Andrea Dömötör

Characteristics and computational processing of zero substantive verbs

Doctoral (PhD) thesis

Theses

Pázmány Péter Catholic University

Faculty of Humanities and Social Sciences

Doctoral School of Linguistics

Doctoral Programme in Language Technology

Supervisor

Prof. Gábor Prószéky

Professor, Doctor of Sciences

Budapest, 2025

1. Aims of the thesis

The dissertation examines the cases of Hungarian copular, locative, and existential sentences where the substantive verb does not (necessarily) have an overt form in the morphologically unmarked case (i.e. indicative mood, present tense, 3rd person singular). This includes the phenomenon of the so-called zero copula, as well as those non-nominal sentences where the verbal element *van* ('be') can be optionally omitted (e.g. *A kulcs a lábtörlő alatt (van).* 'The key (is) under the doormat.'). While the zero copula is a well-known and described phenomenon in Hungarian lingusistic studies, the optional omission of *van* is still rarely mentioned.

The dissertation follows a computational and data-driven approach. It focuses on determining the aspects and conditions of the optional omission of the substantive verb based on corpus data, and discusses the challenges of the automatic processing of sentences without an overt verbal element.

2. Methods and data

The dissertation relies on the aspects of corpus linguistics in both its methods and objectives. The emerge of increasingly diverse and large corpora enables the analysis of previously unimaginable amounts of text data, which contributes to a better understanding of linguistic theories, or even to the discovery of new or yet unexplored phenomena. The use of corpora has now become common in theoretical linguistics (Teubert, 2005), and it is also essential in machine learning-based natural language processing developments.

There are two main approaches of using text data in corpus linguistics. The corpus-based approach tests pre-existing hypotheses on the data by examining frequencies or distributions, while the corpus-driven approach has little or none prior assumptions, all conclusions are drawn from the corpus data (Tognini-Bonelli, 2001). The research presented in the dissertation is largely characterized by the former approach. The corpus-based experiments of the dissertation examine the occurrences of zero substantive verbs, the lexical and distributional characteristics of locative and existential sentences, and the structure (word order variants) of copular sentences, typically based on existing theories.

2.1. Corpora

- **HGC** (Oravecz et al., 2014)

  The Hungarian Gigaword Corpus contains 1.04 billion text words from 6 domains and 5 dialects. The texts were processed by a pipeline containing tokenizer, morphological analyzer and POS-tagger modules. The query interface of the corpus is part of the Sketch Engine framework.

- **Szeged Treebank** (Vincze et al., 2010)

  The Szeged Treebank is the first manually annotated corpus for Hungarian that includes dependency annotations. The original Szeged Treebank consists of 82,000 sentences (1.2 million words), which come from 6 different text types. The most important feature of the corpus from the dissertation's point of view is that it contains empty verb heads to mark, among others, zero copulas.

- **OPUS OpenSubtitles** (Lison és Tiedemann, 2016)

  OPUS OpenSubtitles is a parallel corpus of movie subtitles available in 62 languages, containing a total of 22.1 billion words. Novák et al. (2019) created a lemmatized, morphologically analyzed and word-level matched version of the Hungarian-English part of this corpus, which consists of 644.5 million tokens of English texts and their Hungarian translations. This was used for the creation of training data of the zero copula insertion tool described in Chapter 3.

3. The structure and main theses of the dissertation

The main focus of the dissertation is the conditions of optional omission of the substantive verb, based on real linguistic data acquired from corpora. Besides, the dissertation discusses some syntactic features of nominal sentences and presents an experiment to develop an automatic tool that inserts the zero copula in the input sentence.

Following the introduction, Chapter 2 summarizes the methods and theoretical approaches used in the dissertation. These include the characteristics of corpus-based and corpus-driven methods, machine learning, evaluation metrics, and dependency parsing. The chapter also presents the corpora and text processing tools that were used for the research of the dissertation.

Chapter 3 describes the development of an automatic text processing tool that is capable of recognizing sentences with zero copula and even inserts the zero copula to the correct place in the sentences. The base of the system is a parallel corpus, where sentences with zero copula contain a special tag indicating the place of the copula in the target side of the corpus. Using this as training material, we trained a neural machine translator to insert the zero copula tag into the input sentences. Although the accuracy of the tool did not turn out high enough for corpus building, the research has provided several useful insights.

