A Data-Driven Approach to Idiomaticity in Russian MWEs Based on Experts' Criteria in Theoretical Linguistics

Anonymous ACL submission

Abstract

The article observes data analysis of 285 Russian multi-word expressions (MWEs) based on 15 lexical, grammatical and other criteria described in theoretical books and papers on the concept of idiomaticity. The MWEs were collected from the same theoretical sources as the criteria, and a set of experts in linguistics annotated them with these criteria. The distribution of scores in the annotated dataset shows that there are no absolutely idiomatic expressions, and some expressions are clusters of several MWEs. Lexical criteria are among top-scorers and seem to be the most manifested; grammatical criteria are bound to certain conditions; presence of obsolete words and grammar influence ability of an MWE to be replaced with one word. The analysis can be used to build a novel classification of MWEs and as a method for their automatic extraction.

1 Introduction

Multi-word expressions (MWEs), groups of lexemes occurring in a text and more complex linguistically, than just a free word-group, can confuse automatic text processing in many ways. Even the terminology surrounding them is quite extensive and lacks commonly accepted definitions, hence, often relying on an approach to their treatment. What unites these approaches is understanding that lexemes in MWEs cannot be treated fully as their equals in free word-groups. This property of MWEs is often referred to as *idiomaticity*. But to what extend is idiomaticity manifested in linguistic features of MWEs? The answer to this question can help building a data-driven classification and facilitating their automatic extraction via feature engineering. In our study, we try to oversee one side of it that has not been granted due attention: a data analysis of idiomaticity based on theoretical criteria of MWEhood and expert annotation of a gold standard dataset of Russian MWEs.

First, we describe theoretical and modern applied approaches to classification of MWEs paying more attention to what can be called "data-driven" approaches. Second, we propose a model of 15 linguistic criteria that were derived from theoretical works by linguists. The model encompasses lexical, semantic, grammatical and pragmatic criteria. Third, we take MWEs from works of Russian theorists and label them with these criteria (whether a related feature is manifested in an MWE or not). [1] Fourth, we group the criteria into four categories and perform data analysis on the resulting vectors. Finally, we perform clustering analysis and hypothesize what it shows about the nature of idiomaticity and how it can be extended to an algorithm of MWE extraction.

2 Approaches to Classification of MWEs

(Baldwin and Kim, 2010) underline that MWEs allow to use a comparatively brief lexicon to create nuances of meaning. The lack of freedom in MWEs, or *vice versa* strength of connection, is also referred to as *idiomaticity* – "markedness or deviation from the basic properties of the component lexemes" (Ibid.). Idiomaticity shows in "lexical, syntactic, semantic, pragmatic, and/or statistical levels" (Ibid.).

Statistical criteria of "MWEhood". Statistical methods focus on inferring idiomaticity from co-occurrence of lexemes inside a certain word-group and in free contexts. Among the most common statistical measures used in this task are *mutual information* (Church and Hanks, 1990), *likelihood ratio tests* (Dunning, 1993), *cost criteria* (Kita et al., 1994). Pecina (2008) enlists 55 "lexical association measures used for ranking MWE candidates". The output of these methods is a number that evaluates the strength of idiomaticity. Based on it, MWEs

[1] We use this method of collecting MWEs because we are of the opinion that these examples, suggested by theorists, are a gold standard demonstrating typical features of idiomaticity.

are arranged (ranked), but are hardly classified. *Vice versa* the process usually leads to understanding what type of MWE is better derived with the method. It would not be wrong to state that there is no universal criterion to MWE extraction, and statistical methods are applied to existing classifications.

The term *collocation*, often used in papers describing statistical methods, can be considered a synonym to *MWE*, although Baldwin and Kim (2010) put it that collocations are statistically idiomatic MWEs. In fact, we tend to observe that statistical methods extracting this or that type of MWE are designed without limitations. Hence, any MWE can look statistically significant with a proper measure. Another way to disambiguate between an MWE and collocation is that some collocations lack *non-compositionality* – they are *compositional*, their meaning is easily extracted from lexemes composing them. E.g. *Many thanks!* is a statistically significant proper way of being thankful, but its meaning is clear from the words composing it. The (non-)compositionality cannot be statistically inferred from word frequencies.

In some literature, MWEs are synonymous to *multi-word units* (MWUs), "lexical items that go beyond single word items" (Shin and Chon, 2019), and *words-with-spaces*, "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2002). In our paper, we intentionally do not make any particular distinction between MWEs, collocations, MWUs and words-with-spaces, considering them to be manifestations of the same phenomenon – idiomaticity.

Expert classifications. Baldwin and Kim (2010) suggest a classification that first splits MWEs into two main classes (see fig. 1): institutionalized phrases are collocations proper (statistically common phrases such as "Many thanks!") and lexicalized phrases show idiomaticity to a certain degree and are marked by different features (e.g. decomposable / non-decomposable). We believe that such an approach to classification, although it is supported by examples, cannot be called data-driven; rather, it is an expert view of a complex phenomenon. Also, the bottom level of the obtained hierarchy (VNIC, nominal, VPC, LVC) is based on parts-of-speech analysis of English collocations, hence, binding idiomaticity to one particular linguistic feature. However, in Table 12.2 from the same chapter Baldwin and Kim (2010) approach classification of MWEs from another perspective:

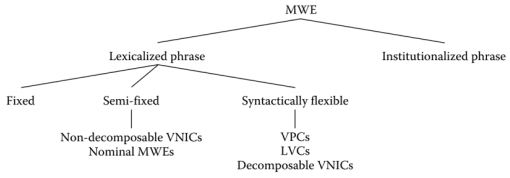

Figure 1: Classification of MWEs by (Baldwin and Kim, 2010). VNIC - verb-noun idiomatic combination; VPC - verb-particle construction; LVC - light-verb construction.

