Multi-word Expressions in the Electronic Dictionary *Tēzaurs*: Integrating Data from Different Sources

Gunta Nešpore-Bērzkalne, Mikus Grasmanis, Ilze Lokmane
Institute of Mathematics and Computer Science, University of Latvia
Raiņa bulvāris 29, Riga, Latvia
gunta.nespore@lumii.lv, mikus.grasmanis@lumii.lv, ilze.lokmane@lu.lv

Abstract

In addition to single word entries, the electronic Latvian dictionary *Tēzaurs* contains more than 76,000 multi-word expressions (MWEs), which are represented in separate entries with the same structure as other entries. A group of MWEs (ca. 5000 entries) was effectively left out of some processes as they could not be automatically interpreted: MWE variants shown in ambiguous parentheses, functional words indicating metadata not separated from lexeme text, etc. These entries mostly originated from printed dictionaries, which has led to repetitive and inconsistent data and required significant manual revision. Efforts are underway to integrate these MWEs into the *Tēzaurs* data network; the problematic entries have been restructured and the principles for data consolidation have been defined.

Keywords: multi-word expressions, lexical variation, condensed headwords, dictionary data integration.

1 Introduction

Tēzaurs[1] is the largest Latvian electronic explanatory dictionary with more than 410,000 entries, which was started as a compilation of nearly 300 pre-existing dictionaries and other sources but has since developed into a multifunctional lexical resource consisting of granular interlinked entities enriched with various kinds of metadata (Grasmanis et al., 2023). Besides entries for single words, *Tēzaurs* also contains ca. 76,000 entries for multi-word expressions (MWEs; dictionary entries that contain more than one orthographic word (Bauer, 2021: 5)). Thus, it is currently the most comprehensive resource of MWEs in Latvian lexicography.

Tēzaurs has evolved over many years as a compilation of numerous valuable but sometimes heterogeneous sources (both printed dictionaries and collections of terms, place names, taxonomies, idioms, etc.) with incomplete and inconsistent guidelines for consolidation and standardization or without any at all. Initially, the inclusion of each new source was purely additive, extending *Tēzaurs* with an additional set of entries mostly only loosely connected with entries coming from other sources. During this time, the *Tēzaurs* platform has also been improved providing new means for data normalization, standardization, and enrichment with additional or meta information, thus enabling a consolidation of information, a portion of which is presented in this paper.

In *Tēzaurs*, most MWEs have been assigned category tags according to their function and level of figurativeness (Rituma et al., 2024): see Table 1. This paper focuses on idiomatic MWEs, the principal sources of which are The Dictionary of Standard Latvian (Ceplītis, 1971–1996) and The Dictionary of Latvian Phraseology (Laua et al., 1996), the latter containing idiomatic MWEs only.

Although the format of MWEs derived from different sources was initially highly heterogeneous, most of the data was unified and integrated into *Tēzaurs* by automatic or semi-automatic methods. However, a group of MWEs (ca. 5000 entries; most of them – phraseological units) required significant manual revision. These entries contained data from printed dictionaries that could not be automatically interpreted: some headwords contained several MWE variants, and a part of them was metadata, not a lexeme (headwords representing multiple MWEs are hereinafter referred to as condensed headwords). The information was repetitive and inconsistent, and therefore difficult to process, as the same MWE in different entries may be represented differently. As a result, a headword could

[1] Available at https://tezaurs.lv or as data from https://repository.clarin.lv/repository/xmlui/handle/20.500.12574/160

appear the following way: ***dūc*** *(arī* ***rūc****)* ***kā*** *(****bišu****)* ***stropā*** *(arī* ***strops****,* ***pa stropu****), retāk* ***dūc*** *(arī* ***rūc****)* ***kā*** *(****bišu****)* ***spietā*** *(arī* ***spiets****)* (lit. '**buzz** (also **hum**) **as in a** (**bee**) **hive** (also **a hive**, **around a hive**), less often **buzz** (also **hum**) **as in a** (**bee**) **swarm** (also **as a swarm**') 'the noise made by many people'.

