarked possessor of a possessive structure; only the broader context can clarify its role.

The instances of the default_NONE-s are more promising: the number of adjectives or participles where we assigned a default label based on the window, but in the second round of the manual annotation their suffixlessness was disambiguated as a NONE is extremely high. Moreover, there were only six cases when the default_NONE tag assigned by the window-based manual annotation was disambiguated as a NOM (meaning that the given adjective or participle proved to be the subject of the sentence), in every other case the default_NONE happened to be a NONE in the end. What does this mean? I give an example in (25): (25a) shows the token to be labelled (*telepített* 'settled') and the following two words (the parsing window). Based on the window one can imagine a sentence where the token in question is a subject or an unmarked possessor; however, when reading the whole sentence (25b), it becomes clear that this participle is the modifier of the noun *családban* 'family.INE'.

(25) a. ***telepített*** *mintegy negyven*
settle-PST.PTCP some forty

label based on the window: telepített telepít IGE._MIB.*default*_NONE

b. *Kétéves koromban elvesztettem anyai nagyszüleimet, s velük együtt a szülőfalumból Magyarországra* ***telepített*** *mintegy negyven* ***családban*** *szinte minden rokonomat.*
At the age of two, I lost my maternal grandparents, and with them, almost all my relatives in about forty ***families forced to move*** *from my home village to Hungary.*
label based on the sentence: telepített telepít IGE._MIB.NONE

To sum up, based on the results discussed above it can be stated that the high precision (shown in Table 2.3) meets our expectations for the process (disambiguating the "suffixlessness" based on the parsing window only). Recall could be improved by automatically assigning every adjective and participle the label NONE instead of inserting an unnecessary intermediate tag, the default_NONE.

DOI: 10.15774/PPKE.BTK.2021.17

2.2.3.2 Evaluating the algorithm's performance

Table 2.5 shows the evaluation of the performance of `Nom-or-What`. Here the results of the algorithm were compared to the manual annotation that used only the information gathered from the parsing window.

Table 2.5. Evaluating the performance of the algorithm on the manual annotation made by considering only the two-word-wide parsing window. False positive results include instances of tokens in a vocative role (34) and inner members of extended named entitites (42) – the algorithm is not prepared to handle these cases.

TP	FP	FN	precision	recall	F-measure
2 112	162	148	92.88%	93.45%	93.16%

Both the precision (96.08%) and the recall (93.45%) are high: `Nom-or-What` is reliable; it disambiguates all that can be disambiguated based on the parsing window, but does not specify roles that should not be specified yet. The high number of TP hits (2 112) is itemized in Table 2.6. As can be seen, 71% of the TP results (1 501) are specific tags, and only 29% of them (611) are default tags. Another value worthy of inspection is the last row of Table 2.7; here are the number of cases when the algorithm underspecifies: it assigns the default label to adjectives and participles instead of providing a NONE. These cases are parallel to the phenomenon seen in Table 2.3 explained in example (25). The 131 instances in Table 2.7 further reinforce the idea that instead of a default_NONE tokens of the category `NPMod` should always be labelled with NONE (if no other case suffix is appropriate).

Table 2.6. True positive results obtained when evaluating the performance of the algorithm. The rows show how many types a given tag appeared correctly.

tag	#
NOM	621
GEN	257
NONE	623
SUFF	247
default_NONE	364
all TP	2 112

DOI: 10.15774/PPKE.BTK.2021.17

Table 2.7. False negative results obtained when evaluating the performance of the algorithm. The rows show how many times the given underspecification took place: for example, the algorithm assigned a SUFF to a given suffixless nominal while the window-based manual annotation specified the role of that given token as a subject (NOM) 13 times.

error type	#
FN	148
SUFF instead of NOM	13
SUFF instead of GEN	4
default_NONE instead of NONE	131

Table 2.8 shows the evaluation of the performance of `Nom-or-What` on the gold standard annotation created by gathering all the information from the whole sentence. As expected (considering the fact that the algorithm works well and its decisions are close to the manual ones based on the window, see the results in Table 2.5), the numbers are close to the ones in Table 2.3. The precision is still high. Underspecification (the high number of FN hits), notwithstanding is a problem here as well (as it was in Table 2.5).

Table 2.8. Evaluating the performance of the algorithm on the manual annotation made by considering the whole sentence. False positive results include instances of tokens in a vocative role (34) and inner members of extended named entitites (42) – the algorithm is not prepared to handle these cases.

TP	**FP**	**FN**	**precision**	**recall**	**F-measure**
1 573	132	717	92.26%	68.69%	78.75%

2.2.4 Widening the window - on the usefulness of the third word

As discussed in sections 2.2.3.1 and 2.2.3.2, the large number of FN hits represents many instances of underspecification: a default tag is assigned to a token (by the algorithm or by the window-based manual annotation), yet its role is specified when the whole sentence is known. The question arises: if we widen the two-token-wide parsing window (and, for example, include the third consecutive word into the analysis), does the number of FN hits decrease? In other words: can we reduce underspecification?

DOI: 10.15774/PPKE.BTK.2021.17

To answer this question, I selected a sample of 50 sentences that contained FN hits (87 altogether). Referring back to Table 2.3, this sample is 13% of the total set of underspecification. Table 2.9 shows the results.

Table 2.9. The table summarizes the results of the manual annotation of a sample of 50 sentences (with 87 underspecified tags in them).

	could be specified to	**remained default**
default_NULL	NULL: 18	38
SUFF	NOM: 7	13
	GEN: 0	11
sum	35	62

As can be seen, approximately 40% of the underspecified tags can be (correctly) specified based on the third token in the parsing window. However, this proportion is not evenly distributed among the different types of underspecification. The subject of the sentence (NOM) could be determined based on this third token in more than half of the cases (7 out of 13); if the unmarked possessor, on the other hand, was not specified based on the original parsing window, could not be specified based on a wider window either (0 out of 11). The role of tokens of the category `NPMod` could be specified based on the third word in half of the cases.

These numbers indicate, that – especially in the case of nouns (or plural adjectives and proper names) – it may worthwile writing rules that rely on the third word in the parsing window to be able to precisely specify the subject role of a given nominal; the recall could be improved (if we scale the above numbers, it can be predicted that it would be increased to ∼83% from the current 73%). However, it would not help the specification of the role of the unmarked possessor.

2.2.5 Summary

In this section I presented the algorithm `Nom-or-What` whose goal was to do a case disambiguation of nouns, adjectives, numerals and participles bearing no overt case suffix, hence automatically labelled as `Nom`. The disambiguation must take place by considering only the morphological annotation of the given token and of the two tokens following it.[10].

[10]The implementation of the algorithm, the test corpus and the annotated corpus is available at the following url: `https://github.com/ppke-nlpg/Nom-or-What` I also attach the programming code of `Nom-or-What` in Appendix A.1 with a short extract of the annotated corpus in Appendix A.3

DOI: 10.15774/PPKE.BTK.2021.17

First, I set up the rules of the disambiguation based on corpus research. Then I implemented the algorithm and evaluated its performance on a test corpus consisting of 1 000 sentences that were manually annotated. The results show that the algorithm performs well, with high precision and recall outcomes. Moreover, by conducting a double manual annotation it became possible to evaluate the theoretical background itself, namely, that the role of a token may be specified with great certainty based on a two-token-wide parsing window and nothing else. The results show that it is indeed possible to specify the role of a suffixless nominal without processing the whole sentence. However, the algorithm is not yet capable of recognising and specifying the vocative role of a nominal, mostly because without parsing the whole sentence and knowing the person and number feature of the verb, for instance, one can only rely on orthography to detect a *vocativus*. Neither can the algorithm detect the members of extended named entities - the recognition of those requires world knowledge. What is more, a bigger issue is the detection of nominal predicates which is impossible without looking further (back or forth) in the sentence.

2.3 `Nom-or-Not`?

In this section I briefly present an upgraded version of `Nom-or-What` called `Nom-or-Not` which was an attempt to combine a function designed to solve the problem of nominal predicates (Dömötör, 2018) and `Nom-or-What`; however, it produced less impressive results than `Nom-or-What`, raising more questions than it answered. This algorithm was published in Ligeti-Nagy et al. (2019) and is the result of joint efforts between my colleagues Andrea Dömötör and Noémi Vadász and me.

2.3.1 Background and motivation

The algorithm described in Dömötör (2018) (named `is-pred`)[11] was also designed to constitute a part of the AnaGramma parser. It follows the principles of the sausage machine model described in section 1. As I presented in section 2.2 `Nom-or-What` was basically the implementation of the first phase of this two-phased parsing model. The second phase carried out by `is-pred` uses the whole left context, the so-called *pool*. Moreover, `is-pred` strongly relies on the output of `Nom-or-What`, as there would be little chance to identify

[11] The algorithm `is-pred` is the work of Andrea Dömötör.

DOI: 10.15774/PPKE.BTK.2021.17

predicative nominals based exclusively on the left context without taking into account the local decisions of the first phase.

In sum, the input of `is-pred` is a sequence that consists of the nominal in question and the part of the sentence that precedes it. The left context of the current word is already analysed and disambiguated by `Nom-or-What` (if possible), thus the algorithm can use various pieces of morphosyntactic information from the pool. The output is a value, similar to trivalent logic: `Pred` if the nominal is obviously a predicate, `Nonpred` if it is obviously not a predicate, and `Undefined` if its syntactic role is still unclear from the given information.

The `is-pred` algorithm achieved high precision on its test, though it has some deficiencies that need improvement. Firstly, its responses are binary which do not complete the analysis in the `Nonpred` cases. Second, `is-pred` only handles the predicative copular clauses, therefore the recognition of nominal predicates in equative sentences is a significant gap that this study intends to fill.

The idea behind this algorithm – called `Nom-or-Not`, referring to its role as a synthesis of its antecedents – was, on the one hand, to merge all working and tested rules of previous algorithms, and on the other hand, to fill as many remaining gaps as possible.

2.3.2 Method

The method of `Nom-or-Not` follows `Nom-or-What` and `is-pred` in being rule-based which means that the algorithm does not use machine learning approaches, but rather it is built on linguistically grounded hand-crafted rules. The main difference among the three is that `Nom-or-Not` merges the two phases of parsing and aims to disambiguate each possible role of suffixless nominals in one step. For this task, it is necessary to use both the window and the pool at the same time, therefore the algorithm operates with both forward- and back-looking rules. In either case, the principal source of information is the morphological annotation with only a small scent of lexical information. That is, the disambiguation of suffixless nominals is carried out primarily based on the syntactic structure.

The algorithm is designed to process sentences annotated by the *emMorph* morphological analyser (Novák, 2003, 2014; Novák et al., 2016), where the token, the lemma and the morphological tags are separated by a /, and the morphological tags are in square brackets (*USA/USA/[/N][NOM]*). The algorithm processes the sentences from left to right, word

DOI: 10.15774/PPKE.BTK.2021.17

by word. The rules are only applied if the token under examination is tagged as `Nom`. As the targeted parsing method has a psycholinguistic motivation, the case disambiguation algorithm first gathers all the information of the given nominal that is deducible from the pool (the collection of the information of the already processed elements). The back-looking rules are used for preliminary disambiguation of predicative nominals (derived from Dömötör, 2018), and they are listed below (the label used to mark pedicative nominals is PRED, all other labels are the same as the ones presented in 2.1.1):

- If there is a non-copular finite verb in the pool → the current token is not PRED
- If there is a nominative in the pool → the current token is PRED, if other cases will be ruled out based on the window, and only NOM and PRED remains as an option
- If the word is the possible head of a DP and there is no nominative in the pool → it is not PRED
  - If proper name → Head of DP
  - If possessive → Head of DP
  - If preceded by a determiner and optionally one or more NP-modifiers → Head of DP
  - If demonstrative pronoun ('this', 'that') → Head of DP

Having exploited the left context the algorithm refines its judgement about the nominal in question using the information gathered from the window. The forward-looking rules are almost the same as the ones displayed in Figures 2.1, 2.2 and 2.3. To illustrate the differences, I present the rules activated when the token in question is a noun, a proper name, a plural adjective or a plural participle in Figure 2.4. Obviously, only those branches will be activated that are relevant taking into consideration the conclusions drawn from the information coming from the pool; and every non-final decision is finalised if the knowledge based on the pool makes it possible to rule out a part of the outcome. (E.g. an edge leads us to a leaf with the tag *nom_or_pred* on it, but the pool already made it clear that the actual token cannot be a PRED, therefore here the tag NOM will be assigned to this token.) As the algorithm does not exploit the whole sentence, cases may remain where no certain decision can be made. We use the following tags for these cases, besides the ones used in the case of `Nom-or-What`:

DOI: 10.15774/PPKE.BTK.2021.17

- *Nom/Pred*: a tag signalling that the given word may either be the subject of the sentence or the nominal predicate
- NONE/*Pred*: a tag signalling that the given word may either be a modifier element in an NP or the nominal predicate of the sentence

The algorithm implemented in Python is available with the test corpus containing the gold standard annotation at `https://github.com/ppke-nlpg/nom-or-not`.

DOI: 10.15774/PPKE.BTK.2021.17

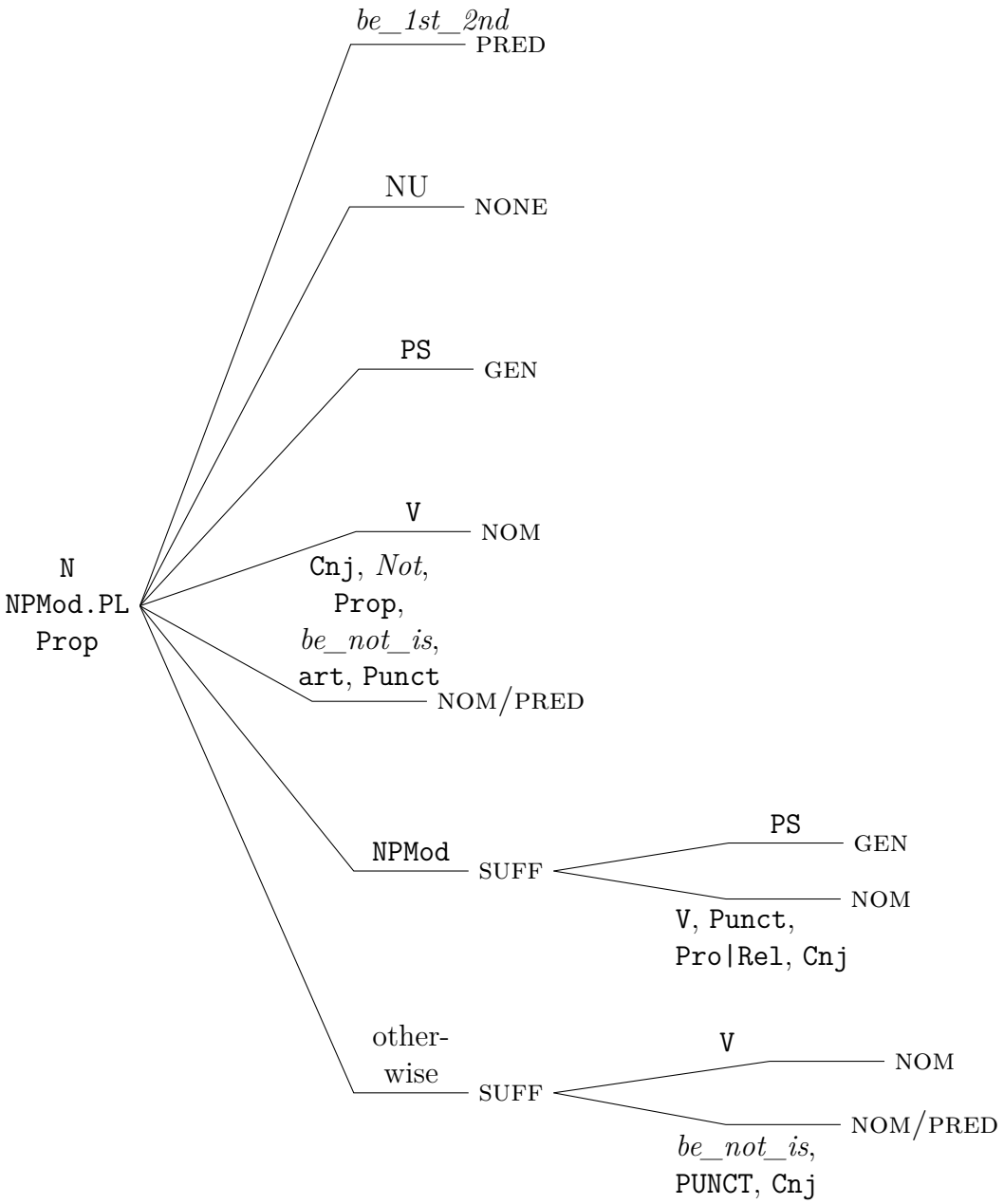

Figure 2.4. Decision tree summarising the rules concerning nouns, proper names, and plural adjectives, numerals and participles. The root of the tree is the POS-tag of the token under examination. The edges on the first level contain information seen on the first element in the parsing window. The edges on the second level contain information seen on the second element in the parsing window. *be_1st_2nd* is a macro for the 1st and 2nd person forms of the copula. *Not* is a macro for negation. *be_not_is* is a macro for any copula except for the singular and plural 3rd person form of *be*.

DOI: 10.15774/PPKE.BTK.2021.17

2.3.3 Results

For the evaluation of the performance of the algorithm we used a randomly composed subcorpus of the Hungarian Gigaword Corpus. The test corpus contains 500 sentences with no restriction to genre, content or quality. We carried out the morphological analysis of the sentences with the *emMorph* tool integrated in `e-magyar` language processing system (Váradi et al., 2018).[12]

The testcorpus contains 2 255 tokens tagged as `Nom` by the morphological analyser. We manually annotated them with tags from the set described above. The output of the algorithm was compared to this gold standard. It is important to note that the human annotation took the whole sentence into consideration and no default tags were allowed (unless the whole sentence was ambiguous). As the algorithm operates without analysing the whole sentence, it accordingly provides ambiguous responses in some cases, meaning that we cannot expect 100% recall. The algorithm was consciously designed to work with high precision instead of high recall.

The evaluation follows the rules described in Table 2.10. The true positive (TP) matches are the correct ones. The erroneous or overspecified results are considered false positives (FP). Finally, we refer to the uncertain (underspecified) responses of the algorithm as false negatives (FN). The results are shown in Table 2.11.

As can be seen, the algorithm achieved moderately good recall and precision lower than expected. We analysed the results in more detail in a confusion matrix (Table 2.12). The rows display the responses of the algorithm, while the columns show the gold standard annotation.

A significant number of the errors (102) is due to an invalid morphological annotation of the surrounding tokens. We eliminated those from the final results.

2.3.4 Discussion

As expected, the algorithm performs with a moderately high recall compared to that of `Nom-or-What` (67.63%) which is thanks to some of the default tags being eliminated from the algorithm. Recall is influenced by the number of false negative hits (361 in the results). Considering that the algorithm does not have the whole sentence available when

[12]The extraction of the sentences was made by Andrea Dömötör, while the manual annotation was carried out by Andrea Dömötör, Noémi Vadász and me.

DOI: 10.15774/PPKE.BTK.2021.17

Table 2.10. Rules of evaluation. The tags in the *result* column are the ones assigned by the algorithm. The tags in the *gold* column are the gold standard annotation.

category	result	gold
TP	NOM	NOM
	GEN	GEN
	PRED	PRED
	NONE	NONE
	SUFF	SUFF
FP	NOM	SUFF
	GEN	SUFF
	NONE	NONE/PRED
	PRED	NONE/PRED
	NOM	NOM/PRED
	PRED	NOM/PRED
	every other non-matching tags	
FN	SUFF	NOM
	SUFF	GEN
	NONE/PRED	NONE
	NONE/PRED	PRED
	NOM/PRED	NOM
	NOM/PRED	PRED

Table 2.11. Test results of the `Nom-or-Not` algorithm evaluated on 500 randomly selected and manually annotated sentences

Precision	Recall	F-measure
77.82%	79.3%	78.55%

Table 2.12. Confusion matrix. The rows refer to the tags assigned by the algorithm. The columns represent the gold standard annotation.

	`Nom`	`Gen`	none	`Pred`	`Voc`	suff	***Nom/Pred***
`Nom`	281	9	54	57	0	0	0
`Gen`	3	229	8	0	0	0	0
none	1	1	811	3	0	0	0
`Pred`	105	0	2	62	0	0	5
`Voc`	0	0	0	0	0	0	0
suff	83	27	79	5	0	0	0
Nom/Pred	143	2	26	69	1	1	0
none*/Pred*	0	0	39	0	0	0	0

DOI: 10.15774/PPKE.BTK.2021.17

deciding, underspecification (resulting in false negative hits) is understandable in many of the cases. These results are not as problematic for the whole parsing task as the false positive ones, since the uncertain tags can still be specified at a later point of parsing with the scanning of further words.

The confusion matrix in Table 2.12 reveals that the majority of FP hits (268) is in connection with NOM or PRED, and more than half of them (162) is caused by a swap of these two tags. This can be explained, on the one hand, with the fact that our rules detecting predicative nominals are highly dependent on our preceding decisions on nominals: if a NOM was found, we assume that no more NOM should be identified. However, our rules do not take clause boundaries into consideration, even though a previously found NOM may be the subject of a clause other than the one under examination. Stopping the backwards-looking rules on clause boundaries is a rather essential issue to solve later. Obviously, any erroneously annotated NOM can lead to further mistakes during the analysis, even within the same clause. On the other hand, transposing NOM with GEN or vice versa is often caused by a verb falsely considered a copular verb. *Lehet* (may be) or *lesz* (will be) are just two examples of verbs that can either be a copular verb or a normal verb. This distinction is not available in their current morphological annotation, therefore the algorithm always assumes them to be a copular verb.

Another source of errors (159 cases) is the undiscovered inner structure of extended named entities and constructions such as (26a) and (26b). Here we assume that there is no case suffix on the first element, therefore a NONE would be the correct tag for it. However, detecting these names is challenging, it was not solved in `Nom-or-What` and nor in `Nom-or-Not`.

(26) a. *elnök* *úr*
president sir
`N.Nom` `N.Nom`
'Mr. President'

b. *Kinaesthetics* *termék*
Kinaesthetics product
`Prop.Nom` `N.Nom`
'the product Kinaesthetics'

DOI: 10.15774/PPKE.BTK.2021.17

Finally, cases like (27a) present a challenge to the algorithm as well: these are some sorts of exclamations without any particular case suffix on them, as they play no role in the sentence. We would assign a NONE tag to them, but their distinction is quite problematic and at the moment unsolved in a sentence.

(27) a. *Támadás!*
attack
N.NOM

'Attack!'

Setting the unsolved problems and all the errors aside, we can see that the algorithm performs well with the unmarked possessor's case and with tokens not bearing any suffix at all (tagged with NONE). With PRED, on the other hand, `Nom-or-Not` is quite uncertain, though it never assigns any GEN or NONE tag to the nominal predicates of a sentence.

A part of the underspecification (FN results) may be solved by inserting a final step at the end of the analysis of each sentence: any verb following the tokens tagged as *Nom/Pred* can clarify its role as NOM.

2.3.5 Conclusion

In this section I presented a rule-based algorithm called `Nom-or-Not` which was meant to be the successor of some related algorithms, each of which were implemented to solve a small part of the complex problem. When designing `Nom-or-Not` I intended to provide an algorithm able to deal with every possible role of suffixless nominals.

I presented the design of the algorithm accompanied by the preliminary results obtained by evaluating the algorithm's performance on a test corpus containing 500 manually annotated sentences. Although I expected a higher precision, the majority of FP results is not a random mistake, but rather a systematic error that can and should be solved by extending my rules or by evaluating the algorithm on a more precisely annotated test corpus. The recall is higher than the expectations, proving that eliminating the default tags of adjectives, participles and numerals results in a better performance.

As can be seen, while `Nom-or-What` proved to be quite a success, though this cannot be said about `Nom-or-Not`. This indicates that resolving the ambiguous cases in the first

DOI: 10.15774/PPKE.BTK.2021.17

phase of parsing is indeed easily manageable, but when the given role (nominal predicates) has a wider context to rely on, hand-crafted rules with local decisions may not be enough.

There are numerous tasks ahead: first, it is necessary to revise the rules concerning predicative nominals, as they seem to generate a significant number of FP results. This may be supported by further studies on Hungarian copular sentences. After inserting a final check in the algorithm that enables it to clarify the role of tokens temporarily annotated with a tag of a default value, `Nom-or-Not` may provide a solution replete with high precision and recall for this case-disambiguation task for Hungarian.

2.4 Summary

This chapter meant to plunge into the computationally significant problems concerning noun phrases in Hungarian. There may be numerous nominal tokens in a sentence that do not bear any overt case suffix (they are suffixless), making it difficult for a parser to specify their exact role in the sentence. This "suffixlessness" may encode several different meanings: it may mark the subject of the sentence, an unmarked possessor in a possessive structure, a vocative role, a nominal followed by its postposition, a modifier of an other nominal, or a predicative nominal. My hypothesis was that most of these roles can be specified during parsing by local decisions without the need of knowing the whole sentence. To test this hypothesis, I prepared a set of hand-crafted rules which only took the information gathered from a two-token-wide forward-looking parsing window into account. After implementing these rules as the so-called `Nom-or-What` algorithm, I tested its performance on a test corpus consisting of 1 000 sentences. The double manual annotation of these sentences made it possible to evaluate not only the performance of the algorithm, but also the basic theoretical idea itself. The results show that on the one hand, the algorithm performs well, it is reliable; on the other hand, it is confirmed that most of these roles that do not require the presence of a case suffix specifically marking them, can be made unambiguous locally without processing the whole sentence.

DOI: 10.15774/PPKE.BTK.2021.17

Chapter 3

Extended named entities

"We know what we are, but know not what we may be"

Hamlet Act 4, Scene 5

3.1 Introduction

This chapter focuses on named entities composed of a proper name and a common noun (the exact term to be clarified later), as they pose a severe challenge to NP-chunking. One of my earlier papers (Ligeti-Nagy, 2015) dealt with the problem of NP-chunking from a linguist's point of view. I intended to examine current morphological tags with regards to their ability to help determine the boundaries of noun phrases. To do this, I collected patterns of morphological tags possibly marking a noun phrase. (The corpus used for this study was Pázmány Corpus: Endrédy, 2016; Endrédy and Prószéky, 2016.) One of my examples showed the following pattern:

`N.NOM N.NOM N.NOM N.NOM N.CAS`

Here, four suffixless nouns (tagged as `nom`) follow each other while the sequence ends with a noun with a case suffix (the makro `CAS` refers to any case ending). Some of the examples I provided for this pattern can be seen in 28a - 28c.

(28) a. *Angela Merkel | Wolfgang Schäuble pénzügyminiszter-rel*
Angela Merkel | Wolfgang Schäuble minister.of.finance-INS

'Angela Merkel [...] with Minister of Finance Wolfgang Schäuble'

DOI: 10.15774/PPKE.BTK.2021.17

b. *Orbán Viktor miniszterelnök | zéró toleranciá-t*
Orbán Viktor prime.minister | zero tolerance-ACC

'prime minister Viktor Orbán [ordered] zero tolerance'

c. *Bundáné Badics Ildikó jegyző | péntek-en*
Bundáné Badics Ildikó notary | Friday-SUP

'notary Ildikó Bundáné Badics [...] on Friday'

A potential dilemma arises when the morphological analysis of the tokens is the same in all the three examples, though the NP boundaries occupy different places. In (28a-28c) a vertical bar (|) marks the boundary between two NPs. To make the difference even more visible, (29a-29c) only shows the morphological tags of the tokens with the NP-boundaries.

(29) a. `N.NOM N.NOM | N.NOM N.NOM N.SUP`

b. `N.NOM N.NOM N.NOM | N.NOM N.SUP`

c. `N.NOM N.NOM N.NOM N.NOM | N.SUP`

With these examples I light is shed on the fact that the morphological information encoded in these tags is not sufficient to determine where an NP might start or end. However, with a slightly modified morphological tagset we may be able to encode every necessary piece of information into the annotation, thus making a rule-based NP-chunking easier. (For some previous attempts on that topic see Ligeti-Nagy, 2015 and Ligeti-Nagy, 2016.)

In this chapter, I focus on one particular phenomenon resulting in examples like (28a-28c): named entities like *Bundáné Badics Ildikó jegyző* 'notary Ildikó Bundáné Badics', *Wolfgang Schäuble pénzügyminiszter* 'minister of finance Wolfgang Schäuble' and *Orbán Viktor miniszterelnök* 'prime minister Viktor Orbán'. In these cases one or more proper nouns are combined with a common noun into one named entity. The problem here is that the relation between the first and the second part of the name is unmarked.

Although I started with examples originating from the field of NP chunking, the whole issue discussed here is rather a task for named entity recognition, NER (see section 3.2.3).

DOI: 10.15774/PPKE.BTK.2021.17

3.1.1 Terminology

The assertion may arise that this structure is nothing more than a compound whose components are written separately. However, referring to this structure as an example for open compounds in Hungarian is a simplification: these structures are certainly some sort of compounds, but more is known about them than just that. Moreover, they are some kinds of named entities, and their first member is a proper name, and their second is a common noun. Hence, a more precise term is needed.

Simon (2013) distinguishes *monomorphemic* (*Charlie*) and *polymorphemic* proper names (*Roosevelt square*). Polymorphemic here means that the named entity consists of more than one word. Although this naming convention might fit into my analysis here, I am going to refer to the structures under investigation as *extended named entities*, or XNEs.[1] This term states nothing more or less about this structure than what we already know: they are named entitites, but somehow more than just simple proper names.

3.2 Background

3.2.1 What this structure is not

There is a structure in Hungarian referred to as *értelmező*, appositional construction. While it is a frequently discussed phenomenon in traditional Hungarian grammars (see for example Jakab, 1977, 1978; Károly, 1958, 1962 etc.), Szőke (2015) provided its first deep analysis in a generative framework. Alberti and Laczkó (2018) distinguish two main types of appositional constructions: attributive apposition (30a) – the structure is called extraposed modifier in traditional Hungarian grammars – and second, identifying apposition (30b). Szőke (2015) even proposes a third category: adverbial apposition (30c). These constructions, especially the first in (30b), are very similar to the so-called designation (31), a type of apposition, in English (Quirk et al., 1985: p. 1310).

(30) a. *a rózsá-k, a piros-ak*
the rose-PL the red-PL
'the roses, the red ones'

[1] The idea came from one of my opponents, Bálint Sass. I am grateful for the suggestion.

DOI: 10.15774/PPKE.BTK.2021.17

b. *a barát-om, Péter*
the friend-PossSg1 Péter
'my friend, Péter'

c. *az iskolá-ban,* ***a portán***
the school-Ine, the
'at the school, at the gate'

(31) *Anna,* ***my best friend****,* was here last night.

Although named entities like *Angela Merkel német kancellár* 'Angela Merkel German chancellor' look very similar to appositive constructions (30b), the difference in their syntactic behaviour is rather salient: while in the case of apposition there is an agreement between the first nominal element, the anchor, and the apposition (32a), in extended named entities only the last noun bears a case suffix (32b, 32c).

(32) a. *Meg.hív-t-am* ***Angela Merkel-t, a német kancellár-t***
Invite-Past-Sg1 Angela Merkel-Acc, the German chancellor-Acc
'I invited Angela Merkel, the German chancellor.'

b. *Meg.hív-t-am* ***Angela Merkel német kancellár-t***
Invite-Past-Sg1 Angela Merkel German chancellor-Acc
'I invited Angela Merkel German chancellor'

c. **Meg.hív-t-am* ***Angela Merkel-t német kancellár-t***
Invite-Past-Sg1 Angela Merkel-Acc German chancellor-Acc
'I invited Angela Merkel German chancellor'

Apart from orthography, the lack of agreement confirms that the named entities in question are not a specific type of appositive construction.

3.2.2 What this structure may be

These kinds of noun phrases are mentioned in one instance in the literature discussing extraposed modifiers: (Károly, 1958: 29) enlists the examples in (33) as the Hungarian

DOI: 10.15774/PPKE.BTK.2021.17

equivalents of *apposition*, a term used in foreign linguistic papers (Károly, 1958: 23). He also argues that they are not extraposed modifiers, but rather nouns with an attributive modifier.

(33) a. *Bodri kutyá-m*
Bodri dog-POSS.SG1
'my dog, Bodri'

b. *a híres Kelemen kovács*
the famous Kelemen blacksmith
'the famous blacksmith Kelemen'

3.2.3 NER as a task in NLP

As the structure under examination is undoubtedly a named entity, I sum up what Named Entity Recognition (NER) is about – primarily relying on the PhD dissertation of Eszter Simon (2013) and some papers in the field of NLP.

NER is the task of automatic identification of selected types of Named Entities (NEs). It plays a crucial role in many NLP tasks, such as machine translation and information extraction. The task itself consists of two substeps: first, locating the named entities in texts, and second, classifying them into pre-defined categories (Simon, 2017).

In early works on the NER problem – for example MUC6, the 6th Message Understanding conference (Sundheim, 1995); or the 2002 and 2003 shared tasks of CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), etc. – names of persons, locations and organisations have been studied the most (while temporal expressions and some numerical expressions have also been included in the discussion). Annotation guidelines avoid providing an exact definition of NEs and simply state that they are "unique identifiers", and rather list examples and counterexamples.

Simon (2013) dedicates a section to the linguistic approach on proper names, more precisely, on the concept of 'unique reference' and how this and other properties contribute to a clear distinction between proper nouns and common nouns (or proper names and common noun phrases). Apart from the unique reference, indivisibility (34, 35) and inflection pattern (36a, 36b) are the two most important pieces of evidence when distin-

DOI: 10.15774/PPKE.BTK.2021.17

guishing a proper name and a common noun phrase (examples taken from Simon, 2013: p. 22).

(34) a. *beautiful King's College*

b. **King's beautiful College*

(35) a. *my son's college*

b. *my son's beautiful college*

(36) a. *Láttam az Egerek és embereket.*
'I saw (Of Mice and Men).Acc'

b. *Láttam az egereket és az embereket.*
'I saw the mice.Acc and the men.Acc'

Lastly, Simon (2013: 23–24.) illustrates the non-compositionality of proper names after dividing them into two: there are phrases which are headed by a common noun and modified by a proper noun (*Roosevelt square*), and phrases consisting of two (or more) proper nouns (*Theodore Roosevelt*). At this point it becomes clear that the phrases I want to investigate here do not belong to this kind of polymorphemic proper names (or at least, not all of them do): *Máris szomszéd* 'neighbour Máris' or *Angela Merkel német kancellár* 'Angela Merkel German chancellor' seems to be the member of the first group of Simon (2013). However, as the author highlights, in the case of *Roosevelt square*, all non-defeasible semantic implications come from the head (*square)*, and so the modifier has no contribution here. Thus if we omit the head (37a), we cannot determine where exactly we should meet: at the pub, at the library, or at the statue? A Hungarian example is shown in (37b), an example taken from Simon (2008): here we do not know the source of the call; it can be the *Bolyai János Gimnázium* (a high school named after Bolyai János), or *a Bolyai utcai presszó* (a pub located in Bolyai street).

(37) a. *Meet me at the Roosevelt!*

b. *A Bolyai-ból hív-nak.*
the Bolyai-Ela call-Pl3

'You have a call from the Bolyai'

DOI: 10.15774/PPKE.BTK.2021.17

If we test *Máris szomszéd* or *Angela Merkel német kancellár*, the conclusion is not the same: if we omit the presumed head, we still know who we invited (38a), or who will step down from her position (38b).

(38) a. *Meghív-tam Máris-t a buli-ba.*
invite-PastSg1 Máris-Acc the party-Ill

'I invited Máris to the party.'

b. *Angela Merkel lemond a hivatal-á-ról.*
Angela Merkel resign the office-PossSg3-Del

'Angela Merkel will step down from her position.'

These examples might indicate that the relation between the two parts of XNEs is closer to some kind of coordination than to a simple head-modifier relation.