	Lexical	Syntactic	Semantic	Pragmatic	Statistical
all aboard	–	–	–	+	+
bus driver	–	–	+	–	+
by and large	–	+	+	–	+
kick the bucket	–	–	+	–	+
look up	–	–	+	–	+
shock and awe	–	–	–	+	+
social butterfly	–	–	+	–	+
take a walk	–	–	+	–	?
to and fro	?	+	–	–	+
traffic light	–	–	+	–	+
eat chocolate	–	–	–	–	–

Figure 2: Classification of MWEs in terms of their idiomaticity, by (Baldwin and Kim, 2010).

they enlist properties of MWEs and annotate several examples with these properties, acquiring a matrix of feature distribution from which they conclude about the probability of "MWEhood", see fig. 2. In our opinion, this approach could be called data-driven, as classes are inferred from annotations. But the examples were few and were chosen so as to demonstrate several cases referring to predesigned classes.

PARSEME [2], a large European project that started in 2014 and focused on annotation of MWEs, split them into similar classes, see Appendix B. Resembling classification in figure 1, this project lays weight on part-of-speech properties of the headword in a phrase and splits the set into distinct subgroups. We tend to believe that such an approach was organizational: it is important to split jobs in large projects. Some other possible drawbacks in it are that it has to group non-standard examples into "other" and that elements in the resulting hierarchy are not of the same level, theoretically. E.g., from the point of view of theoretical linguistics, named entities are represented by proper nouns (if we disregard anaphora) as opposed to all common nouns – at the same time, Noun+Noun compounds are a subgroup of common nouns.

Another approach, yet leading to abandoning all classifications, is found in (Schneider et al., 2014). The authors aimed to annotate a corpus for DiM-

[2] https://typo.uni-konstanz.de/parseme/index.php

SUM (Ibid.), a SemEval task for detecting minimal semantic units and their meanings. They collected a set of classes of idioms in English, that totaled 15 as illustrated in their article. Beside some of the classes, mentioned by Baldwin and Kim (2010), it included named entities, compound words (*motion picture*), support and phrasal verbs (*make decision, cry foul*), coordinated phrases (*cut and dry*), phatic phrases (*You're welcome!*), proverbs (*To each his own*), etc. Annotators in this project only marked what they considered to just be an MWE. It is of interest that the authors did not find any particular POS pattern that could be associated with a collocation type: "Categorizing MWEs by their coarse POS tag sequence, we find only 8 of these patterns that occur more than 100 times" (Schneider et al., 2014).

In CoNLL-U Format from the Universal Dependencies project [3], multi-word tokens, simply, "are indexed with integer ranges like 1-2 or 3-5" (as stated at the given website).

Expert criteria. Another way to build a classification theoretically is to enlist various linguistic criteria according to which something can be defined as an MWE and terminologically label gold examples demonstrating these criteria. Such is an approach by Vinogradov (1977), who adapted a classification by the Swiss scholar Charles Bally and singled out two types of MWEs: less idiomatic combinations and more idiomatic fusions, see Appendix C. Outside the scope of his classification, Vinogradov (1977) placed terminological groups and named entities.

Without building a classification, Manning and Schutze (1999) describe three criteria that characterize collocations: non-compositionality, non-substitutability (lexemes cannot be substituted with synonyms), non-modifiability (lexemes cannot change grammatically). And again several types are mentioned separately: light verbs, verb-particle constructions, proper nouns, terminological expressions.

Cowie and Howarth (1996) suggest the following criteria: familiarity to speakers, ability to be stored in memory as ready-made units, limited and arbitrary variability, opaque semantics.

Tarasevitch (1991) makes use of her own list: stability of use, structural separateness, complexity of meaning, being not built on the generative pattern of free word-groups.

Mel'čuk (1960) considers that idiomaticity influences translation of a phrase or its parts. In a more idiomatic and stable expression it is hard to find an exact match to every lexeme and the whole phrase is easier to translate with a singe word. Baldwin and Kim (2010) also mention pragmatic idiomaticity (being associated with a certain situation), proverbiality (describing a situation of social interest), prosody, but we will leave these criteria outside the scope of our research as they require to go beyond the study of a written text.

In this paragraph, we might have overlooked some criteria, but as far as we know in other works approximately the same criteria repeat.

What approach can be called data-driven? Amin et al. (2021), who call their approach data-driven in the title of their paper, design metrics that help to infer some of the mentioned above criteria for n-grams in a text. A similar scheme is traced in (Rossyaykin and Loukachevitch, 2019) who use a set of statistical, context and distributional measures to infer MWEs from a corpus. Another clustering method, based only on association measure, is found in (Tutubalina, 2015). Nissim and Zaninello (2013) introduce variation patterns as an alternative to association measure. Wahl and Gries (2018), who call their approach "bottom-up", introduce the MERGE algorithm, again as an alternative to association measure (the project is developed by Gries (2022)). Summing up, data-driven are projects in which MWEs are represented as vectors in multi-dimensional space and analysed statistically. Often these vectors are visualised in diagrams to see whether there are clusters that attract more MWEs.

3 Experiment Setup

The intuition lying behind our approach is that the theoretical linguistic expertise about the phenomenon of idiomaticity makes it look like a single unity that can be split into classes, or types of MWEs. However, seeing it as an umbrella term that unites phenomena of different linguistic nature [4] leads us to a hypothesis that we can describe this heterogeneity and outline these phenomena as clusters in a multi-dimensional space, based on vectorization of annotated examples. To visualize these clusters, we propose a data-driven approach. We believe that we do not need a large collection

[3] https://universaldependencies.org/format.html#conll-u-format

[4] E.g., as mentioned, lexical: named entities versus noun compounds; grammatical: diversity of POS-patterns.

for such a task if we have a gold standard that manifests the main features of idiomaticity. Further, in our experiment we propose: a. a set of criteria (features of idiomaticity), b. the gold standard dataset of Russsian MWEs, c. expert annotation, and d. a data analysis of the annotated dataset.