Chapter 4 examines those "zero substantive verbs" in the corpora that do not form a nominal sentence. The chapter analyses three groups of phenomena in detail. Chapter 4.2 focuses on sentences of type *Ott (van) az ajtó.* ('There (is) the door') and seeks answer to the question of why the presence of demonstrative elements referring to place (*itt, ott, hol* 'here, there, where') supports the omission of the locative verb *van*. Chapter 4.3 reveals that the optional omission in certain sentence types does not limit to *van*, it can also happen to its negative version, *nincs/sincs*. This subsection examines the lexical features of sentences of type *(Nincs) semmi gond.* ('(There is) nothing wrong'). Finally, section 4.4 focuses on the omission of the modal auxiliary verb *lehet* ('can/may'), i.e. in which cases is it possible to omit the auxiliary from the auxiliary + infinitive constructions (*Innen jól (lehet) látni.* '(You can) see it from here.').

Chapter 5 deals with some structural (word order) issues of sentences with copula. Chapter 5.1 presents cases where the structure of zero-copula and overt-copula sentences differ from each other, and therefore omitting the copula results in an incorrect sentence. Chapter 5.2 discusses the

question of the position of predicative nominals: in which cases are they typically focused and when is the neutral word order preferred.

Finally, Chapter 6 examines the (auxiliary) verbal features of predicative nominals in the light of two word order issues (the detachment of preverbs and the position of the *-e* interrogative particle), and the adverbial and infinitive arguments of nominals. In the latter case, the chapter also discusses predicative nominals that are already on the path of becoming auxiliary verbs (*kell, muszáj, szabad* 'must, have to, may').

In summary, the main theses of the dissertation are as follows:

1. From an English-Hungarian parallel corpus, I created a Hungarian text corpus in which zero copulas are marked with a special tag. Using this as training data, my colleagues and I trained a neural machine translator to insert the zero copula into the input sentence. This zero copula inserting tool can be useful for computational syntax, as dependency parsing models (e.g., Szeged Treebank) often use empty verb heads for the analysis of nominal sentences without overt copula. The tool we developed achieved an accuracy of nearly 90% on texts from the same source as the training data. However, when tested on other text types, we obtained much lower results. This is due to the fact that the texts in the training corpus are limited to a single genre (movie subtitles) due to the special methodology. In order to improve the general performance of the zero copula insertion tool the training corpus should be extended with texts from other domains.

2. I showed through corpus measurements that the phenomenon of zero substantive verb in Hungarian does not limit to nominal sentences, it also affects other types of substantive verbs that can be optionally omitted in the third person present tense under certain conditions. A significant part of these are idioms (e.g. *rendben/vége/tele van* 'It is all right/over/full', etc.) or genre-specific phenomena (e.g. descriptions in fiction). However, some sentence types containing optionally omitted substantive verbs turned out to be productive, such as sentences containing a demonstrative element referring to a place (*itt/ott/hol* 'here/there/where') or a negative expression (*semmi/sehol* 'nothing/nowhere').

3. My data obtained from HGC support the hypothesis that the verb modifier *itt/ott* ('here/there') weakens the descriptive content of the modified verb, so that it actually becomes a (somewhat more specific) synonym of the verb of existence. É. Kiss (2004) If the modified verb is the substantive verb (*van*) itself, then it becomes semantically redundant and its function reduces to a copula-like inflection-holder. Therefore, if the sentence is in the present tense, 3rd person singular (i.e. there is no marked inflection), the substantive verb can have a zero form. This provides a theoretical explanation for the phenomenon that locative and existential verbs can optionally be omitted from sentences where they are preceded by a pronoun referring to a place (*itt/ott/hol* 'here/there/where').

4. The negative forms of the verbs of existence (*nincs, sincs*) can only be omitted from the sentence if it contains some other expression with a negative meaning (*semmi/sehol/se* 'nothing/nowhere/neither'). Based on corpus data, I showed that in the case of certain *semmi* + noun expressions, the verbless construction is particularly common, for example: *semmi baj/gond/szükség/ok/kétség* ('no problem/trouble/need/doubt').

5. In certain cases, the auxiliary verb (*lehet* 'can/may') can also be omitted from sentences. Based on corpus frequency studies, I came to the conlcusion that the infinitive construction implying modality without an auxiliary verb is characteristic for experient subjects and multi-argument (primarily transitive) verbs, mostly expressing mental or physical perception.

6. Based on corpus data, I showed that the traditional "lexical verb" – "functional element" distinction does not always explain the omission of the verb *van*. In some cases, the same lexical element occurs both with and without a substantive verb. This may be due to part-of-speech ambiguity (*kész, tele*, 'ready, full'), or reference to quantity (*sok, kevés* 'many, few') or state (*rendben* 'fine'). In the case of nominals referring to an environmental situation (*hideg, meleg, sötét* 'cold, warm, dark'), the presence of the verb is necessary when there is no (explicit) subject in the sentence. In the case of *késő* ('late'), a semantic difference can be observed between the constructions with and without a verb: the former usually refers to the time of day, and the latter to a delay.