Moreover, *Máris szomszéd* 'Máris neighbour' and *Angela Merkel német kancellár* 'Angela Merkel German chancellor' themselves are not completely identical either: in the case of the latter, the right part of the construction, *German chancellor* has a unique reference itself, while *neighbour* does not (*Máris neighbour* does).

The second subtask of NER is to categorize NEs into pre-defined categories. HunNER (Simon et al., 2006) is a gold standard named entity annotated corpus of Hungarian. The annotation scheme of the project defines the following types to use:

- person names
- phrases denoting the role of people at an organization, such as *elnök* 'President'
- words of titles, ranks, such as 'Sir'
- organization names
- location names
- brand names
- titles of artworks
- miscellaneous

DOI: 10.15774/PPKE.BTK.2021.17

These categories will come to light later when I try to categorize the results of my corpus queries.

3.3 Method and results

The first step of this study was to gather more examples: the aim is to be equipped with a list of noun phrases with an inner structure similar to that of the multi-word proper names in (28), to draw conclusions from the list and, if possible, to identify the groups of words, which may appear as the first or the last member of polymorphemic proper names. By first member I refer to the proper name – either consisting of one token or more – at the beginning of the noun phrase, and by last member I refer to the common noun ending following the proper nouns, ending the noun phrase.

The corpus used for the study was the syntactically parsed Szeged Treebank 2.0 (Csendes et al., 2005) as I needed manually annotated, gold standard noun phrases.

First, I extracted pairs of words[2] from the corpus that follow each other within the same noun phrase (and the first one is caseless). The list of 9483 word pairs thus extracted had to be cleaned manually. Finally, a collection of a total of 1252 extended named entities was created[3].

The files and the codes can be found at `https://github.com/ppke-nlpg/angela_merkel_and_uncle_jack`.

1252 results mean 974 types. If we observe the lemmas of the names (the tokens *St. Antonio herceg-nek* 'St. Antonio prince-DAT' and *St. Antonio herceg* 'St. Antonio prince' are one type) then we have 902 types. The complete list of the types can be found in the Appendix B.1. The most frequent ones are presented in Table 3.1.

However, the collection of the endings is far more exciting than the list of the whole noun phrases. There are 582 types of these common nouns when inspecting inflected forms, and 455 types of lemmas. The most frequent ones can be seen in Table 3.2, and the whole list is attached in the Appendix B.2.

[2] I started by extracting pairs of words. I soon realized that XNEs with a multiword proper name in them would be missed this way. Therefore, I expended my query to multiple words where all but the last word start with a capital letter.

[3] Examples of strings deleted from the original query list: *ez utóbbi* 'this last.one', *forróság szárazság* 'heat [and] drought', *más forrás-ból* 'other source-DEL' etc.

DOI: 10.15774/PPKE.BTK.2021.17

Table 3.1. Lemmas of the most frequent XNEs in Szeged Treebank 2.0. Column "Type" shows the string, while column "#" represents the number of its occurrences.

Type	**#**
Parsons asszony	13
Varga Mihály pénzügyminiszter	13
Orbán Viktor miniszterelnök	11
Ogilvy elvtárs	10
St. Antonio herceg	10
Fernandez régensherceg	6
Laci atya	6
Pista bácsi	6
St. Antonio főherceg	6
Wirth kapitány	6
Croesus csoport	5
Palmerston tanár	5

Table 3.2. The most frequent endings of XNEs in Szeged Treebank 2.0. Column "Type" shows the word, while column "#" represents the number of its occurrences. NÉVEN is a macro of the words *név, névvel, néven, nevű, cím, című, címmel, címen* meaning 'entitled'. More on this category in section 3.4 below.

Type	**#**
NÉVEN	45
vezérigazgató	32
bácsi	30
cég	29
csoport	27
pénzügyminiszter	27
néni	26
elnök	25
miniszterelnök	22
hírügynökség	21
úr	19
asszony	15
professzor	15

DOI: 10.15774/PPKE.BTK.2021.17

The above results are, nonetheless, only phrases with a proper name and only one common noun at the end. But there are additionally many cases where a modifier is inserted between the proper name and the common noun, for example *Angela Merkel* ***német*** *kancellár* 'Angela Merkel **German** chancellor'. Thus I retrieved these structures from the corpus. The whole list can be found in Appendix B.3, a sample only containing the common noun at the end and its modifier is presented here in Table 3.3.

Table 3.3. A sample of the list containing extended named entities with more than one common noun member, generally a modified common noun, extracted from Szeged Treebank 2.0. The whole list can be found in Appendix B.3.

token	meaning
tanár úr	'Mr. teacher'
operációs rendszer	'operating system'
dominikánus szerzetes	'Dominican monk'
mezőgazdasági miniszter	'Minister of Agriculture'
francia gazdasági és pénzügyminiszter	'French Minister for Economics and Finance'
lengyel elnök	'Polish president'
egykori miniszterelnök	'former prime minister'
világhírű angol matematikus-fizikus	'world famous English mathematician-physicist'
szerb pszichiáter	'Serbian psychiatrist'
18. századi svájci matematikus	'18th century Swiss mathematician'

3.4 Discussion

I manually sorted the endings listed in Appendix B.2 – the most frequent ones of which are presented in Table 3.2 – into the following six categories based on their meaning and function:

1. category of words of the type NÉVEN 'entitled': *név* 'name' (as in *Croesus név* 'Croesus name'), *nevű* 'named', *néven* 'named', *névvel* 'named', *cím* 'title', *című* 'titled', *címmel* 'titled', *címen* 'titled'. *Betű* 'letter' and *szó* 'word' (as in *Komintern szó* 'Komintern word') are also members of this group. They all mark the function or role of the preceeding proper name: in *Living Through Animated Milleniums címmel* 'entitled Living Through Animated Milleniums' the *Living Through Animated Milleniums* is a title, the title of something that is described later.

DOI: 10.15774/PPKE.BTK.2021.17

2. geographical common nouns: *megye* 'county', *utca* 'street', *falu* 'village', *város* 'town', *folyó* 'river' etc.

3. courtesy formulas, how people are addressed, titles, such as *bácsi* 'uncle', *néni* 'aunt', *elvtárs* 'comrade'

4. professions, occupations, ranks, such as *pénzügyminiszter* 'minister of finance', *elnök* 'president', *százados* 'captain', *onkológus* 'oncologist'

5. name of institutions: *bank* 'bank', *olajtársaság* 'oil company', *kutatóintézmény* 'research institution', *pénzintézet* 'financial institution'

6. the name of a product in a brand name – product name combination, such as *noteszgép* 'notebook' in *Lenovo noteszgép* 'Lenovo notebook', or *benzinkút* 'petrol station', *gépkocsi* 'motor car' etc. This category retains a core difference from the above five: the number of its elements is not finite; we can create a brand name - type name for any subject.

One could argue that the link between the proper nouns and the common noun in these named entities is *nevű* 'called' (see example (39)).

(39) a. *Angela Merkel (nevű) német kancellár*
Angela Merkel (called) German chancellor

'German chancellor called Angela Merkel'

b. *OTP (nevű) bank*
OTP (called) bank

'bank called OTP'

It is clear that *nevű* 'called' accurately expresses the relationship here: there is a German chancellor who is called Angela Merkel; and we are talking about a bank called OTP. Therefore XNEs are nothing more than more concise versions of examples (39a) and (39b). For group 6, this liaison word could be *márkájú* 'of the brand'.

If the above idea is correct, *nevű* can be inserted into these phrases (examples (40a) and (40b)).

DOI: 10.15774/PPKE.BTK.2021.17

(40) a. *Tegnap (az) Angela Merkel nevű német kancellár lemond-ott.*
yesterday (the) Angela Merkel called German chancellor resign-PastSg3

'Yesterday German chancellor called Angela Merkel resigned.'

b. *Meglátogat-ott a Máris nevű szomszéd.*
visit-PastSg3 the Máris called neighbour

'I was visited by the neighbour called Máris.'

There seem to be a difference in the two examples: when we know exactly who the owner of the given title is (and only one owner exists), like in the case of a German chancellor, or a Hungarian prime minister, inserting *nevű* is unnecessary and makes the sentence strange, at least.[4] When there are more possible candidates the common noun may refer to (in the case of a neighbour, for example), inserting *nevű* is justifiable, and makes the sentence even more understandable.

The precise inner structure of the XNEs, and the status of the supposed linking word, *nevű*, requires further analysis. Their study is rather theoretical than corpus-based and is beyond the scope of my thesis.

Now let us turn to the list of XNEs that contain some kind of modifier preceding their final element, the common noun. The first thing that becomes clear is that members of the first group (NÉVEN) and of the second (geographical names) are not present in this list: they form a more closed unit where nothing can be inserted in between the proper name and the common noun.

In the case of the other four groups, on the other hand, for one reason or another, one or more modifiers may appear before the final token. The following cases can be distinguished:

1. In many cases it seems that the second part of the name, the common noun, consists of more than one word. Some examples of this kind of complex endings can be found in Table 3.4.

[4]The group of entites the speaker knows well may vary depending on the given culture, economical area etc. Europeans may refer to the president of Bangladesh as *Bangladeshi president called Abdul Hamid*, while Bangladeshi speakers may refer to the German chancellor as *called Angela Merkel*.

DOI: 10.15774/PPKE.BTK.2021.17

2. The modifier of the ending of the XNE may further specify the meaning of that common noun: this is the case with the different types of ministers[5]: *honvédelmi miniszter* 'Minister of Defense', for example. Other examples can be seen in Table 3.5

3. The most populous group includes endings where we define the place of operation (Table 3.6).

4. The fourth group is similar to the third one, but the modifier specifies the origin (or nationality) of the given person, not the place of its operation (*angol újságíró* 'English journalist' vs. *cseh miniszterelnök* 'Czech prime minister'). Examples are in Table 3.7.

5. A modifier can be used to state something about the time of the operation in question: *volt miniszter* 'former minister'. However, this is not a specific duration we give here, nor a precise date, but a relative chronological order. More examples can be found in Table 3.8.

6. The modifier specifies the exact time when the person in question worked, or was born. There is only one example for this in the corpus: *18. századi svácji matematikus* '18th century Swiss mathematician'

7. Finally, we can refer to some other attribute of the given person, such as *feminista írónő* 'feminist authoress'. One can be *világhírű* 'world famous' or *tiszteletbeli* 'honorary'.

Table 3.4. Some examples for complex endings of XNEs

token	**meaning**
operációs rendszer	'operating system'
igazgató főorvos	'medical director'
ügyvezető elnök	'executive chairman'

[5]In case of *pénzügyminiszter* 'minister of finance' or *külügyminiszter* 'minister of foreign affairs' it is only an orthological difference that the modifier narrowing down the meaning of *miniszter* is written in one word with the head. The first part of the compound is the same modifier as *kereskedelmi* in *kereskedelmi miniszter* 'minister of commerce'.

DOI: 10.15774/PPKE.BTK.2021.17

Table 3.5. Some examples for modifiers specifying the ending of XNEs

token	meaning
népjóléti irodavezető	'public welfare office manager'
anyagtervezési főosztályvezető	'head of material design department'
dominikánus szerzetes	'Dominican monk'

Table 3.6. Some examples for modifiers specifying the ending of XNEs

token	meaning
iraki elnök	'President of Iraq'
budapesti ügyvéd	'lawyer in Budapest'
jugoszláv jegybankelnök	'Governor of Yugoslav Central Bank'

Table 3.7. Some examples for modifiers defining the origin of the person in question

token	meaning
spanyol pincér	'Spanish waiter'
osztrák író-rendező	'Austrian writer-director'
szerb pszichiáter	'Serbian psychiatrist'

Table 3.8. Some examples for modifiers defining the time of the operation of the person in question

token	meaning
uralkodó főherceg	'reigning archduke'
egykori miniszterelnök	'former prime minister'
leendő igazgató	'future director'

There are many examples when two (or more) modifiers from the above categories are combined: *világhírű angol matematikus-fizikus* 'worlds famous English mathematician-physicist', *korábbi kereskedelmi igazgató* 'former director of sales'.

3.5 Algorithmic processing of XNEs

Now that we know about this structure and its instances, we can turn to the question of how this phenomenon should be handled in a rule-based parsing process, as in `AnaGramma`.

DOI: 10.15774/PPKE.BTK.2021.17

One could argue that collecting and tagging (e.g. as XNE as a tag)[6] the common nouns in these XNEs will solve the problem: therefore, a new rule can be formulated as (41):

(41) If the algorithm is processing a suffixless proper noun, and a common noun of the category XNE is the first token in the window, then the proper noun in question receives a NONE tag.

The word *waiter* 'pincér' is undoubtedly in the above category (see Table 3.7 for an example). *Angela Merkel*, on the other hand, is undoubtedly a proper name. Let us see the example in 42.

(42) *Angela Merkel pincér helyett a szakács-ot kér-i.*
Angela Merkel waiter instead.of the cook-ACC ask-SG3

'Angela Merkel asks for the cook instead of the waiter.' or
'He/she asks for the cook instead of waiter Angela Merkel.'

If we apply a rule like (41), *Merkel* will be annotated as a modifier of the *pincér* 'waiter', which results in a grammatical sentence (with a non-overt subject), and the second meaning shown in the example. However, we all know (or at least assume) that Angela Merkel is not a waiter here, but the German chancellor, and the grammatical subject of this sentence, and so the first meaning is the correct one. All of these can not be deducted from any morphological information, but rather only from our world knowledge.

Let us assume that there are no sentences like (42); when an XNE-ending follows a proper name then the two always form an XNE. We are still not closer to the solution to the parsing of these instances: apart from the category NÉVEN and from geographical names, every category contains a set of words that can be dynamically extended any time. Providing a list of them is futile.

Based on the above, XNEs should not be processed by a left-to-right, rule-based approach. Their characteristics, on one hand, supports them being taken over from a lexicon containing our world knowledge; on the other hand, a wider context is required to accurately define their inner structure and their boundaries.

[6] I made an attempt to create lists of words of the above categories with the help of a clustering tool built on word embeddings (Siklósi and Novák, 2016b). The lists I created are available at `https://github.com/ppke-nlpg/angela_merkel_and_uncle_jack/tree/master/w2v_try`.

DOI: 10.15774/PPKE.BTK.2021.17

3.6 Summary

In this chapter I highlighted a phenomenon that has not been analysed before comprehensively. It looks similar to extraposed modifiers but it is something different: there are noun phrases consisting of a proper name and one or more common nouns (such as *Angela Merkel kancellár* 'Angela Merkel chancellor'). I collected similar structures from Szeged Treebank 2.0 to be able to define some categories among these phrases. I outlined six main categories: the category of words under the macro NÉVEN 'entitled', geographical names, courtesy formulas, ranks and occupations, names of institutions and the names of products of a given brand. I showed that nothing can be inserted between the two parts of the XNEs of the first two categories. Furthermore, I inspected what kind of words may fit in between the two parts of the other four groups: *Angela Merkel* ***német*** *kancellár* 'Angela Merkel **German** chancellor'. Most frequently we specify the place or time of the operation of the given person with a modifier.

It is undeniable that *Angela Merkel német kancellár* and its companions require further analysis and more attention from linguists and computational linguists alike.

DOI: 10.15774/PPKE.BTK.2021.17

Chapter 4

Locative case suffixes

"All's well that ends well"

A play by William Shakespeare

4.1 Introduction

"Where will I find you? In pants."

One thing is certain: upon seeing a case suffix or a postposition one can be sure that an NP has just ended. But in the following two chapters, especially Chapter 5, I will show how uncertain the above assumption really is. In the subsequent pages I focus on these NP-ending candidates: case suffixes and postpositions (5).

The urge to study case suffixes, more precisely, the role of locative case markers in defining the function of their dependent nominal was triggered by the idea of a question-answering (QA) system for Hungarian. Novák et al. in (2019a) and (2019b) introduce a novel questioning system for Hungarian by presenting the methods of creating proper training data for such a system. The aim of the system is to comprehend texts and to model this comprehension by formulating relevant questions about the sentences. To achieve this, a precisely annotated corpus is inescapable where the annotation contains all the detailed distinctive features necessary for asking appropriate questions.

The research presented in this chapter is part of the initiative process of the project creating the questioning system. The annotation in Hungarian corpora currently available operates with abstract and rather general categories that provide insufficient information for the system. A quite trivial example for this is the case of nouns: no distinction is made between +/- animate nouns, although it would be crucial when formulating a suitable

DOI: 10.15774/PPKE.BTK.2021.17

question about the subject of the sentence: *My friend is coming.* triggers the question *Who is coming?*, but *Winter is coming.* triggers *What is coming?*. The same distinction is needed for other categories of nouns, and for the argument structure of verbs as well: if *My friend* is an agent (*My friend sent an e-mail.*), the question *What did my friend do?* is appropriate, but if it is a patient (*I hit my friend.*), only *What happened to my friend?* will work (but *What happened to me?* will not). A part of the results presented here, and the results of other studies on semantic role labelling are presented in Novák et al. (2019a).

For as I have long been interested in NPs and their delimitation, I have focused on reviewing the annotation of nouns bearing a case suffix, and functioning as adverbial adjuncts (see Ligeti-Nagy and Novák, 2019). The little joke in the title of this section attempts to highlight the diversity in the set of meanings that *bAn* may represent, traditionally called a locative case suffix. (Throughout the next sections I will follow the annotation used by Sass, 2009: when referring to harmonizing case suffixes I use a template, where the vowel's place is filled in with the capital form of the back vowel used there. *bÓl* is the template used for *ból* and *ből*, the elative case suffix.) Staying with the inessive case suffix (*bAn*), examples in (43) illustrate this above mentioned diversity (the list could be further extended). The different functions of the words with *bAn* are demonstrated by the questions they may answer.

(43) a. *Párizs-ban* 'in Paris': *Hol?* 'Where?'

b. *január-ban* 'in January': *Mikor?* 'When?'

c. *szoknyá-ban* 'in skirt': *Milyen ruhában?* 'In what dress?'

d. *írás-ban* 'in writing': *Milyen formában?* 'In what form?'

e. *hármas-ban* 'threesome': *Hányan?* 'How many people?'

It is clear that the case suffix in itself provides insufficient information for questioning or for finding answers to a question. Relying only on the lemma and the morphological information of the final token (*Case:Ine*) it is implausible to identify whether the NP in italics in examples 44a and 44b is an appropriate answer to question *Hol?* 'Where?' - in other words, is *Hol?* 'Where?' a relevant question to this NP?

DOI: 10.15774/PPKE.BTK.2021.17

(44) a. A kórház toxikológiai osztálya több szempontból is érintett lehet *a különleges nappal kapcsolat**ban***(Case:`Ine`).
'The Toxicology Department of the hospital may be affected in several ways ***in*** *connection with the special day.*'

b. Országszerte tömeg várható *a detoxikálók**ban***(Case:`Ine`).
'Nation-wide crowd is expected ***in*** *toxicologies.*'

Thus it is unavoidable to encode such knowledge into the corpus which makes it feasible for a questioning system to learn the difference between *Ubul* öltönyben *ment dolgozni.* 'Ubul went to work *in suits.*' and *Ubul* decemberben *ment dolgozni.* 'Ubul went to work *in December*', or to learn that the token *főiskolá-**n*** (college-`Sup`) 'in college' can be the locative attribute of a verb, but *főiskolá-**ban*** (college-`Ine`) cannot; although both tokens are the combination of the same lemma and a locative case suffix, superlative *On* and inessive *bAn.*

4.2 Literature review

In all the literature dealing with the case system of Hungarian, it has been clearly stated that there is no clear correspondence between case suffixes and sentence roles (sentence functions) (see, for example, Antal, 1961; Kiefer, 2000a). However, the detailed classification of lemmas and the in-depth examination of the lemma and the case suffix were not addressed in either of them.

An initial definition of case endings might be found in Kiefer (2000b):

(45) A case ending is a suffix that can be freely combined with a pronoun, a proper name, an adjective, a numeral, other suffixes, and its ability to be combined is irrespective of the meaning of the noun.

However, as Kiefer in (1987) and (2000b) discusses, the features in (45) are not absolute features of Hungarian case endings. A more accurate and final definition of Hungarian case endings might be the one in (46):

DOI: 10.15774/PPKE.BTK.2021.17

(46) A suffix is a case marker if and only if a nominal bearing this suffix functions as a selected argument of some verb, and the verb requires its argument to bear precisely this suffix.[1]

Based on these criteria, 18 case endings can be identified in Hungarian (see Table 4.1).

Table 4.1. List of Hungarian case endings with an example.

case	**suffix**	**example**
nominativus	-	*ember* 'human'
accusativus	*-t*	*ember-t*
dativus	*-nAk*	*ember-nek*
instrumentalis	*-vAl*	*ember-rel*
causalis-finalis	*-ért*	*ember-ért*
translativus-factivus	*-vÁ*	*ember-ré*
inessivus	*-bAn*	*ember-ben*
superessivus	*-n*	*ember-en*
adessivus	*-nÁl*	*ember-nél*
sublativus	*-rA*	*ember-re*
delativus	*-rÓl*	*ember-ről*
illativus	*-bA*	*ember-be*
elativus	*-bÓl*	*ember-ből*
allativus	*-hOz*	*ember-hez*
ablativus	*-tÓl*	*ember-től*
terminativus	*-ig*	*ember-ig*
formativus	*-ként*	*ember-ként*
superessivus	*-Ul*	*ember-ül*

Among these, we distinguish syntactic (or structural) cases (nominativus, accusativus and dativus) and lexical cases (the others); the latter group cannot be inferred from the syntactic structure and must be defined in the lexical entry of the verb itself. Lexical cases can be further categorized into classes based on the adverbial role they have: we have suffixed nouns marking a place (inessivus, superessivus, adessivus) or directions (sublativus, delativus, illativus, elativus, allativus, ablativus and terminativus). Instrumentalis marks a tool, formativus and superessivus encodes an adverbial of state, causalis-finalis marks a goal, while translativus-factivus is an adverbial of result. Here we encountered cases as abstract case relationship, as a syntactic-semantic category, the syntactic relationship between the verb and its argument; while we have the suffixed nouns and the case endings in (4.1) as a morphological category. It is clear that a case (as a relationship) can

[1]The Hungarian definition of this criterion is "*Valamely toldalék akkor és csakis akkor esetrag, ha a vele toldalékolt főnév leköthetí az igének valamely, alakja szempontjából is meghatározott vonzatát*" (Kiefer, 2000b). The English translation comes from (É. Kiss and Hegedűs, 2021: 22.)

DOI: 10.15774/PPKE.BTK.2021.17

be expressed by several different case endings (and with other NPs like the combination of a noun and a postposition as well), and a case ending might be used to express more than one case. This is the behaviour I want to examine with a specific focus on locative (adverbials of place and direction) cases: what (else) can a noun suffixed with one of the locative case suffixes express?

4.3 Method

The corpus I worked with is the Hungarian subcorpus of the Universal Dependencies corpus (Nivre et al., 2016). This Hungarian UD treebank[2] is derived from the Szeged Dependency Treebank (Vincze et al., 2010). One of the newspaper sections of the latter (*Népszava*) has been semi-automatically converted to the Universal Dependencies scheme followed by a manual correction carried out by linguists. The current study requires a treebank because I only want to extract adjuncts; the syntactic annotation is necessary to filter the words I need.

In Hungarian, locative cases are often handled as three paradigms:

- paradigm of internal locatives: inessive *bAn* (realized as *ban* or *ben*), illative *bA* (realized as *ba* or *be*), and elative *bÓl* (realized as *ból* or *ből*). They all mark an internal place: *a ház**ban*** 'in the house', *a ház**ból*** 'from (inside) the house', *a ház**ba*** 'into the house'.

- paradigm of external locatives, concentrated on a given point: adessive *nÁl* (realized as *nál* or *nél*), allative *hOz* (realized as *hoz*, *hez* or *höz*), ablative *tÓl* (realized as *tól* or *től*). They mark an external place which is connected to a given point, or a person: *a szomszédnál* 'at the neighbour's', *a szomszédhoz* 'to the neighbour's', *a szomszédtól* 'from the neighbour's'.

- paradigm of external, surface-oriented locatives: superessive *On* (realized as *on*, *en* or *ön*), sublative *rA* (realized as *ra* or *re*) and delative *rÓl* (realized as *ról* or *ről*). They are considered to mark a closer connection than the previous ones: *az asztalon* 'on the table', *az asztalról* 'from the (surface of the) table', *az asztalra* 'on to the table'.

[2] https://github.com/UniversalDependencies/UD_Hungarian-Szeged

DOI: 10.15774/PPKE.BTK.2021.17

For this study, every token that was annotated with an `OBL` edge depending on the verb was extracted[3] if it bore one of the above described locative suffixes.[4] A sample of this extraction can be seen in Table 4.2.

Next, I took these nine groups of words and "pre-categorized" each group using the word embedding model of Siklósi and Novák (2016a) which is based on the word2vec models (Mikolov et al., 2013) and uses hierarchical clustering. The input of the clustering is the semantic vector of the words in the groups. The details of the clustering method are described in Siklósi and Novák (2016a) and (2016b). I browsed the results with the help of an online visualisation interface (Novák et al., 2017).

The output of the clustering was a list of groups of words, each group consisting of 3-8, semantically closely related words (a segment of the list of the clustered words with a locative suffix can be seen in Table 4.3). Then I manually cleaned these lists to collect words that could answer the same wh-word – meaning that they occur as the same type of adverbial in the sentence – together in one group. It has to be noted that this phase of the study and of the categorisation did not focus on the stems themselves, but rather on the inflected forms.

[3]The extraction of the tokens in question was carried out by Attila Novák.

[4]The data presented and discussed in this chapter are available at `https://github.com/ppke-nlpg/locatives`.

DOI: 10.15774/PPKE.BTK.2021.17

Table 4.2. A sample of the extraction of nouns bearing a locative case suffix from the Hungarian UD corpus. The first column (*word form*) shows the word forms themselves. The second column (*# total*) presents the total number of occurrences of the given word form with an OBL edge. The third column (*OBL / ALL*) is the proportion of the number of occurrences of the word with an OBL edge and the number of all occurrences of the word. The labels of edges of the given word form are listed in the fourth column (*dep. edges*). The examples are sorted by the frequency of their appearance as an adjunct.

word form	**# total**	OBL / ALL	**dep. edges**
pont-on 'punct-SUP'	610	0.9951	OBL;610;COORD;2;MODE;1
tőkepiac-ban 'capital.market-INE'	307	0.9967	OBL;307;ATT;1
amely-ben 'which–INE'	236	0.9957	OBL;236;LOCY;1
érték-ben 'value-INE'	236	0.9957	OBL;236;COORD;1
devizapiacá-n 'foreign.currency.market.of-SUP'	164	1.0000	OBL;164
tőzsdé-n 'stock.exchange-SUP'	152	0.9934	OBL;152;COORD;1
számá-ban 'number.of-INE'	144	0.9931	OBL;144;COORD;1
köré-ben 'circle.of-INE'	125	0.9842	OBL;125;ATT;1;COORD;1
szint-en 'level-SUP'	118	0.9915	OBL;118;APPEND;1
közlemény-ben 'announcement-INE'	110	1.0000	OBL;110
világ-on 'world-SUP'	109	1.0000	OBL;109
helyzet-ben 'situation-INE'	109	1.0000	OBL;109

DOI: 10.15774/PPKE.BTK.2021.17

Table 4.3. A small sample of the clusters defined by the clustering tool of the word2vec model of Siklósi and Novák (2016a). The words here are all word forms with the case ending *On* on them. The first column is the number of the given cluster, the second column contains the members of the given cluster. For the sake of readability, I do not give a full morphological translation of the Hungarian word forms, I only list the meaning of their lemmas; all of them bear a SUP case ending.

id	**cluster**
0	*százalékán némelyikén többségén autópályáján átkelőhelyén* 'percentage.of' 'some.of.it' 'majority.of.it' 'highway.of' 'crossing.place.of' *pontjain Északnyugat-Magyarországon részein részén partvidékén* 'points.of' 'Northwest-Hungary' 'parts.of' 'part.of' 'coastline.of'
1	*ülésén ülésen közgyűlésen közgyűlésén ülésein napirendjén* 'session.of' 'session' 'general.meeting' 'general.meeting.of' 'sessions.of' 'agenda.of'
2	*tárgyaláson tárgyalásokon megbeszéléseken megbeszélésen* 'conference' 'conferences' 'meetings' 'meeting'
3	*látogatásán találkozójukon találkozóján találkozón* 'visit.of' 'their.meeting' 'meeting.of' 'meeting'
4	*tárlaton kiállításán kiállításon pályázaton árverésen* 'display' 'exhibition.of' 'exhibition' 'tender' 'auction'
5	*konferencián sajtótájékoztatóján tájékoztatón sajtótájékoztatón* 'conference' 'press.conference.of' 'briefing' 'press.conference' *demonstráción gálán esten ünnepségen rendezvényen* 'demonstration' 'gala' 'evening' 'ceremony' 'event'
6	*szerdán pénteken szombaton hétfőn csütörtökön kedden* 'Wednesday' 'Friday' 'Saturday' 'Monday' 'Thursday' 'Tuesday'
7	*Szekszárdon Kaposváron Szombathelyen Székesfehérváron Gödöllőn* 'Szekszárd' 'Kaposvár' 'Szombathely' 'Székesfehérvár' 'Gödöllő' *Szegeden Budapesten* 'Szeged' 'Budapest' (city names)
8	*Kerekegyházán Somogyszobon* 'Kerekegyháza' 'Somogyszob' (city names)
9	*lakóhelyükön helyszínen székhelyen megyeszékhelyen településen* their.residence' 'location' 'seat' 'county seat' 'settlement'
10	*lakótelepen környéken környékén külterületén* 'housing.estate' 'area.of' 'area' 'periphery.of'

DOI: 10.15774/PPKE.BTK.2021.17

The task here can be formulated as such: we are categorising adverbial adjuncts: we have locative adverbials, such as *sarkon* 'on the corner', or *bankban* 'in the bank'; but we have time adverbials as well: *télen* 'in winter', *decemberben* 'in December'. Moreover, there appear to be some period adverbials in the corpus: *hónapra* 'for [5] month' in the sentence *Öt hónapra béreltük a lakást.* 'We rented the apartment for 5 months' etc.

As a second step, I now focused on the lemmas. My goal was to somehow generalise this first categorisation: if a given lemma with a given case suffix functions as a given type of adverbial, is that the general role and feature of that lemma, or does that lemma appear as that type of adverbial only with that specific case suffix? If it seems to be a general role of the lemma, is there any case suffixes triggering some other adverbial role with that lemma? To illustrate this phase of the research I selected some words and present this step with them in Table 4.4 and Table 4.5.

In Table 4.4 I demonstrate the first step of the in-depth examination of lemmas and suffixes. Here I only aimed to find the possible roles (or possible wh-questions) of the lemma-suffix pairs that actually appear in the corpus. This resulted in many empty cells. Next, I tried to fill in the gaps and provide a wh-question for every possible combination of the lemmas and the suffixes.[5] In this way I could define a main category or role of the lemma and specify some exceptional combinations. Table 4.5 shows the results of this second step. In every row there is a lemma followed by a possible main (adverbial) category that it takes in the sentences with the locative case suffixes, in general. Every column represents a suffix. If the word form of the given lemma and a given suffix has a specific adverbial role in the sentence other than the one defined by the main category, then the given cell is not empty but filled with the possible secondary role of the word. Naturally, one word form may have more than one secondary role. Moreover, there are some cases where the combination of the given lemma and the given suffix does not fulfil the main adverbial role of that lemma, but rather only another one. *Időpont* 'time / appointment', for example, is basically a time adverbial with the locative case suffixes: *időpont-ban* 'time-INE' means 'at [that] time', and answers the question *When?*. However,

[5]This goal may seem contradictory to the principle of `AnaGramma` described in section 1.3: we do not want to be able to analyse theoretically existing yet almost never before seen phenomena. However, the UD treebank is a relatively small corpus and by far not representative. For a decent description and categorisation of these locative case suffixes, it is inevitable to piece together the data from the corpus with the results of some linguistic instrospection.

DOI: 10.15774/PPKE.BTK.2021.17

időpont-ról 'appointment-DEL' is rather a place adverbial, answering the question *From where?* (example (47)):

(47) *Siet-ek a fogorvos-hoz. Elkés-ek az időpont-om-ról.*
Hurry-1SG the dentist-ALL. Be.Late-SG1 the appointment-POSS1SG-DEL.
I am in a hurry to the dentist. I am late for my appointment.

Table 4.4. A sample of the table where all the lemmas appearing in the corpus with one or more locative case suffixes are gathered. The columns represent the case suffixes (here only 6 of the 9). One row is one lemma. A non-empty cell represents that the given lemma has an occurrence with the given suffix, and in that case it answers the given wh-word. For example, *címe* 'his/her address' appears with the suffix *On* in the corpus. *Cím-é-n* 'address-POSS3SG-SUP' may be an answer to the question *Where?*, but also to the question *How?*. The empty cells indicate that the given lemma does not appear in the corpus with the given suffix.

lemma	bAn	On	nÁl	bA	rA	hOz
árverés 'auction'		Where? When?			To what? Where?	
címe 'his/her address'		How? Where?				
esztendő 'year'	When?					

Table 4.5. A sample of the table where all the lemmas appearing in the corpus with one or more locative case suffixes are gathered and their adverbial roles are summarised (the whole table is attached in Appendix C.1 as a list.) If a secondary role overwrites the primary one, it is marked here in the table with an asterix.

lemma	main category	bAn	nÁl	On	bÓl	tÓl	rÓl
április 'April'	date						
alkalom 'occasion'	event				cause		
év 'year'	date			thing			thing
óra 'clock / watch'	event						
időpont 'time / appointment'	date						*loc / thing
kör 'circle'	mode	loc					

DOI: 10.15774/PPKE.BTK.2021.17

4.4 Results

4.4.1 Categorisation of nominals with a locative case suffix

Table 4.6 shows the 28 main – and together with the subcategories, a total of 50 – categories into which the lemmas were manually sorted (based on the output of the clustering). The main categories correspond to some semantic category; within this, subcategories usually indicate some suffix-preference within the given category. The categories are as follows:

1. *body*: names of different body parts; a group of them (*bAn-On*, for example *fej* 'head') marks a *body part adverbial* with *bAn* and *On*, and the other members of the paradigms of those, others (*any*, for example *derék* 'waist') may bear any of the suffixes (while being an adverbial of this type).

2. *build=inst*: common nouns denoting a physical building and an institution at the same time (for example *bank* 'bank'). The equation mark (=) in the labels marks the duality of the given words: while they are a *build*, they are also an *inst*.

3. *cause*: causatives, with the proper case suffix they answer to the question *Why?*, e.g. *ok* 'cause'

4. *circumst*: adverbials of circumstances; combined with the proper suffix they answer the question *In what circumstances?*, for example *hátrány* 'drawback'

5. *curr*: currencies, e.g. *forint* 'HUF'

6. *date*: adverbials of time; some of them prefers the case suffix *bAn* (and its paradigm), for example *perc-ben* 'in [that] minute', others prefer *On* (*hét-en* 'this week')

7. *dem*: demonstratives, and words with *-é* possessive suffix, answering the questions *In which?*, *To which?* etc., for example *előbbi* 'the previous one'

8. *direct*: adverbials of direction; among these a complementary distribution of the internal locative case suffixes (*oldalirány-ban* 'laterally') and the external, surface-oriented case suffixes (*délnyugat-on* 'southwest') can be observed as well

9. *event*: adverbials of event; with a proper case suffix they express both time and place: *Találkoztam a barátommal az* ***esküvő-n***. 'I met my friend at the wedding.': *When*

DOI: 10.15774/PPKE.BTK.2021.17

did you meet him? or *Where did you meet him?*. This group can be subcategorised based on the suffix preference of the lemmas: *háború-ban* 'in war' but *tüntetés-en* 'on a demonstration'.

10. *form*: adverbials of form; the answer to question *In what form?*. For example: *szó-ban* 'orally', *papír-on* 'on paper'.

11. *group*: adverbials of a group of people: *család* 'family'. With some suffixes they are locative adverbials answering the question *Where?*; but with other suffixes they are not: *család-on* 'on a family'.