Criteria. The linguistic criteria that we took from the theoretical literature total 15 – see Table 1. To briefly describe them, we grouped them into four categories, resembling the list by Tarasevitch (1991):

- **lexical change** includes synonymizing, translation and insertion of lexemes, i.e. such lexical changes that do not transform a word group into its shorter version,
- **grammatical change** encompasses ability of changes in morphology and syntax,
- **obsolescence** is presence of archaic grammar or vocabulary [5],
- **replacement** shortens a word group, e.g. when a group can be synonymized with one word.

In **replacement**, by the ellipsis, we mean that a word or more can be omitted in an MWE without change of meaning. And portmanteau words are several trailing wordforms "sewn" into one compound wordform. As for translation, it was into English. To check translation word by word, our experts looked for English words (including composite wordforms, e.g. "would like to") with approximately the same denotative meaning that form a meaningful phrase in English corresponding to the Russian expression. E.g. in our example if we translate "белый гриб" word by word, we get *white mushroom* which is a different kind of mushroom. The meanings do not correspond, hence, the phrase cannot be translated word by word.

The criteria were formulated so as to demonstrate idiomaticity, which means that 0 denotes its lack and 1 – its presence. For example, if a word composing a word group can be replaced with a synonym, it means that the group is more or less free: *a tasty / yummy / nice mushroom* as opposed to *a penny bun*, where penny cannot be synonymized. A criterion that we did not include due to this factor is POS pattern: as mentioned, its unclear how it influences idiomaticity.

[5]We included presence of unique words here as well. A unique word remains only in a particular MWE.

Some criteria expose interdependence due to syntactical structure, in particular, government in phrases. If an MWE does not contain government, criteria that separate application to headwords and governed words cannot be applied (this applies to synonymizing and change of grammatical form). And in government, two criteria (v and vi) are mutually exclusive as they are applied either to the headword or to the whole MWE. This means that expressions without government can score 12 points maximum, and expressions with government can score 14 points maximum. Although this can lead to multicollinearity, at this stage of research we prefer to preserve annotation of these criteria due to the following reasons. First, of two dependent criteria, one can be a manifestation of a stronger idiomaticity. For example, if the headword can be replaced with a synonym, it feels that such an MWE is less idiomatic than when just a governed word can be replaced, e.g. *a restless person / girl / man* is (it feels like..) less idiomatic than *deep / late night*. Second, it is of interest with what other criteria these two criteria correlate. When our dataset is extended it may be necessary to separate MWEs with government and, probably, some other grammatical features. For now, in cases when a grammatical criterion is not applicable it was annotated 0. As for interdependence in lexical criteria, it seems that if a Russian MWE can be synonymized with one word in Russian, it is also very likely to be translated with one word into English. However, in the majority if cases (195 out of 285) our experts put different scores for the two criteria.

Also, we excluded one criterion from the initial list: (partial) loss of independent meaning. Most of our annotated examples scored 1 in it. We tend to think that this is the target criterion, and experts hesitated to annotate all MWEs with it, because they expected that a criterion cannot be absolute. Besides it is unclear how to prove that independent meaning is lost. Our final list of criteria includes only those that allow to give a precise instruction on how to annotate an MWE [6]. E.g. *Governed word cannot be replaced with a synonym* is checked with an attempt to change a governed word in an MWE in several possible contexts, e.g. "Я набрал белых грибов." (*I've picked some penny buns.*) does not allow such a change.

An annotated MWE is a vector of zeros and ones, e.g. the vector for "белый гриб" (En. *penny*

[6]They will be discussed in the next paragraph.

bun) is (1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1), see Table 1; its vector sum is 9. The higher is the vector sum of an annotated MWE, the more idiomatic an MWE is.

To our annotation, we also added four linguistic features that are not directly bound to idiomaticity, but are often found as arbitrary criteria: POS-pattern, "Is a sentence?" (whether the MWE contains a subject and/or predicate), headword (if it is not a sentence), phrase structure (e.g. government, agreement, etc.). An example is given in Table 2. These features can be used in further research.

Gold standard of MWEs. For annotation we took 285 examples of MWEs, mainly from (Vinogradov, 1977). The exact sources are enlisted in our GitHub repository at REDACTEDFORANONYMITY. The MWEs were collected without any preselection – all examples that we could find in the papers excluding duplicates. They are also given in the grammatical form given in the source. It should be noted that, where it is possible, scholars put words in their initial form.

Expert annotation. Our annotators, experts with higher education in linguistics, were instructed about the criteria described above. One expert annotated the set of MWEs with one criterion, then another expert looked it through and checked if they agree with the result. Questionable cases were discussed at seminars and led to a final decision about the annotation. However, we expect that the annotation can slightly change due to new arguments about this or that case [7]. In annotation of 285 MWEs with 15 criteria (which totals to 4,275 annotated features), there were 283 registered cases ($\tilde{7}$%) of disagreement between the annotators that were solved via consensus.

Also, although the experts used grammar and other reference books, dictionaries, corpora and web to check their expertise, it is possible that they could overlook something, or still disagree even after the final decision about an MWE was made [8]. Hence, our annotation should be considered as an expert *approximation* of the real world.

To double-check some of the experts' annotation with NLP tools, we used the Russian National Corpus [9]; the criterion "iv. Does not allow insertions of lexemes" was checked with an expression "WORD1 * WORD2" where * shows that any word can be placed inside an MWE. The category **grammatical change** was checked similarly, with * replacing grammatical morphemes. In cases when only one corresponding example was found, we treated it as an occasional creative use of the language and, hence, equal to zero.

4 Data Analysis

We will now describe our resulting dataset and attempt to find patterns in the distribution of annotated scores. Before our analysis, we merged two criteria of grammatical form (v. and vi.) by taking their sum, because they are interdependent. Hence, the number of criteria in our data analysis is 14.

