Classifying Multi-Word Expressions in the Latvian Monolingual Electronic Dictionary Tēzaurs.lv

Laura Rituma, Gunta Nešpore-Bērzkalne, Agute Klints,
Ilze Lokmane, Madara Stāde, Pēteris Paikens
Institute of Mathematics and Computer Science
University of Latvia
Raiņa bulvāris 29, Riga, Latvia
{laura.rituma, gunta.nespore, agute.klints, peteris.paikens}@lumii.lv
ilze.lokmane@lu.lv, madara.stade@gmail.com

Abstract

The electronic dictionary Tēzaurs.lv contains more than 400,000 entries from which 73,000 entries are multi-word expressions (MWEs). Over the past two years, there has been an ongoing division of these MWEs into subgroups (proper names, multi-word terms, taxa, phraseological units, collocations). The article describes the classification of MWEs, focusing on phraseological units (approximately 7,250 entries), as well as on borderline cases of phraseological unit types (phrasemes and idioms) and different MWE groups in general. The division of phraseological units depends on semantic divisibility and figurativeness. In a phraseme, at least one of the constituents retains its literal sense, whereas the meaning of an idiom is not dependent on the literal sense of any of its constituents. As a result, 65919 entries of MWE have been manually classified, and now this information of MWE type is available for the users of the electronic dictionary Tēzaurs.lv.

Keywords: multi-word expression, phraseological unit, idiom, phraseme, semantics.

1 Introduction

Tēzaurs.lv[1] is the largest Latvian electronic explanatory dictionary with more than 400,000 entries. It emerged as a compilation from nearly 300 prior dictionaries and other sources (Grasmanis et al., 2023). Besides entries for single words Tēzaurs.lv also contains approximately 73,000 multi-word expressions (MWEs; dictionary entries that contain more than one orthographic word (Bauer, 2021: 5)) stored as separate entries. Most MWEs are linked to the corresponding word entries or a specific word sense that is included in the MWE. Therefore, dictionary users can either search a specific expression or find it in the matching word entry.

Up until now, all Latvian studies of MWEs have been carried out to accommodate machine translation. A bilingual MWE dictionary has been created, listing the relevant syntactic patterns both in English and the respective Latvian MWEs; this helps obtain syntactic rules for better machine translation (Deksne et al., 2008). Additionally, there have been studies regarding the methods of obtaining MWE lists to improve the quality of translation (Skadiņa, 2016; Mandravickaitė and Krilavičius, 2017) or to expand the dictionary data (Skadiņa, 2018). However, the aim of these studies was not creating a system of MWE classification based on their function and meaning. The MWE lists do not contain sense descriptions and most of the data is not freely accessible. In contrast, Tēzaurs.lv open-access data contains MWE sense descriptions but lacks information on morphological and syntactic structure.

Over the past two years, functionally diverse expressions have been linguistically analyzed and manually sorted into following categories:

- multi-word place names, e.g. *Juglas ezers* 'Jugla Lake', *Egļu ciems* 'Egļu Village';
- taxonomic group names, such as species, families or classes, both international, e.g. *Vultur pryphus*, *Tulipa lanata*, and Latvian, e.g. *aklais dundurs* lit. 'Blind Horse-Fly', *vilnainā tulpe* 'Woolly Tulip';
- complex terms and term candidates, e.g. *centrbēdzes spēks* 'centrifugal force'; *ciešamā kārta* 'passive voice'
- phraseological units, e.g. *mest plinti krūmos* lit. 'to throw the rifle into the bushes' (to give up);

[1]Available interactively at https://tezaurs.lv or as data from https://repository.clarin.lv/repository/xmlui/handle/20.500.12574/104

- collocations, e.g. *pieļaut kļūdu* 'to make a mistake', *apģērba gabals* 'piece of clothing'. We have adopted a rather narrow understanding of collocations, which are statistically significant co-occurences of words outside of all previously mentioned groups. In other words, collocations are fixed word combinations with no semantic reinterpretation (Veisbergs, 2019: 114).

Table 1 shows the total number of MWEs in the dictionary Tēzaurs.lv and the number of MWEs in each category and subcategory.

