Identifying the Most Representative Phraseological Units Using Language Corpora and Artificial Intelligence for Lexicography: The Case of Slovenian Comparative Phrasemes

Matej Meterc[1], Nataša Jakop[2]

[1],[2] ZRC, Fran Ramovš Institute of the Slovenian Language, Novi trg 2, SI-1000 Ljubljana, Slovenia

E-mail: matej.meterc@zrc-sazu.si, natasa.jakop@zrc-sazu.si

Abstract

In preparing phraseological units for the third edition of the *Standard Slovenian Dictionary* (eSSKJ), the authors aimed to identify the most relevant comparative phrasemes in the contemporary standard language using objective corpus-based criteria. A key goal was to determine how representative specific phrasemes and their variants are in actual use. Two lists of the hundred most frequent comparative phrasemes with the structure adjective + *kot* 'as' + noun (e.g., *bel kot sneg* 'white as snow') were extracted from the metaFida v1.0 corpus and CLASSLA-web.sl 1.0 corpora. The twenty most frequent were analyzed in greater detail. The results were compared with the Database of Comparative Phrasemes compiled from older dictionaries and collections, as well as with entries in eSSKJ. Artificial intelligence was also used experimentally to identify representative comparative phrasemes, with up to 80% alignment with expert choices.

Keywords: comparative phrasemes; corpus linguistics; artificial intelligence; lexicography; phraseological minimum

1. Introduction

A key task of modern phraseography is to determine the degree of representativeness of phrasemes and their variants in contemporary language, utilizing special corpus-based methods (Gantar, 2006, 2007; Čermák, 2007; Dobrovol′skij, 2014; Ďurčo, 2014). In the phraseographic process in the general explanatory dictionary eSSKJ, it is essential to determine the representativeness of phrasemes primarily based on corpus data (Meterc & Jakop, 2016). In eSSKJ, which is based on the Gigafida v1.0 corpus (GF), 717 phrasemes have already been published, including 137 comparative phrasemes (CPs). Identifying the most representative CPs of various structures facilitates analysis of their form and meaning and determination of the formal and semantic features relevant for dictionary presentation. For these purposes, the concept of the phraseological minimum is presented

and the potential of using AI is considered to obtain data on the most representative phrasemes.

1.1 Determining Lexicographic Needs: Phraseography for the Standard Slovenian Dictionary (eSSKJ)

Dictionaries apply various criteria for including phraseology. For a phraseme or its variant to be included in eSSKJ, the following conditions must be met 1) one lexical component of the phraseme or its variant must be included in the dictionary, 2) the individual form of the phraseme or its variant must meet the minimum frequency threshold (five attested examples of use from various sources in the GF reference corpus), and 3) the usage examples must be prototypical according to Čermák's criteria (2007: 572–573) and sufficiently informative to determine the meaning of the phraseme. Even though the threshold of five occurrences is relatively low—even considering the quality criteria for examples—a large amount of phraseology can still be included in the dictionary. Therefore, it is important to ensure that no phraseme relevant to contemporary Slovenian, especially those that form part of the standard language, is overlooked. The selection of phrasemes for lexicographic presentation is based on 1) a corpus analysis of word collocations and word sketches, 2) verification of phrasemes containing the target lexical component in dictionaries and collections, and 3) a systematic search for phrasemes in language corpora.

This article focuses on a systematic search for phrasemes by constructing two phraseological minima based on two corpora and by querying and critically evaluating the responses of artificial intelligence (AI), specifically the ChatGPT-4o model (GPT-4o).

Recently, a considerable number of CPs have been included in eSSKJ due to the integration of zoonym entries such as *pes* 'dog', *mačka* 'cat', *koza* 'goat', *ovca* 'sheep', and *kokoš* 'hen'. To determine which (comparative) phrasemes are representative of contemporary Slovenian, this article focuses on the most frequent expressions. However, the research also considers less frequent phrasemes (and their variants) that are present in modern language. It would be difficult to define what counts as "less frequent" without first identifying a list of the most common expressions following a given structural pattern—that is, a phraseological minimum. The lexicographic analysis also relied on the Database of Comparative Phrasemes (DCP; see section 2.1).

To optimally describe the formal and semantic features in the phraseographic process, not only quantitative (especially corpus-based) data but also the structural and semantic properties of CPs must be considered. This article highlights certain specific characteristics of CPs, such as their wide variety of structural patterns, the relatively strong semantic transparency of some CPs, and the polysemy of both their phraseological components and the CPs as a whole.

1.2 The Phraseological Minimum for a Structural Type of Phraseme Based on the Paremiological Minimum

The concept of a phraseological minimum presented in this article is modeled on the concept of a paremiological minimum pioneered by Permjakov (1971), which defines a core set of widely recognized proverbs (or other types of paroemias) in a specific language. Permjakov's study used the PTP (part text presentation) method, filtering 1,491 Russian proverbs and asking respondents to complete missing halves, resulting in a minimum of five hundred proverbs, which was later refined to three hundred for the *Russian–German Dictionary of Proverbs* (Permjakov, 1985). Similar minima have been created for other languages, such as Slovenian (Meterc, 2017) and Croatian (Varga & Babić, 2023). A corpus-based approach to identifying the most frequent Czech proverbs was presented by Čermák (2007), and related research has followed for Slovak and German (Ďurčo, 2014), as well as Slovenian (Meterc, 2017). Paremiological minima are useful in phraseography and paremiography, and they also hold potential for contrastive phraseology and phraseodidactics. The Czech paremiological minimum is presented in *Základní slovník českých přísloví* (Čermák, 2013), and data from the Slovenian minimum have been used in the compilation of eSSKJ and *Slovar pregovorov in sorodnih paremioloških izrazov* (Meterc, 2020–).

Due to the large number of phrasemes and their structural types, the concept of multiple phraseological minima is proposed, each corresponding to a specific structural type. From a lexicographic perspective, the phraseological minimum is useful 1) for selecting the most representative phrasemes, and 2) as an empirical reference point for evaluating less frequent phrasemes, which may also be of interest to phraseography—provided they are attested in use. The interest is in the degree of similarity and difference between the two corpus-based minima, and in how they can help shed light on 1) the material obtained via AI, 2) the entries from existing collections in DCP, and 3) the expressions already prepared for inclusion in eSSKJ.