12. *loc*: a multi-element group; words of various semantic categories appearing as locative adverbials with specific case suffixes. There is a strong suffix preference among them. The subcategories are presented in Table 4.7 in more detail.

13. *loc=who*: names of a predominantly geographic nature which, with certain case suffixes, are similar to the members of the category *org=who* by answering the questions *From who?*, *To whom?* etc. For example *EU-tagállam* 'member state of the EU': *Felkérés érkezett négy* ***EU-tagállam-tól***. 'Requests were received from four Member States.', where the *Who did you recieve the requests from?* is the relevant question.

14. *material*: adverbials of material; primarily bearing the case ending *bÓl, answering the question Of what material?*

15. *meas*: measuring units; they answer to the questions *From how much?* etc. generally together with their modifiers: *öt liter-ből* 'from five litres'.

16. *mode*: adverbials of manner

17. *num* and *num_size*: numerals; the slight difference between the groups becomes clear when adding *rA* or *rÓl* suffix to the lemmas: members of the category *num*, like *négy-re* 'to four', *3500-ra* 'to 3500' answer to the question *To how much?*, while members of the category *num_size*, like *negyedé-re*, 'to a quarter', *tizenötszörösé-re* 'fifteenfold' rather answer the question *How much?*

DOI: 10.15774/PPKE.BTK.2021.17

18. *org*: organizations, different types of companies, firms. Strong suffix preference: *Where – a Gazprom-nál* 'at Gazprom', but **a Gazprom-ban* 'in Gazprom'; while *a cég-ben* 'in the company' and *a cég-nél* 'at the company'

19. *org=who*: organisations and offices answering the question *Who?* when bearing certain suffixes. In these cases, the role of the verb is also significant: if *Jött egy levél a* ***bank-tól*** 'A letter came from the bank', then *Who did the letter come from?*; however, if *Most indultam a* ***bank-tól***. 'I have just started from the bank', *Where have you just started from?*. A suffix preference is also a characteristic of the group.

20. *part*: adverbials of part; although *Where?* and *From where?* etc. are relevant questions of these words, they also answer to the questions *In which part?*, *At which part?* etc. with any of the suffixes.

21. *period*: adverbials of time periods

22. *place*: names of object, places. When bearing the suffixes *nÁl*, *hOz* or *tÓl* their locative adverbial role prevails (in other words, they answer the questions *Where?*, *To where?*, *From where?*), but the suffixes of the paradigm of the internal locatives (*bAn* etc.) and the suffixes of the paradigm of the external, surface-oriented locatives (*On*) complement each other. *Árok-ban* 'in a ditch', *autópályá-n* 'on the highway'.

23. *posi*: names of offices, positions; with suffix preference. The case suffixes *nÁl*, *On* and *rA* do not trigger a locative adverbial role with any of the lemmas in this category.

24. *pov*: adverbials of a point of view – these words bearing the right suffix answer the question *From what point of view?*.

25. *state*: adverbials of state; only when bearing *bAn*, *bA* or *bÓl*.

26. *thing*: the category of words that do not have a specific adverbial role in the sentence but only answer the questions *In what?*, *At what?* etc., generally as the argument of a verb.

27. *way*: words referring to ways or routes. They have this role with every locative suffix except for the ones of the internal locative paradigm.

28. *who*: words answering the questions *In whom?*, *At whom?* etc.

DOI: 10.15774/PPKE.BTK.2021.17

Table 4.6. Main and subcategories of lemmas appearing in the corpus with locative case suffixes. The meaning of the lemmas of the given category bearing the given case suffix is represented by a wh-question the given combination may answer. The table should be read as follows: if the given lemma is the member of a given category (*build=inst*, for example), and it bears a *bA* suffix, then it has an adverbial role in the sentence answering the question *Hova?* 'Where to?'. An empty cell indicates that the given lemma with the given suffix exclusively appears as an argument of a verb thus answering the questions *In what?*, *On what?* etc. (*a* ***bank-on*** *múlik*; 'depends **on the bank**').

	Category		**Example**	**Suffix**								
	main	**sub**		**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
1	body	any	*derék* 'waist'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Hova?* 'To where?'	*Hova?* 'To where?'	*Hova?* 'To where?'	*Honnan?* 'From where?'	*Honnan?* 'From where?'	*Honnan?* 'From where?'
2	body	bAn-On	*fej* 'head'	*Hol?* 'Where?'		*Hol?* 'Where?'	*Hova?* 'To where?		*Hova?* 'To where?	*Honnan?* 'From where?'		*Honnan?* 'From where?'
3	build= inst	bAn	*bank* 'bank'	*Hol?* 'Where?'	*Hol?* 'Where?'		*Hova?* 'To where?'	*Hova?* 'To where?'		*Honnan?* 'From where?'	*Kitől?* 'From who?' *Honnan?* 'From where?'	

DOI: 10.15774/PPKE.BTK.2021.17

	Category		Example	Suffix								
	main	**sub**		**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
4	build= inst	On	*akadémia* 'academy'		*Hol?* 'Where?'	*Hol?* 'Where?'		*Hova?* 'To where?'	*Hova?* 'To where?'		*Kitől?/ Honnan?* 'From who? From where?'	*Honnan?* 'From where?'
5	cause			*Miért?* 'For what?'	*Miért?* 'For what?'	*Miért?* 'For what?'	*Miért?* 'For what?'	*Miért?* 'For what?'	*Miért?* 'For what?'	*Miért?* 'For what?'	*Miért?* 'For what?'	*Miért?* 'For what?'
6	cir-cumst		*hátrány* 'disadvan-tage'	*Milyen körül-mények között?* 'In what circum-stances?'	*Milyen körül-mények között?* 'In what circum-stances?'	*Milyen körül-mények között?* 'In what circum-stances?'	*Milyen helyzetbe?* 'Into what position?'	*Milyen körülmé-nyekbe?* 'To what circum-stances?'	*Milyen helyzetbe?* 'To what situa-tion?'	*Milyen helyzetből?* 'From what situ-ation?'	*Milyen helyzetből?* 'From what situ-ation?'	*Milyen helyzetből?* 'From what situ-ation?'
7	curr		*forint* 'HUF'		*Mennyi-nél?* 'How.much-Ade?'	*Hány _-on?* 'How.much _-Sup?'			*Mennyi-re?* 'How.much-Sub?	*Mennyi-ből?* 'How.much-Ela'	*Mennyi-től?* 'How.much-Abl'	*Mennyi-ről?* 'How.much-Del'
8	date	bAn	*hónap* 'month'	*Mikor?* 'When?'					*Mikor-ra?* 'For when?'		*Mikortól?* 'From when?'	

DOI: 10.15774/PPKE.BTK.2021.17

	Category		Example	Suffix								
	main	**sub**		**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
9	date	On	*kedd* 'Tuesday'			*Mikor?* 'When?'			*Mikorra?* 'For when?'		*Mikortól?* 'From when?'	
10	dem		*előbbi* 'former'	*Melyikben?* 'In which?'	*Melyiknél?* 'At which?'	*Melyiken?* 'On which?'	*Melyikbe?* 'Into which?'	*Melyikhez?* 'To which?'	*Melyikre?* 'To which?'	*Melyikből?* 'From which?'	*Melyiktől?* 'From which?'	*Melyikről?* 'From which?'
11	direct	bAn	*oldalirány* 'side direction'	*Merre?* 'In what direction?' 'Which way?'			*Merre?* 'To what direction?'			*Honnan?* 'From where?'		
12	direct	On	*délnyugat* 'Southwest'			*Hol?* / *Merre?* 'Where?' / 'In what direction?', 'Which way?'			*Merre?* 'To what direction?'			*Honnan?* 'From where?'

DOI: 10.15774/PPKE.BTK.2021.17

	Category		Example	Suffix								
	main	**sub**		**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
13	event	bAn	*háború*	*Mikor? / Hol?* 'When?' / 'Where?'	*Mikor? / Hol?* 'When?' / 'Where?'		*Hova?* 'To where?'	*Hova?* 'To where?'	*Mire? / Mikorra?* 'What. Sub' / 'For when?'	*Honnan?* 'From where?'	*Honnan? / Mikortól?* 'From where?'/'From when?'	
14	event	nÁl-On	*tüntetés* 'demonstration'		*Mikor? / Hol?* 'When?' / 'Where?'	textit-Mikor? / *Hol?* 'When?' / 'Where?'		*Hova?* 'To where?'	*Hova?* 'To where?' / *Mikorra?* 'For when?'		*Honnan?* 'From where?' / *Mikortól?* 'From when?'	*Honnan?* 'From where?'
15	event	On	*olimpia* 'olympics'			textit-Mikor? / *Hol?* 'When?' / 'Where?'			*Hova?* 'To where?' / *Mikorra?* 'For when?'	*Hova? / Mikorra?* 'To where?' / 'For when?'		
16	form	bAn	*szó* 'word'	*Milyen formában?* 'In what form?'			*Milyen formába?* 'Into what form?'			*Milyen formából?* 'From what form?'		

DOI: 10.15774/PPKE.BTK.2021.17

	Category		Example	Suffix								
	main	**sub**		**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
17	form	On	*papír* 'paper'			*Milyen formában?* 'What form. INE?'			*Hova?* 'To where?'			*Honnan?* 'From where?'
18	group		*társaság* 'companionship'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Kiken?* 'On who.PL?'	*Hova?* 'To where?'	*Hova?* 'To where?'	*Kikre?* 'For who.PL?'	*Honnan?* 'From where?'	*Honnan?* 'From where?'	*Kikről?* 'Who. PL3.DEL'
19	loc	any	*szekrény* 'cabinet'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Hova?* 'To where?'	*Hova?* 'To where?'	*Hova?* 'To where?'	*Honnan?* 'From where?'	*Honnan?* 'From where?'	*Honnan?* 'From where?'
20	loc	bAn	*szféra* 'sphere'	*Hol?* 'Where?'			*Hova?* 'To where?'			*Honnan?* 'From where?'		
21	loc	bAn-On	*dokumentum* 'document'	*Hol?* 'Where?'		*Hol?* 'Where?'	*Hova?* 'To where?'		*Hova?* 'To where?'	*Honnan?* 'From where?'		*Honnan?* 'From where?'
22	loc	nÁl	*pék* 'baker'		*Hol?* 'Where?'			*Hova?* 'To where?'			*Honnan?* 'From where?'	

DOI: 10.15774/PPKE.BTK.2021.17

	Category		Example	Suffix								
	main	sub		bAn	nÁl	On	bA	hOz	rA	bÓl	tÓl	rÓl
23	loc	On	*címoldal* 'front page'			*Hol?* 'Where?'			*Hova?* 'To where?'			*Honnan?* 'From where?'
24	loc	city-bAN	*Párizs* 'Paris'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Melyik városon?* 'Which city.SUP?'	*Hova?* 'To where?'	*Hova?* / *Kihez?* 'To where?' / 'To who?'	*Melyik városra?* 'Which city.SUB?'	*Honnan?* 'From where?'	*Honnan?* / *Kitől?* 'From where?' / 'From who?'	*Melyik városról?* 'Which city.DEL?'
25	loc	city-On	*Somogyszob* 'Somogyszob'	*Melyik városban?* 'Which city.INE?'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Melyik városba?* 'Which city.ILL?'	*Hova?* / *Kihez?* 'To where?' / 'To who?'	*Hova?* 'To where?'	*Melyik városból?* 'Which city.ELA?'	*Honnan?* / *Kitől?* 'From where?' / 'From who?'	*Honnan?* 'From where?'
26	loc	country	*Svájc* 'Switzerland'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Melyik országon?* 'Which country.SUP'	*Hova?* 'To where?'	*Hova?* / *Kihez?* 'To where?' / 'To who?'	*Hol?* 'Where?'/ *Melyik országra?* 'Which country.SUB'	*Honnan?* 'From where?'	*Honnan?* / *Kitől?* 'From where?' / 'From who?'	*Melyik országról?* 'Which country.DEL'

DOI: 10.15774/PPKE.BTK.2021.17

	Category		Example	Suffix								
	main	**sub**		**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
27	loc=who		*Hága* 'Hague'	*Hol?* 'Where?'	*Hol?* 'Where?' / *Kinél?* 'At who?'	*Kin?* 'Who.SUP'	*Hova?* 'To where?'	*Hova?* 'To where?' / *Kihez?* 'To who?'	*Kire?* 'Who.SUB'	*Honnan?* 'From where?'	*Kitől?* 'From who?' / *Honnan?* 'From where?'	*Kiről?* 'Who.DEL?'
28	material		*porcelán* 'porcelain'							*Milyen anyagból?* 'What material. ELA?'		
29	meas		*tonna* 'ton'	*Mennyiben?* 'How.much. INE?'	*Mennyinél?* 'How.much. ADE?'	*Mennyin?* 'How.much. SUP?'	*Mennyibe?* 'How.much. ILL?'	*Mennyihez?* 'How.much. ALL?'	*Mennyire?* 'How.much. SUB?'	*Mennyiből?* 'How.much. ELA?'	*Mennyitől?* 'How.much. ABL?	*Mennyiről?* 'How.much. DEL?
30	mode	bAn	*libasor* 'Indian file'			*Hogyan?* 'How?'			*Hogyan?* 'How?'	*Hogyan?* 'How?'		
31	mode	On	*hőfok* 'temperature'	*Hogyan?* 'How?'					*Hogyan?* 'How?'	*Hogyan?* 'How?'		

DOI: 10.15774/PPKE.BTK.2021.17

Category		Example	Suffix								
main	**sub**		**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
32 num		*sok* 'many'	*Men-nyiben?* 'How.much. INE?'	*Men-nyinél?* 'How.much. ADE?'	*Mennyin?* 'How.much. SUP?'	*Men-nyibe?* 'How.much. ILL?'	*Men-nyihez?* 'How.much. ALL?'	*Men-nyire?* 'How.much. SUB?'	*Men-nyiből?* 'How.much. ELA?'	*Men-nyitől?* 'How.much. ABL?'	*Men-nyiről?* 'How.much. DEL?'
33 num_size		*duplája* 'double [of sg]'	*Men-nyiben?* 'How.much. INE?'	*Men-nyinél?* 'How.much. ADE?'	*Mennyin?* 'How.much. SUP?'	*Men-nyibe?* 'How.much. ILL?'	*Men-nyihez?* 'How.much. ALL?'	*Mekko-rára?* 'How.big SUB'	*Men-nyiből?* 'How.much. ELA?'	*Men-nyitől?* 'How.much. ABL?'	*Mekko-ráról?* 'How.big DEL?'
34 org	bAn	*közhivatal* 'public office'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Min?* 'What.SUP?'	*Hova?* 'To where?'	*Hova?* 'To where?'		*Honnan?* 'From where?'	*Honnan?* 'From where?'	
35 org	nÁl	*leányvál-lalat* 'sub-sidiary'		*Hol?* 'Where?'			*Hova?* 'To where?'			*Honnan?* 'From where?'	
36 org=who	any	*kormány* 'govern-ment'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Hova?* 'To where?'	*Hova?* 'To where?'	*Hova?* 'To where?'	*Honnan?* 'From where?'	*Honnan?* 'From where?' / *Kitől?* 'From who?'	*Honnan?* 'From where?'

DOI: 10.15774/PPKE.BTK.2021.17

	Category		Example	Suffix								
	main	**sub**		**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
37	org=who	bAn	*párt* 'party'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Kin?* 'Who.SUP'	*Hova?* 'To where?'	*Hova?* 'To where?' / *Kihez?* 'To who?'	*Kire?* 'Who.SUB' / *Mire?* 'What.SUB?'	*Honnan?* 'From where?'	*Honnan?* 'From where?' / *Kitől?* 'From who?'	*Kiről?* 'Who.DEL?' / *Miről?* 'What.DEL'
38	org=who	nÁl	*adóhatóság* 'tax authority'	*Kiben?* 'Who.INE?'	*Hol?* 'Where?'	*Kin?* 'Who.SUP'	*Kibe?* 'Who.ILL'	*Hova?* 'To where?' / *Kihez?* 'To who?'	*Kire?* 'Who.SUB'	*Kiből?* 'Who.ELA?'	*Honnan?* 'From where?' / *Kitől?* 'Who.ABL?'	*Kiről?* 'Who.DEL?'
39	org=who	On	*alkotmánybíróság*	*Kiben?* 'Who.INE?'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Kibe?* 'Who.ILL'	*Hova?* 'To where?' / *Kihez?* 'To who?'	*Hova?* 'To where?'	*Kiből?* 'Who.ELA?'	*Honnan?* 'From where?' / *Kitől?* 'From who?'	*Honnan?* 'From where?'

DOI: 10.15774/PPKE.BTK.2021.17

	Category		Example	Suffix								
	main	sub		bAn	nÁl	On	bA	hOz	rA	bÓl	tÓl	rÓl
40	part		*széle* 'edge of'	*Hol?* 'Where?' / *Melyik részében?* 'In which part?'	*Hol?* 'Where?' / *Melyik részénél?* 'At which part?'	*Hol?* 'Where?' / *Melyik részén?* 'On which part?'	*Hova?* 'To where?' / *Melyik részébe?* 'Into which part?'	*Hova?* 'To where?' / *Melyik részéhez?* 'To which part?'	*Hova?* 'To where?' / *Melyik részére?* 'Onto which part?'	*Honnan?* 'From where?'/ *Melyik részéből?* 'From which part?'	*Honnan?* 'From where?'/ *Melyik részétől?* 'From which part?'	*Honnan?* 'From where?' / *Melyik részéről?* 'From which part?'
41	period		*félév* 'semester'	*Mikor?* 'When?'					*Mennyi időre?* 'How.much time.SUB?'		*Mikortól?* 'When.ABL'	
42	place	bAn	*intézmény* 'institution'	*Hol?* 'Where?'	*Hol?* 'Where?'		*Hova?* 'To where?'	*Hova?* 'To where?'		*Honnan?* 'From where?'	*Honnan?* 'From where?'	
43	place	On	*kép* 'picture'		*Hol?* 'Where?'	*Hol?* 'Where?'		*Hova?* 'To where?'	*Hova?* 'To where?'		*Honnan?* 'From where?'	*Honnan?* 'From where?'
44	posi	bAn	*munkakör* 'position'	*Hol?* 'Where?'		*Min?* 'What.SUP?'	*Hova?* 'To where?'		*Hova?* 'To where?' /*Mire?* 'What.SUB?'	*Honnan?* 'From where?'		

DOI: 10.15774/PPKE.BTK.2021.17

	Category		Example	Suffix								
	main	**sub**		**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
45	posi	On	*hatalom* 'power'			*Hol?* 'Where?'			*Hova?* 'To where?'			*Honnan?* 'From where?'
46	pov		*szempont* 'aspect'	*Milyen szempontból?* 'From what aspect?'						*Milyen szempontból?* 'From what aspect?'		
47	state		*egyensúly* 'balance'	*Milyen állapotban?* 'In what condition?'	*Mikor?* 'When?'		*Milyen állapotba?* 'Into what condition?'			*Milyen állapotból?* 'From what condition?'		
48	thing		*együttműködés* 'cooperation'									
49	way		*útszakasz*'road section'		*Hol?* 'Where?'	*Hol?* 'Where?'		*Hova?* 'To where?'	*Hova?* 'To where?'		*Honnan?* 'From where?'	*Honnan?* 'From where?'

DOI: 10.15774/PPKE.BTK.2021.17

Category		Example	Suffix								
main	**sub**		**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
50 who		*ellenfél* 'opponent'	*Kiben?* 'Who.INE'	*Kinél?* 'Who.ADE'	*Kin?* 'Who.SUP'	*Kibe?* 'Who.ILL'	*Kihez?* 'Who.ALL'	*Kire?* 'Who.SUB'	*Kiből?* 'Who.ELA'	*Kitől?* 'Who.ABL'	*Kiről?* 'Who.DEL'

DOI: 10.15774/PPKE.BTK.2021.17

4.4.2 The subcategories of the category *loc*

To illustrate subcategories within the above categorisation, Table 4.7 presents lemmas – generally used as adverbials of place – sorted into the category *loc*. The subcategory column (**sub**) indicates a case ending preference in most of the cases: *pék* 'baker' answers the question *Where?* only with the suffix *nÁl*, while *állam* 'state' answers it when bearing *bAn*. Examples *city-bAn*, *city-On* and *country* also separate group members based on some semantic information. In each of the cases in the corpus, country names have a locative adverbial role with the members of the internal locative paradigm and with the external locative suffixes that concentrate around a given point.[6] City names can be divided into two groups: the ones preferring the inessive paradigm (*Esztergom-ban* 'Esztergom-INE') and the ones preferring the superessive paradigm (*Szeged-en* 'Szeged-SUP'). The majority of the settlements inside of the historical borders of Hungary belong to the latter, while the former one includes every clearly foreign settlement. Members of both groups can perform a locative adverbial role with the suffixes of the other external paradigm (*nÁl* etc.). The questions *Melyik város-ban?* 'Which city-INE?', *Melyik város-on?* 'Which city-SUP' and *Melyik ország-on?* 'Which country-SUP?' in the table should not be interpreted as asking for an adverbial of place (for example *Melyik város-ban lak-sz?* 'Which city-INE live-SG2?': 'Which city do you live in?'); they refer to instances such as *Melyik város-ban bíz-ol? – Budapest-ben.* 'Which city-INE trust-SG2? – Budapest-INE': 'Which city do you trust in? In Budapest.'.

[6]Note that although they are not present in the corpus, there are some country names that require the case suffix *On* for being an adverbial of place: *Fülöp-szigetek-en* 'Philippines-SUP', *Izland-on* 'Iceland-SUP'. These are islands, however, not every island has the same behaviour: *Kubá-ban* 'Cuba-INE', *Írország-ban* 'Ireland-INE'.

DOI: 10.15774/PPKE.BTK.2021.17

Table 4.7. Subcategories of words having a locative adverbial role when bearing specific case suffixes. The subcategory column (**sub**) indicates a case ending preference in most of the cases. An example of the given category is presented in the third column, while the next columns show the appropriate questions the given category with the given case ending answers to.

category		**example**	**suffix**		
main	**sub**		**bAn**	**nÁl**	**On**
loc	any	*szekrény* 'cabinet'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Hol?* 'Where?'
loc	bAn	*állam* 'state'	*Hol?* 'Where?'	*Mi-nél?* 'What-ADE?'	*Mi-n?* 'What-SUP?'
loc	nÁl	*pék* 'baker'	*Mi-ben?* 'What-INE?'	*Hol?* 'Where?'	*Mi-n?* 'What-SUP?'
loc	On	*címoldal* 'front page'	*Mi-ben?* 'What-INE?'	*Mi-nél?* 'What-ADE?'	*Hol?* 'Where?'
loc	bAn-On	*könyv* 'book'	*Hol?* 'Where?'	*Mi-nél?* 'What-ADE?'	*Hol?* 'Where?'
loc	city-bAn	*Párizs* 'Paris'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Melyik város-on?* 'Which city-SUP?
loc	city-On	*Miskolc* 'Miskolc'	*Melyik város-ban?* 'Which city-INE?	*Hol?* 'Where?'	*Hol?* 'Where?'
loc	country	*Afganisztán* 'Afghanistan'	*Hol?* 'Where?'	*Hol?* 'Where?'	*Melyik ország-on?* 'Which country-SUP?

DOI: 10.15774/PPKE.BTK.2021.17

4.5 Discussion – generalizations and exceptions

Although the system of main and subcategories presented above may seem like a rather rigid classification, looking at the data – the categorisation of 1 097 lemmas – in more detail, we find many exceptions.

Of course it is also very common for a given word to produce the behaviour corresponding to its *default* category with each of the nine case suffixes. These words are prototypical instances of their categories. The following are some examples:

- body_any: *derék* 'waist'
- build=inst: *ügyészség* 'prosecution'
- date_bAn: *1987* '1987'
- group: *család* 'family'
- org_bAn: *cég* 'firm'
- period: *félév* 'semester'
- place_On: *célállomás* 'destination'
- posi_On: *hatalom* 'power'
- who: *árus* 'vendor'

It is also very common for a word to function as described in its default category in the sentence in most cases (6-7 times out of nine), but in 1-2 cases it has a different behaviour. Some examples are:

- *alkalom* 'occasion': its default category is *event_On*; however, with the case suffix *bÓl* it is a causative: (*cause*).
- *egy* 'one': its default category is *date*; with some suffixes it is an adverbial of manner (*egy-ben* 'one-INE': 'in one piece', or *egy-ből* 'one-ELA': 'directly').
- *eleje* 'beginning.of': its default category is *part*; with some suffixes it functions as an adverbial of time: (*elejé-n* 'beginning.of-SUP': 'at the beginning of', *elejé-től* 'beginning.of-ABL: 'from the beginning of')

DOI: 10.15774/PPKE.BTK.2021.17

- *előadás* 'lecture': it is an event (*event_On*), but when bearing the suffix *bAn* it is an adverbial of time (*date*) answering to the question *When?* or an adverbial of forms (*form*) answering to the question *Milyen formá-ban?* 'What.kind.of form-INE': 'In what form?'.

- *semmi* 'nothing': it is categorized as a *thing*. Its form *semmi-be* 'nothing-ILL', however, also answers to the question *Mennyi-be?* 'How.much-ILL?' of the category *num*; *semmi-ből* 'nothing-ELA' also answers to the question *Honnan?* 'Where from?'.

The above words are just a few of the many examples (all of them can be found in Table 4.7). They illustrate that many lemmas – instead of keeping with a rigid categorisation – need to be treated more flexibly; sometimes completely overwriting the roles provided by the default categorisation, or other times only by supplementing it with more options.

Moreover, in many cases, the 50 categories presented earlier prove to be insufficient: now and again rather than categories, specific questions have to be formed with the given lemma and the given suffix. Some examples:

- The lemma *többség* 'majority' received the tag *thing* as a default category. However, with certain suffixes more complex relevant questions can be formed such as *Milyen arányban?* 'What proportion?' or *Az értintettek mekkora részénél?* 'What proportion of those involved?'.

- *Méter* 'metre' and *kilométer* 'kilometre' are measuring units. The default questions of their default category (*meas*) are satisfying with every case suffix. With the suffix *rA*, however, they can function as adverbials answering the questions *Milyen messze?* or *Milyen messzire?* 'How far?' (while, of course, they can also function as place adverbials answering the question *Where?*).

In addition to the above, it should be mentioned that it is characteristic of some elements of certain categories that we do not ask *about* them, but *with* them, and we ask about their modifying elements.

- It is a typical behaviour of the category of currencies: when bearing the case suffix *rA*, every member has a relevant question of the form: *Hány <pénznem neve>-ra?* 'How.much <name of the currency>-SUB?'. In case of *400* ***forint-ra*** *emelked-ett a benzin ár-a.* '400 HUF-SUB rise-PAST.3SG the petrol price-POSS.3SG.' : 'The price

DOI: 10.15774/PPKE.BTK.2021.17

of petrol rose to 400 HUF.' both *Mennyi-re?* 'How.much-SUB' and *Hány forint-ra?* 'How.much HUF-SUB' are relevant questions.

- The same applies for measurement units. *Hány* ***kilométer-re****?* 'How.many **kilometer-Sub**?': 'How many kilometers away?'; *Hány* ***százalék-on****?* 'How.many **percent-Sup**': 'What percentage is?'; *Hány* ***tonná-tól****?* 'How.many **ton-Abl**?': 'From how many tons?', etc. These are all relevant questions, and this is a proper questioning scheme for all suffixes with this category of lemmas.

Finally, the fact cannot be ignored that there is a category which is not a default tag for any lemma. This one is *cause*, the tag of causatives. I found in the corpus that, although some lemmas have a causative role with some of the locative case suffixes, the tag *cause* is not a default category for any of these lemmas. Some examples illustrate this point:

- *apropó* 'apropos': *thing*, but *apropó-já-n* 'apropos-POSS.3SG-SUP', *apropó-já-ból* 'apropos-POSS.3SG-ELA': 'apropos of sg'
- *cél* 'goal': *loc_any*, but *cél-ból* 'goal-ELA': 'for the purpose'
- *megfontolás* 'deliberation': *thing*, but *megfontolás-ból* 'deliberation-ELA': 'for [this] reason'
- *nyomás* 'pressure': *thing*, but *nyomás-ra* 'pressure-SUB': 'under pressure'

4.6 The results and the QA system

As described in the introduction, the motivation of this study was to provide a useful annotation of adverbials for a QA system, or more precisely, for a corpus annotated with semantic role labels for Hungarian that can be used to train a parser-based system capable of formulating relevant questions about the text it processes. Table 4.6 and the list in Appendix C.1 are the essence and result of this study, though in these forms, they are not useful for being used in an annotation.

As mentioned earlier, in Section 4.5, a properly tagged adverbial in itself does not always provide a complex base for a question. In many cases, it is also required to know the adverbial's modifier and the verb as well. Take, for example, *bank* 'bank' from the

DOI: 10.15774/PPKE.BTK.2021.17

category *build=inst_bAn*. A bank is referred to linguistically as a person when sending an invoice letter, thus we can ask *Who did the letter come from?* in example (48a). But it is referred to as a thing, or more precisely, a place, when it is something we start from: the proper question of *a banktól* in example (48b) is *Where have you just started from?*.

(48) a. *Jö-tt* *egy levél* ***a*** ***bank-tól.***
Come-PAST.SG3 a letter the bank-ABL

'A letter came from the Bank.'

b. *Most indul-ok* ***a*** ***bank-tól.***
Now start-SG1 the bank-ABL

'I have just left the bank.'

These two examples illustrate that we do not wish to simply label these adverbials but to label them so that a verb can meet its desired thematic roles in the sentence with the help of these tags. Table 4.8 from Novák et al. (2019a) summarises the thematic roles used in the description of argument frames. I describe my adverbial categories to meet the annotation in the first column of Table (4.8).

The result of this "conversion" can be seen in Table (4.9). Many questions themselves are now defined in the table of the thematic roles (Table 4.8). Thus I only need to label what thematic role a given adverbial (a noun from a given category bearing a given suffix) can have. The cells indicate what thematic role (from 4.8) the members of a given category can take with the given suffix. Any tag not part of Table 4.8 is a novel proposal and is described in the caption of Table 4.9.

DOI: 10.15774/PPKE.BTK.2021.17

Table 4.8. Thematic roles used in the description of argument structures. The first column shows the annotation itself. The second column contains the name of the given thematic role. The third column presents the typical question(s) of the given role. Finally, an example sentence is given in the fourth column, where the phrase with the given role is in bold. (The order of the thematic roles follows their order in the original table in Novák et al., 2019a.)

Annotation	**Name**	**Question regarding the verb**	**Example**
AG	agent	What is AG doing?	**John** has climbed the tree.
CHAR	characterised	What is characteristic of CHAR?	**Expertise** is an advantage.
ATTR	attribute	–	Expertise is **an advantage**.
EXP	experiencer	How does EXP feel? What has EXP perceived?	**John** has seen a swallow.
PAT	patient	What happened to PAT?	John kissed **Mary**.
PATDST	patient-destination	What happened to PAT.to? Where did PAT get to?	He painted **the wall** green.
TH	theme	–	John relies **on his intuition**.
ST	stimulus	What effect has ST (on EXP)?	John loves **Mary**.
CONT	information content	–	John presented **the plan** to Joe.
REC	recipient	–	John presented the plan **to Joe**. **Mary** received a letter.
RES	result	How did RES come into being?	Mary baked **a cake**.
INS	instrument	What is AG using INS for?	John travels to work **by scooter**.
CAU	causer	What did CAU cause? What was the consequence of CAU?	John was late **because of an accident**.
MOT	motivation	–	John is studying to be **an engineer**.
LOC	location	What happened in/at/on... LOC?	John kissed Mary **in the cinema**.
SRC	source, starting point	–	John came **out of the room**. Mary received a letter **from John**.
DST	destination	How did AG/PAT get to DST?	John went **into the room**.
HOW	mode	–	John **deftly** climbed the tree.
ASPECT	aspect	–	John is doing well **financially**.
ACT	action	–	John wants **to work** from home.

DOI: 10.15774/PPKE.BTK.2021.17

Table 4.9. Table summarising the categories of locative adverbials and their possible thematic roles with the given suffixes. The cells indicate what thematic role (from 4.8) the members of a given category can take with the given suffix. Empty cells indicate a LOC (in columns *bAn*, *nÁl* and *On*), DST (in columns *bA*, *hOz* and *rA*) or SRC (in columns *bÓl*, *tÓl* and *rÓl*) thematic role (they are empty to help focusing on all the other, more interesting roles). LOC-T, DST-T are the time adverbial pairs of LOC and DST answering the question *When?* and *For when?*, respectively. bAn, bA, nÁl etc. as a cell content indicates that the given category with the given suffix does not have a specific thematic role but answers the questions *Mi-ben?* 'What-INE', *Mibe?* 'What-ILL', etc. If this suffix follows a tag (e.g. *WHO-bAn*), then the tag marks the proper wh-question that has to be used (e.g. *Ki-ben?* 'WHO-INE'). Examples are: DEM: demonstrative role (*Which?*); QUANT: quantifiers with the possible questions 'How many?', 'How much?'; FORM: describing some formal properties of the given act; CITY and LAND are tags for names of cities and countries. Any other tag not part of Table 4.8 is a novel proposal. CIRC: describing the circumstances of a given event; DIR: specifying the direction of a given act; MATER: describing the material of the thing in question; STA: giving information about the state of a participant; NUM_SIZE: a numeral denoting a (relative or absolute) size of something; SITU: specifying a situation that is the result or the source of a given action.

category		**case endings**								
main	**suppl**	**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
body	bAn-On		nÁl			hOz			tÓl	
body	any									
build=inst	On	bAn			bA			bÓl		
cause		CAU	CAU	CAU	CAU	CAU	CAU	CAU	CAU	CAU
circumst		CIRC	CIRC	CIRC	SITU	CIRC-hOz	SITU	SITU	SITU	SITU

DOI: 10.15774/PPKE.BTK.2021.17

category		case endings								
main	**suppl**	**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
curr		bAn	QUANT-nÁl	*hány_On* 'how.many_On'	QUANT-bA	hOz	QUANT-rA	QUANT-bÓl	QUANT-tÓl	QUANT-rÓl
date	On	bAn	nÁl	LOC-T	bA	hOz	DST-T	bÓl	SRC-T	rÓl
date	bAn	LOC-T	nÁl	On	bA	hOz	DST-T	bÓl	SRC-T	rÓl
dem		DEM-bAn	DEM-nÁl	DEM-On	DEM-bA	DEM-hOz	DEM-rA	DEM-bÓl	DEM-tÓl	DEM-rÓl
dempron		DEM-bAn	DEM-nÁl	DEM-On	DEM-bA	DEM-hOz	DEM-rA	DEM-bÓl	DEM-tÓl	DEM-rÓl
direct	ine	DIR	nÁl	On	DIR	hOz	rA		tÓl	rÓl
direct	On	bAn	nÁl	LOC / DIR	bA	hOz	DIR	bÓl	tÓl	
event	nÁl-On	bAn	LOC-T/LOC	LOC-T/LOC	bA		DST / DST-T	bÓl	SRC / SRC-T	
event	On	bAn	nÁl	LOC-T/LOC	bA	hOz	DST / DST-T	DST / DST-T	tÓl	rÓl

DOI: 10.15774/PPKE.BTK.2021.17

category		case endings								
main	**suppl**	**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
event	bAn	LOC-T / LOC	LOC-T / LOC	On			rA / DST-T		SRC / SRC-T	rÓl
build=inst	bAn			On			rA			rÓl
form	bAn	FORM-bAn	nÁl	On	FORM-bA	hOz	rA	FORM-bÓl	tÓl	rÓl
form	On	bAn	nÁl	FORM-On	bA	hOz		bÓl	tÓl	
group				kiken			kikre			kikről
loc	nÁl	bAn		On	bA		rA	bÓl		rÓl
loc	any									
loc	bAn		nÁl	On		hOz	rA		tÓl	rÓl
loc	On	bAn	nÁl		bA	hOz		bÓl	tÓl	
loc	bAn-On		nÁl			hOz			tÓl	
loc	city-bAn			CITY-On			CITY-rA			CITY-rÓl
loc	city-On	CITY-bAn			CITY-bA			CITY-bÓl		

DOI: 10.15774/PPKE.BTK.2021.17

category		case endings								
main	**suppl**	**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
loc	country			LAND-On			DST / LAND-rA			LAND-rÓl
loc=who				WHO-On			WHO-rA			WHO-rÓl
material		bAn	nÁl	On	bA	hOz	rA	MATER	tÓl	rÓl
meas		QUANT-bAn	QUANT-nÁl	QUANT-On	QUANT-bA	QUANT-hOz	QUANT-rA	QUANT-bÓl	QUANT-tÓl	QUANT-rÓl
mode	bAn	bAn	nÁl	HOW	bA	hOz	HOW	HOW	tÓl	rÓl
mode	On	HOW	nÁl	On	bA	hOz	HOW	HOW	tÓl	rÓl
num		QUANT-bAn	QUANT-nÁl	QUANT-On	QUANT-bA	QUANT-hOz	QUANT-rA	QUANT-bÓl	QUANT-tÓl	QUANT-rÓl
num_size		QUANT-bAn	QUANT-nÁl	QUANT-On	QUANT-bA	QUANT-hOz	NUM-SIZE-rA	QUANT-bÓl	QUANT-tÓl	NUM-SIZE-rÓl
org	bAn			On			rA			rÓl
org	nÁl	bAn		On	bA		rA	bÓl		rÓl
org=who	any									

DOI: 10.15774/PPKE.BTK.2021.17

category		case endings								
main	**suppl**	**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
org=who	On	WHO-bAn						WHO-bÓl		
org=who	bAn			WHO-On			WHO-rA / rA			WHO-rÓl / rÓl
org=who	nÁl	WHO-bAn		WHO-On	WHO-bA		WHO-rA	WHO-bÓl		WHO-rÓl
part										
period		LOC-T	nÁl	On	bA	hOz	mennyi időre	bÓl	SRC-T	rÓl
place	bAn			On			rA			rÓl
place	On	bAn			bA			bÓl		
way		bAn			bA			bÓl		
posi	On	bAn	nÁl		bA	hOz		bÓl	tÓl	
posi	bAn		nÁl	On		hOz	DST /rA		tÓl	rÓl
pov		AS-PECT	nÁl	On	bA	hOz	rA	AS-PECT	tÓl	rÓl

DOI: 10.15774/PPKE.BTK.2021.17

category		case endings								
main	**suppl**	**bAn**	**nÁl**	**On**	**bA**	**hOz**	**rA**	**bÓl**	**tÓl**	**rÓl**
state		STA	LOC-T	On	STA	hOz	rA	STA	tÓl	rÓl
thing		bAn	nÁl	On	bA	hOz	rA	bÓl	tÓl	rÓl
who		WHO-bAn		WHO-On	WHO-bA		WHO-rA	WHO-bÓl		WHO-rÓl

DOI: 10.15774/PPKE.BTK.2021.17

4.7 Conclusion

In this section I presented – while at the same time keeping in mind the needs of a QA system – that in the case of adverbials in a sentence what kind of annotation would be appropriate when designing a training corpus. In my analysis, I focused 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: the nine members of the internal, the external, and the external, surface-oriented paradigms. The majority of these words functions as an oblique modifier, some of them are arguments.