Certain groups were left outside of this classification, such as expressions in foreign languages (excluding taxa), e.g., *de facto*, *per aspera ad astra*, and abbreviations consisting of multiple tokens, e.g., *t. sk.* 'incl.', *a. god.* 'esteemed', as well as MWEs mentioned in Chapter 5.

This classification provides additional information to the dictionary users regarding MWE functions within the language and promote the study of phraseology in Latvian linguistics. A more formal structure of MWEs is also useful for computational linguistics tasks that involve structured, explicit semantic models, such as semantic parsing and information extraction, controlled natural languages, and structured natural language generation. This is especially relevant in multilingual solutions, as some concepts are expressed as individual word senses in one language and as MWEs in another, necessitating a structured inventory of the applicable MWEs. In addition to the MWE classes, their review improved the overall quality of data, e.g. combining close MWE variants into one dictionary entry. However, we noted that it was often difficult to decide whether two close MWEs are separate and further work is needed to develop objective criteria for this decision.

In this study, we focused on the semantic analysis of phraseological units by separating them into two subgroups depending on the relationship of the words forming the MWEs to the general meaning of the MWE itself (for a more detailed distinction between the two subgroups, phrasemes and idioms, see Chapter 3). The creation of this division is the first step, so that in the future, when the morphosyntactic and lexical variation of these phraseological units, including word order and derivation options (see, e.g. Leseva et al. (2020)), will be analyzed, it would be possible to test the hypothesis that phrasemes are more prone to morphosyntactic and lexical variation than idioms. Other studies also emphasize that decomposable phraseological units tend to be syntactically flexible to some degree (see, e.g., Sag et al. (2002: 5–7)).

Chapter 2 deals with the borderline cases of phraseological units and other MWE groups mentioned above, namely, collocations, taxa, and terms. Chapter 3 outlines the distinction between idioms and phrasemes. Chapter 4 describes the borderline cases involving idioms and phrasemes to show that semantic transparency is essentially scalar. Chapter 5 describes MWE groups that were not included in any of the defined categories. Finally, the last chapter of the article consists of conclusions and future fork for MWE processing.

2 Borderline Cases of the Phraseological Unit and Other MWE Classes

Before creating division of phraseological units, we had to establish terms for defining each MWE group. Difficulties arose when borders between two MWE classes were not that clear and fixed.

In this study, a MWE was classified as a collocation if all of the words that form it are used in their literal sense, i.e., the senses can be found in the dictionary entries of the corresponding words. For example, *izdzert līdz dibenam* lit. 'to drink to the bottom' is a collocation (and not a phraseological unit), since "dibens" 'bottom' has a literal meaning 'lower part (e.g., of a dish)'.

However, during data processing, difficulties arose in separating collocations and phraseological units as latter possess some degree of figurative, transferred or metaphorical meaning (Veisbergs, 2019: 114). Figurativeness fades over time and it is difficult to decide the point at which the use of a word meaning transitions from figurative to literal, therefore to decide whether a MWE has to be classified as a phraseological unit or a collocation. The words that form a MWE are occasionally used in a sense that could be perceived as figurative, but may already be listed in the dictionary as literal, because most language users no longer note the meaning transfer. In that case the MWE is still sorted as a collocation. For example, in the expression *labas acis* lit. 'good eyes', the dictionary entry *acis* 'eyes' lists the meaning of vision without the "figurative" tag. Similarly, the expression *celt trauksmi* 'to raise the alarm' contains the word *celt* 'raise', which has a figurative meaning 'radīt' 'to make' listed in the dictionary without the "figurative" tag. Thus, both

Name of Category	Name of Subcategory	Number of MWEs
complex terms		22,552
multi-word place names		14,733
taxonomic group names		
	International	10,347
	Latvian	7,854
phraseological units		
	phrasemes	2,863
	idioms	4,029
	unclassified phraseological units	358
collocations		3,183
Classified MWEs in total		**65,919**
Unclassified MWEs		5,385
Total number of MWEs in Tēzaurs.lv		**71,304**

Table 1: The number of MWEs sorted into each category and subcategory.

of the mentioned MWEs have been classified as collocations, even though they could also easily be seen as phraseological units, since they do display a certain degree of fading figurativeness.