1.3 Artificial Intelligence as a Source for Acquiring Phraseological and Paremiological Material

AI has been shown to be highly useful in various stages of lexicography, such as determining meaning (Jakubiček & Rundell 2023; de Schryver 2023). Identifying relevant phraseme forms and subsequently classifying their variants is a key prerequisite for beginning phraseographic work. Forms listed in the corpus-driven phraseological minimum and those generated by AI were compared with each other, as well as with data from

eSSKJ and DCP. The reliability of AI responses is important for both identifying relevant forms in the initial stage of (AI-assisted) phraseography and for evaluating AI as a source of information for general users, who may prefer AI-generated answers over dictionary entries. The success of the lists was assessed according to similar criteria as those used to evaluate the effectiveness of AI responses to the question about the best-known and most frequent Slovenian proverbs (Meterc & Mrvič, in press). The interest here is in the typological and formal adequacy of the expressions provided. It is acknowledged that the accuracy of AI responses for Slovenian is likely lower than for English; however, this issue is not addressed in detail in this article.

2. Identifying the Most Representative Slovenian CPs with the Structure Adjective + Conjunction *kot* 'as' + Noun in Various Sources

CPs are fixed multi-word units that follow a comparative structure (typically *X kot/kakor/ko Y* 'X as/like Y') and exhibit varying degrees of idiomaticity. In determining the most representative CPs, one can rely on their productive construction patterns (Kocijan & Librenjak, 2016).

In this study, the analysis was limited to the structure adjective + conjunction *kot* 'as' + noun (abbreviated A + *kot* + N;[1] e.g., *bel kot sneg* 'white as snow'), which is one of the two most frequent construction patterns among the phrasemes included in eSSKJ to date (the other being verbal CPs with the structure verb + conjunction *kot* + noun). The DCP contains mostly verbal CPs, but this category also includes examples originally recorded in the source with an external (non-phraseme-internal) verb *biti* 'be', such as *biti pijan kot čep* 'be drunk as a cork', which could in fact also be classified as adjectival CPs (see Gantar, 2002: 38).

2.1 The Database of Slovenian Comparative Phrasemes (DCP)

DCP has been under development since 2022 at the Department of Lexicology of the Fran Ramovš Institute of the Slovenian Language. The project's aim is to systematically collect, document, and standardize the form and lexicographic treatment of Slovenian CPs, which are dispersed across various linguistic sources—collections, dictionaries, scholarly literature, and corpora—and to incorporate them into a unified, accessible, and scientifically structured database. This database serves as a foundation for phraseological

[1] Slovenian CPs are also lexicalized with the conjunction *kakor* or the colloquial form *ko* (e.g., *hladen kot/kakor/ko led* 'cold as ice'), all of which are also considered in the lexicographic treatment in eSSKJ. However, due to the nature of this research, the analysis was limited to the conjunction *kot*.

analysis of the grammatical, semantic, and pragmatic properties of CPs in contemporary Slovenian.

By 2024, a total of 2,521 CPs had been collected in DCP, of which 282 have the structure A + *kot* + N. The material for DCP comes from several published dictionaries and collections, such as *Hrvatsko-slavenski rječnik poredbenih frazema* (Fink-Arsovski et al., 2006) with 343 Slovenian CPs; *Slovar slovenskih frazemov* (Keber, 2011, 2015) with 1,288 CPs; and the phraseological-paremiological collection *Pregovori in reki na Slovenskem* (Bojc, 1987) with 273 CPs. DCP also includes CPs from more recent sources, such as those incorporated into the database for eSSKJ (137 CPs), and from other collections that contributed 480 CPs. This article assesses whether CPs obtained with the help of AI could also be added to such collections in the future.

DCP is a collection of CPs designed for the systematic analysis of their structure, meaning, and semantic domains. It contains structured data on Slovenian CPs, including 1) the base form of the CP and its variants as found in various sources, 2) a typological classification of their syntactic function (at the phrase or clause level), 3) information on the components of CPs, and 4) data on the thematic domain of the CPs, which covers a range of subject and conceptual areas, such as the human body, social relations and affiliation, psychological and behavioral states, space and environment, nature and living beings, and objects and materials.

This article uses DCP as a support tool for analyzing CPs obtained via AI, focusing in particular on the conventionality and variability of CPs, and the existence of structurally different but similarly motivated CPs (e.g., adverbial vs. adjectival, adverbial vs. verbal, etc.). DCP was also used to analyze the frequency of noun and adjective components in CPs, which helps interpret the phraseological minimum derived from both corpora and AI.

2.2 Determining the Most Frequent CPs Using Language Corpora: Corpus-Driven Phraseological Minima from the metaFida v.1.0 1.0 and CLASSLA-web.sl 1.0 Corpora

Two lists of the hundred most frequent CPs with the structure A + *kot* + N were created from the metaFida v1.0 (MF) and the CLASSLA-web.sl 1.0 (CL) corpora. Hence, the phraseological minimum of CPs from MF is referred to as MFmin and that from CL as CLmin. The twenty most frequent ones were analyzed in greater detail.

2.2.1 The Language Corpora Used

The analysis of phrasemes for eSSKJ is based on Gigafida v1.0 (1.4 billion tokens), which contains 87.9% printed texts (mainly journalistic, with some literary texts) and 12.1%

internet texts (Logar et al., 2013). Additional sources are occasionally used to confirm phraseological meaning or to provide dictionary exemplification, notably MF and CL. These were specifically used to establish two phraseological minima of CPs with the structure A + *kot* + N. metaFida v1.0 (Erjavec 2023) comprising thirty-four corpora, was selected because it partly resembles GF (quite a large share—1.3 out of 6.1 billion tokens—is accounted for by Gigafida 2.0), and also due to its diverse content (e.g., news, academic texts, user-generated content, and literary texts). CLASSLA-web.sl corpus 1.0 (Ljubešić et al., 2024) was used to build the second phraseological minimum. This is a 2.4 billion-token corpus featuring recent online texts. Both corpora are dominated by standard Slovenian, so we consider them to be a sufficiently reliable source for identifying the most representative comparative phrasemes in the standard language, even though they also include nonstandard internet texts and transcribed speech (a considerable proportion of the internet texts from social networks is written in standard Slovenian, as is the transcribed speech).

