Fly, Fly Little Comet!
Exploring Subtoken-Level Metaphorical Patterns in Finnish and Hungarian Texts. New Results from the FiHuCoMet Corpus.

Tímea Borbála Bajzát
Eötvös Loránd University, Doctoral School of Linguistics
bajzat.timi9696@gmail.com

Abstract

The FiHuCoMet Corpus was created to address the gap in the lack of a systematic comparison of metaphor research in Finnish and Hungarian (Bajzát and Simon, 2023). This study aims to: (i) expand the existing quasi-parallel corpus; (ii) explore subtoken-level metaphorical patterns comparatively in the examined languages with rich morphology. The analysis employs a MIPVU-inspired protocol for metaphor identification, the MetaID protocol (Simon et al., 2023). The sub-token level in this study refers to the morphological patterns that can be accessed at the subword level. Although this endeavor is not new, the comparative study conducted on a small-scale corpus has only revealed a few aspects of the potential of comparative metaphor analysis in the context of Finno-Ugric languages selected.

1 Introduction

A noticeable trend in recent years is the research on metaphorical structures, particularly from the perspective of cognitive semantics (Bolognesi and Werkman, 2023; Steen et al., 2010). This trend is evidenced by the significant efforts made over the past two decades to map metaphors as a linguistically accessible phenomenon with a comprehensive, language-specific focus (e.g., Huumo 2019; Máthé 2022). For instance, the development of language-specific adaptations of the MIPVU protocol (Steen et al., 2010), the most accurate and widely-used method for metaphor identification, has yielded numerous results examining the typological features of metaphorical elaborations (Nacey et al., 2019; Bogetić et al., 2019; Marhula and Rosiński, 2019; Urbonaité et al., 2019). However, the Uralic languages were not included in these efforts. This gap was identified by Bajzát and Simon (2023) when they introduced the theoretical and methodological framework of the FiHuCoMet project: the Finnish and Hungarian Comparative Metaphor Research Corpus based on quasi-parallel news texts. Their paper elaborates on the applicability of the adapted Hungarian, morpheme-based version of the MIPVU protocol (Simon et al., 2019, 2023) to the Finnish language. This morpheme-based process of MIPVU is equipped to address the metaphorical potentials that come from the typological features of agglutinative languages.

At the time of the FiHuCoMet project's inception, the corpus consisted of 5,116 tokens, allowing for only a small-scale analysis. Nonetheless, their results suggest relatively similar metaphorical linguistic elaborations in Hungarian and Finnish languages but reveal slight differences, such as variations in the frequencies of metaphorical expressions, metaphorical subtoken-level constructions, and the complexity of argument structures (Bajzát and Simon, 2023).

The method and preliminary results discussed above have inspired us to outline further research questions, which are the focal point of this study. Firstly, this paper introduces the expansion of the FiHuCoMet research corpus. The given study posits that the Hungarian and Finnish corpora exhibit similar metaphorical patterns at the subtoken level, considering their types and proportions.

2 Method

2.1 The Adapted MIPVU Method

The MIPVU method, adapted to Hungarian as the MetaID method, can be consistently applied to annotate metaphorical constructions in agglutinative languages (Simon et al., 2023) from a functional cognitive perspective (Langacker 2008). Due to space limitations, we cannot provide a detailed description of the adaptation process here, but we will highlight the most significant

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 114–119
December 1-3, 2023 ©2023 Association for Computational Linguistics

changes in the following paragraphs (for a thorough discussion, see Simon et al., 2023).

As metaphorization can occur at the subtoken level, the most notable change is that the annotation process is based on morphemes rather than words.

1. Viime kuu-**ssa** niitä oli
last month-INE it-PART be-PST.3SG
60. (Finnish subcorpus)
60
('last month there were 60')

In the first Finnish example above (1), one can observe that time is conceptualized as a place by the highlighted inflectional suffix (*-ssA*, inessive case marking). The cited example represents a very conventional and grammaticalized way to express existence in time linguistically within the Finnish language. In many cases, the morphological units refer to a conceptualization that can be interpreted as an extension of the basic meaning based on similarity.