Due to our criteria being demonstrative of idiomaticity, we expect that the higher is the sum of scores for a given expression, the more idiomatic it is. Figure 3 shows frequency distribution of scores. It is right-skewed: 256 MWEs (90%) score below 7, which is the median in Fig. 3. That supports our idea of two or more trends that break MWEs into mutually exclusive subgroups. On average, an MWE scores 5, which is 36% of the maximum (14). The minimum score is 1, in three expressions: глубокая ночь, неизгладимое впечатление, принять решение (*deep night, memorable impression, to make a decision*). They seem to be typical MWUs, combinations that learners of a language memorize to sound more like native speakers. There are no MWEs that score 11 and above. The only top-scoring expression (10) is диву даться (*to be amazed*). It contains an obsolete wordform and archaic grammar, like four MWEs scoring 9: во свояси, и вся недолга!, очертя голову, спустя рукава (*home, That's it!, headlong, carelessly*). This leads us to an impression that obsolescence correlates with idiomaticity more than others. It appears that the perfect MWE does not exist: some of the criteria either exclude others or make it very difficult to combine them in one expression. Also, two of the top-scorers (диву даться, и вся недолга! are used as clauses in complex and compound sentences. It can be that idiomaticity rises with a more "sentence-like" structure of an expression.

The sum of scores (ranged from the smallest to highest) for every criterion in Figure 4 shows that about nine criteria are below the average ($\tilde{1}$03 scores per a criterion) and only 5 are above. The

[7] It would be impossible to manually test some of the properties in all possible contexts, even using NLP-tools for support.

[8] Many such cases were about the question whether a word is already obsolete.

[9] https://ruscorpora.ru/

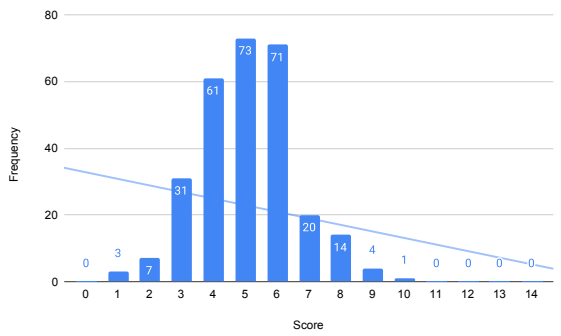

Figure 3: Frequency of scores.

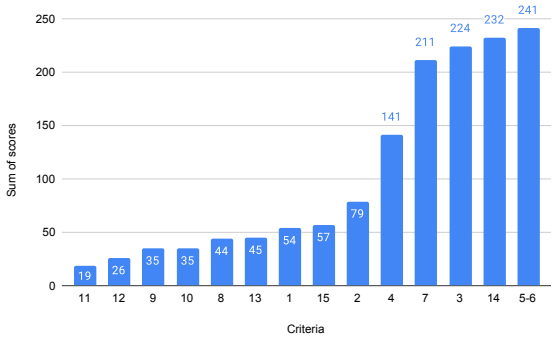

Figure 4: Distribution of scores by criteria.

lowest scoring criterion is ellipsis (19), when an expression allows to omit one or more of its words, and portmanteau (26), when words composing an expression are coined into a single wordform: втирать очки – втирать – очковтирательство (*to tell lies – to lie – lying*). Also, presence of archaic vocabulary and/or grammar. Beside that these can actually be rare features of MWEs, it can be difficult to annotate these criteria. E.g. obsolete and unique words need vast expertise in both old and modern Russian. Also, it can be hard to distinguish between an archaism and a bookish, formal modern word.

Four criteria score very high, and that means that they do not contribute to a stricter division between probable classes (they fill the vector space densely). The top-scorer here is the two joint criteria of change in grammatical form: they encompass 85% (241 MWE) which makes us think that they should be separated on a condition bound to the syntactical structure (phrases with governemtn, a predication center, etc.). Other top-scorers are "Word order never changes" (again a grammatical criterion) – 211 MWE (74%), "Cannot be translated word by word" – 224 (74%), "Can be replaced with one word" – 232 MWEs (81%). It is unclear what unites them beside that they are more manifested in all MWEs, probably, more bound to idiomaticity and, hence, should be used for classifying between MWEs and non-MWEs.

Out of 285 MWEs 154 are unique vectors (54%). It is hard to say whether our annotation exhibits strong general patterns in distribution of all the criteria across the dataset. We will now try to generalize about what happens in each of the defined categories.

Lexical change. This category is the champion in earned scores: 498 (34% of all scores), which is supported by the idea of idiomaticity being more of a lexical nature, influencing primarily the lexical meaning of words. I.e. any candidate for an MWE should score high in it. Some examples of zero-scorers in this group are неизгладимое впечатление, вечерняя газета, хороший тон, положительный тип *memorable impression, evening paper, good manners, positive person* – MWUs composed of a noun and its modifier.

Grammatical change. This category is the second in the earned scores (452, 31%) and can be called influential as well. Although there can be the mentioned bias in selection of examples: those selected by scholars tend to be more fixed structurally, to better demonstrate fixedness. Expressions that score 0 in it are again nouns with modifiers (8 out of 9) (беспробудное пьянство, потрясающее впечатление *deep drinking, stunning impression*). Also, presence of a verb usually makes expressions in Russian as well as in English less fixed grammatically. Compare: пуститься во все тяжкие – во все тяжкие *do whatever it takes – whatever it takes*; the verb *do* can change its grammatical form. And, hence, we deal here with an MWE nested within another expression of a different kind.