Additionally, over time, certain figurative meanings have been preserved only in one expression. For example, the entry *apaļš* 'round' lists a meaning "not having any family", which nowadays is only used in the expression *apaļš bārenis* lit. 'a round orphan'. In such cases, it is advisable to delete this meaning of *apaļš* from the dictionary and sort the MWE as a phraseme.

These issues show that, at times, the line between figurative and direct meanings can be vague – the more frequent and varied the use of a figurative meaning is, the more likely it is that the meaning will lose its figurativeness. Therefore, with certain expressions it is more difficult to discern whether they still count as phraseological units or have already become collocations. In this study, it was decided not to delve into the borderline cases of figurativeness, but instead agree on clear criteria for separation based on dictionary data.

Further difficulties arose from the fact that both terms and taxa can be figurative, e.g., term *auss gliemene* lit. 'ear clam', taxon *atvērtā pērtiķmutīte* lit. 'open monkey-mouth'. Although in Latvian linguistics figurative names are traditionally not recognized as phraseological units, they are essentially idioms, which only differ in their naming function (for a more detailed description of idioms, see Chapter 3). One MWE cannot simultaneously belong to several categories (e.g., term and idiom), so it was decided to classify such cases as terms or taxa despite their figurativeness. In the future, these cases could be re-sorted into further sub-categories.

A distinct group is formed by expressions, that can be used in both literal and figurative sense. Stephen G. Pulman also examines such phraseological units as a special, separate group. He notes that the components of such unit have literal meanings, but that these are not what is involved in their interpretation as a phraseological unit. It is certainly the case that someone unfamiliar with the phraseological unit nevertheless can arrive at an appropriate meaning for it by processing it as a metaphor (Pulman, 1993: 260). For example, expressions *atmest ar roku* lit. 'throw one's hand at something' (to stop, abandon doing something) and *grozīt galvu* lit. 'turn one's head around' (express surprise, concern) can be used in their direct sense to describe a physical action, as well as figuratively. In such cases, the MWE has two meanings: one is direct (categorised as a collocation, given that the expression is also often used in its direct sense) and the other is figurative (categorised as an idiom).

3 Semantic Types of Phraseological Units and Representation in Tēzaurs.lv

Phraseological units are usually expected to comply with three fundamental criteria: they are fixed, consist of multiple words and possess some degree of figurative, transferred or metaphorical meaning (Veisbergs, 2019: 114). In Latvian linguistics, the hyperonymic term 'phraseological unit' encompasses both phrasemes and idioms (Laua, 1992; Skujiņa, 2007), thus the term 'idiom' is used in a narrower sense, as a sub-type of a phraseological

unit.

Semantically, phrasemes are partially compositional and transparent, as one of their components functions in its direct, literal sense, e.g., *domu grauds* 'a grain of thought', where *doma* 'thought' is used in the sense 'the result of thinking', whereas the other component of the phraseme, *grauds* 'grain' in itself does not represent the specific meaning realized in the phraseme. A similar example is *caurs miegs* lit. 'leaky sleep' (fitful, poor sleep), where *miegs* 'sleep' is used in its basic sense, whereas *caurs* 'leaky' acquires the meaning of fitful or poor only in this expression and is not used in the same way in any other distribution. It is generally important that one of the components of a phraseme is used in a literal sense (which can be either basic or secondary) while the other component draws its specific semantic value exclusively from the corresponding MWE.

The meanings of idioms, in turn, are non-transparent, e.g., *karāties mata galā* lit. 'to hang by a thread of hair' (to be in a precarious situation), *kārt zobus vadzī* lit. 'to hang one's teeth on a wedge' (to starve). This means that idioms cannot be worked out by the usual semantic rules (Pulman, 1993: 260).

This distinction is represented in the Tēzaurs.lv entries as well: the phrasemes are linked to the corresponding, literal senses of the used words, e.g. *slinkuma maiss* lit. 'a bag of laziness' (a lazy person) is linked to the basic sense of the word *slinkums* 'laziness'. The same phraseme is also linked to the entry *maiss* 'bag' as a whole (and not to any specific sense) as the word *maiss* does not list a meaning of 'person'.