2. Jelenleg már 2012 halott-**ról**
Currently already 2012 dead-DEL
tudni. (Hungarian subcorpus)
know-INF
('Currently 2012 deaths known')

In the second example (from the Hungarian subcorpus), we can observe a delative case referring not to the spatial position but to the topic of the process of knowing. The inflected noun (*halottról* 'about the dead') belongs to the infinitive (*tudni* 'know') as its argument, and the inflection is used as a case marker, which is obligatory in that specific construction ('know about something') (Sass et al., 2011, p. 171). Since the meaning associated with space is assumed to no longer be activated in such grammaticalized contexts, it is not marked as a metaphorical inflection (Simon et al., 2023). Steen et al. (2010) apply a similar method to handle highly grammaticalized elements.

The method does not attempt to detect the etymological aspects of metaphorization (Steen et al. 2010, p. 33–36). For example, in the context of compounds, the MetaID annotation schema relies on the principle of lexicalization, which is determined based on dictionaries (like the case of the Sesotho language, Seepheephe et al., 2019). If a compound word has not been lexicalized, it is not included as a unified entity in the dictionary we analyzed its component from the aspect of metaphorization. Moreover, only suffixes that do not change the word class of a given word form may receive tags, in line with the original MIPVU method.

Secondly, the modified annotation schema introduces a new set of tags to identify semantic relations based on cognitive grammar categories (Langacker 1987, 2013), with the aim of providing a more precise representation of metaphorical elaboration structures above the words. This approach allows us to detect extended patterns of metaphorization at the clause-level and facilitates cross-linguistic comparisons.

3. A terembem a **sötétség-et** csak
The room-INE the darkness-ACC just
a gyertyák és a mécsesek
the candle-PL and the lantern-PL
fény-e **tör**-te **meg**.
light-POSS.3SG break-PST.3SG PREV
('The darkness in the room was broken only by the light of candles and lanterns)

The third example illustrates one instance of metaphorization in a multi-word expression. The verbal phrase (*törte* 'it broke') is annotated with the label of the metaphor-related expression because it initiates the metaphorical elaboration. At the same time, its arguments (*fénye* 'its light' and *sötétséget* 'darkness') also contribute to the metaphorization process, as we annotated them with the label of the metaphor-related argument (the full list of tags can be seen in Table 1).

The process of annotation is as follows: first, the text is split into morphological units, and then the basic and contextual meanings of the current morphological unit are determined using the dictionary, following the original MIPVU method. If an inter-domain mapping between the primary meaning and the contextual meaning can be assumed, the unit is marked as metaphorical. We annotate semantic relations separately and assess idiomaticity based on collocation (Simon et al. 2023).

The reliability of the MetaID procedure has been previously validated through assessments conducted on Hungarian language corpora (Simon et al., 2023). The kappa-values averaged 0.928 for mtags and 0.923 for mrel. Given that the overall performance of annotators surpasses the 0.8

threshold in kappa statistics (Carletta 1996, p. 252, Artstein–Poesio 2008, p. 22), the initial version of the adapted schema can be deemed reliable (Simon et al., 2023). It is essential to note that an inherent limitation in the current study lies in the absence of a comparable assessment for applying this procedure to the Finnish language yet. We intend to rectify this limitation in the upcoming research period.

2.2 The Brief Overview of the Tag Set

In the following tables (see Table 1 and Table 2) we attempt to introduce briefly the MetaID tag set and their semantics.

Tags	**Function**
MKK	Metaphor-related Expression
dMKK	Direct Metaphor-related Expression
MZ	Metaphor Flag
MKKimp	Metaphor-related Implicit Expression
MKI	Metaphor-related Inflection
MKA	Metaphor-related Argument
MKKomp	Metaphor-related Component
MKKid	Metaphor-related Idiomatic Expression
MKAid	Metaphor-related Idomatic Argument
MKKompid	Metaphor-related Idiomatic Component

Table 1: The tag set for identifying metaphorical structures.