I defined 28 categories – with the subcategories altogether 50 – into which the words meeting the above-described criteria can be sorted. For most words, it is not sufficient to choose a default category; in some cases, with some case endings, words may perform a role different from the one defined by the default category, the labelling of which is also a task. This in-depth categorisation was also conducted and presented here. The list of the 1 097 words examined here and their detailed categorisation is attached in Appendix C.1. The default categories and their proper questions can be found in Table 4.6.

The categorisation presented here provides appropriate features in a train corpus to create a Question-Answer system described in the introduction. However, some open questions remain: in those cases where a combination of a noun and a given case ending can cover more than one possible adverbial function, other clues are required to find the required precise question.

In case of *órá-ra* 'hour/class-SUB' it is also necessary to know the word's modifiers to choose the right question. In example (49a), *Hova?* 'To where?' is the appropriate question of *az órá-ra* 'the class-SUB', however, if the phrase is *néhány órá-ra* 'some hour-SUB', as in example (49b), the right question is *Mennyi idő-re?* 'How.much time-SUB', 'For how long?'.

(49) a. *El-megy-ek* ***az órá-ra***.
Away-go-SG1 the class-SUB.

'I will attend the class.'

DOI: 10.15774/PPKE.BTK.2021.17

b. *El-megy-ek* ***néhány órá-ra***.
Away-go-SG1 some hour-SUB.

'I go away for a couple of hours.'

For the items in some categories, it is therefore imperative to know the modifiers of the words and to label the noun and its modifier together as well. This is the direction in which I would like to take my research further, supplemented by adverbials in the sentence bearing not one of the nine case endings examined so far, but another case ending, such as the 10th locative case suffix, terminativus *-ig*, or the instrumentalis *-vAl*. The next step is, naturally, to measure this categorisation and annotation proposal presented here, to primarily test and evaluate it on the QA system.

There are also many other case suffixes in Hungarian apart from the above examined 9 – most importantly, there is one other locative suffix, the terminative *-ig*, which definitely fits into the study on locative case suffixes. I omitted it from this analysis as its meaning (and the question it answers) is more straightforward than that of the other 9. This *-ig*, and all the other case suffixes must be examined and categorised the same way as the 9 locative ones.

DOI: 10.15774/PPKE.BTK.2021.17

Chapter 5

Postpositions

"As you like it"

A play by William Shakespeare

5.1 Introduction

I started to study postpositions when trying to define the borders of an NP for ANA-GRAMMA (Ligeti-Nagy, 2015, about the parser see 1.3).[1] As I mentioned in 4.1, postpositions play a crucial role in NP-chunking as they unquestionably mark the ending of a noun phrase; see (50) for a very common example. *Után* 'after' is a postposition, and *ebéd után* 'after lunch' is an NP, the end of which is indicated by the postposition. However, as we will see in 5.1.1, one can hardly find two separate linguistic papers arguing for the same categorisation of postpositions – in fact, even encountering the exact same words as postpositions is just as unlikely.

(50) *Ebéd után megy-ek haza.*
Lunch after go-1SG home
'I go home after lunch.'

The diversity in their linguistic perception and their crucial, suffix-like role in sentences make it inevitable to encounter them properly, based on corpus data, always keeping in mind the possible needs of a sentence parser.

[1] The results presented here were partly published in Ligeti-Nagy (2018).

DOI: 10.15774/PPKE.BTK.2021.17

5.1.1 Literature review

There is hardly another word class that, despite its relatively small number of elements, would generate such a long-lasting debate as Hungarian postpositions. It is almost impossible to find two papers arguing the same classification of them. In this section, I attempt to give an overview of what theoretical linguists have said about Hungarian postpositions so far.

In traditional Hungarian grammars (*Magyar Grammatika*, Keszler, 2000) postpositions are described as a group of function words forming a "morphological-like unit" (*morfológiai természetű egység*) with the noun (phrase) preceding them, therefore functioning as case suffixes. The crucial criterion according to these grammars is that i) the postposition directly follows the noun ii) they form a morphological structure together iii) the postposition is usually unstressed.

These grammars categorise postpositions as follows:

(51) a. **Real postpositions**, postpositions taking a caseless noun as a complement

- first, there are "simple" postpositions, without any morphological structure, such as:
  által 'by', *alá* 'to under', *alatt* 'under', *alól* 'from under', *elé* 'to in front of' etc.
- second, there are some postpositions bearing case suffixes, where there is a clear morphological structure, for example:
  ízben 'times', *közben* 'during', *létére* 'despite being' etc.

  Here there is a distinction between postpositions taking a singular noun and postpositions always taking a plural noun, such as *között* 'between', as in *asztalok között* 'between chairs'.

b. **Postpositions taking a noun with a lexical case**: *fogva* + ADE 'because of, due to', or *fogva* + ABL 'beginning from', *képest* + ALL 'compared to, for' etc.

   Magyar Grammatika already states that some of these postpositions taking a noun with a lexical case may appear before the noun, not only after it. There is also a note regarding the overlap between postpositions and verbal prefixes.

DOI: 10.15774/PPKE.BTK.2021.17

c. **Postpositions with a possessive structure**: *alapján* 'based on', *céljából* 'with the aim of', *dacára* 'despite', *ellenére* 'despite' etc.

This categorisation is strongly disputed in papers from a structuralist or generative linguistics background. In the first volume of the series *Strukturális magyar nyelvtan* [A Structural Grammar of Hungarian] (Kiefer, 1992) postpositions are classified into three (and a half) groups:

(52) a. **Case-like postpositions**: the members of this group act like a case suffix. They do not assign a case to the NP, but rather they take a caseless noun phrase complement. Examples:
által 'by', *alá* 'to under',*alatt* 'under', *alól* 'from under', *elé* 'to in front of' etc.

b. **Real postpositions**: elements taking a noun phrase compliment with a lexical case.
Postpositions taking a noun with superessive case suffix:
alul 'below', *át* 'through', *belül* 'inside of' etc.
Postpositions taking a noun with instrumentalis case suffix:
szemben 'opposite to', *együtt* 'together'.

c. **Other postpositions**: postpositions never co-occurring with a personal pronoun or a demonstrative:
végett 'in order to, due to', *közben* 'during', *óta* 'since', *gyanánt* 'as, by way of', *hosszat* 'for'.

d. (Postposition-like elements): a transitional class containing tokens with a possessive case suffix. These are not to be considered postpositions based on the classification rules of this book.
számára 'for', *ellenére* 'despite', *révén* 'by means of', *irányában* 'towards'.

A pure generative approach is still needed. This is provided for example in *The Syntax of Hungarian* by É. Kiss (2002). Here, words earlier described as postpositions are grouped into the following categories:

(53) a. **Adverbs** taking an argument ("postpositions" taking a noun phrase complement with a lexical case):
együtt 'together', *alul* 'below' etc.

DOI: 10.15774/PPKE.BTK.2021.17

b. **Idiomatic participles** taking a noun phrase complement with a lexical case: *nézve* 'regarding', *kezdve* 'beginning from' etc.

c. **Postpositions** taking a caseless noun phrase complement: *alá* 'to under',*alatt* 'under', *alól* 'from under', *elé* 'to in front of' etc.

These three divergent categorisations show how uncertain the definition of postpositions in the Hungarian linguistic literature is. However, one has to refer to the thesis of Éva Dékány as well, which provides a detailed and sorted list of all Hungarian postpositions (Dékány, 2012: 108–109.). Dékány initiated her categorisation from the concept of "dressed" and "naked postpositions". The terms come from Marácz (1986) and were meant to suggest that dressed postpositions have something that naked postpositions don't: the former carry number and agreement suffixes in the presence of a pronoun, whereas the latter do not. The latter group includes postpositions taking a noun with a lexical case, while the former contains those taking caseless complements. Table 5.1 taken from Dékány (2012: 108.) shows "naked postpositions" with the case the given postposition takes, while Table 5.2, also taken from Dékány (2012: 108–109.), presents "dressed postpositions". The inventory of Dékány (2012) is also included in Table 5.3, where I intended to provide a detailed list of the postpositions ever mentioned in the literature (based on the above discussed four items).

DOI: 10.15774/PPKE.BTK.2021.17

Table 5.1. “Naked postpositions” - table slightly modified from Dékány (2012: 108.). The first column shows the postpositions taking a noun with a lexical case; the second column is for their meaning; the third column contains the case they take. The last column of the original table, containing other information on the behaviour of these postpositions, is not shown here. The postpositions in parentheses are excluded from the group of postpositions by Dékány (2012), although they are listed in the table as the original sources of this collection (Kenesei et al., 1997; Asbury, 2008a) do display them.

postposition	**meaning**[2]	case
alul	below, under	superessive
át	through, via, across, over	superessive
belül	inside of	superessive
dacára	despite	(dative)
együtt	together	instrumental
ellenére	despite	(dative)
felül	over, above	superessive
hasonlóan	similarly to	allative
innen	on this side of	superessive
keresztül	through, via, across	superessive
képest	compared to, for	allative
kívül	outside of, apart from	superessive
kívülre	to outside of	superessive
kívülről	from outside	superessive
közel	close to	allative
szembe	to opposite to	instrumental
szemben	opposite to	instrumental
szemből	from opposite to	instrumental
szemközt	opposite to	instrumental
túl	beyond, over	superessive
túlra	to beyond	superessive
túlról	from beyond	superessive
végig	along to the end of	superessive
(*fogva*)	because of, due to	adessive
(*fogva*)	beginning from	ablative
(*kezdve*)	beginning from	ablative
(*kivéve*)	except for	accusative
(*nézve*)	regarding	sublative

DOI: 10.15774/PPKE.BTK.2021.17

Table 5.2. "Dressed postpositions" - table slightly modified from Dékány (2012: 108–109.). The first column shows the postpositions taking a noun without a lexical case; the second column confirms their meaning. The third column of the original table, containing other information on the behaviour of these postpositions, is not shown here. The postpositions in parentheses are excluded from the group of postpositions and handled as possessive marked DP-s by Dékány (2012), although they are listed in the table as the original sources of this collection (Kenesei et al., 1997; Asbury, 2008a) do enlist them.

postposition	**meaning**[3]
alatt	under
alá	to under
alól	from under
által	by
elé	to in front of
ellen	against
ellenére	despite
elől	from in front of
előtt	(at) in front of, before
esetén	in case of
felett/fölött	(at) above
felé	towards
felől	from the direction of
folytán	owing to
fölé	to above
fölül	from above
gyanánt	as, by way of, in lieu of
helyett	instead of
iránt	towards
köré	to around
körül	around
között	between
közé	to between
közül	from between

[3]The representation of the meaning of the postpositions is taken from É. Kiss and Hegedűs (2021).

DOI: 10.15774/PPKE.BTK.2021.17

postposition	**meaning**
mellett	next to, beside
mellé	to next to
mellől	from next to
miatt	because of
mögött	behind
mögé	to behind
mögül	from behind
múlva	in (X time), after (X time)
nélkül	without
nyomán	based on
óta	since
során	during
szerint	according to
tájban / tájt	around (a point in time)
után	behind, after
útján	by way of
végett	in order to, due to
(*javára*)	in favour of
(*kedvéért*)	for the sake of
(*létére*)	despite being
(*részére*)	for (DAT)
(*révén*)	through, by means of
(*számára*)	for (DAT)

DOI: 10.15774/PPKE.BTK.2021.17

The most recent discussion on postpositions and postpositional phrases is the 3rd volume of the Comprehensive Grammar Resources series, *Postpositions and Postpositional Phrases*, edited by Katalin É. Kiss and Veronika Hegedűs (2021). The book partly follows Dékány's (2012) classification and distinguishes two main types of postpositions: case-like and case-assigning postpositions. However, the authors add borderline cases of postpositions: possessive postpositions (*helyében* 'in X's place/shoes', *során* 'during' etc.), which are possessive constructions not yet fully grammaticalized into Ps. They may be regarded as transitional elements. The other group of borderline cases is that of participial postpositions (*kezdve* 'beginning from', *nézve* 'regarding'), which comprise a verbal stem and the *-va/-ve* adverbial participial suffix. A very interesting part of the book (among others) is section 2.2.2.3.8, where the authors discuss the behaviour of case-assigning postpositions. As they show, not all these postpositions behave similarly with respect to certain distributional tests. They sum up the pattern of this behaviour in a table (p. 80). This is something very similar to what I intend to do later in this chapter (see section 5.2)

Of course, there are many other significant papers on the topic of Hungarian postpositions, for example that of Marácz (1986), an old monography of Árpád Sebestyén (1965), or numerous papers from a generative background (É Kiss, 1990; Hegedűs, 2006; Dékány, 2009; Dékány and Hegedűs, 2013). However, the categorisation of postpositions in these papers do fit into one of the above detailed descriptions, therefore it is not necessary to elaborate on all of them.

The main problem of the characterisation of postpositions is rather salient: there is a group of postpositions (let us refer to them as "pure postpositions") that always act like postpositions (I will examine this postposition-like behaviour in 5.2), they bear all the important features of a postposition and finally, they are categorised as postposition by every relevant paper in this field. The rows of these postpositions are highlighted in Table 5.3. On the other hand, there are some words that are postpositions from a certain point of view, and I intended to shed light on this by Table 5.3, but are something different from another point of view. My intention is to capture this continuum: to define the group and the features of typical postpositions and circumscribe the layers surrounding this core containing less typical postpositions.

DOI: 10.15774/PPKE.BTK.2021.17

Table 5.3. List of all the postpositions mentioned in the linguistic studies discussed above. The words in italics are the postpositions. The four columns named MG (Keszler, 2000), Str (Kiefer, 1992), SoH (É. Kiss, 2002) and D.É.K. (Dékány, 2012) indicate the four above-mentioned literature sources and their opinion on the given postpositions. A number indicates that the given word is categorised as a postposition by the author. A letter indicates that the given word is mentioned there, but not as a postposition. Different numbers and different letters refer to different sub-categorisation within the given study. (As an example, the numbers 1, 2 and 3 in the column of MG stand for the three groups of postpositions the author differentiates. The letters a and b in the column of SoH represent the two classes of words often considered a postposition but rejected by the author.) Parentheses show that the given word was not mentioned by the author but it may still fit into one of the categories of the given study. A "-" means that the given paper does not state anything about that word. The table is also available at `https://github.com/ppke-nlpg/postp` along with the other materials of this chapter.

postposition	**MG**	**Str**	**SoH**	**D.É.K.**
alá 'to under'	1	1	1	1
alatt 'under'	1	1	1	1
alól 'from under'	1	1	1	1
által 'by'	1	1	1	1
elé 'to in front of'	1	1	1	1
ellen 'against'	1	1	1	1
elől 'from in front of'	1	1	1	1
előtt '(at) in front of, before'	1	1	1	1
felé 'towards'	1	1	1	1
felől 'from the direction of'	1	1	1	1
fölé 'to above'	1	1	1	1
fölött '(at) above'	1	1	1	1
helyett 'instead of'	1	1	1	1
iránt 'towards'	1	1	1	1
köré 'to around'	1	1	1	1
körül 'around'	1	1	1	1
közé 'to between'	1	1	1	1
között 'between'	1	1	1	1

DOI: 10.15774/PPKE.BTK.2021.17

postposition	**MG**	**Str**	**SoH**	**D.É.K.**
közül 'from between'	1	1	1	1
mellé 'to next to'	1	1	1	1
mellett 'next to, beside'	1	1	1	1
mellől 'from next to'	1	1	1	1
miatt 'because of'	1	1	1	1
mögé 'to behind'	1	1	1	1
mögött 'behind'	1	1	1	1
mögül 'from behind'	1	1	1	1
nélkül 'without'	1	1	1	1
szerint 'according to'	1	1	1	1
után 'behind, after'	1	1	1	1
gyanánt 'as, by way of, in lieu of'	1	3	(1)	1
óta 'since'	1	3	(1)	1
végett 'in order to, due to'	1	3	(3)	1
alul 'below'	2	2	a	2
át 'through'	2	2	a	2
belül 'inside of'	2	2	a	2
együtt 'together'	2	2	a	2
keresztül 'through'	2	2	a	2
kívül 'outside of'	2	2	a	2
túl 'beyond'	2	2	a	2
szembe 'to opposite to'	(2)	(2)	(a)	2
szemben 'opposite to'	2	2	(a)	2
szemből 'from opposite to'	(2)	-	(a)	2
szemközt 'opposite to'	(2)	-	(a)	2
fogva 'as a result of'	2	2	b	a
kezdve 'beginning from'	2	2	b	a
nézve 'regarding'	2	2	b	a

DOI: 10.15774/PPKE.BTK.2021.17

postposition	**MG**	**Str**	**SoH**	**D.É.K.**
felül 'over'	2	-	(a)	2
fölül 'from above'	2	-	(a)	1
közel 'close to'	2	-	(a)	2
túlra 'to beyond'	(2)	-	(a)	2
túlról 'from beyond'	(2)	-	(a)	2
végig 'to the end of'	2	-	(a)	2
javára 'in favour of'	3	a	-	b
kedvéért 'for the sake of'	3	a	-	b
alapján 'based on'	3	a	-	(b)
céljából 'with the aim of'	3	a	-	(b)
dacára 'despite'	3	a	-	(b)
érdekében 'on behalf of'	3	a	-	(b)
értelmében 'in pursuance of'	3	a	-	(b)
esetében 'in case of'	3	a	-	(b)
folyamán 'in the course of'	3	a	-	(b)
irányában 'towards'	3	a	-	(b)
következtében 'following'	3	a	-	(b)
részére 'for'	3	a	-	b
révén 'through, by means of'	3	a	-	b
számára 'for'	3	a	-	b
részéről 'on the part of'	3	a	-	(b)
táján 'around'	3	(a)	-	(b)
ellenére 'despite'	3	a	1	1
esetén 'in case of'	3	a	-	1
nyomán 'based on'	3	a	-	1
során 'during'	3	a	-	1
útján 'by way of'	3	a	-	1
folytán 'owing to'	3	-	-	1

DOI: 10.15774/PPKE.BTK.2021.17

postposition	**MG**	**Str**	**SoH**	**D.É.K.**
múlva 'in (X time), after (X time)'	1	-	-	1
tájban/tájt 'around'	1	-	-	1
ízben 'times'	1	-	-	-
módra 'like'	1	-	-	-
múltán 'in (X time), after (X time)'	1	-	-	-
közben 'during'	1	3	-	-
létére 'despite being'	1	a	-	b
módjára 'like'	1	a	-	-
innen 'on this side of'	2	-	-	2
kívülre 'to outside of'	(2)	-	-	2
kívülről 'from outside'	(2)	-	-	2
hasonlóan 'similarly to'	-	-	-	2
hosszat 'for'	-	3	-	-
képest 'compared to, for'	-	-	(a)	2
kivéve 'except for'	(2)	-	(b)	a

Table 5.3 clearly defines the list of "pure postpositions" (they are highlighted in the table). These 29 words are unquestionably categorised as postpositions by all relevant papers in the field. In the next subsection I examine corpus data to inspect whether their occurrences in texts support their categorisation.

DOI: 10.15774/PPKE.BTK.2021.17

5.2 Postpositions in corpora

In Section 5.1, I stated that the importance of postpositions – from a computational point of view – lies in the fact that they indicate the strict end of a noun phrase in the same way as case suffixes do. However, numerous items in Table 5.3 tend to appear in other positions both before and after the noun phrase. Here in this study, I proceed by keeping only one important criterion: the candidate word has to take a noun phrase, caseless or not, as a complement. This is a further goal of an algorithmic description of the behaviour of as many postposition-like elements as possible. Therefore, I do not wish to include the word *kívülről* in (54b) in my study, but intend to examine that of (54a) (the examples taken from Dékány, 2012: 112[4]).

(54) a. *A hang a szobá-n kívül-ről jött.*
the sound the room-SUP outside.of-DEL come.PST
'The sound came from outside of the room.'

b. *A hang kívül-ről jött.*
the sound outside.of-DEL come.PST.
'The sound came from outside.'

I initiated my search from the list of the words in Table 5.3, but continuously expanded it with other postposition candidates materialized in the queries.

The corpus I used is the 2.0.4. version of the *Hungarian Gigaword Corpus* (Oravecz et al., 2014).

Based on the literature, and on the results of numerous queries, I outlined the following important features of postpositions:

- **position**: by position I mean the preferred ordering of the postposition and its complement regardless of their adjacency.
- the **case-marking** of the complement: at first, to keep the features binary, I started by differentiating between postpositions with a caseless noun and postpositions tak-

[4]Dékány (2012) discussed the examples in 54a and 54b as examples of a postposition being used intransitively as well (54b). In her analysis, *kívülről* in 54b expresses a relation with respect to a deictic centre understood from the context: *here*.

DOI: 10.15774/PPKE.BTK.2021.17

ing a noun with a lexical case. However, later on, it will be necessary to distinguish the postpositions taking a noun with a lexical case by their required case.

- **adjacency**: Are the postposition and the noun always strictly adjacent, or can other tokens intersect them?
- their position in **wh-questions**: does the postposition always follow the wh-word (see example 55a), or can it stay behind (55b)?

(55) a. ***Ki*** ***után*** *jöv-ök?*
Who after come-1SG?
'After whom do I come?'

b. ***Mi-n*** *men-t-él* ***keresztül****?*
What-SUP go-PST-2SG through?
Through what did you go?'

These features are mainly syntactical, and are especially motivated by the computational point of view applied here. However, one has to account for some morphological typicality that does not influence the computational analysis of these words, but it is nonetheless important in order to have a complex view of them.

- **demonstratives**: is the postposition copied onto the demonstrative as well (see example 56a) or is only the case marker copied (56b)?
- person-number **agreement**: when postpositions take a pronominal complement, where does the agreement marker appear? On the postposition itself (example 57a), or on another element (57b)?

(56) a. *az* ***alól*** *a rét* ***alól***
that **below.from** the meadow **below.from**
'from under that meadow'

b. *az-on* *a* *rét-en* ***át***
that-SUP the meadow-SUP **through**
'through that meadow'

DOI: 10.15774/PPKE.BTK.2021.17

(57) a. *alól-**am***
under.from-1SG
'from under me'

b. *rajt-**am** keresztül*
on-1SG through
'through me'

These distributional features are mentioned in one or more papers as the basis of the classification of postpositions, or when necessary but insufficient conditions are to be met by postpositions. The adjacency of the postposition and the noun is mainly studied as a restriction: if a given word can be modified, then it is not a postposition (see for example É. Kiss, 2002: 181–183.). They do not, however, appear together as a compact feature list upon which postpositions are evaluated.

My results are summarised in Table 5.5 which shows the analysis of postpositions based on features of binary values. The meaning of the columns and the binary values they contain are summarized in Table 5.4.

Table 5.4. Features and their binary values when evaluating the behaviour of postpositions in the corpus. The table describes the content of Table 5.5. P stands for postposition.

Column	**the feature**	**if 1**	**if 0**
pos	position of P relative to the noun	P always follows the noun	P may appear before and after the noun
∅	the case of the noun	always caseless noun	P takes a noun with a lexical case
adj	the adjacency of the two words	noun and P always next to each other	other words may appear between P and the noun
wh	the behaviour of P in wh-questions	P follows the wh-word (55a)	P stays behind (55b)
dem	P's behaviour with demonstrative pronouns	P is copied onto the demonstrative (56a)	P is not copied onto the demonstrative (56b)
pers pron	person-number agreement with a personal pronoun	person-number agreement appears on P (57a)	person-number agreement not on P (57b)

DOI: 10.15774/PPKE.BTK.2021.17

It must be noted with regard to the methodology that I needed five counterexamples to prevent a given postposition from receiving a value of 1 for a specific property. For example, if the word appears before a noun five times, then its value for the feature "pos" is 0. Therefore, corpus queries used to build the database presented in Table 5.5 were mainly searches to prove the existence of counterexamples: if the query resulted in four or fewer matches, the given postposition received a value of 1 for the given feature.

The evaluation of the "wh" property based on the corpus is particularly difficult, therefore in most cases, the value was determined based on my linguistic intuition. Cells that contain a ?, indicate that the acceptability of my examples for testing a given property is not entirely certain (see example (58) with the word *innen* 'on this side of'). The same goes for the postposition's behaviour with demonstrative pronouns (see example (59) with the word *módjára* 'like, the way of').

(58) a. *A part megrepedez-ett a folyó-n innen*
the shore crack-PastSg3 the river-Sup on.this.side.of

'The shore was cracked on this side of the river.'

b. *Mi-n innen repedez-ett meg a part?*
what-Sup on.this.side.of crack-PastSg3 Perf the shore

'On this side of what is the shore cracked?'

(59) a. *Viselkedj állat módjára!*
behave animal the.way.of

'Behave like an animal!'

b. *Mi módjára viselked-j-ek?*
what the.way.of behave-Imp-Sg1

'I should behave like what?'

DOI: 10.15774/PPKE.BTK.2021.17

Table 5.5. Listing all the postpositions from the literature and their attribute values. A value of 1 indicates that the given postposition always produces the typical behaviour of postpositions in the syntactic structure under examination. Column *pos* describes the position of the word with regard to the noun. Column ∅ represents whether the postposition always takes a caseless noun. If not, the case it assigns is also marked. Column *adj* delineates the strict adjacency of the two words. Column *wh* represents the word's behaviour in wh-questions. The two columns after the vertical line are the morphological attributes: *dem* details the structure with a demonstrative pronoun, *pers pron* specifies the structure with a personal pronoun. A question mark indicates that I am not entirely certain in the acceptability of my examples for testing the given property. A '-' marks that the given postposition does not appear in the given construction (with a personal pronoun, for example).

postposition	**meaning**	**pos**	**∅**	**adj**	**wh**	**dem**	**pers pron**
alatt	under	1	1	1	1	1	1
alól	from under	1	1	1	1	1	1
ellen	against	1	1	1	1	1	1
elől	from in front of	1	1	1	1	1	1
előtt	in front of	1	1	1	1	1	1
felé	towards	1	1	1	1	1	1
felől	from the direction of	1	1	1	1	1	1
fölé	to above	1	1	1	1	1	1
fölött	(at) above	1	1	1	1	1	1
helyett	instead of	1	1	1	1	1	1
iránt	towards	1	1	1	1	1	1
köré	to around	1	1	1	1	1	1
körül	around	1	1	1	1	1	1
közé	to between	1	1	1	1	1	1
között	between	1	1	1	1	1	1
közül	from between	1	1	1	1	1	1
mellé	to next to	1	1	1	1	1	1
mellett	next to, beside	1	1	1	1	1	1
mellől	from next to	1	1	1	1	1	1
miatt	because of	1	1	1	1	1	1
mögé	to behind	1	1	1	1	1	1

DOI: 10.15774/PPKE.BTK.2021.17

postposition	meaning	pos	∅	adj	wh	dem	pers pron
mögött	behind	1	1	1	1	1	1
mögül	from behind	1	1	1	1	1	1
nélkül	without	1	1	1	1	1	1
szerint	according to	1	1	1	1	1	1
után	after	1	1	1	1	1	1
által	by	1	0:SUP	1	1	1	1
érdekében	on behalf of	1	0:DAT	1	1	0	1
esetében	in case of	1	0:DAT	1	1	0	1
fölül	from above	1	0:SUP	1	1	0	1
részére	for	1	0:DAT	1	1	0	1
részéről	on the part of	1	0:DAT	1	1	0	1
során	during	1	0:DAT	1	1	0	1
számára	for	1	0:DAT	1	1	0	1
fogva	from (time)	1	0:ABL	1	1	0	0
nézve	regarding	1	0:SUB	1	1	0	0
alapján	based on	1	0:DAT	1	1	0/1	-
céljából	with the aim of	1	0:DAT	1	1	0	-
ellenére	despite	1	0:DAT	1	1	0	-
értelmében	in pursuance of	1	0:DAT	1	1	0	-
esetén	in case of	1	0:DAT	1	1	0	-
folyamán	in the course of	1	0:DAT	1	1	0	-
kívülről	from outside	1	0:SUP	1	1	0	-
következtében	following	1	0:DAT	1	1	0	-
nyomán	based on	1	0:DAT	1	1	0	-
révén	by means of	1	0:DAT	1	1	0	-
útján	by way of	1	0:DAT	1	1	0	-
folytán	as a consequence of	1	0:DAT	1	?	0	-
közben	during	1	1	1	1	1	-
múltán	after (time)	1	1	1	1	0	-
óta	since	1	1	1	1	1	-
végett	with the aim of	1	1	1	1	-	1

DOI: 10.15774/PPKE.BTK.2021.17

postposition	meaning	pos	∅	adj	wh	dem	pers pron
gyanánt	as	1	1	1	-	-	-
hosszat	for	1	1	1	-	-	-
ízben	times	1	1	1	-	-	-
létére	despite being	1	1	1	-	-	-
módjára	way of	1	1	1	-	0?	-
módra	mode of	1	1	1	-	-	-
múlva	after (time)	1	1	1	-	-	-
táján	around	1	1	1	1 -?	-?	-
tájban/tájt	around (time)	1	1	1	-?	-?	-
irányában	towards	1	0:DAT	0	1	0	1
javára	in favour of	1	0:DAT	0	1	0	1
kedvéért	for the sake of	1	0:DAT	0	1	0	1
alul	below	1	0:SUP	0	1	0	0
képest	compared to	1	0:ALL	0	1	0	0
kívülre	to outside	1	0:SUP	0	1	0	0
túlra	to beyond	1	0:SUP	0	1	0	0
túlról	from beyond	1	0:SUP	0	1	0	0
hasonlóan	similarly	0	0:ALL	0	1	0	0
kivéve	except for	0	0:ACC	0	1	0	0
kívül	outside	0	0:SUP	0	1	0	0
szembe	to opposite to	0	0:INS	0	1	0	0
szemben	opposite to	0	0:INS	0	1	0	0
szemközt	opposite to	0	0:INS	0	1	0	0
felül	over	1	0:SUP	0	0	0	0
alá	to under	0	0:DAT	1	1	1	1
elé	to in front of	0	0:DAT	1	1	1	1
kezdve	beginning from	0	0:ABL	1	1	0	0
dacára	despite	0	0:DAT	1	1	0	-
belül	inside of	?	0:SUP	1	1	0	0
át	through	0	0:SUP	0	0	0	0

DOI: 10.15774/PPKE.BTK.2021.17

postposition	meaning	pos	∅	adj	wh	dem	pers pron
együtt	together	0	0:INS	0	0	0	0
keresztül	through	0	0:SUP	0	0	0	0
közel	close to	0	0:ALL	0	0	0	0
túl	beyond	0	0:SUP	0	0	0	0
végig	to the end of	0	0:SUP	0	0	0	0
innen	on this side of	0	0:SUP	0	?	0	0
szemből	from opposite to					0	0

The first thing to see in Table 5.5 is that "−" is very frequent in columns 4-6, indicating that the given postposition does not appear in the structure under examination. For example, the word *gyanánt* 'as' cannot be connected to a personal pronoun, therefore in the last position of the vector of *gyanánt* a "−" can be seen.

If we concentrate on the postpositions with a vector containing only 1 values, the group of typical postpositions is outlined: these are almost completely identical to the group of "pure postpositions" (5.1.1); the words categorised as postpositions by every linguistic paper I mentioned. I call this group typical postpositions as they match with "pure postpositions", although Kiefer (1992) called them case-like postpositions because he considered case assigning postpositions being the typical ones.

The exceptions here are, on the one hand, the postpositions whose base form is homonymous to the one attached to a third person singular personal pronoun (see examples (60a) and (60b)). These tokens may appear in front of the noun as well (example (61)). Another exception is *által* 'by', which sometimes - mainly in the literary subcorpus, but also in a small number in the personal subcorpus - takes a noun with a lexical case (example (62)).

(60) a. *elé*
to.in.front.of
'to in front of'

b. *elé*
to.in.front.of.Sg3
'to in front of him/her'

DOI: 10.15774/PPKE.BTK.2021.17

(61) *... szólt elé a kocsisnak.*
... said to.in.front.of the driver.DAT.
'... said to the driver.'

(62) *Természetesen szintén szigorúan tévén által.*
Naturally as.well strictly television.SUP by.
'Naturally, also strictly by television.'

In sum, we can say that the corpus queries confirmed the typical behaviour of the words that are uniformly regarded as postpositions by linguistic studies.

If we omit the values of the properties from 4-6, another significant group emerges: the words which received a value of 1 for the first three, syntactic properties. Since the last three features cannot be fully applied to these words as conditions, omitting those should not be a problem in an algorithmic processing of these tokens. However, the value 1 in the first three cells of the vector of these words indicates that it would be worthwile to annotate them in the corpus as postpositions, since they always take the final position in a noun phrase, strictly following a noun without a lexical case.

The odd one out in the table is *szemből* 'opposite from': this word should not be considered a postposition in any way since it does not earn values for the key syntactic properties as it does not occur in such syntactic structures.