Currently, efforts are underway to fully integrate these MWEs into *Tēzaurs*, i. e., to rewrite condensed headwords in a consistent and machine-interpretable form by separating each MWE variant into its own database record, as well as deduplicate and consolidate the information about them. This process requires substantial manual effort but constitutes a necessary step to ensure that all MWEs are fully accessible for searching, processing, and research. Moreover, the manual review of the data helps to pinpoint and resolve additional issues related to MWE description and representation.

The next chapters outline the structure of MWE entries in *Tēzaurs* (2), the expansion of the initial headwords (3), and how these entries are consolidated and integrated into *Tēzaurs* (4). The last chapter gives conclusions and an insight into the future work for MWE representation.

2 MWE Entry Structure in *Tēzaurs*

The MWEs in *Tēzaurs* have separate entries that can contain the same elements as single-word entries: MWE lexical variants; tags of style and usage context; sense definitions; a section for adding usage examples from the Latvian National Corpora Collection (Saulīte et al., 2022). Additionally, MWE entries contain MWE category tags (see Table 1). As structurally independent units, MWE entries can be linked with other entries in *Tēzaurs*, including entries of their component-words (this is how MWEs are displayed in single-word entries: the entry of the respective word shows all MWEs that the word is linked with), as well as entries involved in the Latvian WordNet showing semantic relations between MWEs and other lexical units (Klints et al., 2024). They can also be linked with the Princeton English WordNet (Fellbaum, 1998). However, MWEs in *Tēzaurs* lack explicit information on their morphological and syntactic structure. An experiment for creating semiautomatic syntactic and structural annotation has been commenced, but the results are not publicly available yet.

Name of Category	No of MWEs
complex terms	23,777
taxonomic group names	18,200
multi-word proper nouns	16,993
phraseological units	8,953
collocations	3,065
foreign language MWEs	1,151
unclassified MWEs	4,348
Total number of MWEs	**76,467**

Table 1: The MWEs categories in *Tēzaurs* and the number of MWEs in them.

3 The Expansion of Condensed MWE Headwords

A list of condensed headwords was extracted, namely those containing (1) parentheses and (2) specific functional words. This list comprised ca. 5,000 entries. The following paragraphs outline the use of parentheses and functional words in the condensed headwords.

Ambiguous parentheses: Parentheses may have several functions that are difficult to interpret automatically: (1) the text in parentheses is optional, e.g., *dūc kā* ***(bišu)*** *stropā* (lit. 'buzzes as in a (**bee**) hive'); (2) the text in parentheses (without the functional word) or one of the words listed in parentheses may replace the text preceding the parentheses, e.g., *dūc (arī* ***rūc****) kā stropā* (lit. 'buzzes (also **hum**) as in a hive'); (3) the pronoun in parentheses is not a part of the lexeme but indicates the free arguments or modifiers of the MWE (as described in (Odijk and Kroon, 2024)), e.g., *izliet* ***(kāda)*** *asinis* (lit. 'to shed (**someone's**) blood') 'to kill'; *atvērt durvis* ***(kādam, kaut kam)*** (lit. 'to open the door (for **someone, something**') 'to create favorable conditions'; any word from a certain lexical group can take the place of the pronoun.

Functional words: Include *arī* 'also', *biežāk* 'more commonly', *retāk* 'less commonly'. E.g., *ņemt acis pirkstos* ***(retāk*** *rokā),* ***arī*** *turēt acis pirkstos* (lit. 'to take eyes in one's fingers (**less often** in hand), **also** to hold eyes in one's fingers') 'to look closely'.

The text of a lexeme is not always clearly discernible from functional words and it is not always possible to determine which portion of the MWE the functional word or the text in parentheses pertains to.