Tags	**Function**
Tr (trajector)	It indicates the primary focal participant of the clause (Langacker 2008, p. 70–73)
Lm (landmark)	It signals the secondary focal participants of the scenario (Langacker 2008, p. 70–73)
Ela (elaboration)	Elaboration marks a non-specified elaborative operation
Poss (possessive)	This tag marks the possessive relation
Expm (explicating metaphorical meaning)	It signals the expressions used as a direct metaphor (MZ + dMKK).
R (unspecified semantic relation)	The label is used when two components of a multi-word unit move away from each other

Table 2: The relation set for identifying metaphorical structures.

3 The Project Infrastructure

3.1 The FiHuCoMet Research Corpus

In the process of building the FiHuCoMet research corpus, a fundamental organizational principle was followed: the incorporation of quasi-parallel texts in both its Hungarian and Finnish subcorpora. Here, 'quasi-parallel' does not mean processing identical source texts in both languages. Instead, it refers to processing texts with nearly identical content (report the same events), primarily comprising international political news obtained from online portals. These methodological choices were made to reduce potential biases in content and stylistic aspects, thus enhancing the objectivity of the studies (Bajzát and Simon, 2023). However, in the initial stage of corpus building, the subcorpora were relatively small, totaling 5,116 tokens (words) across both languages. The expanded corpus now contains 10,652 tokens. The principle of quasi-parallelism has been maintained during expansion. Nevertheless, thematically, the corpus has diversified and is no longer exclusively comprised of political news articles. The Hungarian-language news texts were drawn from Telex (Telex), while the Finnish-language news texts were collected from Helsingin Sanomat (Uutiset | HS.fi). The texts chosen for inclusion in the corpus were stored without their headlines and leads, as these structural elements are often duplicated within the main body of the text. Each of the Finnish and Hungarian subcorpora consists of 15 texts. As mentioned earlier, the sampling process was conducted in two stages. The first sampling took place in February 2022, while the second sampling was carried out in September 2023. The subcorpora can be categorized into the following major thematic units: international political news (F: 1,537 tokens; H: 3455 tokens), scientific and technological news (F: 1,114 tokens; H: 400 tokens), reports on natural disasters (F: 1,179 tokens; H: 932 tokens), news related to armed conflicts (F: 679 tokens; H: 679 tokens), and criminal news (F: 319 tokens; H: 358 tokens). As a result, the Hungarian subcorpus contains a total of 4,828 tokens, while the Finnish subcorpus comprises 5,824 tokens. Although the corpus of 10,652 words is relatively small for an extensive corpus linguistic study, the human capacity for manual annotation at this stage of the research did not allow for the processing of a larger sample. The present study provides exploratory feedback on the trends identified in the first phase of corpus building.

3.2 The Tools of Annotation

To annotate the Hungarian subcorpus, The Concise Dictionary of Hungarian (Pusztai ed. in chief, 2003) was employed. For the Finnish texts, The Dictionary of Contemporary Finnish (Kielitoimiston sanakirja) (Institute for the Languages of Finland 2022) was chosen to determine the default and contextual meanings of the expressions found. To measure idiomaticity in complex structures, a computational measuring tool was used, the word sketch browser of the Hungarian Web 2020 corpus (huTenTen20) and the Finnish Web 2014 corpus (fiTenTen), which provided association scores for collocations. Idiomaticity was evaluated using the LogDice typicality score (Rychlý, 2008), where a higher score (above 8.00) indicates a stronger association between the node and collocation candidates. In such cases, the method employed the MKKid tag to annotate metaphor-related idiomatic expressions and the MKKaid tag to denote their argument structure, or MKKompid when applicable (Bajzát and Simon, 2023, Simon et al., 2023).