A subgroup of case assigning postpositions is also clearly outlined: they are the ones receiving only 0-s. Their common feature is that they can be examined from every aspect, meaning that they do occur in every syntactic structure in which typical postpositions do, however, they behave differently. In their annotation it would be beneficial to stick to their adverb-like character; they should be annotated as adverbs taking an argument which may precede or follow them and may appear further away in the text. It has to be noted that while "pure postpositions" (which are almost the same as the members of "dressed" postpositions) seem to be very similar based on these distributional properties, case assigning (or "naked") postpositions are not: a subgroup of them is the group with the vector 0 0 0 0 0 0, but others appear in the company of other postpositions. Their heterogenity, however, is also mentioned in É. Kiss and Hegedűs (2021), as I mentioned in 5.1.1.

DOI: 10.15774/PPKE.BTK.2021.17

The most interesting elements in Table 5.5 are the ones represented by a vector beginning with 1 0 1 (disregarding the values in the 4-6. cells): these are postpositions that always strictly follow a noun with a lexical case. Nevertheless, they are closer to typical postpositions than adverbs. Their analysis must differ from typical postpositions, and from that of the adverbs as well. The analysis cannot be the same as the one of typical postpositions because of the noun bearing a case suffix; with typical postpositions, the noun is caseless. The adverb type postpositions are also different because they are looking for a complement in the sentence, whereas this kind of postposition does not, their argument always precedes them. Here, this group will be referred to as postpositions taking an argument. To all appearances, they are much closer to typical postpositions in their behaviour than adverbs taking an argument. The problem lies in them requiring a lexical case on the noun. When analysing the connection between a noun and a typical postposition, AnaGramma follows a simple algorithm: when arriving at a noun without a lexical case, it looks one (or two) tokens to the right; if it finds a postposition on the first position after the noun, it immediately connects the two words, and thus they may be the argument of another word together later. In other words, they will be a *supply* fulfilling a *demand*. This is not much different from the connection between a lemma and a case suffix. In the case of adverbs taking an argument, the adverb has a *demand*, which is going to be fulfilled by the *supply* provided by a noun phrase bearing the required case suffix. However, the analysis of the elements with the vector 1 0 1 cannot follow either of the above-mentioned algorithms: the case suffix on the noun preceding the postposition with the vector 1 0 1 indicates the end of a noun phrase, and therefore the algorithm should not check the first token to the right. On the other hand, these elements cannot be handled as adverbs taking an argument, as they do not search for their arguments anywhere in the sentence: it is always immediately before them.

It has to be noted that this category (1 0 1 * * *) almost exlusively comprises postpositions with a clear possessive structure taking a noun with the dative suffix. Two of the three participial postpositions (see É. Kiss and Hegedűs, 2021) can also be described with this vector: *fogva* 'from (x) time' and *nézve* 'regarding'.

A cell in Table 5.5 requires further explanation: *alapján* 'based on' received a value of 0/1 for the feature "dem". This was necessary to indicate that while other postpositions appear in one or the other way with demonstrative pronouns (see examples (56a), (56b)),

DOI: 10.15774/PPKE.BTK.2021.17

alapján 'based on' occurs in both structures in the corpus (63a, 63b). This is the only postposition with this capability.

(63) a. *az alapján az idézet alapján*
that based.on the quote based.on
'based on that quote'

b. *annak az élettani tapasztalatnak alapján*
that.DAT the physiological experience.DAT based.on
'based on that physiological experience'

There is one other group worthy of note: * 0 0 1 0 0.[5] Words here take a case-marked noun which is not always strictly adjacent to them, but they do follow the wh-question (as opposed to the case assigning postpositions with a 0 0 0 0 0 0 vector). Words of this category are *kívül* 'outside', *túlra* 'to beyond', *szembe* 'to opposite to', among others. This group seems to be a transitional class between postpositions that are always behind the noun (1 0 1 * * *) and case assigning, postpositions appearing further away (0 0 0 0 0 0) as if they were in the middle of a process where the postposition gradually departs from its strictly adjacent position on the right side of the noun (or an adverb gradually approaches the noun until it is as closely attached to it as a case suffix – typical postpositions). This is, again, a very promising research question: is there a scale from typical postpositions to adverbs where the above mentioned classes occupy different positions? Is the scale defined by the closeness or attachedness of the postposition with regard to the noun?

Table 5.6 shows tokens often appearing in our corpus queries with postposition-like behaviour, some of which have already been mentioned as postposition-candidates in Ligeti-Nagy (2015). It appears that strict adjacency is a common feature of them; furthermore, almost every one of them appears exclusively in a noun phrase ending position, after the noun. Therefore, they are close relatives of the group of postpositions taking an argument. The difference between the two groups is that these postposition-candidates have a more complex morphological structure; they contain a possessive case marker or an essive case suffix. Their syntactic analysis is not much different from what their detailed

[5] Bálint Sass, one of my opponents drew my attention to this group after applying automatic clustering on my dataset.

DOI: 10.15774/PPKE.BTK.2021.17

Table 5.6. List of postposition-like elements and their feature vectors which emerged in my corpus queries.

postposition	**meaning**	pos	∅	**adj**	**wh**	**dem**	**pers pron**
címén	in the name of	1	0	1	1	0	-
eltérően	differently	0	0	1	1	0	0
fényében	in view of	1	0	1	1	0	-
függően	depending on	1	0	1	1	0	0
hiányában[6]	in default of	1	0	1	1	0	-
idején	in	1	0	1	1	0	-
jegyében	in spirit of	1	0	1	1	0	-
keretében	within the framework of	1	0	1	1	0	-
kezdődően	beginning from	1	0	1	1	0	0
köszönhetően	thanks to	0	0	1	1	0	0
követően	following	1	0	1	1	0	0
megegyezően	same way as	1	0	1	1	0	0
megelőzően	previous to	1	0	1	1	0	0
megfelelően	accordingly	0	0	1	1	0	0
terén	in the field of	1	0	1	1	0	-
ürügyén	under cover of	1	0	1	1	0	-
vonatkozóan	with respect to	1	0	1	1	0	0

morphological analysis would activate, but their meaning and their exclusive occurrence in this typical position of postpositional elements justify their inclusion in the group of postpositions.

Based on the aforementioned results, three major groups of postposition-like elements are outlined. During the parsing process, considering AnaGramma's left-to-right approach within the *supply-and-demand* framework, the following algorithms are thought to be feasible:

- typical postpositions: these words can always be found directly after a noun without a lexical case, that is currently tagged as `NOM` as a default. The essence of their processing is the following: arriving at the noun without a lexical case, looking at the elements in the *window*, the parser sees them, and so, without any further analysis, the two words are linked, and further, they are involved in the syntactic analysis together, similarly to nouns with a lexical case suffix.

  My suggestion is to use POSTP as their POS-tag – as it is already the case with most of the words in this group. The members of this group are the words with a vector beginning with 1 1 1 in Table 5.5.

DOI: 10.15774/PPKE.BTK.2021.17

- for words not at all typical (compared to the typical ones): these words always take a noun with a lexical case as an argument; they can appear in front of the noun as well, and other elements can appear between them. They received a value of 0 in every cell in Table 5.5. My suggestion is to use ADV as their POS-tag, and provide them with a feature to indicate which lexical case they take. This will be their *demand* during the parsing process.
- for postpositions taking a noun with a lexical case: although always occurring strictly after the noun, these words differ from typical postpositions in that they take a noun with a lexical case. In Table 5.5 and Table 5.6 their vectors start with 1 0 1. Their algorithmic analysis requires further study; however, I suppose they can be linked to the noun at a very early stage of the parsing process.

5.3 Summary

In this section a detailed, corpus-driven analysis of the Hungarian postposition-like elements was presented. First I collected, compared and partly unified the diverse categorisation of postpositions in the linguistic literature. Then I examined how six features mentioned in linguistic literature, used for the distinction of the distinct categories of postpositions, characterise these words when studied in a corpus. As can be seen in Table 5.5, the numerous postposition-candidates can be clearly arranged based on these features. I found typical and less typical elements, the algorithmic analysis of which must be different. Three main groups were outlined: prototypical postpositions, postpositions taking a noun with a lexical case but always following the noun, and words always taking a noun with a lexical case and appearing before or after the noun.

There are always new candidates popping up that should be inserted somewhere into the framework and there are some postposition candidates that seem to attract more attention, such as *köszönhetően* 'thanks.to.sg', the 'postpositionality' of which was also mentioned by Pomázi (2018). Though this was not discussed earlier in linguistic papers as a postposition, it did however appear in my searches as something to be analysed as a postposition.

DOI: 10.15774/PPKE.BTK.2021.17

Moreover, *híján* 'in the absence of' has been discussed in a paper of Katalin É. Kiss (2020) after I conducted my study on postpositions – this word's behaviour in the corpus should also be examined later.

In Chapter 4 I applied the results of a word embedding model to classify nouns with a locative case suffix on them. This method could be also used for the categorisation of postpositional phrases: an appropriate question could be formed that these phrases will answer and then it would be clearly stated whether a given postpositional phrase is an appropriate answer to a given question or not. This categorisation is the next step in my research on Hungarian postpositions.

DOI: 10.15774/PPKE.BTK.2021.17

Chapter 6

Conclusion

This thesis focused on some central linguistic phenomena related to noun phrases in Hungarian. What is common in these phenomena is that they proved to be significant in the parsing process of Hungarian texts within the framework of a parser called `AnaGramma`, the aim of which was to model human sentence processing by parsing the text word by word, from left to right (see 1.3). The research presented in the previous sections is not only based on a computational approach, but more importantly, it is corpus-driven.

The following issues were addressed:

- When nothing marks the end of the noun phrase – the cases of "suffixlessness" and their role in the parsing process (Chapter 2). I designed 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 accuracy. 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: Table 2.5 showed that 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%).

DOI: 10.15774/PPKE.BTK.2021.17

  - 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%, Table 2.3). `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.

- A problem from inside of a noun phrase: I highlighted 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'; Chapter 3). I collected similar structures from a syntactically annotated corpus to be able to define some categories among these phrases. Furthermore, I inspected 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 (section 3.4): 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 showed 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 distinguished 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

DOI: 10.15774/PPKE.BTK.2021.17

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.

- Locative case suffixes: categorisation with respect to adverbial roles in a sentence (Chapter 4). In this section I presented an annotation that would be appropriate when designing a teaching corpus for a Question-Answering system. I focused 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 defined 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 sorted 1 097 lemmas into these categories (Appendix C.1).
  - In addition to the default categories of the lemmas, I further specified the behaviour of the lemmas with regard to the nine locative case suffixes: I defined what additional adverbial role the lemma may have with a given suffix (in addition or instead of its default role).
- Postpositions in Hungarian: In this section a detailed, corpus-driven analysis of Hungarian postposition-like elements was presented. I collected, compared and unified the diverse categorisation of postpositions in the linguistic literature. Then I examined 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 systematised the main linguistic sources' opinion on postpositions (Table 5.3).

DOI: 10.15774/PPKE.BTK.2021.17

– I defined 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 proved that *szemből* 'from oppose to', though it is considered a postposition in many papers, is actually not a postposition at all. Its behaviour simply could not be evaluated based on the six features because it does not appear in postpositional places in the corpus.

– I outlined 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 refined 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).

DOI: 10.15774/PPKE.BTK.2021.17

The list of interesting phenomena of NPs in Hungarian, of course, could be further expanded. I mentioned some possible research questions in sections 2.3.5, 4.7 and 5.3. 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.

DOI: 10.15774/PPKE.BTK.2021.17

Appendix A

The Nom-or-What algorithm

A.1 Python implementation of the Nom-or-What algorithm

```
import re


def format_sents(s):
    """ Takes a string (a sentence) and returns it as a list of token/lemma/
    annotation. Also appends end of sentence characters to each sentence.
    """
    endofsent = ['#', '#', '#']
    sents = []
    sent = s.strip().split(' ')
    for word in sent:
        sents.append(word.split('/'))
    sents.append(endofsent)
    sents.append(endofsent)
    return sents


def check_macro(macro_name, anal):
    """
    Checks if a macro is valid for a given token.
    param:
    macro_name: name of the macro (to be looked up in macros.yml)
    anal: token and annotation to be checked
    returns:
    true, if the given macro contains the given annot.
    false, othewise
    """
    macro_res = False
    if macros['macros'][macro_name]['type'] == 'list':
```

```
        if anal[0] in macros['macros'][macro_name]['value']:
            macro_res = True
        else:
            macro_res = False
    elif macros['macros'][macro_name]['type'] == 'regex':
        if macros['macros'][macro_name]['regexp_type'] == 'search':
            macro_res = re.search(macros['macros'][macro_name]['value'], anal[2])
    elif macros['macros'][macro_name]['type'] == 'ends':
        macro_res = anal[2].endswith(macros['macros'][macro_name]['value'])
    elif macros['macros'][macro_name]['type'] == 'complex':
        if macros['macros'][macro_name]['compl_type'] == 'and':
            all_true = True
            for macro in macros['macros'][macro_name]['sub_macros']:
                if not check_macro(macro, anal):
                    all_true = False
                    break
            macro_res = all_true
        elif macros['macros'][macro_name]['compl_type'] == 'or':
            one_true = False
            for macro in macros['macros'][macro_name]['sub_macros']:
                if check_macro(macro, anal):
                    one_true = True
                    break
            macro_res = one_true
    elif macros['macros'][macro_name]['type'] == 'neg':
        macro_res = not check_macro(macros['macros'][macro_name]['sub_macro'], \
        anal)
    return macro_res


def NUM_rules(window, curr_POS):
    """
    Applies the rules of numerals to a given token.
    param:
    window: list of two tuples, containing the first and second token/lemma/
    annotation after the given word.
    curr_POS: the annnotation of the currently analysed token
    return:
    curr_POS, probably changed during the process
    """
    first_right_word, first_right_lemma, first_right_annot = window[0]
    second_right_word, second_right_lemma, second_right_annot = window[1]
    if re.search("FN|SZN|MN|NU", first_right_annot):  # if the next token is
    # a nominal, or a postposition, the word gets a "none"
        curr_POS = curr_POS.replace('NOM', 'none')
    elif check_macro('not_kop_v', window[0]):  # if the next token is verb,
    # but not a copula, the word is a nominative
        curr_POS = curr_POS.replace('NOM', 'nom')
    elif check_macro('def_art', window[0]): # if the next token is a definite
    # article or an adverb, the word is a nominative
        curr_POS = curr_POS.replace('NOM', 'nom')
```

```
    else:
        curr_POS = curr_POS.replace('NOM', 'defnone')
        if check_macro('PUNCT', window[1]) or check_macro('def_art', window[1]) \
        or re.search('HA', second_right_annot) or check_macro('V', window[1]):
            curr_POS = curr_POS.replace('defnone', 'none')

    return curr_POS


def ADJ_rules(window, curr_POS):
    """
    Applies the rules of NP modifiers to a given token.
    param:
    window: list of two tuples, containing the first and second token/lemma/
    annotation after the given word.
    curr_POS: the annnotation of the currently analysed token
    return:
    curr_POS, probably changed during the process
    """
    first_right_word, first_right_lemma, first_right_annot = window[0]
    second_right_word, second_right_lemma, second_right_annot = window[1]
    if check_macro('NUs', window[0]):  # if the next token is postposition(al
    like element), the word gets a 'none'
        curr_POS = curr_POS.replace('NOM', 'none')
    elif check_macro('not_kop_v', window[0]):  # if the next token is a verb,
    # but not a copula, the word # is a nominative
        curr_POS = curr_POS.replace('NOM', 'nom')
    elif check_macro('szn', window[0]):  # before a numeral: default value
        curr_POS = curr_POS.replace('NOM', 'suff')
        if check_macro('PSE', window[1]):  # if the second token has a poss. suff.
            curr_POS = curr_POS.replace('suff', 'gen')
        else:  # otherwise
            curr_POS = curr_POS.replace('suff', 'nom')
    elif check_macro('full_stop', window[0]):  # if the next token is a fix outsider,
    # this word must be the end of an NP,thus 'nom'
        curr_POS = curr_POS.replace('NOM', 'nom')
    else:  # otherwise
        curr_POS = curr_POS.replace('NOM', 'defnone')
        if check_macro('full_stop', window[1]) or check_macro('V', window[1]):
        # if there is a punct.mark in the window
            curr_POS = curr_POS.replace('defnone', 'none')

    return curr_POS


def NOUN_rules(window, curr_POS, token):
    """
    Applies the rules of NP modifiers to a given token.
    param:
    window: list of two tuples, containing the first and second token/lemma/
    annotation after the given word.
```

DOI: 10.15774/PPKE.BTK.2021.17

```
    curr_POS: the annnotation of the currently analysed token
    token: a list containing the current token, lemma, annotation
    return:
    curr_POS, probably changed during the process
    """
    first_right_word, first_right_lemma, first_right_annot = window[0]
    second_right_word, second_right_lemma, second_right_annot = window[1]
    if check_macro('NUs', window[0]):  # if the next token is postposition(al
    # like element), the word gets a 'none'
        curr_POS = curr_POS.replace('NOM', 'none')
    elif check_macro('not_gen', token):
        curr_POS = curr_POS.replace('NOM', 'nom')
    elif check_macro('PSE', window[0]):
        curr_POS = curr_POS.replace('NOM', 'gen')
    elif check_macro('not_kop_v', window[0]) or check_macro('pl', window[0]):
        curr_POS = curr_POS.replace('NOM', 'nom')
    elif check_macro('TULN', window[0]):
        curr_POS = curr_POS.replace('NOM', 'suff')
        if not check_macro('cimunevu', window[1]):
            curr_POS = curr_POS.replace('suff', 'nom')
    elif check_macro('NPMod', window[0]):
        curr_POS = curr_POS.replace('NOM', 'suff')
        if check_macro('PSE', window[1]) and not re.search('PSe3', curr_POS):
            curr_POS = curr_POS.replace('suff', 'gen')
        elif check_macro('V', window[1]) or check_macro('PUNCT', window[1]) \
        or check_macro('relpron', window[1]) or re.search('PSe3', curr_POS):
            curr_POS = curr_POS.replace('suff', 'nom')
    elif check_macro('full_stop', window[0]) or first_right_word == ':' or \
    first_right_word == '!':
        curr_POS = curr_POS.replace('NOM', 'nom')
    else:
        curr_POS = curr_POS.replace('NOM', 'suff')
        if check_macro('V', window[1]) or check_macro('PUNCT', window[1]) or \
        check_macro('relpron', window[1]) or re.search('PSe3', curr_POS) or \
        second_right_word == "#":
            curr_POS = curr_POS.replace('suff', 'nom')

    return curr_POS


def nom_or_what(s, macro):
    """
    Function disambiguating suffixless nominals in a sentence.
    param:
    s: sentence as a string
    macro: macros loaded from a .yml file
    return:
    new_sent: a list of the tokens of the sentence with the novel tags (word/lemma/
    annotation)
    to_write_later: a list that stores tuples which contain the window of each
    suffixless nominal and their new annotation; this is to help creating the
```

```
    annotation file
    """
    global macros
    macros = macro
    sent = format_sents(s)
    new_sent = []  # to store a sentence with the novel tags
    to_write_later = []  # to store the NOM token and the window to create the
    annotation file later
    for i in range(len(sent)):
        token = sent[i]
        curr_word = token[0]
        curr_lemma = token[1]
        curr_POS = token[2]
        if curr_word != "#":  # if the given token is not the end of a sentence
            window = sent[i + 1:i + 3]
            if check_macro('NOM', token):  # if the given token is a suffixless
            # nominal
                if check_macro('Noun_tree', token):  # if the given token is
                # a noun, or a plural adjective, participle
                    curr_POS = NOUN_rules(window, curr_POS, token)
                elif check_macro('Adj_tree', token):  # if the given token is
                # a singular adjective or participle
                    curr_POS = ADJ_rules(window, curr_POS)
                elif check_macro('Num_tree', token):  # if the given token is
                # a numeral
                    curr_POS = NUM_rules(window, curr_POS)
                new_token = curr_word + ' ' + curr_lemma + ' ' + curr_POS
                to_write_later.append((window, new_token))  # a tuple of the
                # window of the given token and the full (novel) annotation
                # of the token
        new_sent.append(curr_word + '/' + curr_lemma + '/' + curr_POS)
    return (new_sent, to_write_later)


def write_to_annot_file(new_sent, to_write, outp, i):
    """
    Function to write the sentences into a file preparing it for the manual
    annotation.
    param:
    new_sent: A sentence as a list of tokens as word/lemma/tag.
    to_write: A list containing tuples of the windows and the annotations of the
    suffixless nominals.
    outp: The output file.
    i: The number of the current sentence.
    """
    outp.writelines(str(i) + '. ' + ' '.join([token.split('/')[0] for token \
    in new_sent]) + '\n')
    [outp.writelines(('-' + nom.split()[0] + ' ' + window[0][0] + ' ' + \
    window[1][0] + '\n', nom + '\n', nom + '\n', nom + '\n')) for(window, nom) \
    in to_write]
    outp.writelines(('\n', '\n'))
```

DOI: 10.15774/PPKE.BTK.2021.17

A.2 Macros used in the `Nom-or-What` algorithm

```
# YAML configuration file for Nom-or-What algorithm

macros:
    nu_mn:
        type: list
        regexp_type: None
        value: [ 'alatti', 'általi', 'elleni', 'előli', 'előtti',
                 'felőli', 'fölötti', 'helyetti', 'iránti',
                 'képesti', 'körüli', 'közötti', 'melletti',
                 'mellőli', 'miatti', 'mögötti', 'nélküli', 'szerinti',
                 'végetti', 'utáni', 'közti' ]
    not_gen:
        type: list
        regexp_type: None
        value: ['mindez', 'Mindez', 'mindaz', 'Mindaz', 'ez', 'Ez',
                 'az', 'Az', 'emez', 'Emez', 'amaz', 'Amaz', 'aki', 'Aki',
                 'ami', 'Ami']
    NPMod:
        type: regex
        regexp_type: search
        value: 'MN|SZN|MIB|MIF|MIA'
    NPMod_PL:
        type: complex
        compl_type: and
        sub_macros: [NPMod, pl]
    FN:
        type: regex
        regexp_type: search
        value: 'FN'
    TULN:
        type: regex
        regexp_type: search
        value: 'TULN'
    DetNM:
        type: regex
        regexp_type: search
        value: 'DET_NM'
    Noun_tree:
        type: complex
        compl_type: or
        sub_macros: [FN, NPMod_PL, TULN, DetNM]
    Adj:
        type: regex
        regexp_type: search
        value: 'MN|MIB|MIF|MIA|OKEP'
    pl:
        type: regex
        regexp_type: search
        value: 'PL'
```

DOI: 10.15774/PPKE.BTK.2021.17

```
not_pl:
    type: neg
    sub_macro: pl
Adj_tree:
    type: complex
    compl_type: and
    sub_macros: [Adj, not_pl]
Num_tree:
    type: regex
    regexp_type: search
    value: 'SZN'
NOM:
    type: ends
    regexp_type: NONE
    value: 'NOM'
IGE:
    type: regex
    regexp_type: search
    value: 'IGE'
PART:
    type: regex
    regexp_type: search
    value: 'MIB|MIA|MIF|OKEP'
not_PART:
    type: neg
    sub_macro: PART
V:
    type: complex
    compl_type: and
    sub_macros: [IGE, not_PART]
van:
    type: lemma
    value: 'van'
kop_v:
    type: complex
    compl_type: and
    sub_macros: [V, van]
not_kop:
    type: neg
    sub_macro: van
not_kop_v:
    type: complex
    compl_type: and
    sub_macros: [V, not_kop]
art:
    type: list
    value: ['a', 'az']
det:
    type: regex
    regexp_type: search
    value: 'DET'
```

DOI: 10.15774/PPKE.BTK.2021.17

```
def_art:
    type: complex
    compl_type: and
    sub_macros: [art, det]
NU:
    type: regex
    regexp_type: search
    value: 'NU'
cimunevu:
    type: list
    value: ['című', 'nevű', 'címen', 'néven', 'címmel', 'névvel']
NUs:
    type: complex
    compl_type: or
    sub_macros: [NU, nu_mn, cimunevu]
szn:
    type: regex
    regexp_type: search
    value: 'SZN'
PSE:
    type: regex
    regexp_type: search
    value: 'PSe'
PUNCT:
    type: regex
    regexp_type: search
    value: 'SPUNCT'
koto:
    type: list
    value: ['is', 'sem', 'nem', 'pedig']
nm:
    type: regex
    regexp_type: search
    value: 'NM'
full_stop:
    type: complex
    compl_type: or
    sub_macros: [koto, PUNCT, def_art, nm]
relpron:
    type: list
    value: ['aki', 'ami', 'hogy', 'de']
casesuff:
    type: regex
    regexp_type: search
    value: 'SUB|INE|SUP|ILL|FAC|ACC|DAT|ILL|ABL|INS|ALL|ELA|DEL|ADE'
```

DOI: 10.15774/PPKE.BTK.2021.17

A.3 Extract of the annotated test corpus

```
1: Magyarországról az elsö térképet Bakócz_Tamás érsek Lázár nevű titkára
készítette , valószínűleg az 1510-es években .
-első térképet Bakócz_Tamás
első első MN.NONE
első első MN.default_NONE
első első MN.default_NONE
-Bakócz_Tamás érsek Lázár
Bakócz_Tamás Bakócz_Tamás TULN.XNE
Bakócz_Tamás Bakócz_Tamás TULN.XNE
Bakócz_Tamás Bakócz_Tamás TULN.SUFF
-érsek Lázár nevű
érsek érsek FN.GEN
érsek érsek FN.SUFF
érsek érsek FN.SUFF
-Lázár nevű titkára
Lázár Lázár TULN.NONE
Lázár Lázár TULN.NONE
Lázár Lázár TULN.NONE
-nevű titkára készítette
nevű nevű MN.NONE
nevű nevű MN.NONE
nevű nevű MN.NONE
-titkára készítette ,
titkára titkár FN.PSe3.NOM
titkára titkár FN.PSe3.NOM
titkára titkár FN.PSe3.NOM
-1510-es években .
1510-es 1510-es MN.NONE
1510-es 1510-es MN.NONE
1510-es 1510-es MN.NONE

2: Ha jót cselekszem.

3: A délutáni napban képzelődhettem .
-délutáni napban képzelődhettem
délutáni délutáni MN.NONE
délutáni délutáni MN.NONE
délutáni délutáni MN.NONE

4: Okait e látszattartalomnak nem elsősorban a tehetetlenségben találnánk meg
, hanem az ideológiai megkötöttségekben , a sajtó függetlenségét és szabadságát
is korlátozó központi vezérlésben .
-ideológiai megkötöttségekben ,
ideológiai ideológiai MN.NONE
ideológiai ideológiai MN.default_NONE
```

DOI: 10.15774/PPKE.BTK.2021.17

```
ideológiai ideológiai MN.default_NONE
-sajtó függetlenségét és
sajtó sajtó FN.GEN
sajtó sajtó FN.GEN
sajtó sajtó FN.GEN
-korlátozó központi vezérlésben
korlátozó korlátoz IGE._OKEP.NONE
korlátozó korlátoz IGE._OKEP.default_NONE
korlátozó korlátoz IGE._OKEP.default_NONE
-központi vezérlésben .
központi központi MN.NONE
központi központi MN.NONE
központi központi MN.NONE

5: Sokszor én is belecsúszok abba a hibába .
-én is belecsúszok
én én FN_NM.e1.NOM
én én FN_NM.e1.NOM
én én FN_NM.e1.NOM

6: Több levelet írt a másik honti jóbarátnak és munkatársnak , Pajor_István
1848_49-es nemzetőrnek , megyei tisztviselőnek , írónak és költőnek is .
-Több levelet írt
Több sok SZN._FOK.NONE
Több sok SZN._FOK.NONE
Több sok SZN._FOK.NONE
-másik honti jóbarátnak
másik másik MN_NM.NONE
másik másik MN_NM.default_NONE
másik másik MN_NM.default_NONE
-honti jóbarátnak és
honti honti MN.NONE
honti honti MN.default_NONE
honti honti MN.default_NONE
-Pajor_István 1848_49-es nemzetőrnek
Pajor_István Pajor_István TULN.XNE
Pajor_István Pajor_István TULN.XNE
Pajor_István Pajor_István TULN.SUFF
-1848_49-es nemzetőrnek ,
1848_49-es 1848_49-es MN.NONE
1848_49-es 1848_49-es MN.default_NONE
1848_49-es 1848_49-es MN.default_NONE
-megyei tisztviselőnek ,
megyei megyei MN.NONE
megyei megyei MN.default_NONE
megyei megyei MN.default_NONE
```

DOI: 10.15774/PPKE.BTK.2021.17

```
7: ( Gördeszkások élesen , suhantak el az Ikrek mellett , karcolva össze a
decens szürke márványlapokat .
-Gördeszkások élesen ,
Gördeszkások gördeszkás FN.PL.NOM
Gördeszkások gördeszkás FN.PL.SUFF
Gördeszkások gördeszkás FN.PL.SUFF
-Ikrek mellett ,
Ikrek iker FN.PL.NONE
Ikrek iker FN.PL.NONE
Ikrek iker FN.PL.NONE
-decens szürke márványlapokat
decens decens MN.NONE
decens decens MN.default_NONE
decens decens MN.default_NONE
-szürke márványlapokat .
szürke szürke MN.NONE
szürke szürke MN.NONE
szürke szürke MN.NONE

8: A patak tőlük keletre húzódott .
-patak tőlük keletre
patak patak FN.NOM
patak patak FN.SUFF
patak patak FN.NOM

9: A " sziget " elnevezés arra utal , .
-sziget " elnevezés
sziget sziget FN.XNE
sziget sziget FN.XNE
sziget sziget FN.SUFF
-elnevezés arra utal
elnevezés elnevezés FN.NOM
elnevezés elnevezés FN.NOM
elnevezés elnevezés FN.NOM

10: Azzal érveltem ,

11: Egy magyar paraszt rudazókötélre akasztva lóg a saját diófáján .
-Egy magyar paraszt
Egy egy SZN.NONE
Egy egy SZN.NONE
Egy egy SZN.NONE
-magyar paraszt rudazókötélre
magyar magyar MN.NONE
magyar magyar MN.default_NONE
magyar magyar MN.default_NONE
```

DOI: 10.15774/PPKE.BTK.2021.17

```
-paraszt rudazókötélre akasztva
paraszt paraszt MN.false_pos
paraszt paraszt MN.false_pos
paraszt paraszt MN.NONE
-saját diófáján .
saját saját MN.NONE
saját saját MN.NONE
saját saját MN.NONE
```

DOI: 10.15774/PPKE.BTK.2021.17

Appendix B

Extended named entities

B.1 List of lemmas of extended named entities in Szeged Treebank 2.0

13 Parsons asszony
13 Varga Mihály pénzügyminiszter
11 Orbán Viktor miniszterelnök
10 Ogilvy elvtárs
10 St. Antonio herceg
8 Fernandez régensherceg
6 Laci atya
6 Pista bácsi
6 St. Antonio főherceg
6 Wirth kapitány
5 Croesus csoport
5 Palmerston tanár
5 Pataki úr
5 Szabó József gyerekhősről
4 Allianz biztosító
4 Anati professzor
4 Dow Jones hírügynökség
4 Fernandez régens
4 Giglioli professzor
4 Járai Zsigmond pénzügyminiszter
4 Java alkalmazások
4 Jean-Marie Messier vezérigazgatót
4 José pincér
4 Martonyi János külügyminiszter
4 Moody's hitelminősítő
4 NTFS állományrendszer
4 Reuters hírügynökség
4 Ring néni
4 Torgyán József miniszter
4 Zoli bácsi
3 Áder János házelnök
3 ADSL technológia
3 Apache webkiszolgáló
3 Aral benzinkutat
3 Borókai Gábor kormányszóvivő
3 Csáky külügyminiszter
3 Családnet program
3 DJIA index
3 Ellesley doktor
3 Gabi bácsi
3 Generali Vienna csoport
3 Gesztenyefa kávéház
3 Hutchins fűtő
3 II. János Pál pápa
3 Interfax hírügynökség
3 Kapitány úr
3 Laci bácsi
3 Marika néni
3 Nasdaq piac
3 Nikkei225 index
3 Oracle adatbázis-kezelő
3 Orbán Viktor kormányfő
3 Patkó bácsi
3 Pentium 4 processzor
3 Pjandzs folyóba
3 Pjandzs határátkelőhelyről
3 Ron Sommer vezérigazgató
3 Rotch Energy Limited céget
3 Sugár András vezérigazgató
3 Viktor Csernomirgyin exkormányfő
3 Vivendi Universal médiacsoport
3 WAP böngésző

DOI: 10.15774/PPKE.BTK.2021.17

3 Zsóka néni
2 Abu Sayyaf terroristacsoportot
2 Ági néni
2 alezredes úr
2 Altera chipgyártó
2 Andrássy gróftól
2 Árkád üzletházban
2 Aspetti della morte címen
2 Avanti cégtől
2 AXA biztosító
2 Balázs Imi bácsinál
2 Barbai úr
2 BB részvénycsomagot
2 Bécs Budapest - Wien Budapest 2000 címmel
2 Bentley ProjectWise rendszerre
2 Berkeley egyetemen
2 BMC szoftverszolgáltató
2 Buda-Cash Signal néven
2 Bükk hegységben
2 Búzás Gergely muzeológustól
2 Cegetel mobilszolgáltató
2 Celeron processzor
2 Charing Cross állomásra
2 Chirac elnök
2 Cisco Expo 2000 konferencián
2 Clinton elnök
2 Cogne cég
2 Computex 2002 kiállításon
2 Credit Suisse csoport
2 Csáky István külügyminiszter
2 Don Raffaello pizzériába
2 Dreina templomba
2 Enterprise kiszolgálóra
2 Ernő bá
2 Erzsi néniékhez
2 Eurohypo néven
2 Fernandez herceg
2 Fernandez néven
2 Fernandez di St. Antonio régensherceggel
2 Fertőboz állomásig
2 Frischmann Gábor elnöktől
2 Gabi szappant
2 Gabi termékcsaládnál
2 Gánt-Kő és Tőzeg Kft. néven
2 Gary Kasparov sakkvilágbajnokot
2 Gemma sorort
2 Géza bácsi
2 Gold Partner minősítést
2 HBW Express takarékpénztártól
2 Hedvig nővér
2 Hertha-Barcelona mérkőzést
2 Heves megyében
2 HLA Global néven
2 Horváth and Partner céget
2 Hűséges Almák néven
2 HVB Croatia néven
2 HVB Hungary Rt. néven
2 Hyundai cég
2 IKB Deutsche Industriebank AG bankkal
2 Intel-PRO/100+ hálózatadapterről
2 Izrael állam
2 Izrael az otthonunk névvel
2 Janesch Péter építésszel
2 Járai Zsigmond MNB-elnök
2 Jelcin elnök
2 Józsi bácsi
2 Jukosz olajtársaság
2 Jurij Szkuratov főügyészt
2 Kannibál Béby néven
2 Kassa-Nagyszalánc-Velejte vasútvonalat
2 Kata nénihez
2 Kerepesi temetőben
2 KfW bankkal
2 Kien Giang tartomány
2 Kőbánya-felső állomásra
2 Kodaj Károly rendszermérnöktől
2 Kolláth György vezérigazgató
2 Komintern szó
2 Kovács Attila elnök
2 Krőzsel Károly ezredes
2 László Csaba pénzügyminiszter
2 Lemberg megyében
2 Lotus LearningSpace szoftverrel
2 Macintosh géphez
2 Majláth országbírót
2 Massachusetts államban
2 MasterCard kártyával
2 Mazlum-Der alapítványnak
2 Mercedes-Benz gépkocsi
2 Michele nevet
2 Mihail Kaszjanov pénzügyminiszter
2 Múzeumok majálisa címmel
2 Napkelet királyfi
2 Nasdaq Composite index
2 Nasdaq Europe néven