Obsolescence. As we mentioned earlier, this category seldom earned a positive decision from our annotators: 209 MWEs (73%) score zero in it. There are only 9 collocations that score 3 in it (e.g. и вся недолга!; не до жиру, быть бы живу; ничтоже сумняшеся *end of story, survive before thrive; nothing doubting*) and some other. And the very same MWEs mainly score 0 or 1 out of 5 in the category **replacement**, but total 7, 8 and 9 scores for all the criteria (with 5 being the average score). Although it is probably clear without the data, but our experiment supports it that this

category is a strong marker of idiomaticity. To add, obsolete words make it harder to modify an expression; archaic grammar hinders morphological and other changes in different contexts. Hence, it hinders replacement as well.

Replacement. This category seems to positively correlate with **grammatical** and **lexical change**, but negatively correlates with **obsolescence**. The example in Table 1 demonstrates it. 143 expressions (50%) score 0 or 1 both in **obsolescence** and **replacement** which might mean that these two categories are not crucially important in formation of an WME, but they probably point at two distinct sub-classes.

5 Clustering

Figure 6 visualizes the distribution of MWEs in our resulting matrix with the help of tSNE algorithm. It shows that our data are not homogeneous with, probably, three or more major types of MWEs (multitudes with high density). And there appear to be two trends in distribution of vectors, that result in the ⋂ shape of the vector plot. However, tSNE cannot be used as a clustering technique. In this section, we propose two approaches to clustering MWEs: quartiles and hierarchical agglomerative clustering with 4 and 7 cluster.

5.1 Clustering into Quartiles

Our first task is to analyze whether the binary matrix of annotated MWEs can provide insights on types of MWEs that can be inferred based on how *strongly* idiomaticity shows in them. For this, we ranged our dataset by the sum of scores that each MWE got and divided it into quartiles to compare extreme quartiles in our range. The resulting bounds are: (0.999, 4.0] < (4.0, 5.0] < (5.0, 6.0] < (6.0, 10.0] with the lowest score being 1 and highest – 10. The number of MWEs in each quartile is: 111, 72, 66, 36 [10] (see Fig 5). The upper quartile with 36 MWEs naturally includes expressions that manifest maximum features of idiomaticity and contains only 12.6% of the dataset. Hence, it is hard for expressions to strengthen their idiomaticity. The lower quartile contains 38.9% of expressions among which there are many word groups of Adjective+Noun and Verb+Noun and variants of an expression given with a forward slash:

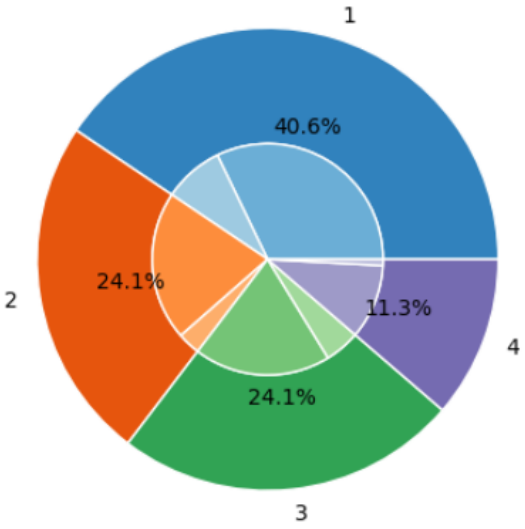

Figure 5: Distribution of idioms in quartiles. The outer circle: quartiles. The inner circle: idioms (larger sets) versus non-idioms.

глубокая (Adj. *deep*) ночь (Noun *night*) / осень / старость / печаль / мысль
вымыть (Verb *to wash*) / намылить голову (Noun *head*)
страх (Noun *fear*) / тоска / досада / злость / ужас / зависть / смех / раздумье / охота берет (Verb *to take*)

Then we checked in several dictionaries of Russian idioms which MWEs in our dataset are listed as idioms or phraseological units. It appears that 226 MWEs are idioms. The rest are mainly terminology and word groups with variation as shown in the examples above. We calculated that in the quartiles the number of non-idioms is: 29 (26% of MWEs in this quartile), 11 (15%), 17 (26%), 3 (8%), corresponding to lower, two medium and upper quartiles. Hence, the lower quartile contains 48.3% of all non-idioms, and the upper quartile - only 5% of them. [11] We would like to conclude here that, in our annotated dataset, idiomaticity does strengthen with the score of an MWE going up and the criteria prove effective in detection of idioms.

6 Hierarchical Agglomerative Clustering

Our second task was to cluster MWEs based on the distance between their vectors in the n-dimensional

[10] Such unequal distribution is due to large sets of MWEs having the same score.

[11] The median quartiles swap in range (18.3% and 28.3%), probably due to noise.

space to see whether any particular features are highlighted in a cluster. As criteria 5 and 6 are interdependent, we merged them by summing the score of the two (the result is also binary: 0 or 1). We applied several clustering techniques (DBSCAN, BIRCH, and Hierarchical Agglomerative Clustering) and found the latter most informative on our dataset. [12] The dendrogram in Figure 7 illustrates Hierarchical Agglomerative Clustering performed on the matrix [13]. With the distance between vectors approaching 8 there are 4 aggregated clusters, containing 99 (35%), 68 (24%), 93 (33%), 25 (8%) of MWEs correspondingly. The smallest cluster has one obvious distinguishing feature: all but one expressions in it are Noun+Adjective, although such word-groups are found in other clusters as well. And this cluster forms earlier than others, at the distance of approximately 4. At the distance of about 6, there are 7 clusters. One of them, containing 27 MWEs (10%), accumulates expressions that cannot change grammatically, except one MWE. And these MWEs seldom allow other changes. As for other clusters, so far, we cannot make any conclusions about them. We believe, they require more data and analysis for a steadfast conclusion.

7 Conclusion

The paper describes an attempt to search for theoretical grounds in the notion of idiomaticity with the help of linguistic annotation and data analysis of the gold standard of Russian MWEs. We have proposed a model of 15 criteria that were grouped into four categories based on linguistic analysis. We annotated a corpus of 285 Russian MWEs found in the same theoretical books from which we took the criteria. Our analysis revealed several trends.