2.2.2 Constructing the Two Phraseological Minima

The corpus search was limited to adjectival CPs with the structure A + *kot* + N. The retrieved concordances were then sorted by frequency, taking into account the lemmas of adjectives and nouns. The exported lists were manually reviewed, and (frequent) non-phraseological comparative structures were removed (e.g., *zaposlen kot učitelj* 'employed as a teacher'). In the creation of MFmin, there were approximately 1,630 such cases up to a frequency of twenty-nine (at which the hundred most frequent phraseme forms were recorded), and about 540 such cases up to a frequency of nineteen during the creation of CLmin.

The frequencies and the resulting order of expressions in the minimum should be taken with some caution due to duplicated concordances, non-prototypical examples, certain lemmatization issues (e.g., confusion between formally identical adjectival and adverbial CPs, such as *tih* vs. *tiho kot miška* 'quiet as a mouse'), and, in some cases, equal numbers of occurrences for two or more expressions. Nevertheless, in constructing the phraseological minimum, the aim was to provide a general orientation regarding how frequently a phraseme appears. Some variants of the same CP were also listed separately in the minimum. For further use it would be necessary to group the variants under a common phraseological lemma.

2.2.3 Comparing the Structure of the Minima and Other Observations

In MFmin, the hundred most frequent CPs have frequencies ranging from 667 to twenty-nine, and, in CLmin, from 361 to nineteen. The two minima, derived from different language corpora, share 87 identical expressions out of a hundred. The twenty-three

expressions that appear only in MFmin include CPs that do not have negligible frequencies in the CL corpus, such as *okrogel kot žoga* 'round as a ball' (ranked eighty-second in MFmin with a frequency of thirty-five; frequency in CL: eleven concordances). A similar pattern is observed with the twenty-three CPs found only in CLmin.

Corpus-derived minima often include different forms of the same phraseme or its conventional variants, even though not to the same extent in both minima. Both lists include the phraseme *simpl kot pasulj* 'simple as beans', which, due to the Anglicism *simpl* (standard Slovenian: *enostaven*, *preprost*), carries a colloquial tone and is typical of informal communication. In CLmin, which contains more texts from informal online communication, it is ranked nineteenth, whereas in MFmin, which also includes internet texts but to a lesser extent, it is ranked eightieth. Interestingly, its variants in standard language—*enostaven kot pasulj* and *preprost kot pasulj*—appear only in CLmin, but not in MFmin.

The top twenty CPs from both lists are presented by frequency and compared below. Each expression is accompanied by its frequency ranking, with the number of occurrences in parentheses for each corpus. Each CP is translated into English only once—if it appears in both lists, the translation is provided only on the MFmin list. For each expression, its ranking in the other corpus is also provided, along with an indication of whether it is documented in DCP or not (DCP:/).

MFmin	CLmin
1. (667) *čist kot solza* 'clean as a tear' (9th CLmin, DCP)	1. (361) *trd kot kamen* (2nd MFmin, DCP)
2. (370) *trd kot kamen* 'hard as a rock' (1st CLmin, DCP)	2. (318) *drag kot žafran* (5th MFmin, DCP)
3. (366) *star kot človeštvo* 'old as humanity' (7th CLmin, DCP)	3. (209) *oster kot britev* (6th MFmin, DCP)
4. (345) *zdrav kot dren* 'healthy as dogwood' (6th CLmin, DCP)	4. (206) *suh kot poper* (15th MFmin, DCP)
5. (306) *drag kot žafran* 'expensive as saffron' (2nd CLmin, DCP)	5. (189) *bel kot sneg* (8th MFmin, DCP)
6. (304) *oster kot britev* 'sharp as a razor' (3rd CLmin, DCP)	6. (175) *zdrav kot dren* (4th MFmin, DCP)
7. (301) *besen kot ris* 'furious as a lynx'* (22nd CLmin, DCP)	7. (169) *star kot človeštvo* (3rd MFmin, DCP)
8. (253) *bel kot sneg* 'white as snow' (5th CLmin, DCP)	8. (167) *zdrav kot riba* (11th MFmin, DCP)

9. (240) *trden kot skala* 'solid as a rock' (10th CLmin, DCP)	9. (162) *čist kot solza* (1st MFmin, DCP)
10. (221) *svoboden kot ptica* 'free as a bird' (12th CLmin, DCP:/)	10. (131) *trden kot skala* (9th MFmin, DCP)
11. (220) *zdrav kot riba* 'healthy as a fish' (8th CLmin, DCP)	11. (129) *rdeč kot kri* (14th MFmin, DCP)
12. (196) *napet kot struna* 'tense as a string' (15th CLmin, DCP)	12. (129) *svoboden kot ptica* (10th MFmin, DCP:/)
13. (186) *lačen kot volk* 'hungry as a wolf' (14th CLmin, DCP)	13. (126) *dober kot kruh* (18th MFmin, DCP)
14. (170) *rdeč kot kri* 'red as blood' (11th CLmin, DCP)	14. (124) *lačen kot volk* (13th MFmin, DCP)
15. (167) *suh kot poper* 'dry as pepper' (4th CLmin, DCP)	15. (120) *napet kot struna* (12th MFmin, DCP)
16. (158) *hladen kot špricer* 'cold as a wine spritzer' (17th CLmin, DCP)	16. (118) *star kot Zemlja/zemlja* (17th MFmin, DCP)
17. (151) *star kot Zemlja/zemlja* 'old as the Earth/soil' (16th CLmin, DCP)	17. (113) *hladen kot špricer* (16th MFmin, DCP)
18. (149) *dober kot kruh* 'good as bread' (13th CLmin, DCP)	18. (104) *trd kot beton* 'hard as concrete'* (28th MFmin, DCP)
19. (137) *črn kot oglje* 'black as coal'* (29th CLmin, DCP)	19. (96) *simpl kot pasulj* 'simple as beans'* (DCP: /)
20. (137) *jezen kot ris** 'angry as a lynx' (42nd CLmin, DCP)	20. (89) *črn kot noč* 'black as night'* (25th MFmin, DCP)

Table 1: The top twenty CPs of MFmin and CLmin

Seventeen out of the top twenty expressions (85%) in one minimum also appear among the top twenty in the other. The six expressions that are found only in the top twenty of one of the minima are marked with an asterisk (*); all of them appear in the other minimum as well, but they ranked lower than twentieth (their positions are provided in parentheses). One CP that is not included in DCP but appears in both minima is *svoboden kot ptica* 'free as a bird'; in CLmin, there is one additional example not present in DCP: *simpl kot pasulj* 'simple as beans'. This means that more than 92% of the CPs from both minima are included in DCP. The variants *jezen kot ris* 'angry as a lynx' and *besen kot ris* 'furious as a lynx' were found among the top twenty CPs in MFmin.