DOI: 10.15774/PPKE.BTK.2021.17

2 Netanjahu miniszterelnök
2 Panasonic noteszgépet
2 Pec városban
2 Pentium processzor
2 Perner Ferenc professzorral
2 Piquemal tábornok
2 PlayStation 2 játékkonzolból
2 Pólus bevásárlóközpontokat
2 Pong néven
2 Private Storage Utility néven
2 Professional Services Partner fokozattal
2 QoS szolgáltatásokat
2 Radzeer cirkálót
2 Region Kontakt néven
2 Reichardt család
2 Rolling Stones hivatalos pornográfusa címmel
2 Rupert Murdoch médiamágnást
2 SAP szoftvergyártó
2 SAP szoftverház
2 Seattle Coffee kávébárláncot
2 Shell kútnál
2 SITA hírügynökség
2 Staatsdruckereit és a Dorotheum aukciósházat
2 Standard and Poor's hitelminősítő
2 Swiss Re viszontbiztosítónál
2 Számalk Kereskedőház Rt. néven
2 Száraz Laci atyához
2 Szent Mihály arkangyal
2 Teleki Pál miniszterelnök
2 Tesco áruházban
2 Theodore Roosevelt repülőgép-hordozót
2 Trastevere negyedben
2 Trebitsch úron
2 Tüske Borika nénivel
2 TV Sziget néven
2 Unicredito Italiano bankházzal
2 V-fon névvel
2 Vastag Sapka címmel
2 Vég Tibor ügyvéd
2 Viktor Juscsenko kormányfő
2 Visa Classic hitelkártya
2 Volvo szakműhelyt
2 Vörös Félhold szervezetnek
2 Warins kalózt
2 Warins Bob néven
2 Wright fivérek
2 XML formátumot
2 Zaporozsje megyeszékhelyen
2 Zetor traktorgyárat
1 A osztály
1 Abbey National bankot
1 Abbott Laboratories gyógyszergyártó
1 ABN Amro bank
1 ABN Amro pénzintézet
1 ABN Amro Holding pénzintézet
1 Acesa fizetőút-üzemeltető
1 Aczél Gábor igazgató
1 Ádám József klubigazgató
1 Adidas póló
1 Adobe felirat
1 ADSL internetkapcsolat
1 ADSL kapcsolat
1 ADSL modem
1 Advair asztmagyógyszer
1 Aegon biztosító
1 Aeroflot légitársaság
1 Agnelli család
1 Agrana konszern
1 Aigner Szilárd meteorológus
1 Airbus A380 repülőgépek
1 Albert Sabin virológus
1 Alfa projekt
1 Allegra antiallergén
1 Allianz biztosítótársaság
1 Allianz csoport
1 Almira székestrónváros
1 Alvarec elnök
1 Alvarez elnök
1 Amadinda együttessel
1 Ambien altatótabletta
1 Anatolij Kinah kormányfő
1 Anett néni
1 Antenna Hungária csoport
1 Anti Blocking System néven
1 António Felizardo vezérigazgató-helyettestől
1 AOLTV szolgáltatás
1 AP hírügynökség
1 APA hírügynökség
1 Apache webszerver
1 Apolló űrhajó
1 Appliance Components Companies részvénytársaságnak
1 Aquila közszolgáltató

DOI: 10.15774/PPKE.BTK.2021.17

1 Arató Zsolt sajtófőnök
1 ArchLine szoftverről
1 Árkád bevásárlóközpont
1 Arkan kapitányt
1 Art-Show stúdió
1 Athenaeum nyomda
1 ATI kártya
1 ATM/Frame Relay gerinchálózat
1 Auchan áruház
1 Audi modellt
1 Audi AG luxusautó-gyártó
1 Axa biztosítótársaság
1 AXA Colonia biztosító
1 BA légitársaság
1 Baath párt
1 Bács-Kiskun megyében
1 Bálint Sándor őrnagy
1 Bankgesellschaft pénzintézet
1 Baross tér 13. szám
1 Bartha András elnök-vezérigazgató
1 Bathó Ferenc főcsoportfőnök
1 BBL-ING bank
1 Bea néni
1 Beatles együttes
1 Beck sörgyár
1 Beck György vezérigazgató
1 Bedő Erik vezérigazgató
1 Bel Paese sajtot
1 Bellus néni
1 Benjamin Netanjahu miniszterelnök
1 Benkő Gábor képviselő
1 Béres csoport
1 Bergbauholding ÖBAG bányavállalkozás
1 Bernáth János polgármester
1 Bertelsmann cég
1 Bertelsmann család
1 Bertelsmann médiacsoport
1 BETA hírügynökség
1 Bill Clinton elnök
1 Bloomberg hírügynökség
1 Bloomsbury negyedben
1 BMW cég
1 BMW márka
1 Boeing utasszállítót
1 Bokros pénzügyminiszter
1 Borisz Jelcin elnök
1 Borókai László ügyvéd
1 BorsodChem közgyűlés
1 Bory Jenő szobrászművész
1 Böszörményi László igazgató
1 Bozóky Imre MLSZ-elnök
1 Bozsik névtől
1 Braun Róbert vezérigazgató-helyettes
1 Bricostore barkácsáruház
1 Bristol-Myers Squibb gyógyszergyár
1 British Nuclear Fuels Plc. társasággal
1 Brown úr
1 Bruno Bulthé vizsgálóbíró
1 BSkyB médiavállalkozás
1 Buda Béla pszichiáter
1 Byrd kapitánnyal
1 CandA áruházlánc
1 Candy márka
1 Carlos Ghosn vezérigazgató
1 Carly Fiorina elnök-vezérigazgató
1 Carrefour hipermarketlánc
1 Center pálya
1 CEZ áramszolgáltató
1 Charles Schwab brókercég
1 Christian Noyer alelnök
1 Chuck Yeager pilóta
1 Ciano külügyminiszter
1 Cirrus Maestro kártyát
1 Cisco eszközök
1 Cisco rendszer
1 CiscoWorks 2000 hálózatmenedzsment
1 Citroën személyautók
1 Claas betakarítógépek
1 Clearstream Banking AG letétkezelő
1 Coca-Cola üveg
1 Com-Ware kiadó
1 Commodore 64 számítógép
1 Cora hipermarket
1 Corvin-Szigony beruházás
1 Credit Suisse bankcsoport
1 Croesus név
1 Crown farmernadrág
1 Családnet PC-program
1 Csányi Klára főosztályvezető
1 Cseh-Szombathy László családszociológus
1 Csisztu Zsuzsa sportriporter
1 Csonka Dezső rendszermérnök
1 CTK hírügynökség
1 Czikó Gábor elnök-vezérigazgató
1 D csoport
1 D meghajtó

DOI: 10.15774/PPKE.BTK.2021.17

1 Daewoo FSO autógyár
1 Daewoo Securities értékforgalmazó
1 Daewoo-FSO autógyár
1 DaimlerChrysler autógyár
1 Damjanich zászlóalj
1 Daniel T. arap Moi elnök
1 Datamedia közvélemény-kutató
1 David Wilby tábornok
1 DAX részvényindex
1 Deák László elnök-vezérigazgató
1 Deceban Traian Remes pénzügyminiszter
1 Delaware államban
1 Delhaize szupermarketlánc
1 Dell gépet
1 Dell számítógépgyártó
1 Dentsu reklámcég
1 Dessewffy utcák
1 Dicső Anikó riporter
1 DJIA részvényindex
1 Dobóczky András vezérigazgató
1 Dobson Tibor szóvivő
1 Dow index
1 Dubcek család
1 Duna Plaza bevásárlóközpont
1 Edgardo Boeninger szenátor
1 Edge név
1 Edward Rydz-Smygly marsall
1 El Nino hurrikán
1 Eli Lilly gyógyszergyártó
1 Emil atya
1 Enron energia-nagykereskedő
1 Eravis részvények
1 Erdész András elnök-vezérigazgató
1 Eros isten
1 Erzsébet tér
1 Euronext tőzsdevállalkozás
1 Ewering szenátorné
1 Excel 2000 kurzus
1 Fájlok törlése parancsgomb
1 Family sörözőbe
1 Fanta üdítőt
1 FAT állományrendszerrel
1 Fathom Technology társaságba
1 Fed elnök
1 Fehér Márta filozófus
1 Fejér megyében
1 Fejes Pista bácsi
1 Fekete Lajos tanár
1 Feldmájer Péter ügyvéd
1 Feleséged őnagysága
1 Felkészítő Napok rendezvénysorozat
1 Ferenc császár
1 Ferenczy kiadó
1 Fernandez bácsi
1 FHB jelzáloglevél
1 FIAT csoport
1 Finnair légitársaság
1 Fitch hitelminősítő
1 FlatStack technológia
1 FlatStack vezérlőmodul
1 Fodor Lajos vezérezredes
1 Fortune magazin
1 Frei elnök
1 Fritz Schaudinn zoológus
1 FTSK Excelsior SE hegymászó
1 Fuji Heavy Industries csoport
1 Fülöp Ferenc klubigazgató
1 Fundamenta lakás-takarékpénztár
1 Furmann Imre ügyvéd
1 Futó Iván főszerkesztő
1 Gabriella novíciát
1 Gabriella soror
1 Gaidosch Tamás menedzser
1 Garadnai Róbert adattárház-menedzsertől
1 Garamvölgyi László ezredes
1 Gaskó úr
1 Gdansk város
1 Gémesi István dandártábornok
1 Gente hetilap
1 Georges Mathé onkológus
1 Gerard Mortier igazgató
1 Gerhard Domagk patológus
1 Gerhardt Ferenc vezérigazgató
1 Globus közgyűlés
1 Gold Record kiadó
1 Gombperec úr
1 Gomperez hídlakót
1 Göncz Árpád államfő
1 Gondvána szuperkontinens
1 Gordon Brown pénzügyminiszter
1 GPRS technológiára
1 Greenwood tanár
1 GSM hálózat
1 GSM/DCS infrastruktúra
1 GTS Central Europe cégcsoport
1 Guardian napilap

DOI: 10.15774/PPKE.BTK.2021.17

1 Gy. Zs. monogram
1 Győri Gábor rendszermérnök
1 Gyula bácsi
1 H. János poggyászkocsi-kezelő
1 Halász Sándor vezérigazgató-helyettes
1 Halbauer János ügyvezető
1 Halmos Gábor vezérigazgató
1 Hans Eichel pénzügyminiszter
1 Harkányi Gábor projektvezető
1 Harvard Mark I. számítógép
1 Head sportszergyártó
1 Hedvig soror
1 Heinz Sundt TA-vezérigazgató
1 Helga Rabl-Stadler asszonnyal
1 Hellen Omwando elemző
1 Helsingin Sanomat napilap
1 Hernádi Miklós szociológus
1 Heusch kapitányt
1 Hezbollah szervezet
1 Hilton szálloda
1 Homoktövis lakópark
1 Hónig Péter vezérigazgatót
1 Horn Gyula miniszterelnök
1 Horthy kormányzó
1 Horváth Gergely vezérigazgató
1 Horváth Iván ügyvéd
1 Horváth Mária kötvényes
1 Hrisztyenko miniszterelnök-helyettes
1 HTC-Jorgosz botrány
1 HTC-Jorgosz cégcsoport
1 HTML nyelvet
1 Hyundai konszern
1 I betűvel
1 IBM processzor
1 IIS webszerverek
1 Ikarus buszok
1 Illés Antal ügyvezető
1 Illés Zoltán Fidesz-alelnök
1 INA hírügynökség
1 Infineon chipgyártó
1 ING bankcsoport
1 ING Groep NV bankcsoport
1 Inktomi cég
1 Intel cégnek
1 Intel félvezetőgyártó
1 Intel technológia
1 Intel Pentium III processzor
1 Intel Pentium-III processzor
1 IntelliEye technológia
1 InterCity szerelvények
1 Interseas Editions kiadó
1 Investicna a Rozvojova Banka pénzintézetre
1 Investkredit csoport
1 IPX/SPX protokollrendszer
1 IRB bank
1 IRB hitelintézet
1 IRB pénzintézet
1 Iris karakterfelismerő
1 Istar processzorok
1 Isten szót
1 IT szektor
1 Itar-Tassz hírügynökség
1 Iza néni
1 Jackson tábornok
1 Jaki szerepjáték
1 Jan Kavan külügyminiszter
1 Jancsó Péter Graboplast-vezér
1 Janne Haaland Matlary szociológus
1 János bácsi
1 János Pál pápa
1 Jászai Pál történész
1 Játékvilág Magyarországon címmel
1 Java futtatókörnyezet
1 Java technológiákkal
1 Javier Solana NATO-főtitkár
1 Jean Lamierre elnök
1 Jean Lemierre bankelnök
1 Jean-Pierre Chevenement belügyminiszter
1 Jelcin család
1 Jelena Kondakova űrhajósnő
1 Jervelino miniszter
1 Jil Sander divatmárkák
1 Johannes Dietz ÖIAG-igazgató
1 John Sidgmore elnök-vezérigazgató
1 Jolantha soror
1 Jorge Faliba atya
1 Joshua Hallford elemző
1 Josihiro Nakama professzor
1 Juci nénihez
1 Julianus barát
1 Jupiter Media Metrix kutatóintézet
1 Jürgen Schrempp vezérigazgató
1 Jurij Balujevszkij vezérezredes
1 Jurij Lvov pénzügyminiszter-helyettes
1 Jutka néni

DOI: 10.15774/PPKE.BTK.2021.17

1 K+F pénzalapok
1 Kádár János Miklós festőművész
1 Karl Landsteiner patológus
1 Karl-Heinz Grasser pénzügyminiszter
1 KBC Bancassurance bankcsoport
1 Kerekes László ügyvezető
1 KFKI cégcsoporton
1 KFKI csoport
1 Királyi városok programban
1 Kirch csoport
1 Kisbán Eszter táplálkozáskutató
1 Klapka György lövészdandár
1 KLM légitársaság
1 Kocsi Tibor újságírót
1 Kodak RFS 3600 filmszkenner
1 Kodak RFS3600 filmszkenner
1 Kofi Annan ENSZ-főtitkár
1 Kökényesi Antal ezredes
1 Kolompos együttes
1 Komáromi Endre ezredes
1 Komercní Banka pénzintézetet
1 Kommerszant-Vlaszty hetilap
1 KOÓS ZOLTÁN ügyvéd
1 Kossuth LK sportlövész
1 Kőszeghy Béla projektmenedzser
1 Kovács vívócsalád
1 Kovács Gyula építőmérnök
1 Krieger Erzsébet kötvényes
1 Krisán Attila ezredes
1 Kukes kisváros
1 Laci bá
1 LaciKomercní Banka pénzintézetet
1 Lafarge cég
1 Lakatos Antal elnök
1 Lauda Air légitársaságot
1 Lebegy tábornok
1 Legeza Péter vezérigazgató-helyettes
1 Lexmark International nyomtatógyártó
1 Libahara faluból
1 Lindsay Owen-Jones elnök-vezérigazgató
1 Linux rendszerek
1 Lion diszkóba
1 Lionel Jospin kormányfő
1 Lloyds pénzintézetnek
1 Lóczy Lajos geológus
1 Lőkösháza megálló
1 London város
1 Lovas fivérek
1 Lowell Edwards elektromérnök
1 Lucent Technologies távközlésieszközgyártó
1 Lufthansa repülőtársaságot
1 Lukoil olajvállalat
1 Lurgi Lentjes Bischoff cég
1 MA 2000 csomag
1 Mabetex cég
1 Machintosh számítógépek
1 MagicGate kapu
1 Magyar Bálint pártelnök
1 Magyar Kódex könyvsorozat
1 Marek Belka pénzügyminiszter
1 Marián Béla szociológus
1 Markovszky vezérigazgatót
1 Marriott International Inc. szállodalánc
1 Marton zenekar
1 MasterCard jutalom-hitelkártya
1 Mastercard International hitelkártyacég
1 MasterCard International hitelkártyakibocsátó
1 Mazeikiu Nafta olajfinomító
1 MCI vállalat
1 Media Convergence Server sorozat
1 Mellwill százados
1 Menahem Begin miniszterelnök
1 Menatep bank
1 Merck and Co. gyógyszeróriás
1 Merkúr bolygó
1 Merrill Lynch bankház
1 Meta Group Inc. e-business
1 Metro hírújság
1 Metrotech irodaház
1 Michael Frenzel vezérigazgató
1 Microsoft termékek
1 Microsoft Office alkönyvtár
1 Mike Jackson tábornok
1 Milka csokoládé
1 Mir űrállomásról
1 Mirabilis cégtől
1 Miska bácsi
1 Missura Gábor MNB-szóvivő
1 Misura Gábor közgazdász
1 MKB Visa kártyával
1 Molnár Miklós történész
1 Moneda kormánypalota
1 Morgan Stanley brókerház
1 Mycal áruházlánc

DOI: 10.15774/PPKE.BTK.2021.17

1 N betűvel
1 Nádasy Zoltán vezérigazgató-helyettes
1 Nagy Gergely üzletág-igazgató
1 Nagy Imre polgármester
1 Nagy Tibor dandártábornok
1 Navaz Sarif miniszterelnöknek
1 Négyesi János üzletág-igazgató
1 Németh família
1 Németh Miklós miniszterelnök
1 Nemzeti parkok menüpont
1 Netfinity kiszolgálócsalád
1 Netfinity Director szoftver
1 NetIQ cég
1 Netscape Enterprise webszerver
1 New Europe Exchange tőzsdén
1 New Holland cég
1 New York város
1 Nexus reklámügynökség
1 Nicholas Scheele elnök
1 Nissan autógyár
1 Noël Forgeard vezérigazgató
1 Nokia cég
1 Northstar processzorok
1 Norton család
1 Novotel-Palace szálloda
1 NTFS 5 állományrendszer
1 NTT DoCoMo mobilszolgáltató
1 NTV tévétársaság
1 Obrandza falvakból
1 Öcs-Pula útvonalon
1 Old Trafford stadion
1 Oldsmobile cég
1 OM AB tőzsdeüzemeltető
1 On-line szó
1 Opel autógyár
1 Opel rendőrautó
1 Oracle termékek
1 Orange mobilszolgáltató
1 Orchidea hálózat
1 OSI modell
1 Oskar Lafontaine pártelnök
1 őzse Lászlóné jegyző
1 Palm gépek
1 Palm V készülékek
1 Pánczél Barabás kézilabdaedző
1 Pánczél Barabás tréner
1 Pantel-GTS konzorcium
1 PAP hírügynökség
1 Pap László professzor
1 Pap Lászlóné tanárnő
1 Papelaco cég
1 Parker tengernagy
1 Parsons gyerekek
1 Pause gomb
1 PBX központokra
1 Pearson kiadócsoport
1 Penta csoport
1 Penthouse szexmagazin
1 Pentium chip
1 Pentium III processzor
1 pénztáros néni
1 Perner professzor
1 Pernod-Ricard SA italcég
1 Peters fűtőmester
1 Peugeot autógyár
1 Peugeot Assistance segélyszolgálathoz
1 Piquemal hadseregtábornok
1 Piturca kapitány
1 PKN Orlen olajvállalatról
1 Pollino miniszter
1 Pollino tűzoltó
1 PP modul
1 Preininger Ambrus altábornagy
1 ProcterandGamble óriáscég
1 Próféta gúnynéven
1 Progress adatbázis-kezelő
1 Prozac antidepresszáns
1 Prudential biztosító
1 Putyin elnök
1 QMS Qcolor színkorrekció
1 R. László mozdonyvezető
1 Rác fürdő
1 Raiffeisen Zentralbank csoport
1 Raiffeisen Zentralbank konszern
1 Rajfu Szupeszu szekta
1 Rambus memóriát
1 Ravasz István hadtörténész
1 Redilco cég
1 Regina Mundi monostor
1 Rejtő E. Tibor vezérigazgató
1 Remington Rand cég
1 Renault autógyár
1 Répássy Róbert frakcióvezető-helyettes
1 Ribbentrop-Csáky találkozón
1 Richard Ellesley doktor
1 Richard Schenz OMV-vezér

DOI: 10.15774/PPKE.BTK.2021.17

1 Riverbank név
1 Röber edző
1 Robert White sebész
1 Robin kakas
1 Robin segélyhajó
1 Roger testvér
1 Roper Starch Worldwide kutatóintézet
1 Royal Dutch Shell csoport
1 RTL csoport
1 Rudolf Virchow patológus
1 Ryanair légitársaság
1 RZB konszern
1 S. József rendőr
1 SAM 46 hangszóróp ár
1 Sampras-Agassi összecsapás
1 Samsung monitor
1 Sándor pápa
1 Sant Ubaldo kolostorból
1 SANWorks menedzsmentszoftver
1 SAP szoftveróriás
1 SAPARD iroda
1 Sátoraljaújhely-Széphalom útra
1 Schering-Plough gyógyszergyártó
1 Schering-Plough Corp. gyógyszergyártó
1 Sealand hercegség
1 Seattle Filmworks cég
1 Secure Digital Music Initiative technológia
1 Shell töltőállomáson
1 Shell Smart program
1 Simon tanár
1 Sinclair ZX80 gép
1 Sipos István vezérigazgató
1 Skodák százados
1 Slovnaft anyavállalat
1 Smith öregfiú
1 Société Générale bank
1 Société Générale nagybank
1 Solaris platform
1 Sony HDD-magnók
1 Sony Life Insurance Co. biztosítótársaság
1 Sopron-Szombathely vonalszakasz
1 Sor Libchavy gyár
1 Sound Blaster Pro hangkártya
1 St. Antonio hercegnő
1 St. Antonio de Vicenzo Y Galapagos főherceg
1 Standard Edition szoftverek
1 Start gomb
1 Start menü
1 Sterbinszky Amália királynőig
1 Stet Hellas mobilszolgáltató
1 Steyr céget
1 StorageWorks háttértár-alrendszer
1 Strasszer Tibor ügyvéd
1 Stumpf István miniszter
1 Surányi jegybankelnök
1 Szabó miniszter
1 Szabó Ferenc ezredes
1 Szabó Lajos hivatalvezető
1 Szacsvay László színművész
1 Szarvas Ferenc vezérigazgató
1 Százados úr
1 Százszorszép együttes
1 Szelényi professzor
1 Szent Anna kápolna
1 Sziget fesztivál
1 Szilassy Nelly zongoraművész
1 Szinán pasa
1 Szingapúr folyó
1 Szkander bég
1 Szlobodan Milosevics elnök
1 Sztálin elvtárs
1 T-Mobile International mobilszolgáltatónak
1 Takács Albert alkotmányjogász
1 Takács Károly vezérigazgató
1 Tamás Gáspár Miklós filozófus
1 TAPI interfész
1 Tatár György talmudista
1 Tázló József rendszermérnök
1 Teleki miniszterelnök
1 Tempo üzlethálózat
1 Tengernagy úr
1 Tesco-Global áruházlánc
1 Tessin kanton
1 Thatcher asszony
1 Thomas Klestil államfő
1 Thomas Lynch elnökhelyettes
1 Thomson Multimedia csoport
1 Thrust SSC rakétaautó
1 Thuega AG szolgáltatócégben
1 Tillotson elvtárs
1 Time Warner részleg
1 Timi néni
1 Tiran vállalat

DOI: 10.15774/PPKE.BTK.2021.17

1 Tiso elnök
1 Tiszt elvtárs
1 Tito marsall
1 TLD webszerverek
1 Todt szervezet
1 Tokaj hegység
1 Tony Blair miniszterelnök
1 Torgyán házaspár
1 Törőcsik Jenő elnök
1 Trabant gépkocsik
1 Transport Layer Security technológia
1 Tredi céggel
1 Trigon bank
1 Tristan Tzara költő
1 TRW autóalkatrész-gyártóra
1 Tui név
1 UBS bankcsoport
1 Ulpius nagyapa
1 Ulrich Hartmann vezérigazgató
1 UNC formátumban
1 Unicode szabvány
1 Unicredito bank
1 UniCredito Italiano bankcsoport
1 Universal stúdió
1 US Airways Inc. légitársaság
1 Václav Havel államfő
1 Vadsu őfelségét
1 Vagit Alekperov Lukoil-elnök
1 Vaio SR17 noteszgéppel
1 Vajda Mihály filozófus
1 Váradi József vezérigazgató
1 Varga úr
1 Vasszosz Lisszaridesz pártelnök
1 Vaszil Rohovij miniszterelnök-helyettes
1 VDSL frekvenciaspektrum
1 Velentzei Vladimir gyárigazgató
1 Vénusz napraforgó-étolaj
1 Vénusz napraforgóolaj
1 Viktor Vekselberg vezérigazgatóra
1 Visa kártya
1 Visa kártyatársaság
1 Visa International kártyák
1 Visa Smart Partner program
1 Vitai Attila vezérigazgató
1 Vivendi cégcsoport
1 Vivendi csoport
1 Vivendi Telecom Hungary csoport
1 Vivendi Universal médiaóriás
1 Vivendi Universal SA médiacsoport
1 Vizi professzor
1 Vízi E. Szilveszter akadémikus
1 Víziszony története címet
1 Vlagyimir Putyin elnök
1 VNU kiadóvállalat
1 VoIP technológiákon
1 Volksbanken csoport
1 Voxline név
1 VUB bank
1 Wellcome Trust magánalapítvány
1 Wienerberger építőanyag-gyártó
1 Wieslaw Kaczmarek tárcavezető
1 Wim Duisenberg bankelnök
1 Windows 2000 kiszolgáló
1 Windows 2000 kiszolgálócsalád
1 Wintach teszt
1 Winterthur biztosítótársaság
1 Wired Equivalent Privacy titkosítással
1 Wirth parancsnokot
1 WMS technológiát
1 WorldCom márkanéven
1 Wright testvérek
1 Wright testvérpár
1 Yellow Pages szolgáltatás
1 You And I együttes
1 Zacher Gábor főorvos
1 Zagrosz hegység
1 ZIL teherautó
1 Zoli bá
1 Zoran Zsivkovics polgármester
1 Zsitkovszky Sándorné rendőr
1 Zsuzsa néni
1 Zycie napilapnak

DOI: 10.15774/PPKE.BTK.2021.17

B.2 List of the endings of extended named entities in Szeged Treebank 2.0

45 néven
32 vezérigazgató
30 bácsi
29 cég
27 csoport
27 pénzügyminiszter
26 néni
25 elnök
22 miniszterelnök
21 hírügynökség
19 úr
15 asszony
15 professzor
13 bank
13 elvtárs
13 processzor
12 címen
12 herceg
12 technológia
11 külügyminiszter
10 atya
10 kapitány
10 régensherceg
9 biztosító
9 index
8 miniszter
8 pénzintézet
8 tanár
8 ügyvéd
7 autógyár
7 család
7 elnök-vezérigazgató
7 ezredes
7 főherceg
7 hitelminősítő
7 kártya
7 kormányfő
7 légitársaság
6 állományrendszer
6 állomás
6 bankcsoport
6 megyében
6 mobilszolgáltató
6 program
6 tábornok
5 állam
5 együttes
5 gép
5 gyerekhősről
5 médiacsoport
5 pápa
5 rendszermérnök
5 soror
5 szoftver
5 város
5 vezérigazgató-helyettes
4 adatbázis-kezelő
4 alkalmazások
4 bá
4 bevásárlóközpont
4 biztosítótársaság
4 cégcsoport
4 doktor
4 folyó
4 gyógyszergyártó
4 hegység
4 kiadó
4 konszern
4 pincér
4 régens
4 rendszer
4 szervezet
4 szolgáltatás
4 webszerver
3 államfő
3 áruház
3 áruházlánc
3 bankház
3 benzinkutat
3 böngészőnek
3 chipgyártó
3 exkormányfő
3 filozófus
3 fivérek
3 formátumban
3 fűtő
3 gépkocsi
3 határátkelőhelyről
3 házelnök
3 igazgató
3 kávéház
3 kiszolgáló
3 kormányszóvivő
3 napilap
3 negyedben
3 noteszgépet
3 pártelnök
3 patológus
3 piac
3 polgármester
3 számítógép
3 szociológus
3 ügyvezető
3 webkiszolgáló
2 szó
2 társaság
2 alapítványnak
2 arkangyal
2 aukciósházat
2 bankelnök
2 betűvel
2 cirkálót
2 dandártábornok
2 egyetemen
2 elemző
2 építésszel
2 filmszkenner
2 fokozat
2 főügyészt
2 gomb
2 gróftól
2 hálózat
2 hálózatadapterről
2 hetilap
2 hitelkártya
2 játékkonzolból
2 kalózt
2 kápolna
2 kávébárláncot
2 kiállításon
2 királyfi
2 kiszolgálócsalád

DOI: 10.15774/PPKE.BTK.2021.17

2 klubigazgató
2 konferencián
2 kötvényes
2 közgyűlés
2 kutatóintézet
2 kútnál
2 márka
2 marsall
2 médiamágnást
2 megyeszékhelyen
2 mérkőzést
2 miniszterelnök-helyettes
2 minősítést
2 MNB-elnök
2 modell
2 muzeológustól
2 nővér
2 olajtársaság
2 olajvállalat
2 országbírót
2 pizzériába
2 rendőr
2 repülőgép-hordozót
2 részvénycsomagot
2 részvényindex
2 sakkvilágbajnokot
2 stúdió
2 szakműhelyt
2 szálloda
2 szappant
2 százados
2 szoftvergyártó
2 szoftverház
2 szoftverszolgáltató
2 szót
2 takarékpénztártól
2 tartomány
2 temetőben
2 templomba
2 termékcsaládnál
2 termékek
2 terroristacsoportot
2 testvér
2 történész
2 traktorgyárat
2 üzletág-igazgató
2 üzletházban
2 vállalat
2 vasútvonalat
2 vezérezredes
2 viszontbiztosítónál
1 őfelsége
1 őnagysága
1 adattárház-menedzsertől
1 akadémikus
1 alelnök
1 alkönyvtár
1 alkotmányjogász
1 altábornagy
1 altatótabletta
1 antiallergén
1 antidepresszáns
1 anyavállalat
1 áramszolgáltató
1 asztmagyógyszer
1 autóalkatrész-gyártóra
1 bányavállalkozás
1 barát
1 barkácsáruház
1 bég
1 belügyminiszter
1 beruházás
1 betakarítógépek
1 bolygó
1 botrány
1 brókercég
1 brókerház
1 buszok
1 chip
1 családszociológus
1 császár
1 csokoládé
1 csomag
1 diszkóba
1 divatmárkák
1 e-business
1 edző
1 elektromérnök
1 elnökhelyettes
1 energia-nagykereskedő
1 ENSZ-főtitkár
1 építőanyag-gyártó
1 építőmérnök
1 értékforgalmazó
1 eszközök
1 falu
1 család
1 farmernadrág
1 felirat
1 félvezetőgyártó
1 festőművész
1 fesztivál
1 Fidesz-alelnök
1 fizetőút-üzemeltető
1 főcsoportfőnök
1 főorvos
1 főosztályvezető
1 főszerkesztő
1 frakcióvezető-helyettes
1 frekvenciaspektrum
1 fürdő
1 fűtőmester
1 futtatókörnyezet
1 geológus
1 gerinchálózat
1 Graboplast-vezér
1 gúnynéven
1 gyár
1 gyárigazgató
1 gyerekek
1 gyógyszergyár
1 gyógyszeróriás
1 hadseregtábornok
1 hadtörténész
1 hálózatmenedzsment
1 hangkártya
1 hangszórópár
1 háttértár-alrendszer
1 házaspár
1 HDD-magnók
1 hegymászó
1 hercegnő
1 hercegség
1 hídlakót
1 hipermarket
1 hipermarketlánc
1 hírújság
1 hitelintézet
1 hitelkártya-kibocsátó
1 hitelkártyacég
1 hivatalvezető
1 hurrikán
1 infrastruktúra
1 interfész

DOI: 10.15774/PPKE.BTK.2021.17

1 internetkapcsolat
1 iroda
1 irodaház
1 isten
1 italcég
1 jegybankelnök
1 jegyző
1 jelzáloglevél
1 jutalom-hitelkártya
1 kakas
1 kanton
1 kapcsolat
1 kapu
1 karakterfelismerő
1 kártyatársaság
1 képviselő
1 készülékek
1 kézilabdaedző
1 kiadócsoport
1 kiadóvállalat
1 királynőig
1 kisváros
1 kolostorból
1 költő
1 könyvsorozat
1 konzorcium
1 kormánypalota
1 kormányzó
1 közgazdász
1 központokra
1 közszolgáltató
1 közvélemény-kutató
1 kurzus
1 lakás-takarékpénztár
1 lakópark
1 letétkezelő
1 lövészdandár
1 Lukoil-elnök
1 luxusautó-gyártó
1 magánalapítvány
1 magazin
1 márkanéven
1 médiaóriás
1 médiavállalkozás
1 megálló
1 meghajtó
1 memóriát
1 menedzser
1 menedzsmentszoftver
1 menü
1 menüpont
1 meteorológus
1 MLSZ-elnök
1 MNB-szóvivő
1 modem
1 modul
1 monitor
1 monogram
1 monostor
1 mozdonyvezető
1 nagyapa
1 nagybank
1 napraforgó-étolaj
1 napraforgóolaj
1 NATO-főtitkár
1 novíciát
1 nyelvet
1 nyomda
1 nyomtatógyártó
1 ÖIAG-igazgató
1 olajfinomító
1 OMV-vezér
1 onkológus
1 öregfiú
1 óriáscég
1 őrnagy
1 összecsapás
1 osztály
1 pálya
1 parancsgomb
1 parancsnokot
1 párt
1 pasa
1 PC-program
1 pénzalapok
1 pénzügyminiszter-helyettes
1 pilóta
1 platform
1 poggyászkocsi-kezelő
1 póló
1 projekt
1 projektmenedzser
1 projektvezető
1 protokollrendszer
1 pszichiáter
1 rakétaautó
1 reklámcég
1 reklámügynökség
1 rendezvénysorozat
1 rendőrautó
1 repülőgépek
1 repülőtársaságot
1 részleg
1 részvények
1 részvénytársaságnak
1 riporter
1 sajtófőnök
1 sajtot
1 sebész
1 segélyhajó
1 segélyszolgálathoz
1 sörgyár
1 sorozat
1 söröző be
1 sportlövész
1 sportriporter
1 sportszergyártó
1 stadion
1 szabvány
1 szállodalánc
1 szám
1 számítógépgyártó
1 székestrónváros
1 szekta
1 szektor
1 személyautók
1 szenátor
1 szenátorné
1 szerelvények
1 szerepjáték
1 szexmagazin
1 színkorrekció
1 színművész
1 szobrászművész
1 szoftveróriás
1 szolgáltatócégben
1 szóvivő
1 szuperkontinens
1 szupermarketlánc
1 TA-vezérigazgató
1 találkozón
1 talmudista
1 tanárnő

DOI: 10.15774/PPKE.BTK.2021.17

1 táplálkozáskutató
1 tárcavezető
1 távközlésieszköz-gyártó
1 teherautó
1 tengernagy
1 tér
1 testvérpár
1 teszt
1 tévétársaság
1 titkosítással
1 töltőállomáson
1 tőzsdén
1 tőzsdeüzemeltető
1 tőzsdevállalkozás
1 tréner
1 tűzoltó
1 üdítőt
1 újságírót
1 űrállomásról
1 űrhajó
1 űrhajósnő
1 utasszállítót
1 utcák
1 útra
1 útvonalon
1 üveg
1 üzlethálózat
1 vezérlőmodul
1 virológus
1 vívócsalád
1 vizsgálóbíró
1 vonalszakasz
1 zászlóalj
1 zenekar
1 zongoraművész
1 zoológus

B.3 Extended named entities with a modifier before the common noun at the end of the phrase (from Szeged Treebank 2.0)