Lexical fixedness is found in the most MWEs. It can be a target category and an idiomaticity test when MWEs are extracted. A better proof requires to compare it to non-MWEs.

Either scholars tend to choose MWEs that are more fixed grammatically, or, like lexical fixedness, it is a stable property of all idiomatic expressions. At least, in our dataset it is nearly as frequent as lexical fixedness. Also, some grammatical criteria depend on the syntactical structure of expressions, and, hence, cause multicollinearity among vectors. It maybe necessary to cluster and classify MWEs without grammatical features or inside a syntactical grammatical class that they belong to.

The idiomaticity criteria mentioned in the theoretical literature can be hard to prove and annotate, cf. identifying obsolete words. In assigning them to certain expressions, linguists have to lean on their personal expertise and intuition rather than on formal parameters. This may also be the reason why automatic extraction of MWEs is so disputable.

Our annotation of 285 Russian MWEs exhibits certain correlations. E.g. presence of archaic words and grammar and the property of an expression to be shortened or replaced by a single word seem to oppose each other and manifest two different classes of idiomatic expressions.

There are no absolutely idiomatic expressions that correspond to all the theoretical criteria of idiomaticity.

What is determined as an MWE can be several expressions with a different degree of idiomaticity.

Clustering MWEs into quartiles results in singling out Adjective+Noun and Verb+Noun word groups as less idiomatic. The Hierarchical Agglomerative Clustering also clusters Noun+Adjective MWEs together.

Future work. Future work requires a larger annotated collection. We can see several more ways of making it partially automatic. E.g. impossibility of translating an MWE word by word as well as translation with one word can be checked in parallel corpora. We plan on extension of methods to check criteria with analysis of corpora, e.g. vary morphological patterns that we used in the Russian National Corpus. Also, the criteria should be checked in free word groups to prove that these are actually criteria of idiomaticity. Furthermore, the stated correlations require a quantitative analysis. And, finally, the next big step is to make a cluster analysis to infer types of MWEs.

8 Limitations

Our work is limited by the choice of the language for annotation and by the number of annotated MWEs.

The choice of the language in the current study is conditioned by our team being experts in the Russian language. However, we believe that our

[12] With DBSCAN and the parameters $e = 0.1$ and $m = 5$, we got 19 small clusters and 132 cases of noise. Changing parameters would only increase noise or resulting in one supercluster. BIRCH resembles results of the plain Hierarchical Agglomerative Clustering which we report here.

[13] Clustering is performed with the Scikit-learn library (Pedregosa et al., 2011); number of clusters is given as 0; the linkage algorithm is Ward.

results can be applied to other languages. As we mentioned, some of the described features can be already observed in English, e.g. verbs in MWEs make expressions less fixed grammatically. Our annotated dataset contains some translation into English, and also examples can be taken from the studied English literature to create a similar corpus. A further extension can lie in application of NLP-tools to automatic annotation of such corpora as PARSEME collections with the criteria that we described.

The number of annotated MWEs is low compared to crowd-annotated projects and automatically extracted datasets. We would like to underline that our annotation is an expert one. At this point of hypothesis testing, it cannot easily attract a larger expert effort like in PARSEME, before the theoretical grounds for it are tested. Small collections serve as a marker of a phenomenon. E.g. in (Mihalcea and Strapparava, 2005) a sample of 200 one-liners is chosen for manual verification of noise in data.

9 Ethics Statement

In our research, we use publicly available data with due regard to copyright law. We also make our annotated dataset and software public in our GitHub repository, adhering to open science principles.

References

Miriam Amin, Peter Fankhauser, Marc Kupietz, and Roman Schneider. 2021. Data-driven identification of idioms in song lyrics. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 13–22.

Aishwarya Ashok, Ramez Elmasri, and Ganapathy Natarajan. 2019. Comparing different word embeddings for multiword expression identification. In *Natural Language Processing and Information Systems: 24th International Conference on Applications of Natural Language to Information Systems, NLDB 2019, Salford, UK, June 26–28, 2019, Proceedings 24*, pages 295–302. Springer.

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, and Josef Van Genabith. 2010. Automatic extraction of arabic multiword expressions. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 19–27.

Timothy Baldwin and Su Nam Kim. 2010. *Multiword Expressions*. Chapman and Hall/CRC.

Verginica Barbu Mititelu, Ivelina Stoyanova, Svetlozara Leseva, Maria Mitrofan, Tsvetana Dimitrova, and Maria Todorova. 2019. Hear about verbal multiword expressions in the Bulgarian and the Romanian wordnets straight from the horse's mouth. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 2–12, Florence, Italy. Association for Computational Linguistics.

Archna Bhatia, Paul Cook, Shiva Taslimipoor, Marcos Garcia, and Carlos Ramisch, editors. 2022. *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*. European Language Resources Association, Marseille, France.

Pawel Chrzkaszcz. 2016. Extraction and recognition of Polish multiword expressions using Wikipedia and finite-state automata. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 96–106, Berlin, Germany. Association for Computational Linguistics.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Anthony P Cowie and Peter Howarth. 1996. Phraseological competence and written proficiency. *British studies in applied linguistics*, 11:80–93.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Beatriz Fisas, Luis Espinosa Anke, Joan Codina-Filbá, and Leo Wanner. 2020. CollFrEn: Rich bilingual English–French collocation resource. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 1–12, online. Association for Computational Linguistics.

N Grégoire, S Evert, and B Krenn. 2008. Proceedings of the lrec workshop: Towards a shared task for multiword expressions (mwe 2008).

Stefan Th Gries. 2022. Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach. *Lexis. Journal in English Lexicology*, (19).

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Uxoa Iñurrieta, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka, and Kepa Sarasola. 2017. Rule-based translation of Spanish verb-noun combinations into Basque. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 149–154,

Valencia, Spain. Association for Computational Linguistics.