The minima also had to be reviewed due to lemmatization in adjectives that appear in different nominative forms (adjective doublets); for example, *močen* and *močan* 'strong' (the latter being the non-preferred variant). The nominative form *močan kot medved* appears in MF in twenty examples and in CL in fourteen examples. The nominative form *močen kot medved* appears only once in MF and not at all in CL. Both forms are lemmatized in the corpus under *močen*, although the dominant nominative form in the phraseme is *močan* (e.g., *močan kot medved* 'strong as a bear').

2.3 Determining the Most Frequent CPs with the Structure A + kot + N Using Artificial Intelligence and Language Corpora

GPT-4o was used to investigate the most representative phrasemes with the structure A + *kot* + N. On the same date (May 15th), two identical questions were posed in Slovenian in two separate chats, using the following prompt:

> Please provide a list of the 20 most common comparative phrasemes in Slovenian that have a structure where the first word is an adjective, the second word is *kot*, and the third is a noun. (OpenAI 2025)

The responses are analyzed using data from language corpora and DCP.

2.3.1 Verifying the Relevance of Forty CPs from Two GPT-4o Responses: Relation to Phraseological Minima and Other Observations

Table 2 presents two lists of AI-generated responses. Each CP is annotated with the following information: overlap between Lists A and B, inclusion in MFmin and CLmin, presence in DCP, and the number of adjectival (A) and nominal (N) components found in the structure A + *kot* + N in DCP. In the table, "DCP" indicates that the CP is recorded in DCP in the same form, "A" shows how many CPs in DCP share the same adjectival component, and "N" indicates how many share the same nominal component. The label "/" indicates that no examples were found.

List A	List B
1. *hladen kot led* 'cold as ice' B1; MFmin: 26th, CLmin: 30th; DCP, A: 4 N: 2	1. *hladen kot led* 'cold as ice' A1; MFmin: 26th, CLmin: 30th; DCP, A: 4 N: 2
2. *trmast kot osel* 'stubborn as a donkey' B9; MFmin: 96th, CLmin:/; DCP, A: 3 N: 1	2. *lačen kot volk* 'hungry as a wolf' A9; MFmin: 13th, CLmin: 14th; DCP, A: 2 N: 1
3. *počasen kot polž* 'slow as a snail'	3. *pameten kot lisica* 'clever as a fox'

B4; MFmin: 58th, CLmin: 49th; DCP, A: 1 N: 1	A/; MFmin:/, CLmin:/; DCP:/, A:/ N: 1
4. *gladek kot svila* 'smooth as silk' B19; MFmin:/, CLmin: 61st; DCP:/, A: 2 N: 1	4. *počasen kot polž* 'slow as a snail' A3; MFmin: 58th, CLmin: 49th; DCP, A: 1 N: 1
5. *lep kot slika* 'pretty as a picture' B11; MFmin: 49th, CLmin: 39th; DCP, A: 5 N: 1	5. *težak kot svinec* 'heavy as lead' A/; *težek kot svinec* in MFmin: 37th, CLmin: 40th; DCP: *težek kot svinec*, A:/ N: 1
6. *čist kot solza* 'clear as a tear' B8; MFmin: 1st, CLmin: 9th; DCP, A: 4 N: 1	6. *močan kot medved* 'strong as a bear' A17; *močen kot medved* in MLmin:/, CLmin: 100th; DCP, A: 4 N: 2
7. *grd kot smrtni greh* 'ugly as a mortal sin' B/; MFmin:/, CLmin:/; DCP, A: 5 N: 1	7. *hiter kot blisk* 'fast as lightning' A/; MFmin: 23rd, CLmin: 23rd; DCP, A: 7 N: 1
8. *slep kot krt* 'blind as a mole' B/; MFmin:/, CLmin:/; DCP:/, A: 1 N:/	8. *čist kot solza* 'clear as a tear' A6; MFmin: 1st, CLmin: 9th; DCP, A: 4 N: 1
9. *lačen kot volk* 'hungry as a wolf' B2; MFmin: 13th, CLmin: 14th; DCP, A: 2 N: 1	9. *trmast kot osel* 'stubborn as a donkey' A2; MFmin: 96th, CLmin:/; DCP, A: 3 N: 1
10. *tiho kot miška* 'quietly as a mouse' B/; MFmin:/, CLmin:/; DCP, N: 1	10. *priden kot čebela* 'hardworking as a bee' A/; MFmin: 46th, CLmin: 36th; DCP, A: 4 N: 3
11. *zvest kot pes* 'faithful as a dog' B20; MFmin: 72nd, CLmin: 50th; DCP, A: 1 N: 12	11. *lep kot slika* 'pretty as a picture' A5; MFmin: 49th, CLmin: 39th; DCP, A: 5 N: 1
12. *bogat kot Krez* 'rich as Croesus' B/; (MFmin:/, CLmin:/; DCP, A: 1 N: 1	12. *ubog kot miš* 'poor as a mouse' A/; MFmin: /, CLmin:/; DCP:/, A:/ N: 2
13. *pameten kot knjiga* 'clever as a book' B/; MFmin:/, CLmin:/; DCP:/, A:/ N:/	13. *pijan kot čep* 'drunk as a cork' A/; MFmin: 22nd, CLmin: 35th; DCP, A: 15 N: 2
14. *črn kot noč* 'black as night' B18; MFmin: 25th, CLmin: 20th; DCP, A: 9 N: 2	14. *glasen kot trobenta* 'loud as a trumpet' A/; MFmin:/, CLmin:/; DCP:/, A: 1 N:/
15. *mrtev kot kamen* 'dead as a stone' B/; MFmin:/, CLmin:/; DCP:/, A:/ N: 5	15. *šibek kot muha* 'weak as a fly' A/; MFmin:/, CLmin:/; DCP:/, A:/ N: 1
16. *bister kot biser* 'bright as a pearl' B/; MFmin:/, CLmin:/; DCP:/, A:/ N: 1	16. *suh kot trska* 'thin as a splinter' A/; MFmin: 91st, CLmin:/; DCP, A: 8 N: 1
17. *močan kot medved* 'strong as a bear' B6; *močen kot medved* in MLmin:/, CLmin: 100; DCP: *močan kot medved*, A: 4 N: 2	17. *bel kot sneg* 'white as snow' A/; MFmin: 8, CLmin: 5; DCP, A: 12 N: 1