Zoltán nevű ismerősöm
Gál Zoltán nevezetű haveromnak
Galambos tanár úr
Ford márkájú mikrobusszal
Paul Boutin webes tervező és tanácsadó
Radó András termelési és logisztikai vezérigazgató-helyettes
Gergely László anyagtervezési főosztályvezető
Szűcs Vince informatikai igazgató
Linux operációs rendszeren
Solaris operációs rendszeren
Bartolits István elnökhelyettesi tanácsadó
Volvo gyártmányú járműre
Ottó dominikánus szerzetes
John Chadwick egyesült királyságbeli tanácsadó , IOUG-tag
Pozsonyi Gábor consulting partner
Berki Endre korábbi vezérigazgató-helyettes
Graur Tamás korábbi kereskedelmi igazgató
Pokorni Zoltán oktatási miniszter
Dászkál János és Mészárosné Székely Éva kártyaszolgálati szakértő
Krisán Attila határőrségi szóvivő
Radovan Sztojkovszki macedón tárca nélküli miniszter
Glatz Ferenc akadémiai elnök
Bülent Ecevit török kormányfő
Helmut Kohl nyugatnémet és Németh Miklós magyar kormányfő
Kang Kjung Szik korábbi pénzügyminisztert
Kim In Ho egykori elnöki tanácsadót
Farit Gazizullin privatizációs miniszter
Sylvia Ann Hewlett amerikai feminista írónő
Christiaan Barnard dél-afrikai sebész

DOI: 10.15774/PPKE.BTK.2021.17

Stephen W. Hawking világhírű angol matematikus-fizikus
Leonhard Euler 18. századi svájci matematikus
Guillaume Appollinaire lengyel származású francia költő
George Orwell angol (újság)író
Rada Ivekovics szerb pszichiáter
Miroszlav Nikolics belgrádi szociológus-irodalmár
Torgyán József kisgazda pártelnököt , földművelésügyi és vidékfejlesztési minisztert
Zamecsnik Péter budapesti ügyvéd
Szabó Albert szélsőjobboldali pártvezető
Szaddám Huszein iraki elnöknek
Eduard Kukan szlovák külügyminiszter
Nagy Józsefné népjóléti irodavezető
Freili Géza igazgató főorvos
Farkas Kálmánné hatvani polgármester
Szabó János honvédelmi miniszter
Erdei Zoltán ügyvezető elnök
Jiri Rusnok cseh pénzügyminiszter
Csillag István leendő gazdasági miniszter
Elek István tiszteletbeli főkonzulhoz
Civin János korábbi elnök-vezérigazgatót
Jacques Chirac francia köztársasági elnök
Ovidiu Musetescu román privatizációs miniszter
Wieslaw Kaczmarek lengyel kincstárügyi miniszter
Bohuslav Sobotka cseh pénzügyminiszter
Matolcsy György gazdasági miniszter
Boros Imre PHARE-ügyekkel megbízott tárca nélküli miniszter
Laurent Fabius francia gazdasági és pénzügyminiszter
Gordon Brown brit pénzügyminiszter
Kurt Biedenkopf szász miniszterelnök
Szaid Barkat mezőgazdasági miniszter
Donald Evans amerikai kereskedelmi miniszter
Hans Eichel német pénzügyminiszter
Vlagyimir Putyin orosz elnök
Domingo Cavallo gazdasági miniszter
Gerhard Schröder német kancellár
Giuliano Amato olasz miniszterelnök
Matolcsy György gazdasági miniszter
Alekszej Kudrin orosz pénzügyminiszter
Aleksander Kwasniewski lengyel elnök
Eduardo Duhelde argentin elnöknek
Hans Niessl burgenlandi tartományi elöljáró
Mladjan Dinkic jugoszláv jegybankelnök
Didier Reynders belga pénzügyminiszter
Leszek Miller kijelölt lengyel miniszterelnök
Aleksander Kwasniewski lengyel elnök
Mihail Kaszjanov orosz kormányfő
Anatolij Kinah ukrán kormányfő
Katona Kálmán közlekedési miniszter

DOI: 10.15774/PPKE.BTK.2021.17

Tatár Attila tűzoltó ezredes
Bárdossy László egykori miniszterelnöknek
Kapolyi László egykori ipari miniszter
Békesi László volt pénzügyminiszter
Gerhard Tötschinger osztrák író-rendező
Franjo Tudjman köztársasági elnököt
Zeman cseh kormányfő
Milos Zeman cseh miniszterelnök
Wilson Hutchins amerikai fűtő
José Pombio spanyol pincér
St. Antonio uralkodó főherceg
St. Antonio uralkodó herceg
Alvarez volt elnököt
Bárányné Sülle Gabriella stratégiai és üzletpolitikai igazgató
Torgyán József földművelésügyi miniszter
Lágler Katalin ügyvezető igazgató
Pepó Pál környezetvédelmi miniszterhez
Kucsis Gyula Zala megyei ügyész
Martonyi János magyar külügyminiszter
Mihail Kaszjanov orosz kormányfő
Alvarez volt elnököt
Zoltán nevű ismerősöm
Gál Zoltán nevezetű haveromnak
Galambos tanár úr
Ford márkájú mikrobusszal
Paul Boutin webes tervező és tanácsadó
Radó András termelési és logisztikai vezérigazgató-helyettes
Gergely László anyagtervezési főosztályvezető
Szűcs Vince informatikai igazgató
Linux operációs rendszeren
Solaris operációs rendszeren
Bartolits István elnökhelyettesi tanácsadó
Volvo gyártmányú járműre
Ottó dominikánus szerzetes
John Chadwick egyesült királyságbeli tanácsadó
Pozsonyi Gábor consulting partner
Berki Endre korábbi vezérigazgató-helyettes
Graur Tamás korábbi kereskedelmi igazgató
Pokorni Zoltán oktatási miniszter
Dászkál János és Mészárosné Székely Éva kártyaszolgálati szakértő
Krisán Attila határőrségi szóvivő
Radovan Sztojkovszki macedón tárca nélküli miniszter
Glatz Ferenc akadémiai elnök
Bülent Ecevit török kormányfő
Helmut Kohl nyugatnémet és Németh Miklós magyar kormányfő
Kang Kjung Szik korábbi pénzügyminisztert
Kim In Ho egykori elnöki tanácsadót
Farit Gazizullin privatizációs miniszter

DOI: 10.15774/PPKE.BTK.2021.17

Sylvia Ann Hewlett amerikai feminista írónő
Christiaan Barnard dél-afrikai sebész
Stephen W. Hawking világhírű angol matematikus-fizikus
Leonhard Euler 18. századi svájci matematikus
Guillaume Appollinaire lengyel származású francia költő
George Orwell angol (újság)író
Rada Ivekovics szerb pszichiáter
Miroszlav Nikolics belgrádi szociológus-irodalmár
Torgyán József kisgazda pártelnököt , földművelésügyi és vidékfejlesztési minisztert
Zamecsnik Péter budapesti ügyvéd
Szabó Albert szélsőjobboldali pártvezető
Szaddám Huszein iraki elnöknek
Eduard Kukan szlovák külügyminiszter
Nagy Józsefné népjóléti irodavezető
Freili Géza igazgató főorvos
Farkas Kálmánné hatvani polgármester
Szabó János honvédelmi miniszter
Erdei Zoltán ügyvezető elnök
Jiri Rusnok cseh pénzügyminiszter
Csillag István leendő gazdasági miniszter
Elek István tiszteletbeli főkonzulhoz
Civin János korábbi elnök-vezérigazgatót
Jacques Chirac francia köztársasági elnök
Ovidiu Musetescu román privatizációs miniszter
Wieslaw Kaczmarek lengyel kincstárügyi miniszter
Bohuslav Sobotka cseh pénzügyminiszter
Matolcsy György gazdasági miniszter
Boros Imre PHARE-ügyekkel megbízott tárca nélküli miniszter
Laurent Fabius francia gazdasági és pénzügyminiszter
Gordon Brown brit pénzügyminiszter
Kurt Biedenkopf szász miniszterelnök
Szaid Barkat mezőgazdasági miniszter
Donald Evans amerikai kereskedelmi miniszter
Hans Eichel német pénzügyminiszter
Vlagyimir Putyin orosz elnök
Domingo Cavallo gazdasági miniszter
Gerhard Schröder német kancellár
Giuliano Amato olasz miniszterelnök
Matolcsy György gazdasági miniszter
Alekszej Kudrin orosz pénzügyminiszter
Aleksander Kwasniewski lengyel elnök
Eduardo Duhelde argentin elnöknek
Hans Niessl burgenlandi tartományi elöljáró
Mladjan Dinkic jugoszláv jegybankelnök
Didier Reynders belga pénzügyminiszter
Leszek Miller kijelölt lengyel miniszterelnök
Aleksander Kwasniewski lengyel elnök
Mihail Kaszjanov orosz kormányfő

DOI: 10.15774/PPKE.BTK.2021.17

Anatolij Kinah ukrán kormányfő
Katona Kálmán közlekedési miniszter
Tatár Attila tűzoltó ezredes
Bárdossy László egykori miniszterelnöknek
Kapolyi László egykori ipari miniszter
Békesi László volt pénzügyminiszter
Gerhard Tötschinger osztrák író-rendező
Franjo Tudjman köztársasági elnököt
Zeman cseh kormányfő
Milos Zeman cseh miniszterelnök
Wilson Hutchins amerikai fűtő
José Pombio spanyol pincér
St. Antonio uralkodó főherceg
St. Antonio uralkodó herceg
Alvarez volt elnököt
Bárányné Sülle Gabriella stratégiai és üzletpolitikai igazgató
Torgyán József földművelésügyi miniszter
Lágler Katalin ügyvezető igazgató
Pepó Pál környezetvédelmi miniszterhez
Kucsis Gyula Zala megyei ügyés
Martonyi János magyar külügyminiszter
Mihail Kaszjanov orosz kormányfő
Alvarez volt elnököt

DOI: 10.15774/PPKE.BTK.2021.17

Appendix C

Locative case suffixes

C.1 Table of words with a locative case suffix appearing in the Hungarian UD corpus

1-: date_On
11-: date_On – nÁl:date_On, event
11-e: date_On
16-a: date_On
18-a: date_On
1942-: date_bAn
1948-: date_bAn
1949-: date_bAn
1950-: date_bAn
1955-: date_bAn
1957-: date_bAn
1964-: date_bAn
1967-: date_bAn
1979-: date_bAn
1980-: date_bAn
1981-: date_bAn
1982-: date_bAn
1990-: date_bAn
1991-: date_bAn
1992-: date_bAn
1994-: date_bAn
1995-: date_bAn
1997-: date_bAn
1998-: date_bAn
1999-: date_bAn
1-je: date_On
2:0: thing – nÁl:Mikor?, Milyen állásnál?; rA:Milyen arányban?; rÓl:Milyen állásról?
2000-: date_bAn
2001-: date_bAn
2002-: date_bAn
21-e: date_On
2-2-: thing – nÁl:Mikor?, Milyen állásnál?; rA:Milyen arányban?; rÓl:Milyen állásról?, Milyen helyzetből?
23-a: date_On
26-a: date_On
27-e: date_On
29-e: date_On
2-a: date_On
30-a: date_On
3500-: num
4-5-e: date_On
4-e : date_On
65-47: thing – nÁl:Mikor?, Milyen állásnál?; rA:Milyen arányban?; rÓl:Milyen állásról?

DOI: 10.15774/PPKE.BTK.2021.17

67-56: thing – nÁl:Mikor?,
Milyen állásnál?; rA:Milyen arányban?;
rÓl:Milyen állásról?
6-a : date_On
76,8: num
7-e: date_On
8-a: date_On
97-: date_bAn
abház: who
abortusz: thing
adásvétel: thing
adóhatóság: org=who_nÁl
Afganisztán: loc_country
ág: loc_any
Agassi: who
ágazat: thing bAn:loc_any
akadályoztatás: thing
akadémia: build=inst_On
akció: thing
alapja: mode_bAn –
On:mode_On, Mi alapján?;
rA:thing;
bÓl:thing
alapkőletétel: event_On
albizottság: posi_bAn
alkalmazott: who
alkalom: event_On – hOz:thing;
bÓl:cause, Milyen alkalomból?
alkohol: thing
alkotás: thing
alkotmány: loc_bAn
alkotmánybíróság: org=who_On
állam: loc_bAn – rA:thing
állapot: state
állás: posi_bAn – bAn:mode;
nÁl:posi_bAn, date;
bA:mode
álláspont: thing
állítása: thing
álmatlanság: thing
Altenkirchen: loc_city-bAn
amely: thing
amilyen: thing
Amszterdam: loc_city-bAn
Ankara: loc_city-bAn
április: date_bAn
apropója: thing – On:cause;
rA:cause;
bÓl:cause
ár: thing
arany: thing
áremelkedés: thing
árok: place_bAn
árus: who
árverés: event_On
ásványkincs: thing
aszfalt: loc_On
Athén: loc_city-bAn
átkelőhely: place_On
átlag: thing
átszámozás: thing
átszervezés: thing
átütemezés: thing
augusztus: date_bAn
aukció: event_On
autó: place_bAn
autóbaleset: event_bAn
bAn:event_bAn, thing;
bA:thing; bÓl:thing
autópálya: place_On
az: dempron – rA:dempron, direct
Ázsia: place_bAn
bajbajutott: who
bajnokság: loc_bAn
Bakony: place_bAn
baleset: thing – bAn:thing; event
nÁl:thing, loc_any;
hOz:thing, loc_any

DOI: 10.15774/PPKE.BTK.2021.17

baloldal: place_On
bank: build=inst_bAn – rA:build=inst_bAn, thing
bankautomata: place_bAn – nÁl:place_bAn, thing
barát: who
bár: place_bAn
bársonyszék: loc_any
Batumi: loc_city-bAn
Bécs: loc_city-bAn
befejezés: thing
befektetés: thing
belföld: loc_On
Belgrád: loc_city-bAn
bemutató: event_On
beosztás: posi_bAn
Berlin: loc_city-bAn
beruházás: thing – tÓl:thing, event
beszéd: event_nÁl-On – bAn:event_nÁl-On, loc_any; bÓl:event_nÁl-On, loc_any
beszélgetés: thing
bevétel: thing
bevezetés: thing
bezárás: thing – nÁl:event_nÁl-On
bírálat: thing
bizottság: posi_bAn
biztonság: thing
biztosítás: thing
BL-: loc_bAn
bocsánatkérés: thing
Bodmér: who
bojkottálása: thing
bolgár: who
Bor: loc_city-bAn
bőr : body_bAn-On
börze: thing
bravúr: thing
Brazília: loc_country
Brüsszel: loc_city-bAn
Budapest: loc_city-On
büdzsé: thing
buli: event_On – bAn:event_bAn; bA:event_bAn; bÓl:event_bAn
bűnbocsánat: thing
Bundesliga: loc_bAn
Bundestag: build=inst_bAn
busz: thing
Carmen: loc_bAn
cég: org_bAn
cégcsoport: org_bAn
cégegyesülés: thing
cél: loc_any – rA:loc_any, thing; bÓl:cause
célállomás: place_On
ceremónia: event_On – rA:event_On, loc
cikk: loc_bAn-On
ciklus: date_bAn
cím: loc_bAn – On:loc_bAn, mode
címoldala: loc_On
Cottbus: loc_city-bAn
család: group
csatározások: circumst – tÓl:thing; rÓl:thing
csatorna: loc_any
Csecsenföld: place_On
csevegés: thing
csigatempó: mode_bAn – rA:thing; bÓl:thing
csökkentés: thing
csomagterv: thing
csoport: group
csúcs: thing nÁl:thing;loc
csúcsidő: date_bAn
csúcstalálkozó: event_On – rA:event_On, thing
csütörtök: date_On
cukoripar: loc_bAn

DOI: 10.15774/PPKE.BTK.2021.17

Daewoo: org_nÁl
Daewooé: dem
Daewoo-megállapodás : thing
Dagesztán: loc_country
dalszínház: place_bAn
darab: thing
dátum: thing
december: date_bAn
Dél-Korea: loc_country
délnyugat: direct_On
demokrácia: thing
demonstráció: event_On –
On:event_On, loc
derék: body_any – On:body_any, part
diáklétszám: thing
díler: who
dízelolaj: thing
dogma: thing
dokumentum: loc_bAn-On
dollár: curr – bAn:curr, Hány dollárban?;
On:curr, thing
dömping: thing
döntő: event_bAn
Dosztojevszkij-rendezés: thing
Dubcek: who
dugó: loc_bAn
duplája: num_size
eb: thing
EBESZ-csúcs: event_On –
On:event_On, loc
Eb-középdöntő: event_bAn –
bAn:event_bAn, thing
Eb: event_On – On:event_On, loc
eddigi: dem
edzés: event_On
egész: thing – bAn:thing, mode
egy: date_bAn – bAn:mode;bA:mode;
rA:date_bAn; bÓl:date_bAn, mode
egyeduralom: thing
egyensúly: state
egyesülések: thing
egyetem: build=inst_On –
On:build=inst_On, event
egyforma: thing – rA:Milyen részekre?
egymás: who
egynémelyike: dem
egység: thing – bÓl:thing, group
egységes: mode_On – rA:mode_On,
Milyenre?;
bÓl:thing
együttes: place_bAn
együttműködés: thing
eladás: thing
él: part – bÓl:part, mode
eleje: part – On:part, date;
rA:part, date;
tÓl:part, date
élelmiszeripar: loc_bAn
elem: thing – bAn:thing, mode
élet: loc_bAn – bAn:loc_bAn, date
életmód: thing
elhelyezés: thing
elismerés: thing
eljárás: thing
ellátás: thing
ellenfél: who
ellentét: thing
ellenzék: loc_bAn
elnök: who
elnökség: org=who_bAn
elnökválasztás: event_On
elnyerés: thing
előadás : event_On – bAn:date, form
előbbi: dem
előirányzat : thing
előkészítés : thing
előny: thing
Előszállás: loc_city-On

DOI: 10.15774/PPKE.BTK.2021.17

elosztás: thing
elseje: date_On
elszállítás: thing
elszigetelés: thing
eltörlés: thing
elv: pov
ember: who
emberrablás: thing
emelés: thing
én: who
energiaáramlás: thing
engedmény: thing
ennyi: num
épület: place_bAn
érdek: thing – bAn:cause, pov;
bÓl:cause
érdekeltség: thing
érdeklődés: thing – rA: cause
eredmény: thing
érés: thing
erőfeszítés: thing
Erste: org=who_nÁl
érték: thing – On:thing, Mennyiért?
értékelés: thing
értékesítése: thing
értéktőzsde : loc_On
értelem: pov
érthetőség: thing
eset: thing – bAn:thing, pov;
On:thing, date;
hOz:thing, loc
est: event_On
Észak-Kaukázus: place_bAn
Északnyugat-Magyarország: place_On
esztendő: date_bAn
étterem: place_bAn
EU-: org=who_bAn
Európa: place_bAn
Európa-bajnokság: event_On
EU-tagállam: loc=who
év: date_bAn
évad: date_bAn
évezred: date_bAn
évfordulója: date_On
évjárat: thing
évtized: date_bAn
ez: dempron – rA:dempron, direct
ezüst: material – bA:material, mennyibe?
fa: material
fájdalmas: thing
fajta: thing
FÁK-piac: loc_On
falu: loc_any
február: date_bAn
fedélzet: loc_On
fedezés: thing
fedőnév : thing –
On:mode, Milyen fedőnéven?
fegyházbüntetés: thing – rA:thing, mire
Fehér-csoport: group
fej: body_bAn-On –
bÓl:body_bAn-On, mode
Fejér: place_bAn
fejlesztés: thing
fék: thing – On:thing, loc
feladatellátás: thing
felállítás: thing
feldarabolás: thing
fele: date_bAn – nÁl:Mikor?;
On:date_bAn, date_On;
rA:date_bAn, num_size;
rÓl:date_bAn, num_size
felelősség: thing
félév: period
félidő: date_bAn – nÁl: Mikor?
fellépés: event_nÁl-On
félpálya: place_On
felszín: place_On

DOI: 10.15774/PPKE.BTK.2021.17

feltöltés: thing
felvétel: thing
fennakadás: thing
fenyő: place_On
férfi: who
Fidesz-birodalom: thing – bAn:thing, loc
figyelem: thing
film: thing – bAn:thing, loc
filmszemle: event_On
filmvászon: loc_On
filozófia: thing
finanszírozás: thing
finomító: place_bAn
Fiorentina: place_bAn
Fiume: loc_city-bAn
fő: meas – bAn:meas, Hány főben?;
nÁl:meas, Hány főnél?;
rA:meas, Hány főre?;
bÓl:meas, Hány főből?
fogadócsoport: group
foglalkoztatás: thing
fogság: loc_bAn
fogyasztóiár-változás: thing
főhadiszállás: place_On
főiskola: build=inst_On –
On:build=inst_On, date
fokozás: thing
földrengés: thing – tÓl:thing, date
folyamat: thing
folytatás: thing
főműsoridő: loc_bAn – bAn:loc_bAn, date
forduló: thing – bAn:thing, date
fordulópont: place_On
forint: curr – nÁl:curr, Hány forintnál?;
On:curr, thing;
rA:curr, Hány forintra?;
rÓl:curr, Hány forintról?
forma: form_bAn
formaság: thing
forrás: thing
fórum: loc_On
fős: meas – rA:meas, Mekkorára?, Hány fősre?
fotó: loc_On – rA:thing
főutak: way
főútvonalak: way
főváros: place_bAn
front: place_On
fűtőolaj: thing
fuvarokmány: loc_bAn-On
gála: event_On
garázs: place_bAn
gárda: group
gazda: who
gazdaság: thing – bAn:thing, loc
Gazprom: org_nÁl
gépkocsi: thing – bA:thing, loc
gépkocsivezető: who
globalizáció: thing
Gnjilane: loc_city-bAn
Gödöllő : loc_city-On
gondolkodás: thing
gondozás: thing – bAn:thing, form
gorillacsalád: group
Groznij: loc_city-bAn
gyakorlatok: thing – rA:thing, event
gyár: place_bAn
győzelem: thing
gyűjtemény: loc_bAn
háború: event_bAn –
bAn:event_bAn, circumst;
rÓl:thing
Hága: loc=who
hajléktalan: who
hajnal: date_bAn
hajó: thing – rÓl:loc
hajrá: date_bAn
haláleset: thing

DOI: 10.15774/PPKE.BTK.2021.17

hallata: thing – On:thing, Mi előzte meg, hogy az legyen, ami van?
hálózat: loc_any
Hamburg: loc_city-bAn
hangjegy: thing
hangzás: thing
Hannover: loc_city-bAn
háromszorosa: num_size
használat: thing
hatalom: posi_On
hatály: thing
hatálybalépés: thing – nÁl:event_On; tÓl:event_On
határ: loc_any
határozat: thing
hatás: thing – rA:cause
hatáskör: thing – bAn:loc
hatékonyság: thing
hatóság: org=who_nÁl
hátrány: circumst – On:thing; tÓl:thing
ház: loc_any – On:loc_any, thing
Ház: build=inst_bAn
hazahozatal: thing
hazugság: thing
hegy: place_On
hely: place_On – bAn:loc_any; bÓl:mode
helyreállítás: thing
helyszín: place_On
helytállás: thing
helyzet: thing – bAn:thing, state
hét: date_On – rA:date_On, period
hétfő: date_On
hétvége: date_On – rA:date_On, period
hiba: thing
hipnózis: state
história: thing
hitel: thing
hitelkeret: thing
hivatal: posi_bAn
hó: circumst – nÁl:thing; On:loc_any; bA:loc_any, thing; hOz:thing; rA:loc_any; bÓl:thing; tÓl:thing; rÓl:thing
hóbucka: thing
hócsapda: thing
hőfok: mode_On – On:mode_On, Mekkora hőfokon?; rA:thing, bÓl:thing
holttest: thing
hómentesítés: thing
hónap: date_bAn – rA:date_bAn, period
honfitárs: who
Horvátország: loc_country
hős: who
hótorlasz: thing
idő: date_bAn – bAn:date_bAn, mode; rA:date_bAn, period, mode; bÓl:date_bAn, loc_any
időpont: date_bAn – nÁl:mikor? rÓl:date_bAn; loc
időszak: date_bAn
igazgatótanács: group
igazolás: thing
illusztrálás: thing
indok: thing
indoklás: thing – bAn:thing, form, loc; bÓl:thing, loc
Indonézia: loc_country
infláció: thing
inflációkiegyenlítés: thing
interjú: loc_bAn
intézmény: place_bAn
irány: direct_bAn – bA:direct_bAn, Milyen irányba?
irányítás: thing
írás: form_bAn

DOI: 10.15774/PPKE.BTK.2021.17

irattár: place_bAn
iskolabezárás: thing
ismeretlenség: thing
Isztambul: loc_city-bAn
ítélet: form_bAn – nÁl:form_bAn, event;
rA:form_bAn, event
ítélethirdetés: thing – rA:thing, event
íz: thing – bAn:Hányszor?, Milyen ízben?
Izmit: loc_city-bAn
január: date_bAn
jármű: thing – hOz:thing, loc
járőr: who
játék: thing
játékos: who
játékrész: date_bAn
játékvezető: who
javaslat: thing – bAn:loc
jegyzék: thing – bAn:loc
jelen: date_bAn – rA:thing;
bÓl:date_bAn, loc_any
tÓl:thing
jelenlét: thing – bAn:thing, circumst
jelentés: thing – bAn:thing, loc
jog: thing – On:thing, form
jogcím: mode_On –
On:mode_On, Milyen jogcímen?;
rA:thing; bÓl:thing
jóváhagyás: thing
jövedelem: thing
jövő: date_bAn - bA:date_bAn, loc_any;
bÓl:date_bAn, loc_any;
tÓl:thing
jövőkép : thing
július: date_bAn
Kamaraszínház: loc_bAn
kamatadó: thing
kamatfizetés: thing
kamera: thing
kamion: thing
kancelláré: dem
Kandahár: loc_city-bAn
kanton: place_bAn - rA:thing
káosz: circumst - On:thing; tÓl:thing;
rÓl:thing
kapcsolat: thing - bAn:pov, thing
Kaposvár: loc_city-On
kapu: loc_any
kar: body_bAn-On
kár: thing
kassza: place_bAn
katalógus: loc_bAn
kátyú: thing - bÓl:thing, loc
kávébár: place_bAn
kávészünet: date_bAn
kávézás: thing
kedd: date_On
kegyetlenkedés: thing
kémhistória: thing
kényszer: thing - bÓl:cause
kényszerűség: thing - bÓl:cause
kép: place_On
képernyő: thing
képviselő: who
kérdés: thing - rA:thing, cause
Kerekegyháza: loc_city-bAn
keréknyom: thing
kereskedelem: loc_bAn
kereslet: thing
keret: thing - bAn:mode,
Minek a keretében?
kerület: place_bAn
készenlétihitel-megállapodás: thing
készítés: event_nÁl-On
kéz: body_any - On:body_any, mode
kezelés: thing
Kft.-: org=who_bAn
kg-: meas - bAn:meas,
Milyen kategóriában?

DOI: 10.15774/PPKE.BTK.2021.17

kiadás: form_bAn
kiadvány: loc_bAn
kialakítás: thing
kialakulás: thing
kialakuló: thing bAn:milyen_állapotváltozáson_megy_át?
kiállítás: event_On
kibocsátás: thing
kielégítés: thing
kiépítés: thing
kifejtés: thing
kilométer: meas - bAn:meas, Hány kilométerben?, thing; nÁl:meas, loc_any; hOz:meas, loc_any; rA:meas, Milyen messzire?, Hol?/Milyen messze?; rÓl:meas, Milyen messziről?
Kína: loc_country
kinnlevőség : thing
kisbank: place_bAn
kisebbség: state - nÁl:loc_any
kiskereskedelem: loc_bAn
kisváros: place_bAn
kitörés: thing
kő: thing - bA:thing, loc
köbméter: meas - bAn:meas, Hány köbméterben?, thing; rA:meas, Hány köbméterre?
kókuszszőnyeg: thing
Köln: loc_city-bAn
Kolozsvár: loc_city-On
költség: thing
költségvetés: thing
kóma: state nÁl:thing
koncentráció: thing
koncert: event_On
konferencia: event_On
konfliktus: thing
konjunktúra-időszak: thing
konjunktúrateszt: thing
konkurencia: thing
konstrukció: form_bAn
konszolidálódás: thing
kontinens: place_On
kontinenstalálkozó: thing
könyv: loc_bAn
kőolaj: thing
kor: date_bAn - bA:date_bAn, loc_any; bÓl:date_bAn, loc_any
kör: mode_bAn - bAn:mode_bAn, loc_any; nÁl:loc_any; On:loc_any; bA:mode_bAn, Hogyan?; rA:Hány részre?, Hány körre?; bÓl:loc_any
kórház: build=inst_bAn
kormány: org=who_any
kormányjavaslat: thing
kormányoldal: loc_On
környék: place_On
korona: thing - bA:thing, loc
korosztály: group
korrektség: thing
korú: who
körülmény: thing
körzet: place_bAn
koszorú: thing - On:thing, loc
Koszovó: loc_country
kötelmei: thing
Kouchner: who
következmény: thing
közakarat: thing
közállapot: thing
közbeszéd: thing
közeg: place_bAn
közele: loc_bAn

DOI: 10.15774/PPKE.BTK.2021.17

közeljövő: date_bAn - rA:thing; tÓl:thing
közelmúlt: date_bAn - rA:thing;
bÓl:loc_any; tÓl:thing
közepe: part - On:part, date;
rA:part, date;
tÓl:part, date
közgyűlés: event_On
közhivatal: org_bAn
közhivatalnok: who
közlekedés: thing
közlemény: thing bAn:loc
közmegelégedés: thing
közoktatás: loc_bAn
közpénz: thing
köztársaság: place_bAn - bAn:place_bAn,
date
külföld: loc_On
külterület: loc_On
külügyminisztérium: org=who_bAn
külvilág: thing
láb: body_any
labda: thing
lakás: place_bAn
lakó: who
lakóhely: place_On
lakótelep: place_On
láng: circumst - nÁl:thing; On:thing;
bA:thing; hOz:thing; rA:thing;
tÓl:thing; rÓl:thing
Laptárs: loc_nÁl - hOz:loc_nÁl, thing
látogatás: event_On
látogató: who
LB-: loc
leányvállalat: org_nÁl
leértékelődés: thing
lefolytatás: thing
légkör: circumst - bÓl:loc_any; tÓl:thing;
rÓl:thing
légsúly: thing
lehetőség: thing
lélek: thing - On:thing, body
lemez: thing
lemondás: thing
Lengyelország: loc_country
lény: thing
lépcső: place_On
lépés: thing - bAn: Hány lépésben?
letartóztatás: thing - bAn:thing, loc_any;
nÁl:thing, loc_any;
hOz:thing, loc_any;
bÓl:thing, loc_any
levegő : loc_bAn-On
levél: form_bAn - On:form_bAn, loc_any
levélpapír: loc_On
libasor: mode_bAn - bA:mode_bAn,
Hogyan?;
rA:thing; bÓl:loc_any
Lipótmező: loc_On
lista: thing - On:thing;loc
lökés: thing
lopás: thing
LRI-: org_nÁl
Luxemburg: loc_country
Macedónia: loc_country
maga: who - tÓl:who, mode
magaslat: place_On
Maglód: loc_city-bAn
magnószalag: thing
Magyarország: loc_country
május: date_bAn
makacsság: thing
maradás: thing
március: date_bAn
márka: thing
márkanév: thing
Martonyi-vizit: event_nÁl-On
másik: who
másodfok: loc_On - On:loc_On, event

DOI: 10.15774/PPKE.BTK.2021.17

másodperc: date_bAn
matematika: thing
mater: loc_bAn
McDonald's: build=inst_bAn
McDonald'sé : dem
meccs: event_On
megbeszélés: event_On
megbízás: thing - bÓl: Milyen alapon?, cause
megfontolás: thing - bÓl: cause, Milyen megfontolásból?
meghívás: thing
megkeresés: thing - rA:thing, cause
meglepetés: mode_bAn
megmentés: thing
megmutatkozás: thing
megoldás: thing
megosztás: thing
megpróbáltatás: thing
megromlás: thing
megsemmisítés: thing
megspórolás: thing
megszületés: thing
megszüntetés: thing
megteremtés: thing
megvalósítás: thing
megvétel: thing
megyeszékhely: place_On
mellékútvonal: way
mely: dempron
mélység: place_bAn
menet: thing
merénylet: thing
méret: thing
mérkőzés: event_On
mérleg: thing
mérték: thing - bAn:thing, meas
mese: thing
méter: meas - bAn:meas, Hány méterben?, thing; nÁl:meas, loc_any; hOz:meas, loc_any; rA:meas, Milyen messzire?, Hol?/Milyen messze?; rÓl:meas, Milyen messziről?
mezőny : place_bAn
Mezőnyjáték : thing
mi: thing
midibuszai: thing - bA:thing, loc
millió: thing
mindaz: dempron
minden: thing
minimálbér: thing
minimálzene : thing
miniszterelnök: who
minisztérium: loc=who
minta: thing - rA:thing, mode
Mitrohin-akták: loc_bAn
mivolta: thing
MLSZ-: org_bAn
mód: mode_bAn - On:mode_On; bÓl:thing
módosítás: thing
módszer: thing
mondat: thing
Montreal: loc_city-bAn
Moszkva: loc_city-bAn
mozivászon: loc_On
mű: thing
műfaj: thing - bAn:thing, pov
műfordítása: form_bAn
működés: thing
múlt: date_bAn - bA:loc_any; rA:thing; bÓl:loc_any; tÓl:thing
munka: thing
munkahely: place_On
munkakör: posi_bAn
munkanap: date_On

DOI: 10.15774/PPKE.BTK.2021.17

munkásság: thing
műsor: loc_any - rA:loc_any, thing
műterem: place_bAn
művészet: thing
nagy: dem - bAn:dem, mode;
On:dem, mode;
rA:dem, num_size
nagypolitika: thing
nagyság: thing
nagyságrend: thing
nap: date_On - On:date_On, loc_any;
rA:date_On, period
napirend: thing - rÓl:thing, loc
NATO-: org_bAn
negyed: place_bAn - bAn:place_bAn,
date;
rA:place_bAn, date;
tÓl:place_bAn, date
negyedév: date_bAn
némaság: thing
némelyik: dem
német: who
Németország: loc_country
nép: who
Népszava: place_bAn
népszerűség: thing
név: thing - bAn:thing, mode,
Kinek a nevében?;
On:mode, Milyen néven?
névérték: thing
Nis: loc_city-bAn
nők : who
növekedés: state
növelés: thing
november: date_bAn
NSZK-: loc_country - bAn:loc_country,
date
nyár: date_On
nyelv: thing
nyilatkozat: thing - bAn:thing, loc
nyilvánosság: thing
nyomás: thing - rA:thing, cause
nyomdok:
nyomorúság: thing
Nyugat: loc_On
nyugat: direct_On
nyújtás: thing
ő(k): who
ok: thing - On:cause;
bÓl:cause, Milyen okból?
október: date_bAn
olajár: thing - On: Mennyiért?
olajtartály: thing
olasz: who
oldal: part
oldalirány: direct_bAn
bAn:direct_bAn;milyen_irányban?
oldalvonal: place_On
olimpia: event_On
öltöző: place_bAn
ön: who
önkormányzat: org=who
önkorrekció: thing
önmaga: who - bAn:who, pov
Operaház: build=inst_bAn
operaszínpad: loc_On
óra: event_On - rA:event_On, period
Orahovac: loc_city-bAn
Oroszország: loc_country
Országgyűlés: org=who_bAn
országhatár: thing
ország: place_bAn
országoké: dem
orvos: who - bÓl:who, thing
összeállítás: thing - bAn:thing, mode
összecsapás: thing
összehasonlítás: thing