The annotation was carried out using the WebAnno surface, designed by the CLARIN Research Infrastructure for Language Resources and Technology (Castilho et al., 2016). This platform allows for more transparent tagging of semantic relations and facilitates collaboration.

It's important to note that Hungarian texts were annotated by a native speaker, whereas Finnish texts were annotated by a non-native speaker with an upper-intermediate level of Finnish. Presenting this as a pilot study, we aim to inspire future collaborations between research communities, enabling cross-linguistic metaphor analysis with native speakers.

4 The Results

Figure 1 illustrates the proportion of tokens annotated with metaphorical expression labels across the entire corpus, with each column representing a news text. The most significant difference between Finnish and Hungarian news is noticeable in the 4th, 5th, 9th, 10th, 11th, and 12th pair of text. Generally, the extent of metaphorical marking in each subcorpus was similar in both Hungarian and Finnish. However, a noticeable difference in text length was observed in the case of two radically different text pairs (4th and 11th). In the fourth pair of texts, the Finnish text was relatively short (Finnish: 55 tokens; Hungarian: 211), while in the eleventh pair, the Hungarian text was significantly shorter than the Finnish one (Hungarian: 167 tokens, Finnish: 867 tokens). This confirms the observation that differences in text length can lead to a significant difference in their metaphorical markedness potential in 'quasi' parallel corpora.

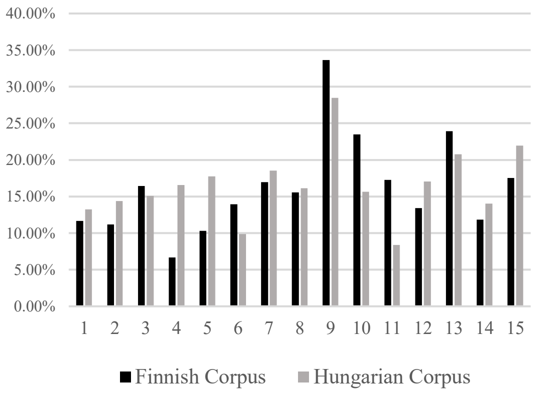

Figure 1: Relative frequencies of mtags in the subcorpora

For the other subcorpora, the length of the texts was relatively balanced. Furthermore, in nine subcorpora, the results indicate that Hungarian texts tend to have a slightly higher proportion of metaphorical tags compared to Finnish texts. The variance in sample sizes may lead to not only more frequent but also linguistically more complex metaphorical structures in Hungarian online news. These observations are in line with the results of the previous study (Bajzát and Simon, 2023). However, this can be nuanced by the fact that a higher proportion was also found in Finnish texts. Additionally, the slight differences can also be caused by the potential stylistic motivation. The higher occurrence and greater elaboration of metaphorical structures suggest that the speaker represented the events in a more sophisticated manner with greater stylistic potential.

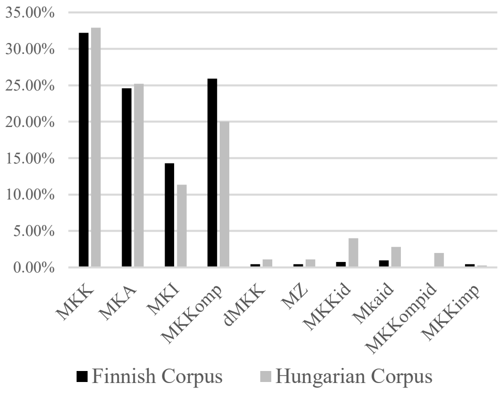

Figure 2: The proportions of mtags

Figure 2 illustrates the frequency of identified metaphorical units within the Finnish and Hungarian subcorpora. A notable difference is the higher prevalence of metaphorical elaboration at the morpheme level in the Finnish subcorpus, as indicated by the MKI bar. The varying proportions of MKKomp tags also demonstrate that the Finnish subcorpus exhibits more distinctive metaphorical elaborative operations. While the number of metaphorical expressions (MKK) and the examination of arguments do not show significant differences, it is observed that idiomaticity was more prevalent in Hungarian texts. This observation can further support the hypothesis of higher stylistic markedness in this subcorpus in terms of metaphorical constructions.