18. *preprost kot pasulj* 'simple as beans' B/; MFmin:/, CLmin: 63rd; DCP, A: 1 N: 1	18. *črn kot noč* 'black as night' A14; MFmin: 25th, CLmin: 20th; DCP, A: 9 N: 2
19. *slab kot mušji drek* 'bad as fly shit' B/; MFmin:/, CLmin:/; DCP:/, A:/ N:/	19. *gladek kot svila* 'smooth as silk' A4; MFmin:/, CLmin: 61st; DCP:/, A: 2 N: 1
20. *hiter kot strela* 'fast as lightning' B/; MFmin: 31st, CLmin: 25th; DCP, A: 7 N: 1	20. *zvest kot pes* 'faithful as a dog' A11; MFmin: 72nd, CLmin: 50th; DCP, A: 1 N: 12

Table 2: Lists of AI-generated responses

As shown in Table 2, ten expressions from one list also appear on the other, meaning the two lists overlap by 50%. One form appearing on only one list (*hiter kot strela* 'fast as lightning') is actually a variant of a CP on the other list (*hiter kot blisk* 'fast as lightning'). On the list A, eight expressions are found in both minima, and four appear in at least one of them. On the list B, twelve are present in both minima, and four appear in one. DCP confirms 70% of CPs from each list. In DCP, the most frequent adjectives are *pijan* 'drunk' in fifteen CPs and *bel* 'white' in twelve CPs; the most frequent noun compared is *pes* 'dog' in eleven CPs, followed by *kamen* 'stone' in five. All these components also appear in the CPs listed above: *pes* is found on both lists, *pijan* and *bel* on the list B, and *kamen* on the list A. Each list contains six expressions not confirmed by DCP due to morphological or lexical variants (e.g., *težak kot svinec*, *pameten kot lisica*) or their non-established phraseological status (e.g., *bister kot biser*). In borderline cases, the evaluation is supported by data from additional sources. In cases in which there is no confirmation in the minimum sets, collections, and even the distribution of individual components within a comparative structure, AI has either generated an incorrect structure (e.g., *slab kot mušji drek*) or an expression not attested as a CP in standard Slovenian (e.g., *slep kot krt*).

Of interest was the extent to which the expressions listed in the AI responses are conventional idiomatic expressions (i.e., CPs) and whether they are of the appropriate type, as specified in the prompt to AI. To confirm the conventionality and idiomaticity of the expressions listed, examples of their use were examined in both language corpora. Expressions that appeared in at least one of the minima (CLmin or MFmin) already meet the frequency criterion, and their idiomaticity was verified through usage examples while creating the minima.

In terms of expressions that do not appear in the minima, there are nine on the list A. Of these, corpus data confirm the conventionality and idiomaticity of six expressions because they occur repeatedly in the corpora—for example, *slep kot krt* 'blind as a mole' (MF: sixteen examples; CL: eleven examples). The conventionality or idiomaticity of two expressions cannot be confirmed because there are no usage examples: *pameten kot knjiga*

'clever as a book' and *slab kot mušji drek* 'bad as fly shit'. There are no CPs with the structure *pameten kot N* in DCP, though the component *knjiga* 'book' is part of established phrasemes, such as *biti kot odprta knjiga* 'be like an open book'. The structure *slab kot N* is likewise not recorded in DCP, although the component *drek* 'shit' is productive in phraseme formation as a comparative element (e.g., *vreden kot pasji drek* 'worthless as dog shit'). How can this be explained? When generating adjectival CPs, AI may have drawn from established expression elements that are part of other well-known and conventional phraseological structures/cores in Slovenian, with the comparative noun component following the conjunction *kot* seemingly acting as the key trigger for generating new potentially phraseological combinations. This reflects a typical generative (compositional) strategy of language models (Hupkes et al., 2020), which combine frequent elements and patterns from established linguistic fragments—including fragments of phrasemes—into new word combinations, even when those combinations are not empirically attested as phrasemes in actual usage. This highlights the need for caution when considering such AI-generated expressions as phraseological units: their form may resemble that of phrasemes but, without verifiable idiomatic status, they remain outside the phraseological inventory of a language. This phenomenon could be described as a "hallucinated phraseme."

List B contains six expressions that do not appear in either of the two minima. Among them, corpus evidence and attestation in DCP confirm the conventionality and idiomaticity of the following two expressions: *težak kot svinec* 'heavy as lead' (MF: twelve examples, CL: four examples) and *močan kot medved* 'strong as a bear' (MF: twenty-one, CL: fourteen). DCP contains confirmed phrase-forming potential for comparative structures with *težak kot N* (N = *cent* 'hundredweight', *beton* 'concrete', *slon* 'elephant', *svinec* 'lead', *kamen* 'stone'), whereas the variant form *težek kot N* is not attested in DCP. In the second case, both morphological variants of the adjective (*močen* and *močan*) have confirmed phrase-forming potential in DCP, although they differ slightly in the range of noun components that fill the comparative slot. On the other hand, the noun *svinec* 'lead' (according to DCP data) appears in CPs exclusively with the adjective *težak*, whereas the comparative structure with the noun *medved* 'bear' is more open and appears with adjectives such as *močan* 'strong', *zaščiten* 'protected', and *kosmat* 'hairy'. From this it can be concluded that productivity in phraseme formation is neither automatic nor predictable for all adjectival morphological doublets. In dictionaries, this cannot be presented automatically at the level of the phraseological lemma (e.g., *močan/močen kot medved*). Moreover, the lexical filling of the comparative component (the noun) is not unlimited because comparative structures with specific adjectives show varying degrees of openness in this regard.