Unlike phrasemes, idioms should be linked to entries as a whole (and not separate word senses), e.g. the idiom *cieta galva* lit. 'a hard head' has two meanings: 1) difficulty learning, remembering, and 2) a stubborn, rebelious character; this idiom is linked to both entries, *galva* 'head', and *ciets* 'hard'.

4 Borderline Cases of Phraseological Unit Classification

To some extent, the separation of phrasemes and idioms is linked to the notion of idiom decomposability mentioned in linguistic literature (Sag et al., 2002: 5) which demonstrates how the overall sense of a given idiom is related to its parts. Although we use a similar approach, it does not easily provide a simple and indisputable division into categories, since phraseological units are very diverse both formally and semantically. One could agree with the view that MWEs have varying degrees of semantic transparency and should be described with reference to a semantic scale ranging from totally transparent in meaning to completely opaque (Parra Escartín et al., 2013: 346). However, there is no consensus on how many intermediate sections and corresponding types would exist on such a scale.

In this study, problems arose when a phraseological unit is decomposable in principle -– each word meaning can be discerned – but some of them are used figuratively. For example, in the expression *aizlaist vējā* lit. 'to let loose in the wind', to squander (classified as an idiom), the locative *vējā* 'wind' has a listed figurative sense 'a way in which (something) disappears, ceases to exist', so it can be used in different distributions, whereas the meaning *aizlaist* 'let loose' is used in its literal sense: 'to let something go by acting passively'. This expression cannot be classified as a phraseme since other components of a phraseme acquire figurative meanings only in that specific combination. In this expression, both components retain their own meanings – literal for one and figurative for the other – therefore it is classified as an idiom. In such cases, there are two potential solutions: to define subtypes for idioms, or to introduce a third group of phraseological units that is neither a phraseme nor an idiom.

Even though the degrees of semantic transparency and semantic types of phraseological units are still under study, from the perspective of data processing, separating phraseological units from other MWE groups and dividing them into at least two subtypes provides significant benefits, since this data will be available for further research as a separate group.

5 MWEs Not Included in The Existing Classification

A small part (7.5%) of the existing MWEs within Tēzaurs.lv have not been categorized yet. This is either because they cannot be assigned to any of the existing MWE categories, or because some entries have been listed as MWEs by mistake. The classification of these MWEs will be addressed in future work.

Firstly, there are naming units that are difficult to fit into any of the current categories, such as mytho-

logical entities (*Meža māte* lit. 'Forest Mother'), names of dances and games (*vistiņu ķeršana* 'tag, catchers'), old names for months (*lapu mēnesis* lit. 'leaf month', May), names for fingers (*garais Ancis* lit. 'Long Ancis', middle finger), etc.

Secondly, there are names that contain nomenclature words, for example, *ātrvilciens Eurostar* 'high-speed train Eurostar', *operētājsistēma UNIX* 'operating system UNIX'. Based on Tēzaurs.lv principles, such entries should not count as MWEs and the lemma should only consist of the proper name.

Furthermore, in many cases, the names of food dishes have not been classified at the moment. This thematically and semantically varied group has been set aside for future research and testing of more fine-grained classification, since they often belong to one or more overlapping categories. For instance, certain dish names can be idioms and food technology terms (*viltotais zaķis*, lit. 'mock rabbit', meatloaf), idioms but not terms (*ērzeļa pauti* lit. 'stallion's testicles', deep-fried balls of batter), terms and phrasemes (*smilšu mīkla*, lit. 'sand dough', shortcrust pastry), phrasemes but not terms (*aklā putra* lit. 'blind porridge', porridge with no fat), as well as collocations that can either be terms (*rauga mīkla*, yeast dough) or not (*balta putra* lit. 'white porridge', milk porridge).

6 Conclusions and Future Work

Firstly, extensive work has been carried out to sort various MWEs into distinct categories, during which it was concluded that the existing system of classification does not cover all types of MWEs in Tēzaurs.lv; there are certain groups (e.g., abbreviations, certain naming units and dish names) that remain unsorted. This, in turn, shows the need for additional MWE categories. The results of this work are integrated in the relevant entries of the dictionary and are accessible to all its users.