To transform condensed headwords into a set of single-lemma lexemes (variants of MWE), a format for describing lexeme variants needed to be selected. The lexeme variants already present in *Tēzaurs* (in both single-word and MWE entries) are represented as lists: each variant is stored in separate database records. However, previously the number of variants per lexeme has typically been small (up to three), whereas the condensed headwords can give a substantially larger number of variants. Presenting these as simple lists may not be user-friendly. An alternative approach was also considered, where MWE variants were represented by other, unambiguously interpretable condensed headwords (formulae). However, implementing such a solution would require substantial changes to the entire *Tēzaurs* system to apply the new format to the existing entries. Therefore, it was decided to start with explicitly expanding the condensed headwords into a list of lexeme variants for each entry. This approach also enables a more thorough analysis of existing data and better decisions regarding future MWE representation.

The condensed headwords were expanded as follows:

- Each MWE variant was moved to a separate record, with most parentheses removed; the only remaining (now unambiguous) parentheses indicate the free arguments or modifiers of the MWE;
- Functional words were converted into tags indicating frequency of use (except *arī* 'also', which gives no additional information).

Thus, the previously mentioned example ***ņemt acis pirkstos*** *(retāk* ***rokā****), arī* ***turēt acis pirkstos*** provides three variants: (1) *ņemt acis pirkstos*, (2) *ņemt acis rokā*, and (3) *turēt acis pirkstos*; furthermore, a tag is added to the second variant to indicate its less frequent use.

The structure of the other previously mentioned example ***dūc*** *(arī* ***rūc****)* ***kā*** *(****bišu****)* ***stropā*** *(arī* ***strops****,* ***pa stropu****), retāk* ***dūc*** *(arī* ***rūc****)* ***kā*** *(****bišu****)* ***spietā*** *(arī* ***spiets****)* is more complex; it is depicted in Figure 1. The formula starts with a black arrow on the left, and ends with another arrow on the right. Each possible path in this formula describes one variant of the MWE. The convergence of all possible paths provides the set of possible MWE variants. The *arī* 'also' clauses are depicted as branches, the *retāk* 'less often' clause is coloured in gray. The optional clause is depicted as a path skipping the optional word.

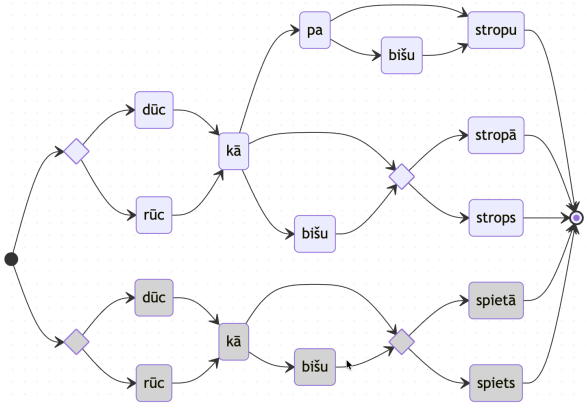

Figure 1: Structure of the MWE *dūc kā bišu stropā* (lit. 'buzzes as in a bee hive) 'the noise made by many people' with variants

4 Integration of the MWE Entries into the *Tēzaurs* Data

To integrate an MWE entry into the *Tēzaurs* data, an essential step is to ensure that the entry is linked to all its component entries. This enables dictionary users to navigate from the MWE entry to the corresponding single-word entries and back, as well as to find the MWE in the entries of its components.

After expanding the condensed headwords into several simple lexemes, it becomes apparent that many lexemes repeatedly occur in multiple entries. Our goal is to describe each lexeme only once; thus, we developed a consolidation methodology and process, as depicted in Figure 2. In brief, the process can be explained as *expand-deduplicate-combine-compose*. If, in addition to the lemma, all other parts of the lexemes also match, the consolidation occurs immediately. However, for many lexemes with the same lemma, deviations are noticeable in both the set of given senses and the additional information. At first glance, the deviations are partly significant and partly sporadic. In the following subsections, we describe which types of information deviate more frequently and how such cases were handled.