In addition to these CPs, two forms can be conditionally interpreted as rare variants of established CPs: *pameten kot lisica* 'clever as a fox' (MF: none, CL: one) as a variant of the phraseme *zvit kot lisica* 'cunning as a fox' (MF: sixty-three, CL: twenty-seven), and *ubog*

kot miš ‘poor as a mouse’ (MF: one, CL: none) as a variant of *reven kot cerkvena miš* ‘poor as a church mouse’ (MF: c. 360, CL: c. 120), which differs in structure. DCP includes *zvit kot lisica*, *ubog kot cerkvena miš*, and *reven kot cerkvena miš*, but no CPs with the adjective *pameten* ‘smart’. The noun *lisica* ‘fox’ does not appear with other adjectival CPs in DCP, whereas the component *miš* ‘mouse’ is more productive in phraseme formation, also appearing in adjectival CPs such as *tih kot miš* ‘quiet as a mouse’. From this it can be inferred that AI, following our instructions (A + *kot* + N), generated examples in which, in the case of *ubog kot miš*, one of the obligatory components of the original phraseme (*cerkven* ‘church’) was omitted, and in the second case (*pameten kot lisica*), a conventional component was replaced by a non-conventional one based on semantic similarity (*pameten* ‘clever’ vs. *zvit* ‘cunning’). The latter may also be due to the influence of English (*clever as a fox*), which plays an important role in generating such results. Large language models like ChatGPT have been trained on data from numerous languages, but English accounts for most of those data. As a result, the model often internally interprets the input in English, processes it based on English language patterns, and then generates the output in the target language (Wendler et al., 2024).

The conventionality and idiomaticity of two forms cannot be confirmed because there is no evidence for them in the corpora or DCP, nor do they appear to be rare variants of any other established phraseme: *glasen kot trobenta* ‘loud as a trumpet’ and *šibek kot muha* ‘weak as a fly’. DCP contains no CPs with the component *trobenta* ‘trumpet’; with *glasen* ‘loud’, there is one adjectival CP (*glasen kot Čič* ‘loud as an Istro-Romanian’). In contrast, *muha* ‘fly’ as a comparative element is more productive in phrase formation, appearing in CPs such as *pijan kot muha* ‘drunk as a fly’.

The second criterion of interest was whether the listed phrasemes truly represent CPs with the structure A + *kot* + N. Three expressions do not meet this criterion. One of them is *tiho kot miška* ‘quietly as a mouse’, which is an adverbial CP with the same motivation as the adjectival CP *tih kot miška* ‘quiet as a mouse’, the latter appearing in both minima (seventy-fifth place in MFmin and ninety-fourth in CLmin). The adverbial CP is also attested in use, with approximately three hundred examples in MF and around two hundred in CL—suggesting that it is very likely among the most frequent CPs with the structure adverb + *kot* + noun. It is also recorded in DCP.

List A also includes forms with an additional adjectival component: *grd kot smrtni greh* ‘ugly as a mortal sin’ and *slab kot mušji drek* ‘bad as fly shit’. For the former, frequency and idiomaticity can be confirmed through usage examples (fifty-six examples in MF and twenty-eight in CL), whereas the latter is not found in the corpora or DCP, and its typological validity therefore cannot be confirmed. All forms on List B—both those whose phraseological status can or cannot be confirmed through corpus data or DCP attestations—at least formally correspond to the structure A + *kot* + N.

Considering both levels of typological validity, list A is successful in 85% of cases (seventeen out of twenty expressions), and list B in 90% (eighteen out of twenty).

Phrasemes that appear in DCP as well as in all the sources examined (MF, CL, List A, and List B), and thus form the very core of the representative CPs with the structure A + *kot* + N, include *čist kot solza* 'clean as a tear', *črn kot noč* 'black as night', *hladen kot led* 'cold as ice', *lačen kot volk* 'hungry as a wolf', *lep kot slika* 'pretty as a picture', *počasen kot polž* 'slow as a snail', and *zvest kot pes* 'faithful as a dog'.

2.3.2 Evaluating the Representativeness of Phrasemes from the Perspective of Form as Confirmed in Contemporary Use

In addition to assessing whether the AI-listed forms are truly phrasemes, the focus was also on whether they are sufficiently representative forms of phrasemes. These are all typologically valid forms that appear in at least one of the minima (with a corpus frequency of at least twenty-nine occurrences in MFmin and nineteen in CLmin). Among the forms not included in the minima, all those are considered representative that appear at least five times in the given corpora (the dictionary threshold for eSSKJ) and that are the most frequent forms of a specific phraseme—for example, *slep kot krt* 'blind as a mole' (MF: sixteen occurrences, CL: eleven, whereas DCP lists only *slep kot kura* 'blind as a hen'); *bogat kot Krez* 'rich as Croesus' (MF: eleven, CL: eight, also in DCP); and *mrtev kot kamen* 'dead as a stone' (MF: ten, CL: three, with DCP listing *hladen/mrzel/trd/gluh/težek kot kamen* 'cool/cold/hard/deaf/heavy as a stone'). Other expressions are considered representative if they are only slightly less frequent than the primary variant of the phraseme; for instance, *preprost kot pasulj* 'simple as beans' (MF: twenty-five, CL: thirty-one, also in DCP), which is included only in CLmin, compared to the slightly more frequent variants *enostaven kot pasulj* (in CLmin only; MF: thirty-eight, CL: eighty-two, not in DCP) and *simpl kot pasulj* (included in both minima: MF: thirty-eight, CL: ninety-six, not in DCP).

Searches in the MF and CL language corpora revealed that the following AI-generated expressions are not representative:

1. Those that occur in the corpora with at least one example but are significantly less frequent (or even marginal in frequency) compared to the most common form of the same phraseme. For example, on list A: *bister kot biser* 'clear as a pearl' (MF: one, CL: none; absent in DCP) as a variant of the CP *čist kot kristal* 'clear as crystal' (MF: twenty, CL: eighteen, present in DCP); on list B: *pameten kot lisica* 'clever as a fox' (MF: none, CL: one) as a variant of *zvit kot lisica* 'cunning as a fox' (MF: sixty-three, CL: twenty-seven), and *ubog kot miš* 'poor as a mouse' (MF: one, CL:

none) as a variant of the CP with a different structure: *reven kot cerkvena miš* 'poor as a church mouse' (MF: c. 360, CL: c. 120).

2. Those that are completely absent from contemporary corpus texts: on list A, *pameten kot knjiga* 'clever as a book', and, on list B, *glasen kot trobenta* 'loud as a trumpet' and *šibek kot muha* 'weak as a fly'.