Kenji Kita, Yasuhiko Kato, Takashi Omoto, and Yoneo Yano. 1994. A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing*, 1(1):21–33.

Khang Nhut Lam, Feras Al Tarouti, and Jugal Kalita. 2015. Phrase translation using a bilingual dictionary and n-gram data: A case study from Vietnamese to English. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 65–69, Denver, Colorado. Association for Computational Linguistics.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

I.A. Mel'čuk. 1960. On the terms "stability" and "idiomaticity" (o terminakh "ustoichivost" i "idiomatichnost"). *Voprosy Yazykoznaniya*, 4:73–80.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538.

Malvina Nissim and Andrea Zaninello. 2013. Modeling the internal variability of multiword expressions through a pattern-based method. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(2):1–26.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, volume 2008, pages 54–61. Marrakech,[s. p.].

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Scott Songlin Piao, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for mwe extraction. *Computer Speech & Language*, 19(4):378–397.

PO Rossyaykin and NV Loukachevitch. 2019. Measure clustering approach to mwe extraction. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 562–575.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T Mordowanec, Henrietta Conrad, and Noah A Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus.

Dongkwang Shin and Yuah Chon. 2019. A multiword unit analysis coca multiword unit list 20 and collogram. *Journal of Asia TEFL*, 16:608–623.

Ranka Stanković, Jelena Mitrović, Danka Jokić, and Cvetana Krstev. 2020. Multi-word expressions for abusive speech detection in Serbian. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 74–84, online. Association for Computational Linguistics.

Maria Tarasevitch. 1991. Soviet phraseology: Problems in the analysis and teaching of idioms. *Linguistics and Language Pedagogy: The State of the Art*, pages 484–488.

Elena Tutubalina. 2015. Clustering-based approach to multiword expression extraction and ranking. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 39–43.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

V.V. Vinogradov. 1977. Ob osnovnikh tipakh phraseologicheskikh edinits v russkom yazyke (on the main types of phraseological units in the russian language). In *Izbranniye Trudy: Lexicologia i Lexikographia (Selected Works on Lexicology and Lexicography*, pages 140–161.

Alexander Wahl and Stefan Th Gries. 2018. Multi-word expressions: A novel computational approach to their bottom-up statistical extraction. *Lexical collocation analysis: advances and applications*, pages 85–109.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2019. Ilfhocail: A lexicon of Irish MWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 162–168, Florence, Italy. Association for Computational Linguistics.

A Approaches to MWE Processing

Works, summing up recent approaches to MWE processing, are published every now and then (Pearce, 2002; Piao et al., 2005; Constant

et al., 2017; Ashok et al., 2019). And, probably, the main world event in this topic is the annual ACL-affiliated workshop on MWEs. In recent years, we observe a trend to evaluate methods for particular languages, e.g. LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008) addressed the issue in English, German, Czech, Estonian and other languages (Grégoire et al., 2008). Later appeared Arabic (Attia et al., 2010), Russian (Tutubalina, 2015), Polish (Chrzkaszcz, 2016), Spanish and Basque (Iñurrieta et al., 2017), Irish (Walsh et al., 2019), Bulgarian and Romanian (Barbu Mititelu et al., 2019), Serbian (Stanković et al., 2020) and many others. Currently, there is also a trend to discuss translation and alignment (Lam et al., 2015; Fisas et al., 2020; Han et al., 2020). The last LREC workshop on MWEs (Bhatia et al., 2022) addressed heuristic and machine-learning approaches to their detection, tool-kits for annotation, low-resource corpora, figurative language, etc. However, discussions about the nature of MWEs remain.

B PARSEME Classification of MWEs

- Nominal MWEs
  - Multiword named entities
  - NN compounds
  - Other nominal MWEs
- Verbal MWEs
  - Phrasal verbs
  - Light verb constructions
  - VP idioms
  - Other verbal MWEs
- Prepositional MWEs
- Adjectival MWEs
- MWEs of other categories
- Proverbs

C Vinogradov's Classification of MWEs

- combinations (*to conclude an agreement*)
- fusions
  - with archaic words – *to eke out*
  - with archaic grammatical forms – *hither and thither*
  - that were changed so that they do not resemble lexemes from which they were composed – *lo and behold* (*lo* from *look*)
  - complete loss of initial meaning – *caught red-handed* (initially meant catching someone who hunted an animal they were not allowed to hunt)

D Example Annotation of 15 Criteria and Four Linguistic Features

Table 1 enlists the 15 criteria and provides an example annotation of one MWE. Table 2 adds four features that are not part of the criteria.

E Clustering Visualizations

We use t-SNE algorithm (Van der Maaten and Hinton, 2008) implemented in the Scikit-learn library (Pedregosa et al., 2011) to visualize the resulting 15*285 table in 2D, Figure 6. The algorithm loses much information about distances between vectors in the multidimensional space. Hence, this visualisation can only be used for hypothesizing about our data.

We also apply Hierarchical Agglomerative Clustering, cf. Figure 7.

MWE	Criterion	Annotation
Белый гриб (penny bun)	**Lexical change**	
	i. Governed words cannot be replaced with a synonym	1
	ii. Headword cannot be replaced with a synonym	1
	iii. Cannot be translated word by word	1
	iv. Does not allow insertions of lexemes	1
	Grammatical change	
	v. Never changes grammatical form	0
	vi. Only headword changes grammatical form	0
	vii. Word order seldom or never changes	1
	Obsolescence	
	viii. Contains lexical archaisms	0
	ix. Contains unique lexemes	0
	x. Archaic syntax and/or morphology	0
	Replacement	
	xi. Allows ellipsis	1 белый
	xii. Allows portmanteau words	0
	xiii. Can be replaced with headword	1
	xiv. Can be replaced with one word	1 боровик
	xv. Can be translated with one word	1 cep, porcino

Table 1: Example annotation of 15 criteria.