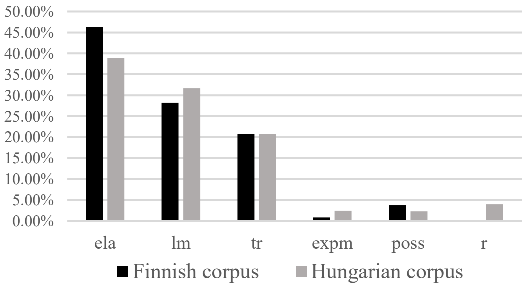

Figure 3: The proportions of mrel labels

Figure 3. illustrates the overall frequencies of labels assigned to the relations among metaphorical expressions in the Finnish and Hungarian subcorpora. The data reaffirm that the Finnish subcorpus has a higher proportion of metaphorical elaborative operations. A higher proportion of possessive metaphorical relations are also more characteristic of the Finnish material.

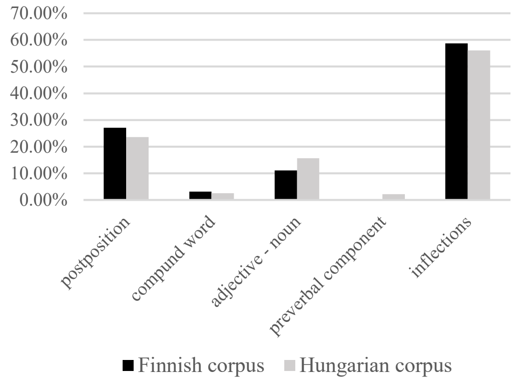

Figure 4: The distribution of elaborative relations

In Figure 4, we can observe slight differences in the distribution of semantic relations within metaphorical expressions, which highlight language-specific tendencies and patterns of morphological elaborative operations in metaphorization. In the Finnish subcorpus, a higher proportion of postpositions was measured (F: 27.11%; H: 23.62%) but the number of metaphorical adjective structures is lower in the Finnish texts (F: 11.12%; H: 15.62%). The inflections initiated the metaphorization are frequent in both languages.

5 Summary and Future Perspectives

The study aimed to report the latest findings from an ongoing project. Although the current FiHuCoMet corpus is still relatively small, it has more than doubled in annotated text volume compared to the previous phase. Recent results, particularly the analysis of subtoken-level metaphorization operations, confirm that while there are similarities in elaboration patterns between the two corpora, language-specific differences seem to be important as well. Looking ahead, it is justified to expand the corpus further and include texts of various types from multiple sources in parallel corpus building. Additionally, extending the metaphor identification map to include other Finno-Ugric languages is advisable for more comprehensive insights into comparative metaphor identification in these languages.

References

Ron Artstein and Massimo Poesio. 2008. *Inter-coder agreement for computational linguistics. Computational Linguistics 34:4*, pages 555–596. http://dx.doi.org/10.1162/coli.07-034-R2

Tímea B. Bajzát and Gábor Simon. 2023. Family relationship or family resemblance? A case study of comparative metaphor analysis in Finnish and Hungarian news texts. Under review.

Jean Carletta. 1996. *Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics 22:2*, pages 249–254.