These forms were also searched for—both forms absent from corpora and with very low frequency (a single occurrence)—in DCP, but they were not found in identical form there either. The possibility that some of them are very rare variants of phrasemes already confirmed by at least one example (examples under point 1) or variants of as yet unconfirmed phrasemes (examples under point 2) cannot be ruled out, and it is also very likely that some of them were simply fabricated (AI hallucinations).

Taking into account the non-representative forms from points 1 and 2 (as well as the typologically inappropriate forms from the list A discussed earlier), the success rate of AI's list A is 75% (fifteen out of twenty expressions), and the list B scores 80% (sixteen out of twenty).

2.3.3 The Possibility of Generating Larger Sets of Relevant Phrasemes: Toward an AI- and Corpus-Driven Phraseological Minimum

There is not enough room in this article to analyze longer lists from AI responses or a larger number of AI responses. Nevertheless, the fact that the two AI-generated lists of expressions presented above differ, each introducing new relevant expressions, already suggests that repeated identical queries could yield even more relevant results from AI. The same question presented above was submitted to GPT-4o nine more times on various days up to May 27th, 2025. Also considering the first of the two original lists obtained from AI on May 15th, this yields two hundred expressions listed. Among them, seventy-nine were unique and thirty-six appeared two or more times. Among the more frequently mentioned expressions, those included in the corpus-based phraseological minima predominate. An exception due to irrelevance of the form is the expression *pameten kot lisica* 'clever as a fox', which appeared eight times, although just one example of this form was found in actual use. This illustrates the significant influence of translation from English in answer generation (see above). The results from this multi-day querying—like the results obtained in queries about the most relevant Slovenian proverbs (Meterc & Mrvič, in press)—are more relevant than asking AI for a long list (e.g., 150 expressions) in a single chat session. With repeated querying, one could create a relatively extensive AI-driven phraseological minimum, which would nevertheless still need to be verified using a language corpus.

3. Conclusion: AI- and Corpus-Based Phraseological Minimum as an Empirical Starting Point of Modern Phraseography

This article presented the creation of two phraseological minima for CPs with the structure adjective + *kot* + noun and it proposed ways to improve and refine them (e.g., by addressing corpus lemmatization issues and grouping variants of the same CP). The two minima derived from different corpora largely overlap (by 87%), which confirms the existence of a reliable core set of CPs with the structure A + *kot* + N. This core is particularly valuable for lexicographic work because it is empirically verified through multiple corpora. Using the data from these minima, the performance of GPT-4o's responses was able to be evaluated. It can be concluded that the concept of a phraseological minimum for a phraseme structure can serve as a valuable addition to the phraseographic process, which, in general dictionaries, primarily involves the analysis of individual phrasemes based on single-word headwords.

AI-generated answers about the most representative Slovenian CPs with the structure A + *kot* + N are somewhat reliable for non-linguist users, but they may also include expressions that are not attested in Slovenian, making them less dependable than dictionary-based data. However, these outputs hold greater potential for lexicographers, who can use them as sets of "likely candidates" and then evaluate their validity using language corpora. The two twenty-item AI lists were estimated as being 85% to 90% typologically appropriate and 75% to 80% appropriate in terms of formal correctness and contemporary usage. Interestingly, the results for CPs are comparable to those from a similar study of Slovenian proverbs (Meterc & Mrvič, in press), in which 80% to 100% of items were typologically appropriate and 70% to 80% formally appropriate. Both cases show that AI, when used iteratively and in combination with corpus validation, can support the construction of a new type of AI- and corpus-based paremiological or phraseological minimum.

Among the fifty-one CPs already prepared for the eSSKJ dictionary database, twenty appear in both minima and twenty-seven appear in at least one. Ten of these were also found in both AI-generated lists: *čist kot solza* 'clean as a tear', *gladek kot svila* 'smooth as silk', *lačen kot volk* 'hungry as a wolf', *preprost kot pasulj* 'simple as beans', *priden kot čebela* 'hard-working as a bee', *slep kot krt* 'blind as a mole', *suh kot trska* 'thin as a splinter', *težek kot svinec* 'heavy as lead', *trmast kot osel* 'stubborn as a mule', and *zvest kot pes* 'loyal as a dog'. This indicates that, despite the currently low number of A + *kot* + N phrasemes in the dictionary, a surprisingly high proportion appear in both MFmin and CLmin (c. 40% in both, 53% in at least one), as well as among AI results (c. 20%).

Considering that around a quarter of the top hundred CPs from both minima (twenty-two from MFmin, twenty-five from CLmin) are already included in the dictionary database, it

can be expected that the full core set will eventually be represented in eSSKJ. In addition, the dictionary will feature other CPs that are not among the most frequent but are still sufficiently common to be lexicographically relevant (e.g., *ponosen kot pav* 'proud as a peacock', which just meets the inclusion threshold with five corpus attestations in GF).

The forms found in MFmin and CLmin usually correspond to the canonical forms presented in eSSKJ. AI-generated expressions that were confirmed in corpora also tend to match these canonical forms. An interesting example is *rdeč kot rak* 'red as a crab', which appears in both minima (MFmin: ninety-eighth, CLmin: sixty-ninth), but is represented in eSSKJ as a variant of the phraseme *rdeč kot kuhan rak* 'red as a cooked crab', with the meanings: 1. very red; 2. deeply blushing, flushed with anger or excitement.

This study demonstrates how identifying the core of the most representative CPs with a specific structure—through corpus and/or AI-driven methods—supports modern phraseography by helping determine which forms are the most relevant (canonical and frequent variants) and which are less so. Moreover, this approach allows more precise semantic analysis of frequent phrasemes, which can guide the interpretation of less frequent but structurally similar CPs. For example, the polysemous phraseme *čist kot solza* 'clean as a tear' provides an exemplary case for the semantic analysis of other *čist kot N* expressions with different motivations and lower frequencies.

Information about the most representative CPs with the same or similar structure allows lexicographers to view a phraseme within a broader phraseological landscape, thus facilitating the analysis of its formal and semantic properties. In the future, further research will be needed to explore the potential of developing specialized (phraseological) dictionaries based on such phraseological minima. Based on the analysis in the article, we assess that within the framework of phraseography, language corpora remain the central and most reliable tool, continuing to improve with the increasing volume and variety of corpus types. Therefore, at least at present, artificial intelligence does not represent an alternative that could replace or substitute them, but rather offers a research-interesting complement, providing additional information that is useful to verify with language corpora.