DOI: 10.15774/PPKE.BTK.2021.17

összeköttetés: thing - bAn:thing,
Milyen kapcsolatban?
összesség: pov
összetett: thing
ősz : date_On
osztály: date_bAn -
bAn:date_bAn, loc_any, thing;
nÁl:loc_any; bA:date_bAn, loc_any;
hOz:date_bAn, loc_any;
bÓl:date_bAn, loc_any;
tÓl:date_bAn, kitől?
otthon: place_bAn - rA:loc_any;
rÓl:loc_any
otthonmaradás: thing
pálya: place_On
pályázat: thing
páncélterem: place_bAn
papír: form_On - On:form_On, loc_any
papírpohár: thing
Párizs: loc_city-bAn
parkoló: place_bAn
parlament: build=inst_bAn
pártállás: thing
párt: org=who_bAn
partvidék: place_On
példány: meas - bAn:meas,
Hány példányban?
péntek: date_On
pénz: thing
pénzszűke: thing
per: thing
perc: date_bAn - rA:period
periódus: date_bAn - rA:period
perújrafelvétel: thing
piac: place_On
Pilisszántó: loc_city-On
pillanat: date_bAn - rA:period
pódium: place_On
polgármester: who
polgár: who
politika: thing
pont: loc_any
porcelán: material
porond: loc_On
Portó: loc_city-bAn
poszt: posi_On
pótlék: thing
pozíció: posi_bAn
Pozsony: loc_city-bAn
Pristina: loc_city-bAn
privatizáció: thing
privatizálás: thing
produkció: thing - bAn:thing, form
program: thing - bAn:thing, loc;
bA:thing,loc;
rÓl:thing, loc
publikálás: thing
puszta: loc_any
rádió: place_bAn
rajzfilm: thing
rangsor: loc_bAn
rasszizmus: thing
reálérték: thing - On:thing, Mennyiért?
reálgazdaság: thing
reálkeresetek: thing
reformintézkedés: thing
rekordja: thing
rendelkezés: thing
rendezés: thing
rendezvény: event_On
rendőr: who
rendőrjárőr : who
rendőrség: build=inst_On
rendszer: thing bAn:thing, circumst;
bÓl:thing, loc
repülőtér: place_On
restitúció: thing

DOI: 10.15774/PPKE.BTK.2021.17

rész: loc_any -
bAn:loc_any, date, mode;
On:loc_any, part
részegység: org_bAn
részesülő: who
részleg: place_On - tÓl:loc, Honnan?
részlet: thing - rÓl:thing, Miről?
részvényopció: thing - hOz:thing, Mihez?
részvétel: thing - rA:thing, Mire?
ring: place_bAn
ritmus: mode_bAn - bÓl:thing
roma: who - rÓl:who, Kiről?
rom: loc_nÁl
rövidzárlat: thing - rA:thing, Mire?
rt-: org=who_bAn
rubel: curr - On:curr, thing;
rA:curr, Hány rubelre?
Saarbrücken: loc_city-bAn
sajtótájékoztató: event_On
sakkasztal: place_On
sakkélet: thing
sakkolimpia: event_On
sarok: loc_any
Seattle: loc_city-bAn
segélyakciók: thing
semmi: thing - bA:num;
bÓl:thing, loc_any
semmitmondás: thing
sietség: thing
siker: thing
sír: loc_any - On:loc_any, thing
sítalp: thing On:mode
Skopje: loc_city-bAn
sok: num - bAn:num, Mennyire?
Somogyszob: loc_city-On
sor: loc_any
sorozat: date_bAn - bAn:date_bAn, mode
sorrend: form_bAn - bA:form_bAn, thing
sors: thing
Spanyolország: loc_country
Starbucks: build=inst_bAn
Starbucks-kávé: thing
stúdió: place_bAn
Svájc: loc_country
szabadrúgás: thing - nÁl:event;
hOz:thing, loc; bÓl:mode
szabadság: thing - On:thing, loc
szabály: thing
Szahara: place_bAn
szakítás: thing
szálloda: place_bAn
szállodaszobája: place_bAn
szám: thing - bAn:thing, loc
számítógép: thing
szavazat: thing
század: date_bAn
százalék: meas - bAn:meas,
Hány százalékban?;
On:meas, Hány százalékon?, loc_any;
rA:meas, Hány százalékra?;
rÓl:meas, Hány százalékról?
Szeged: loc_city-On
szék: posi_bAn
Székesfehérvár: loc_city-On
székhely: place_On
Szekszárd: loc_city-On
szekta: thing
széle: part
szem: body_any - rA:body_any, mode
személyiség: thing
szeminárium: thing - rA:thing, event
szempont: pov
szeptember: date_bAn
szerda: date_On
szerelőakna: place_bAn
szerep: thing - bAn:thing, mode_bAn
szerkesztőség: place_bAn - hOz:loc_any
szerv: org_nÁl

DOI: 10.15774/PPKE.BTK.2021.17

szervezés: thing - bAn:thing, mode,
Kinek a szervezésében?
szervezet: org=who_bAn
szerzemény: thing
szerződés: loc_any - nÁl:event;
hOz:thing; tÓl:thing, event
szett: date_bAn
szezon: date_bAn
szféra: loc_bAn
sziget: place_On - rA:place_On, thing
szín: loc_On
színhely: place_On
színmű: thing
színpad: place_On - rA:place_On, thing
szint: place_On
szintér: loc_On
színvallás: thing
szív: body_any - hOz:body_any, thing;
bÓl:body_any, mode;
rÓl:thing
szlovének: who
szó: form_bAn
szoba: place_bAn - rA:place_bAn,
loc_any
szobor: thing - rÓl:thing, Miről?
szokás: thing
szombat: date_On
Szombathely: loc_city-On
szomszéd: who
Szöul: loc=who
szövetség: thing
Szovjetunió: loc_country
sztár: who
sztrájkhullám: thing
szupermarket: place_bAn
tag: who
tagállam: loc_bAn
tájékoztató: event_On
Tajvan: loc_country
találkozó: event_On
talp: body_bAn-On - On:body_bAn-On,
mode; rA:body_bAn-On, part
támadások: thing
támogatás: thing
tanácskozás: thing
tárgyalás: event_On
tárlat: place_On - On:place_On, event
társadalom: group
társaság: group - bAn:group, circumst
tartalom: thing
tartás: thing
tartomány: place_bAn
táv: thing - nÁl:loc; On:loc;
hOz:loc; rA:period
tavalyi: dem
tavasz: date_On
távozás: thing
tejipar: loc_bAn
tekintet: thing
tél: date_On
telefon: thing - bAn:loc; nÁl:loc;
On:mode; bA:thing, loc;
hOz:loc; bÓl:loc
település: place_On
teljesítménye: thing
tengelyszélesség: place_On
tenyér: body_bAn-On
tér: way
térd: body_bAn-On -
On:body_bAn-On, mode;
rA:body_bAn-On, mode
térfél: place_On
termelés: thing
terminál: place_On
térség: place_bAn
térszintje: thing
terület: place_On
terv: thing

DOI: 10.15774/PPKE.BTK.2021.17

test: thing
testület: org_bAn
tetőszerkezet: thing
tevékenység: thing
Tevje-alakítása: thing
tézis: thing - bAn:loc;
nÁl:loc; bÓl:loc
Thaiföld: loc_country
Tiszabecs: loc_city-On
tisztázás: thing - rA:cause
tiszt: posi_On
tisztség: posi_bAn
titok: thing – bAn:circumst
tizede: num_size
tizenötszöröse: num_size
több: num
többség: thing -
bAn: Milyen arányban?;
nÁl: Mennyinél?;
Az értintettek mekkora részénél?
többsége: part - bAn:Milyen arányban?;
bA:thing
tőke: thing
tőkeemelés: thing
tőketartalék: thing
tőketörlesztés: thing
Tokió: loc_city-bAn
tömkelege:
tonna: meas –
bAn:meas, Hány tonnában?; thing;
rA:meas, Hány tonnára?
törlesztés: thing
torna: event_On
Törökország: loc_country
torta: loc_any – hOz:loc_any, thing;
bÓl:thing
történet: thing – bAn:thing, loc, event
törvény: loc_bAn
továbbítása: thing
tőzsde: place_On
Trafó: place_bAn
traktor: thing
transzport: thing
triumfálás: thing
tulajdona: thing – bA:thing, Kihez?
túldimenzionálás: thing
tüntetés: event_nÁl-On
udvar: loc_any
ügy: thing
ügyészség: build=inst_bAn
újságcikk: thing
újságok: place_bAn
ülés: event_nÁl-On
ülésterem: place_bAn
unió: thing
universitas: loc_bAn
ünnepek: thing
ünnepség: event_On
út: loc_any
utasfülke: place_bAn
utasítás: thing – rA:thing, cause
utca: way – bAn:loc_any
ütés: thing
úthálózat: thing
útja: loc_any
utóbbi: dem
utód: who
utódállamai: loc_bAn
útszakasz: way
útvesztő: place_bAn -
bÓl:place_bAn, thing
üzem: build=inst_bAn
üzemanyagárak: thing
vacsora: thing
vagyon: thing
választmány: thing
váll: body_any
vállalatok: org_bAn

DOI: 10.15774/PPKE.BTK.2021.17

vállalkozás: thing
válogatott: group - bA:group, thing
valóság: thing
válság: thing – bAn:thing, state;
bA:thing, state;
bÓl:thing, state
változás: thing
valuta: thing
várakozás: thing
város: place_bAn
várt: thing
vásárló: who
vasútvonal: way
Vatikán: loc_On
Vb-címmérkőzés: event_On
védekezés: thing – bAn:thing, loc;
nÁl:thing, event
vég: part - On:part, date;
rA:part, date;
tÓl:part, date
végelszámolás: thing
végrehajtás: thing
vendégjáték: thing
Ventspils: loc_city-bAn
versengés: thing
verseny: event_On –
bAn:event_On, circumst;
nÁl:event_On, thing
versenynap: date_On
versenyző: who – bÓl:who, thing
veszély: thing
vészhelyzet: event_bAn – bA:thing;
bÓl:event_bAn, thing; tÓl:thing;
rÓl:thing
vétel: thing
vezeték: thing
vezetés: thing - bAn:thing, posi
vezető: who
vidék: loc_On - tÓl:thing
videotéka: build=inst_bAn – rA:thing;
rÓl:thing
világ: loc_bAn – On:loc_any
világháló: thing
világkiállítás: event_On
világpiac: loc_On
világviszonylat: pov
virradó: thing – rA:Mikor?, date_On
viselkedés: thing
visszahelyezés: thing
viszony: thing – bAn:thing, loc;
Milyen kapcsolatban?
vita: event_nÁl-On – bAn:event_nÁl-On,
event_bAn
vizsgálat: thing
Vjahirev: who
Welteke: who
Würzburg: loc_city-bAn
zárójel: thing – bAn:thing, mode
zászló: loc_any – bÓl:thing
zene: thing
Zeneakadémia: build=inst_On
zöme: part – bAn:Milyen arányban?;
bA:thing
zseb: loc_bAn – bÓl:loc_bAn, mode

DOI: 10.15774/PPKE.BTK.2021.17

References

Abney, S. P. (1992). Parsing by Chunks. In R. C. Berwick, S. P. Abney, and C. Tenny (Eds.) *Principle-Based Parsing: Computation and Psycholinguistics*, (pp. 257–278). Dordrecht: Springer Netherlands.

Alberti, G., and Laczkó, T. (2018). *Syntax of Hungarian: Nouns and Noun Phrases*. Comprehensive Grammar Resources. Amsterdam University Press.
URL https://books.google.hu/books?id=9v_uAQAACAAJ

Antal, L. (1961). A magyar esetrendszer [The Hungarian case system]. *Nyelvtudományi Értekezések*, *29*.

Asbury, A. (2008a). Marking of semantic roles in Hungarian morphosyntax. In *Approaches to Hungarian 10: Papers from the Veszprém Conference*, (pp. 9–30). Akadémiai Kiadó.

Asbury, A. (2008b). *The Morphosyntax of Case and Adpositions*. PhD thesis, University of Utrecht.

Bartos, H. (2001). Mutató névmási módosítók a magyarban: egyezés vagy osztozás? [Demonstrative modifiers in Hungarian: agreement or feature sharing?]. *Újabb tanulmányok a strukturális magyar nyelvtan és a nyelvtörténet körébol. Kiefer Ferenc tiszteletére barátai és tanítványai [Recent studies in Hungarian structural grammar and diachroic linguistics. In honour of Ferenc Kiefer, from his friends and students]*, (pp. 19–41).

Csendes, D., Csirik, J., and Gyimóthy, T. (2004). The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In P. Sojka, I. Kopeček, and K. Pala (Eds.) *Text, Speech and Dialogue*, (pp. 41–47). Berlin, Heidelberg: Springer Berlin Heidelberg.

Csendes, D., Csirik, J., Gyimóthy, T., and Kocsor, A. (2005). The Szeged Treebank. In V. Matousek, and et al. (Eds.) *Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005, LNAI 3658*, (pp. 123–131). Springer Verlag.

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

Dékány, É. (2009). The nanosyntax of Hungarian postpositions. In *Tromsø Working Papers on Language and Linguistics: Nordlyd 36.1*, vol. 36, (pp. 45–71). Norway: Septentrio Academic Publishing.

Dékány, É., and Hegedűs, V. (2013). Word order variation in Hungarian PPs. In *Approaches to Hungarian 14: Papers from the 2013 Piliscsaba conference*, (pp. 95–120).

DOI: 10.15774/PPKE.BTK.2021.17

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Domokos, L., and Benedek, M. (Eds.) (1795). *Magyar Grammatika, mellyet készített Debreczenbenn egy Magyar Társaság. A Magyar Hírmondó íróinak költségével, Alberti betőivel [Hungarian grammar prepared by a Hungarian Society in Debrecen. Financed by the writers of the Magyar Hírmondó]*. Bécs.

Dömötör, A. (2017). Hány VAN nincs? A létige - zéró váltakozás korpuszvezérelt vizsgálata [How many IS is not there? Corpus-driven study of copula - zero alternation]. In Zs. Ludányi (Ed.) *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2017: XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia*, (pp. 28–38).

Dömötör, A. (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, and T. E. Gráczi (Eds.) *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2018*, (pp. 3–10).

É Kiss, K. (1990). Mi tartozik a névutók osztályába? [What belongs to the category of postpositions?]. *Magyar Nyelvjárások*, *37*, 167–172.

É. Kiss, K. (2000). The Hungarian NP is like the English NP. In *Approaches to Hungarian VII.*, (pp. 119–150). Szeged: University of Szeged Press.

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

É. Kiss, K. (2020). Nem-részvétel vagy nemlét: Két abesszívuszi névutó a magyarban [Non- participation vs. non-existence: two abessive postpositions in Hungarian]. *Nyelvtudományi Közlemények*, *116*, 139–153.

É. Kiss, K., and Hegedűs, V. (Eds.) (2021). *Syntax of Hungarian: Postpositions and Postpositional Phrases*. Comprehensive Grammar Resources. Amsterdam University Press.

Endrédy, I. (2014). Corpus driven research: ideas and attempts. In *PhD Proceedings Annual Issues of the Doctoral School - 2014. Faculty of Information Technology and Bionics, Pázmány Péter Catholic University*, (pp. 137–140).

Endrédy, I., and Indig, B. (2015). HunTag3: a general-purpose, modular sequential tagger – chunking phrases in English and maximal NPs and NER for Hungarian. In *7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC '15)*, (pp. 213–218). Poznań, Poland: Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu.

Endrédy, I., and Novák, A. (2013). More Effective Boilerplate Removal - the GoldMiner Algorithm. *Polytech. Open Libr. Int. Bull. Inf. Technol. Sci.*, *48*, 79–83.

Endrédy, I. (2016). *Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz [Language technology algorithms for automatic corpus building and more precise data processing]*. PhD thesis, Pázmány Péter Catholic University Faculty of Information Technology and Bionics, Roska Tamás Doctoral School of Sciences and Technology.

DOI: 10.15774/PPKE.BTK.2021.17

Endrédy, I., and Prószéky, G. (2016). A Pázmány Korpusz [The Pázmány Korpusz]. *Nyelvtudományi Közlemények*, *121*, 191–205.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA).
URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/109.pdf

Fogarasi, J. (1843). *Müvelt magyar nyelvtan*. Pest: Heckenast.

Frazier, L., and Fodor, J. D. (1978). The Sausage Machine: A New Two-stage Parsing Model. *Cognition*, *6*(4), 291–325.

Földi, J. (1912). *Földi János magyar grammatikája (1790)*. No. 28 in Régi Magyar Könyvtár. Budapest: MTA. Published by Károly Gulyás.

Galgóczi, G. (1848). *Magyar nyelvtan*. Pest.

Gyarmathy, S. (1794). *Okoskodva tanító magyar nyelvmester [Hungarian language master teaching argumentatively]*. Kolozsvár.

Hegedűs, V. (2006). Hungarian spatial PPs. In *Nordlyd: Tromsø University Working Papers in Linguistics*, vol. 33, (pp. 220–233). Norway: Septentrio Academic Publishing.

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

Hong Shen and Anoop Sarkar (2005). Voting Between Multiple Data Representations for Text Chunking. In *Canadian Conference on AI*.

Indig, B., Laki, L. J., and Prószéky, G. (2016a). Mozaik nyelvmodell az AnaGramma elemzőhöz. In A. Tanács, V. Varga, and V. Vincze (Eds.) *XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016)*, (p. 260–270). Szegedi Tudományegyetem, Szeged: Szegedi Tudományegyetem.
URL http://real.mtak.hu/34384/

Indig, B., Vadász, N., and Kalivoda, Á. (2016b). Decreasing Entropy: How Wide to Open the Window? In C. Martín-Vide, T. Mizuki, and M. A. Vega-Rodríguez (Eds.) *Theory and Practice of Natural Computing: 5th International Conference, TPNC 2016, Sendai, Japan, December 12-13, 2016, Proceedings*, (p. 137–148). Springer International Publishing, Cham: Springer International Publishing.

Jakab, I. (1977). Az értelmező és az értelmezett szószerkezeti viszonya. *Magyar Nyelvőr*, *101*, 9–19.

Jakab, I. (1978). Igen, az értelmező mellérendelés. *Magyar Nyelvőr*, *102*, 293–298.

Károly, S. (1958). Az értelmező és az értelmezői mondat a magyarban. *Nyelvtudományi Értekezések*, *16*.

Károly, S. (1962). Az értelmező jelző. In J. Tompa (Ed.) *A mai magyar nyelv rendszere II.*, (pp. 295–314). Budapest: Akadémiai Kiadó.

DOI: 10.15774/PPKE.BTK.2021.17

Kenesei, I., Fenyvesi, A., and Vago, R. M. (1997). *Hungarian*. Descriptive Grammar Series. London: Routledge.

Keszler, B. (Ed.) (2000). *Magyar grammatika [Hungarian grammar]*. Nemzeti Tankönyvkiadó. Original document in Hungarian.

Kiefer, F. (1987). The cases of Hungarian nouns. *Acta Linguistica Academiae Scientiarum Hungaricae*, *37*(1/4), 93–101.
URL http://www.jstor.org/stable/44362762

Kiefer, F. (1992). *Strukturális magyar nyelvtan: Mondattan [A Structural Grammar of Hungarian: Syntax]*. Strukturális magyar nyelvtan [A Structural Grammar of Hungarian]. Akadémiai Kiadó. Original document in Hungarian.
URL https://books.google.hu/books?id=ht6hQgAACAAJ

Kiefer, F. (2000a). A ragozás. In F. Kiefer (Ed.) *Strukturális magyar nyelvtan 3. Morfológia*, (pp. 569–618). Akadémiai Kiadó.

Kiefer, F. (2000b). *Strukturális magyar nyelvtan: Morfológia [A Structural Grammar of Hungarian: Morphology]*. Strukturális magyar nyelvtan [A Structural Grammar of Hungarian]. Akadémiai Kiadó. Original document in Hungarian.

Komáromi Csipkés, G. (2008). *Hungaria illustrata. A magyar nyelv magyarázata [Explanation of Hungarian]*. No. 228 in A Magyar Nyelvtudományi Társaság Kiadványai. Budapest. Facsimile publication of the original document with translation. Translated with introduction and notes by Zsuzsa C. Vladár. Edited by Éva Zsilinszky.

Kornai, A. (1985). The internal structure of noun phrases. In *Approaches to Hungarian I.*, (pp. 79–92). Szeged: University of Szeged Press.

Kövesdi, P. (2010). *Elementa Linguae Hungaricae. A magyar nyelv alapjai [The basics of Hungarian]*. No. 232 in A Magyar Nyelvtudományi Társaság Kiadványai. Budapest. Facsimile publication of the original document with translation. Translated with introduction and notes by Zsuzsa C. Vladár. Edited by Éva Zsilinszky.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.
URL http://dx.doi.org/10.1093/bioinformatics/btz682

Ligeti-Nagy, N. (2015). Szövegkorpuszok pontosabb annotációja gépi elemzéshez [Improving Corpus Annotation for Automatic Parsing]. In A. Benő, E. Fazekas, and B. Zsemlyei (Eds.) *Többnyelvűség és kommunikáció Kelet-Közép-Európában. XXIV. Magyar Alkalmazott Nyelvészeti Kongresszus [Multilingualism and Communication in Eastern-Central-Europe. Proceedings of the 24th Congress of Hungarian Applied Linguistics]*, (pp. 421–429). Kolozsvár: Erdélyi Múzeum-Egyesület.

Ligeti-Nagy, N. (2016). A főnévi csoportok és ami utánuk marad: Automatikus szintagmakinyerés magyar nyelvű szövegekből [Noun Phrases and What They Leave Behind: Automatic Phrase Extraction from Hungarian Corpora]. In A. Á. Reményi, Cs. Sárdi, and Zs. Tóth (Eds.) *Távlatok a mai magyar alkalmazott nyelvészetben*, (pp. 249–260). Budapest: Tinta Könyvkiadó.

DOI: 10.15774/PPKE.BTK.2021.17

Ligeti-Nagy, N. (2018). Névutók, előre! Korpuszvezérelt elemzés a névutószerű elemekről [Postpositions, come forward! Corpus-driven study on postposition-like elements]. In V. Vincze (Ed.) *XIV. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 52–63).

Ligeti-Nagy, N., Dömötör, A., and Vadász, N. (2019). What does the Nom say? An algorithm for case disambiguation in Hungarian. In *The 5th International Workshop on Computational Linguistics for Uralic Languages. Proceedings of the Workshop*, (pp. 27–41). Tartu, Estonia: ACL SIG for Uralic Languages.
URL https://www.aclweb.org/anthology/W19-0303

Ligeti-Nagy, N., and Novák, A. (2019). Hol ugat a kutya? Örömében. Helyhatározói esetragos névszók pontosabb annotációja [Where does the dog bark? In his joy. More accurate annotation of adverbial nominals]. In *XV. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 225–234). Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport.

Ligeti-Nagy, N., Vadász, N., Dömötör, A., and Indig, B. (2018). Nulla vagy semmi? Esetegyértelműsítés az ablakban [Zero or Nothing? Case Disambiguation in the Window]. In V. Vincze (Ed.) *XIV. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 25–37).

Marácz, L. (1986). Dressed or Naked: the Case of the PP in Hungarian. In *Topic, Focus and Configurationality. Papers from the 6th Groningen Grammar Talks, Groningen, 1984*, (pp. 227–252).

Mihăltz, M. (2011). Magyar NP-felismerők összehasonlítása [The comparison of Hungarian NP chunkers]. In A. Tanács, and V. Vincze (Eds.) *VIII. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 333–335).

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, (pp. 1–12).

Nemeskey, D. M. (2020). *Natural Language Processing methods for Language Modeling*. PhD thesis, Eötvös Loránd University.

Nemeskey, D. M. (2021). Introducing `huBERT`. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, (pp. 3–14). Szeged.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In N. (Conference Chair). Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (Eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Novák, A. (2003). Milyen a jó Humor? [What is Good Humor Like?]. In *I. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 138–144). Szeged: SZTE.

Novák, A. (2014). A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In N. Calzolari, 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*

DOI: 10.15774/PPKE.BTK.2021.17

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

Novák, A., Laki, L. J., Novák, B., Dömötör, A., Ligeti-Nagy, N., and Kalivoda, Á. (2019a). Creation of a corpus with semantic role labels for Hungarian. In A. Friedrich, D. Zeyrek, and J. Hoek (Eds.) *Proceedings of the 13th Linguistic Annotation Workshop, LAW@ACL 2019, Florence, Italy, August 1, 2019*, (pp. 220–229). Association for Computational Linguistics.
URL https://doi.org/10.18653/v1/w19-4026

Novák, A., Laki, L. J., Novák, B., Dömötör, A., Ligeti-Nagy, N., and Kalivoda, Á. (2019b). Egy magyar nyelvű kérdezőrendszer [A Questioning System for Hungarian]. In *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019)*. Szeged: SZTE. Original document in Hungarian.

Novák, A., Siklósi, B., and Oravecz, C. (2016). A New Integrated Open-source Morphological Analyzer for Hungarian. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (Eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Novák, A., Siklósi, B., and Wenszky, N. (2017). Szóbeágyazási modellek vizualizációjára és böngészésére szolgáló webes felület [Online interface for the visualization and browsing of word embedding models]. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 355–362).

Oravecz, Cs., Váradi, T., and Sass, B. (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).

Pereszlényi, P. (2006). *Grammatica Lingvae Ungaricae. A magyar nyelv grammatikája [Grammar of Hungarian]*. No. 226 in A Magyar Nyelvtudományi Társaság Kiadványai. Budapest. Facsimile publication of the original document with translation. Translated with introduction and notes by Zsuzsa C. Vladár. Edited by Éva Zsilinszky.

Pomázi, B. (2018). Az *N-nak köszönhetően* mint névutói szerkezet [*N-nak köszönhetően* as a postpositional phrase]. In *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből*, (pp. 58–69).

Prószéky, G., and Indig, B. (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, G., Indig, B., Miháltz, M., and Sass, B. (2014). Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé [Towards a psycholinguistically motivated parser]. In *X. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 79–87). Szegedi Tudományegyetem.

DOI: 10.15774/PPKE.BTK.2021.17

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

Prószéky, G., Tihanyi, L., and Ugray, G. (2004). Moose: a robust high-performance parser and generator. In J. Hutchins (Ed.) *Proceedings of the 9th EAMT Conference. La Valletta: Foundation for International Studies*, (pp. 138–142).

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Ramshaw, L., and Marcus, M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.
URL https://www.aclweb.org/anthology/W95-0107

Recski, G. (2010a). Főnévi csoportok azonosítása szabályalapú és hibrid módszerekkel. In A. Tanács, and V. Vincze (Eds.) *VII. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 333–341).

Recski, G. (2010b). *NP Chunking in Hungarian*. Master thesis, Eötvös Loránd University of Science.

Recski, G. (2014). Hungarian noun phrase extraction using rule-based and hybrid methods. *Acta Cybernetica*, *21*(3), 461–479.
URL https://cyber.bibl.u-szeged.hu/index.php/actcybern/article/view/3855

Recski, G., Rung, A., Zséder, A., and Kornai, A. (2010). NP alignment in bilingual corpora. In N. Calzolari (Ed.) *Proc. 7th International Conference on Language Resources and Evaluation (LREC'10)*.
URL http://eprints.sztaki.hu/6420/

Recski, G., and Varga, D. (2012). Magyar főnévi csoportok azonosítása. *Általános Nyelvészeti Tanulmányok*, *24*.

Recski, G. A., Varga, D., Zséder, A., and Kornai, A. (2009). Főnévi csoportok azonosítása magyar-angol párhuzamos korpuszban. In A. Tanács, D. Szauter, and V. Vincze (Eds.) *VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2009)*, (pp. 3–13). Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport.
URL http://eprints.sztaki.hu/6293/

Riedl, S. (1866). *Kisebb magyar nyelvtan [A small Hungarian grammar]*. Pest: Pfeifer Ferdinánd.

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *RASLAN*.

Révai, M. (1806). *Elaboratior Grammatica Hungarica*. Pest.

Sarma, B., and Barman, A. (2015). A Comprehensive Survey of Noun Phrase Chunking in Natural Languages. *International Journal of Engineering Research and Technical Research*, *V4*.

DOI: 10.15774/PPKE.BTK.2021.17

Sass, B. (2009). "Mazsola" - eszköz a magyar igék bvítményszerkezetének vizsgálatára. In T. Váradi (Ed.) *Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia eladásaiból*, (pp. 117–129). Budapest: MTA Nyelvtudományi Intézet.

Sebestyén, Á. (1965). *A magyar nyelv névutórendszere [The postpositional system of Hungarian]*. Budapest: Akadémiai Kiadó.

Siklósi, B., and Novák, A. (2016a). Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra [Using word embedding models for lexical categorization]. In A. Tanács, V. Varga, and V. Vincze (Eds.) *XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016)*, (pp. 3–14).

Siklósi, B., and Novák, A. (2016b). Közeli rokonunk, az autó [Our close relatives, cars]. In A. Tanács, V. Varga, and V. Vincze (Eds.) *XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016)*, (pp. 27–36).

Simon, E. (2008). Nyelvészeti problémák a tulajdonnév-felismerés területén [Linguistic issues in Named Entity Recognition]. In B. Sinkovics (Ed.) *LingDok 7. Nyelvész-doktoranduszok dolgozatai*, (pp. 181–196). Szeged: Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola.

Simon, E. (2013). *Approaches to Hungarian Named Entity Recognition*. PhD thesis, PhD School in Psychology – Cognitive Science. Budapest University of Technology and Economics.

Simon, E. (2017). The Definition of Named Entities. In B. Gyuris, K. Mády, and G. Recski (Eds.) *K + K = 120. Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*. Budapest: MTA Nyelvtudományi Intézet (Research Institute for Linguistics, Hungarian Academy of Sciences). Http://nytud.hu/kk120.

Simon, E., Farkas, R., Halácsy, P., Sass, B., Szarvas, G., and Varga, D. (2006). A HunNER korpusz. In *V. Magyar Számítógépes Nyelvészeti Konferencia*.

Sundheim, B. (1995). Appendix C: Named Entity Task Definition (v2.1). In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.
URL https://www.aclweb.org/anthology/M95-1024

Szabolcsi, A. (1981). The Possessive Construction in Hungarian: A Configurational Category in a Non-configurational Language. *Acta Linguistica Academiae Scientiarum Hungaricae*, *31*(1/4), 261–289.
URL http://www.jstor.org/stable/44310514

Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., and Csirik, J. (2006). A highly accurate Named Entity corpus for Hungarian. In *5th International Conference on Language Resources and Evaluation, LREC 2006 ; Conference date: 22-05-2006 Through 28-05-2006*, (pp. 1957–1960).

Szenczi Molnár, A. (2004). *Novae Grammaticae Ungaricae libri duo. Új magyar grammatika két könyvben [New Hungarian grammar in two books]*. No. 222 in A Magyar Nyelvtudományi Társaság Kiadványai. Budapest. Facsimile publication of the original document with translation. Translated with introduction and notes by Zsuzsa C. Vladár. Edited by Éva Zsilinszky.

DOI: 10.15774/PPKE.BTK.2021.17

Szőke, B. (2015). *Az értelmezős szerkezetek vizsgálata a magyar nyelvben [Analysis of appositional constructions in Hungarian]*. Ph.D. thesis, University of Szeged Faculty of Arts, Doctoral School in Linguistics.

Szvorényi, J. (1866). *Magyar nyelvtan tanodai s magánhasználatra*. Pest.

Tjong Kim Sang, E. F., and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, (pp. 142–147). URL https://www.aclweb.org/anthology/W03-0419

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. URL https://www.aclweb.org/anthology/W02-2024

Tjong Kim Sang, Erik. F., and Buchholz, S. (2000). Introduction to the CoNLL-2000 Shared Task: Chunking. *CoRR, cs.CL/0009008*. URL https://arxiv.org/abs/cs/0009008

Vadász, N. (2017). Anaforafeloldás menet közben – névmások egy pszicholingvisztikailag motivált elemzőben [Anaphora resolution on the fly – pronouns in a psycholinguistically motivated parser]. In Zs. Ludányi (Ed.) *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2017. XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia*, (pp. 192–205).

Vadász, N., and Indig, B. (2017). A birtokos esete az ablakkal. In G. Scheibl (Ed.) *LingDok: nyelvész-doktoranduszok dolgozatai*. Szegedi Tudományegyetem. Megjelenés alatt.

Vadász, N., Kalivoda, Á., and Indig, B. (2017). Ablak által világosan – Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017)*, (pp. 3–12). Szegedi Tudományegyetem Informatikai Intézet, Szeged: Szegedi Tudományegyetem Informatikai Intézet. URL http://rgai.inf.u-szeged.hu/project/mszny2017/files/kotet.pdf

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

Váradi, T., and Gábor, K. (2004). A magyar Alexin fejlesztéséről [On developing the Hungarian Intex module]. In Z. Alexin (Ed.) *II. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 3–10).

Varga, D. (2012). *Natural Language Processing of Large Parallel Corpora*. PhD thesis, Eötvös Loránd University.

Varga, D., and Simon, E. (2006). Magyar nyelvű tulajdonnév-felismerés maximum entrópia módszerrel [Hungarian Named Entity Recognition with a Maximum Entropy Approach]. In Z. Alexin, and D. Csendes (Eds.) *IV. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 32–38).

Verseghy, F. (1816). *Analyticae Institutionum Linguae Hungaricae*. Buda.

DOI: 10.15774/PPKE.BTK.2021.17

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., and Csirik, J. (2010). Hungarian Dependency Treebank. In N. Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias (Eds.) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Váradi, T. (2002). The Hungarian National Corpus. In *Proceedings of the 3rd LREC Conference*, (pp. 385–389). Las Palmas, Spain.

Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholtz, I., Novák, A., Indig, B., Prószéky, G., and Vincze, V. (2017). Az e-magyar digitális nyelvfeldolgozó rendszer [E-magyar – A Digital Language Processing System]. In V. Vincze (Ed.) *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*, (pp. 49–60).

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B., Farkas, R., and Vincze, V. (2018). E-magyar – A Digital Language Processing System. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga (Eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA).

Zeliger, E. (Ed.) (1989). *Sylvester János latin-magyar nyelvtana (1539) [Hungarian-Latin grammar of János Sylvester].* Budapest.

DOI: 10.15774/PPKE.BTK.2021.17

Összefoglaló – Abstract in Hungarian

Disszertációmban négy, a magyar főnévi csoportok végződésével kapcsolatos nyelvi jelenséget vizsgáltam, melyeknek közös tulajdonsága, hogy magyar nyelvű szövegek számítógépes elemzése során kezelésük valamilyen szempontból kritikusnak bizonyult. A dolgozat alcíme a számítógépes megközelítést vetíti előre, de kiemelendő, hogy az itt bemutatott kutatásokat a korpusz-vezéreltség is jellemzi.

A számítógépes megközelítés kiindulópontja egy elemző, az `AnaGramma` (Prószéky and 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 részkutatások mind az `AnaGramma` elveit szem előtt tartva készültek.

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 következő jelenségekkel foglalkoztam:

- 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, mint amilyen az *Angela Merkel kancellár*.
- A főnévi csoportok végét jelző elemek:
  - Helyhatározói esetragok: a mondatban betöltött szabad határozói szerepük alapján történő kategorizációt mutattam be.
  - A magyar névutók: a szakirodalom olykor egymásnak ellentmondó szempontrendszerének bemutatása után hat disztribúciós tulajdonság alapján csoportosítottam a névutókat.

DOI: 10.15774/PPKE.BTK.2021.17

Noha az itt felsorolt témák igen különbözőek, mind a vizsgálatukat igénylő módszerek, mind az általuk érintett nyelvi szint szempontjából, kutatásuk és mélyebb megértésük minden számítógépes elemző számára esszenciális.

A dolgozatot egy bevezetővel kezdtem, majd ismertettem a kutatásom motivációját. Ennek keretében került sor az `AnaGramma` alapelveinek és működésének a bemutatására, majd az általam használt korpuszok bemutatására (1. fejezet). Ezt követően először az esetrag nélküli névszók vizsgálatát ismertettem. Bemutattam az ezek mondatbeli szerepének egyértelműsítésére írt algoritmusom tervezésének és implementálásának folyamatát (2. fejezet). Az itt bemutatott kutatás során ébredtem rá az úgynevezett *extended named entitites* (XNE) fontosságára, így a dolgozat 3. fejezete ezekre koncentrált. Végül rátértem azokra a morfémákra, amelyek minden kétséget kizáróan jelzik egy főnévi csoport végét: az esetragok (4. fejezet) és a névutók (5. fejezet). Végül a konklúzióban (6. fejezet) az eddigiek összefoglalásán túl pontokba szedve felsoroltam kutatásom legfontosabb eredményeit.