Secondly, certain borderline cases between different MWE categories were observed. A part of these cases stems from the fact that figurativeness is also used in term creation, and currently they are sorted in the category of terms. Other borderline cases arise when frequently used figurative senses gradually become literal and thus cause difficulties to distinguish phraseological units from collocations that do not contain figurative meanings.

Future work includes combining MWE variants in one entry and the continued analysis of morphosyntactic and lexical variations of phraseological units, e.g., the expression *Kā putns gaisā* lit. 'like a bird in air' can vary as *kā putns kokā* lit. 'like a bird in a tree', and *kā putns zara galā* lit. 'like a bird at the end of a branch'. All variants have the same syntactic structure and meaning (to be without obligations, worries or cares). Determining variants is also related to distinguishing between the fixed components of a phraseological unit and its characteristic environment, which is not a part of the unit itself. For example, the phraseological unit *gaiss tāds, ka cirvi var pakārt* lit. '(one) could hang an axe in this air' is a phraseme, but in certain environments it can appear simply as *cirvi var pakārt* lit. '(one) could hang an axe here', (a feeling of stuffiness indoors). Thus, a phraseme can be reduced and subsequently become an idiom. After collecting such variants, we will test the hypothesis of whether phrasemes are lexically and syntactically more flexible than idioms.

Acknowledgments

This work was supported by the Latvian Council of Science project "Advancing Latvian computational lexical resources for natural language understanding and generation" (LZP2022/1-0443) in synergy with the State Research Programme project LATE (VPP-LETONIKA-2021/1-0006). We also thank the anonymous reviewers for their input in improving this paper.

References

Laurie Bauer. 2021. *An Introduction to English Lexicology*. Edinburgh University Press, Edinburgh.

Daiga Deksne, Raivis Skadiņš, and Inguna Skadiņa. 2008. Dictionary of multiword expressions for translation into highly inflected languages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Mikus Grasmanis, Pēteris Paikens, Lauma Pretkalniņa, Laura Rituma, Laine Strankale, Artūrs Znotiņš, and Normunds Grūzītis. 2023. Tēzaurs.lv – the experience of building a multifunctional lexical resource. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, pages 400–418. Lexical Computing CZ s.r.o.

Alīse Laua. 1992. *Latviešu valodas frazeoloģija*. Zvaigzne, Rīga.

Svetlozara Leseva, Verginica Barbu Mititelu, and Ivelina Stoyanova. 2020. It takes two to tango – towards a

multilingual MWE resource. In *Proceedings of the 4th International Conference on Computational Linguistics in Bulgaria (CLIB 2020)*, pages 101–111, Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.

Justina Mandravickaitė and Tomas Krilavičius. 2017. Identification of multiword expressions for Latvian and Lithuanian: Hybrid approach. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 97–101, Valencia, Spain. Association for Computational Linguistics.

Carla Parra Escartín, Gyri Smørdal Losnegaard, Gunn Inger Lyse Samdal, and Pedro Patiño García. 2013. Representing multiword expressions in lexical and terminological resources: An analysis for natural language processing purposes. In *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.*, pages 338–357, Ljubljana/Tallinn. Institute for Applied Slovene Studies/Eesti Keele Instituut.

Stephen G. Pulman. 1993. The recognition and interpretation of idioms. In C. Cacciari and P. Tabossi, editors, *Idioms: Processing, Structure, and Interpretation*, Laurence Earlbaum Cognitive Science Monographs, pages 249–270. Lawrence Erlbaum Asociates, New Jersey.

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*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Inguna Skadiņa. 2016. Multi-word expressions in english-latvian machine translation. *Baltic Journal of Modern Computing*, 4(4).

Inguna Skadiņa. 2018. Looking for a needle in a haystack: Semi-automatic creation of a latvian multiword dictionary from small monolingual corpora. In *Proceedings of the 18th EURALEX International Congress*, pages 255–265.

Valentīna Skujiņa, editor. 2007. *Valodniecības pamatterminu skaidrojošā vārdnīca*. Valsts valodas aģentūra, Rīga.

Andrejs Veisbergs. 2019. The fuzzy concept of idiom and what it might mean for bilingual dictionaries. *Baltic Journal of English Language, Literature and Culture*, 9:111–129.