4. Acknowledgements

The research presented in this article was conducted as part of the research program Slovenian Language in Synchrony and Diachrony (P6-0038) and the research project Language, Culture, and Values: The Economic Image of Everyday Life in Folkloric Patterns (J6-50197).

5. References

Bojc, E. (1987). *Pregovori in reki na Slovenskem.* Ljubljana: Državna založba Slovenije.

Čermák, F. (2007). *Frazeologie a idiomatika česká a obecná.* Prague: Univerzita Karlova v Praze, Karolinum.

Čermák, F. (2013). *Základní slovník českých přísloví.* Prague: Lidové noviny.

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4), pp. 355–387.

Dobrovol'skij, D. (2014). The Use of Corpora in Bilingual Phraseography. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress. The User in Focus.* Bolzano: Institute for Specialised Communication and Multilingualism, pp. 867–885.

Ďurčo, P. (2014). Empirical Research and Paremiological Minimum. In H. Hrisztova-Gotthardt & M.A. Varga (eds.) *Introduction to Paremiology: A Comprehensive Guide to Proverb Studies.* Warsaw: Versita, pp. 183–205.

Erjavec, T. (2023). Corpus of Combined Slovenian Corpora metaFida 1.0, Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042. Available at: http://hdl.handle.net/11356/1775.

eSSKJ: Slovar slovenskega knjižnega jezika. Accessed at: www.fran.si (July 2025).

Fink-Arsovski, Ž., Kržišnik, E., Ribarova, S., Dunkova, T., Kabanova, N., Trostinska, R., Mironova Blažina, I., Spagińska-Pruszak, A., Vidović Bolt, I., Sesar, D., Dobríková, M., & Kursar, M. (2006). *Hrvatsko-slavenski rječnik poredbenih frazema.* Zagreb: Knjigra.

Gantar, P. (2002). Temeljne prvine zasnove frazeološkega slovarja. *Slavistična revija* 50(1), pp. 29–49.

Gantar, P. (2006). Corpus Approach in Phraseology and Dictionary Applications. *Slavistična revija* 54(1), pp. 161–162.

Gantar, P. (2007). *Stalne besedne zveze v slovenščini. Korpusni pristop.* Ljubljana: Založba ZRC, ZRC SAZU.

Hupkes, D., Dankers, V., Mul, M., & Bruniet, E. (2020). Compositionality Decomposed: How Do Neural Networks Generalise? *Journal of Artificial Intelligence Research* 67, pp. 757–795.

Jakubíček, M., & Rundell, M. (2023). The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography? In: *Electronic Lexicography in the 21st Century (eLex 2023). Proceedings of the eLex 2023* Conference. Brno: Lexical Computing CZ, pp. 518–533.

Keber, J. (2011, 2015). *Slovar slovenskih frazemov.* Ljubljana: Založba ZRC, ZRC SAZU.

Kocijan, K., & Librenjak, S. (2016). Comparative Idioms in Croatian: MWU Approach. In G. Corpas Pastor (ed.) *Computerised and Corpus-Based Approaches to Phraseology: Monolingual and Multilingual Perspectives.* Geneva: Editions Tradulex, pp. 523–532.

Ljubešić, N., Rupnik, P., & Kuzman, T. (2024). *Slovenian Web Corpus CLASSLA-web.sl 1.0,* Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042. Available at: http://hdl.handle.net/11356/1882.

Logar, N., Erjavec, T., Krek, S., Grčar, M., & Holozan, P. (2013). *Written Corpus ccGigafida 1.0*, Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042. Available at: http://hdl.handle.net/11356/1035.

Meterc, M., & Jakop, N. (2016). Lexikografické spracovanie frazeologických variantov v novom slovníku slovinského spisovného jazyka. In M. Lišková (ed.) *Akademický slovník současné češtiny a software pro jeho tvorbu aneb Slovníky a jejich uživatelé v 21. století: sborník abstraktů z workshopu, Praha, 29.–30. listopadu 2016*. Prague: Ústav pro jazyk český AV ČR, pp. 55–56.

Meterc, M. (2017). *Paremiološki optimum: najbolj poznani in pogosti pregovori ter sorodne paremije v slovenščini.* Ljubljana: Založba ZRC, ZRC SAZU.

Meterc, M. (2020–). *Slovar pregovorov in sorodnih paremioloških izrazov.* Available at: www.fran.si.

Meterc, M., & Mrvič, R. (in press). The Best-Known and Most Frequent Slovenian Proverbs, Listed by ChatGPT-4o: The Possibility to Create an AI-Supported/Based Paremiological minimum. *Linguistica.*

OpenAI (2025). *ChatGPT-4o (version from May 2024) [Large Language Model].* Available at: https://chat.openai.com (accessed July 9th).

Permjakov, G.L. (1971). *Paremiologicheskiy eksperiment: materialy dlya paremiologicheskogo minimuma.* Moscow: Nauka.

Permjakov, G. (1985). *300 obshcheupotrebitel'nykh russkikh poslovits i pogovorok (dlya govoryashchikh na nemetskom yazyke).* Moscow: Russkiy yazyk.

Varga, M.A., & Babić, S. (2023). Kroatische Sprichwortvarianten bei der Erstellung des kroatischen parömiologischen Thesaurus. *Yearbook of Phraseology* 14(1), pp. 147–164.

Wendler, C., Veselovski, V., Monea, G., & West, R. (2024). Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 1. Konstanz: KOPS Universität Konstanz, pp. 15366–15394.

List of Abbreviations

A + *kot* + N = CP with the structure adjective + conjunction *kot* 'as' + noun
AI = artificial intelligence
CL = CLASSLA-web.sl corpus 1.0
CLmin = phraseological minimum of comparative phrasemes from the CL corpus
CP = comparative phraseme
DCP = Database of Comparative Phrasemes
eSSKJ = *Standard Slovenian Dictionary*, third edition

GF = Gigafida 1.0 corpus
GPT-4o = ChatGPT-4o model
MF = metaFida v1.0 corpus
MFmin = phraseological minimum of comparative phrasemes from the MF corpus
N = noun