MWE	POS-pattern	Is a sentence?	Headword	Phrase structure
Белый гриб (porcini mushroom)	Adj.+Noun	No	гриб	Agreement

Table 2: Example annotation of four linguistic features.

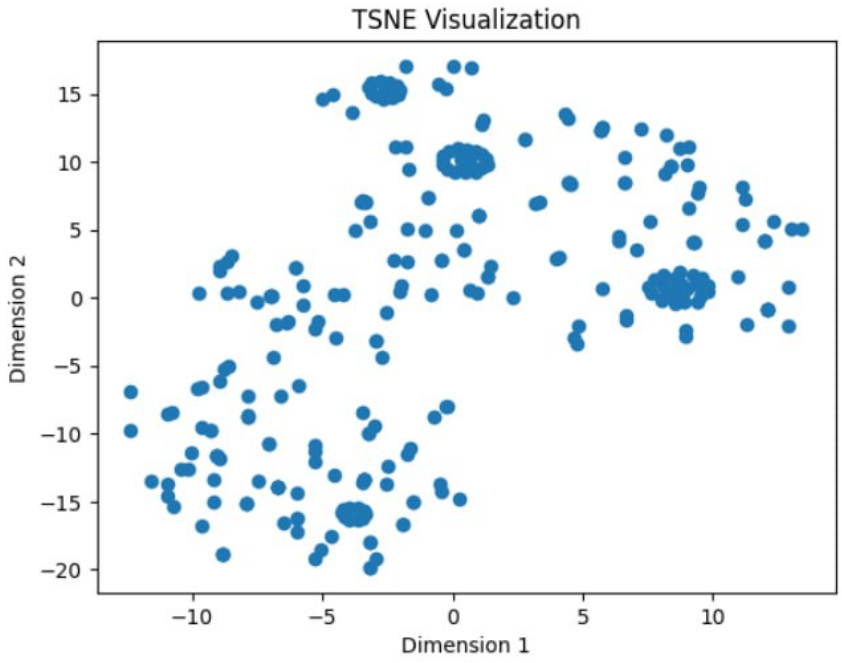

Figure 6: t-SNE 2D scatter plot of annotated MWEs.

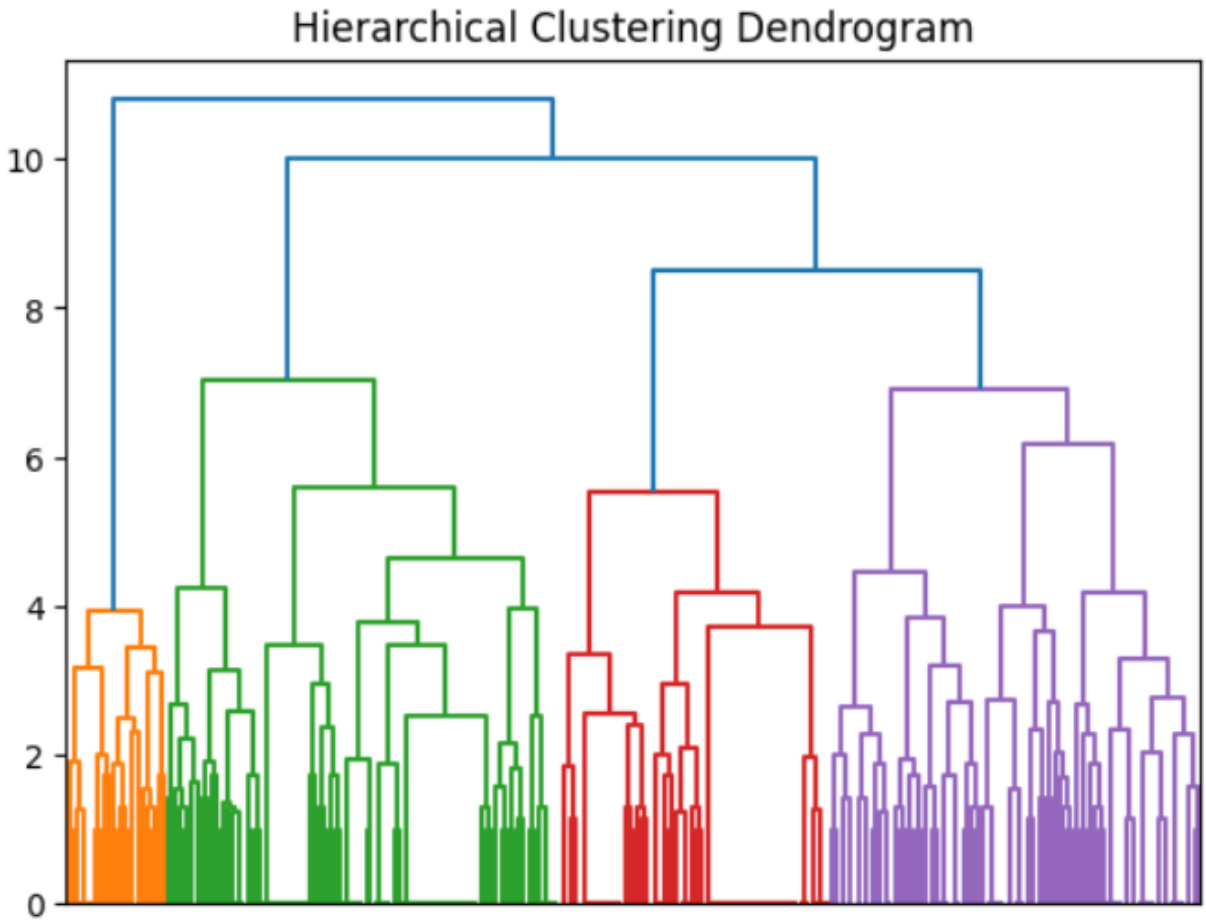

Figure 7: Hierarchical Agglomerative Clustering of MWEs.