4.1 Different Sets of MWE Variants

Various entries that show the same lexeme may contain different sets of lexeme variants. As seen in Figure 2, the lexeme *sist nost* (lit. 'to beat down') is mentioned in all initial entries (data from different entries shown in different colours), unlike other

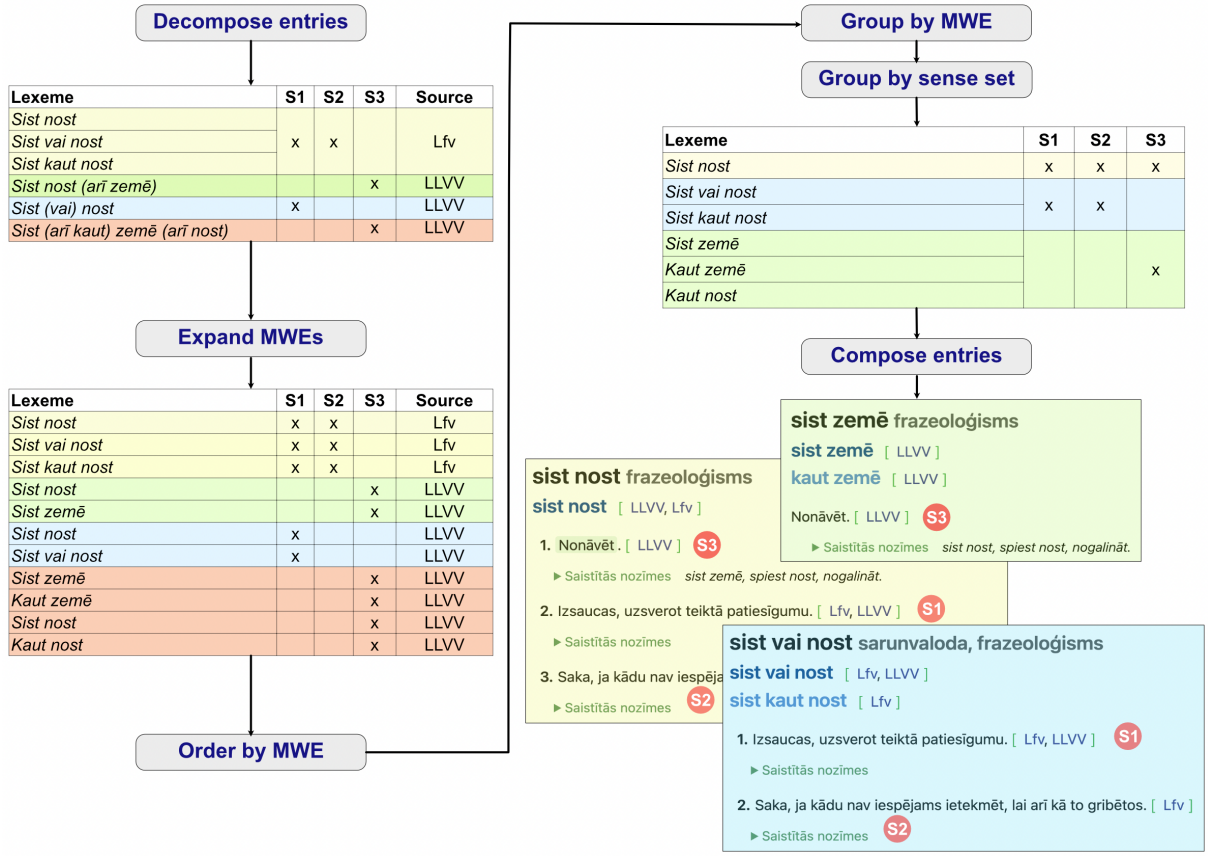

Figure 2: Processing of MWEs with overlapping data

variants. If these entries have an identical set of senses, in the final entry we include all lexeme variants with sources indicated for each of them. If the sets of senses differ between entries, the course of action described in the following section is taken.

4.2 Different Set of Senses

Entries from different sources may have varying levels of sense granularity and different sets of senses. This includes whether or not the literal meaning of an MWE is indicated (in the example in Figure 2, the literal sense is 'to kill' (S3), which is not given in the entry from (Laua et al., 1996)).

We group MWE variants into entries according to the set of senses (usually retaining the finest granularity) with the source indicated for each. Synonymous senses of different entries are linked with the synonymy link, which links the MWE with other synset elements of the corresponding sense, if there are any. As seen in Figure 2, after sorting MWE lexemes by sets of senses, the four initial entries were recompiled into three, where each MWE lexeme is only described once, providing all information corresponding to this variation.