Tuomas Huumo. 2019. Why monday is in front tuesday: On the uses of English and Finnish FRONT adpositions in SEQUENCE metaphors of time. *Linguistics 57:3*. pages 607–652. https://doi.org/10.1515/ling-2019-0010

Ksenija Bogetić, Andrijana Broćić, Katarina Rasulić. 2019. Linguistic metaphor identification in Serbian. In *Metaphor identification in multiple languages: MIPVU around the world*, John Benjamins, Amsterdam, pages 203–226. http://dx.doi.org/10.1075/celcr.22

Marianna, Bolognesi and Ana Werkman Horvat. 2023. *The metaphor compass. Directions for metaphor research in language, cognition, communication, and creativity.* Routledge, London, and New Work. https://doi.org/10.4324/9781003041221

Richard Eckart Castilho de, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevich, Anette Frank and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84. https://www.aclweb.org/anthology/W16-4011

Kielitoimiston sanakirja. 2022. Kotimaisten kielten keskus, Helsinki, online publication. URN:NBN:fi:kotus-201433. https://www.kielitoimistonsanakirja.fi.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar. Volume I theoretical prerequisites*, Stanford University Press, Stanford–California. https://doi.org/10.1515/9780804764469

Ronald W. Langacker. 2008. *Cognitive grammar: A basic introduction*. Oxford University Press, Oxford. DOI:10.1017/S0022226709005799

Langacker, Rondald W. 2013. Essentials of cognitive grammar. Oxford University Press, Oxford.

Joanna Marhula and Maciej Rosiński. 2019. Linguistic metaphor identification in Polish. In *Metaphor Identification in Multiple Languages: MIPVU around the world*, John Benjamins, Amsterdam, pages 183–202. http://dx.doi.org/10.1075/celcr.22

Zsuzsa Máthé. 2022. Space, time and transience. *Argumentum 18*, pages 273–286. 10.34103/ARGUMENTUM/2022/15

Susan Nacey, Aletta G. Dorst, Tina Krennmayr & Gudrun Reijnierse W. (eds.). 2019. *Metaphor identification in multiple languages: MIPVU around the world*. John Benjamins, Amsterdam, and Philadelphia. https://doi.org/10.1075/celcr.22

Ferenc Pusztai (eds.). 2003. *Magyar értelmező kéziszótár*, Akadémiai Kiadó, Budapest.

Pavel Rychlý. 2008. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing. 6–9. Brno*. 13.pdf (muni.cz)

Nts'oeu Raphael Seepheephe, Beatrice Ekanjume-Ilongo, Motlalepula and Raphael Thuube. 2019. Linguistic metaphor identification in Sesotho. In *Metaphor identification in multiple languages: MIPVU around the world*, John Benjamins, Amsterdam, pages 267–287. http://dx.doi.org/10.1075/celcr.22

Bálint Sass., Tamás Váradi, Júlia Pajzs and Margit Kiss. 2011. Magyar igei szerkezetek. *A leggyakoribb vonzatok és szókapcsolatok tára*. Tinta Könyvkiadó, Budapest.

Gábor Simon, Tímea Bajzát, Júlia Ballagó, Zsuzsanna Havasi, Mira Roskó and Eszter Szlávich. 2019. Metaforaazonosítás magyar nyelv szövegekben: egy módszer adaptálásáról. *Magyar Nyelvőr 143: 2*, pages 223–247. http://nyelvor.c3.hu/period/1432/143208.pdf

Gábor Simon, Tímea B. Bajzát, Júlia Ballagó, Zsuzsanna Havasi, Emese K. Molnár and Eszter Szlávich. 2023. When MIPVU goes to No Man's Land: A New Language Resource for Hybrid, Morpheme-based Metaphor *Identification in Hungarian. Language Resources and Evaluation*. In press.

Gerard Steen, Aletta G. Dorst, Berenike Herrmann, Anna Kaal, Tina Krennmayr and Trijntje Pasma. 2010. *A Method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins, Amsterdam, John Benjamins. http://dx.doi.org/10.1075/celcr.14

Justina Urbonaitė, Inesa Šeškauskienė, Jurga Cibulskienė. 2019. Linguistic metaphor identification in Lithuanian. In *Metaphor identification in multiple languages: MIPVU around the world*, John Benjamins, Amsterdam, pages 159–181. http://dx.doi.org/10.1075/celcr.22