7. Regarding the structure of sentences with copula, I proved using questionnaire and corpus-based methods that nominal predicates, which are usually in verb-modifier position, may be focused if their meaning can be defined as multi-valued, i.e. they can activate predicates with alternative meanings. Typical multi-valued predicates are colors, numeral expressions, and nominal predicates denoting origin, community membership, or occupation. In contrast, binary-meaning adjectives (which do not have alternative values other than their opposite) are almost never observable in focus position. This confirms the theory of Rooth (1985) and Krifka (2008), according to which focusing indicates the presence of relevant alternatives.

8. I demonstrated the (initial) process of grammaticalization of the copula as a verb inflection with a corpus-driven study of the inflected forms of *szabad* ('may/can'). *Szabad* can be found in three different inflected forms in standard Hungarian (*szabadna, szabadott, szabadjon*). In the case of the conditional mood, the inflected form proved to be much more preferred than the nominal + copula construction, based on the corpus data.

4. Relevant publications

Dömötör Andrea, Yang Zijian Győző és Novák Attila. Much Ado About Nothing – Identification of Zero Copulas in Hungarian Using an NMT Model. In: Calzolari N. et al. szerk. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille. European Language Resources Association, 2020, 4802–4810.

Dömötör Andrea, Yang Zijian Győző és Novák Attila. Nesze semmi, fogd meg jól! Zéró kopulák automatikus felismerése neurális gépi fordítással. In: Berend Gábor, Gosztolya Gábor és Vincze Veronika szerk. *XVI. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged. Szegedi Tudományegyetem TTIK, Informatikai Intézet, 2020, 385–398.

Dömötör Andrea. Syntax is clearer on the other side - Using parallel corpus to extract monolingual data. In: Candito M. et al. szerk. P*roceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, Paris. Association for Computational Linguistics, 2019, 118–125.

Dömötör Andrea. Fókusz vagy igemódosító? In: Scheibl György szerk. *Lingdok 17.: Nyelvészdoktoranduszok dolgozatai*, Szeged. Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola, 2018, 147–158.

Dömötör Andrea. Nem mind VP, ami állít - A névszói állítmány azonosítása számítógépes elemzőben. In: Ludányi Zsófia, Krepsz Valéria és Gráczi Tekla Etelka szerk. *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2018: XII. Alkalmazott Nyelvészeti Doktoranduszkonferencia*, Budapest. MTA Nyelvtudományi Intézet, 2018, 3–10.

Dömötör Andrea. Hány VAN nincs? A létige – zéró váltakozás korpuszvezérelt vizsgálata. In: Ludányi Zsófia szerk. *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2017: XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia*, Budapest. MTA Nyelvtudományi Intézet, 2017, 28–38.

Hivatkozások

É. Kiss Katalin. Egy igekötőelmélet vázlata. *Magyar nyelv*, 2004, 100(1):15–43.

Krifka Manfred. Basic notions of information structure. *Acta Linguistica Hungarica*, 2008, 55 (3–4):243–276.

Lison Pierre és Tiedemann Jörg. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: Chair) Nicoletta Calzolari (Conference, Choukri Khalid, Declerck Thierry, Goggi Sara, Grobelnik Marko, Maegaard Bente, Mariani Joseph, Mazo Helene, Moreno Asuncion, Odijk Jan és Piperidis Stelios szerk. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA), may, 2016.

Novák Attila, Laki László János és Novák Borbála. Mit hozott édesapám? Döntést – Idiomatikus és félig kompozicionális magyar igei szerkezetek azonosítása párhuzamos korpuszból. In: *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019)*, Szeged. Szeged University, 2019, 63–71.

Oravecz Csaba, Váradi Tamás és Sass Bálint. The Hungarian Gigaword Corpus. In: Calzolari Nicoletta és et al. szerk. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Izland. European Language Resources Association (ELRA), 2014, 1719–1723.

Rooth Mats. *Association with focus*. Doktori disszertáció, University of Massachusetts, Amherst, 1985.

Teubert Wolfgang. My version of corpus linguistics. *International Journal of Corpus Linguistics*, 2005, 10(1):1–13.

Tognini-Bonelli Elena. *Corpus Linguistics at Work*. John Benjamins, 2001.

Vincze Veronika, Szauter Dóra, Almási Attila, Móra György, Alexin Zoltán és Csirik János. Hungarian Dependency Treebank. In: *Proceedings of LREC 2010*, Valletta, Malta. ELRA, May, 2010.