4.3 How Many Entries?

Considering the previously described problems and solutions, a crucial question arises: whether MWEs with similar wording and identical meanings should be represented in a single entry or across multiple entries? What is the optimal solution between two extreme approaches: (1) considering each MWE with an even slightly different form to be independent and thus represented in a separate entry, or 2) grouping all MWEs that share the same senses in a single entry, where other metadata, e.g. register, could be specified separately for each variant? For example, one may consider the MWEs *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'): they have the same syntactic structure and meaning 'to be somewhere for a very short time', and even express similar imagery; thus it can be questioned whether they should be in one entry or two. Furthermore, how many entries should there be for the MWEs *apsildīt kājas* (lit. 'to warm up the feet'), *apsildīt degunu* (lit. 'to warm up the nose'), and *ap-*

sildīt nagus (lit. 'to warm up the nails'), which all share the same structure and meaning, i.e., 'to be somewhere for a very short time', but use different imagery?

Ermakova et al. (2022) proposes representing MWE variants in one or two entries, or a specific entry group depending on the number and variety of the MWEs (Hub-node model of MWE entries), whereas (Gantar and Krek, 2022) distinguish MWE variants from independent MWEs primarily on the basis of their inventory of senses and number and sequence of components. In the PARSEME (Savary et al., 2026) annotation guidelines[2], an MWE is regarded as a sequence of lexicalised components; if one of the components is replaced, the resulting expression is treated as a different MWE (it should be noted, however, that the objective of PARSEME is the annotation of MWEs in corpora rather than their lexicographic description).

In connection with various lexical resources, the representation and encoding of MWE variants, as well as the best version of the lemma/canonical form have been widely discussed e.g., (Villavicencio et al., 2004; Odijk and Kroon, 2024; Skoumalová et al., 2024; Al-Haj et al., 2014; Vrbinc and Vrbinc, 2016). However, these studies rarely provide explicit criteria for distinguishing MWE variants from independent MWEs. Implicitly, it could be assumed that expressions falling outside the scope of described variation should be treated as independent MWEs.

It is clear that in the future, *Tēzaurs* will need more refined guidelines for distinguishing MWE variants from independent MWEs. Developing and applying such guidelines, however, constitutes a separate step of MWE processing, as they should be applied to the entire *Tēzaurs* data rather than only to the dataset discussed in this paper. At the present stage, our objective is to ensure that the newly integrated MWEs are incorporated consistently into the existing structure of *Tēzaurs*.

Therefore, whenever possible, we currently retain the division of MWEs into entries as inherited from the sources; however, decisions can not be postponed in cases where MWE division into entries differs across sources.

In these cases, we try to follow the Latvian lexicographic tradition, which is reflected by the principal MWE sources (Laua et al., 1996; Ceplītis, 1971–1996) as well. Regarding idiomatic MWEs, the common practice (Laua, 1992) in Latvian lexicography is to group MWEs with an identical meaning (set of senses), similar form and imagery into a single entry and view them as MWE variants. MWEs with identical or close meaning, similar form, but different imagery, on the other hand, are split into separate entries and viewed as synonyms.

In practice, the distinction between variants and synonyms is not clear-cut, as the closeness of imagery is difficult to assess. In borderline cases we maintain a more fine-grained entry division and create synonym links between the identical MWE senses. This enables the dictionary users to still access MWEs with identical meanings all at once – either in a single entry or through synonymy links.

5 Conclusions and Future Work

Extensive work has been carried out to integrate ca. 5,000 idiomatic MWEs into the dictionary and make them more accessible to users. Principles have been developed for consolidating data from diverse sources. Manual review of entries has improved the overall quality of the data.

Future work includes identifying more user-friendly methods of representing MWEs with numerous variants and refining the criteria for division of similar MWEs into entries. We also plan to introduce syntactic annotation for MWEs, with first experiments already underway.

Acknowledgments

This work was supported by State Research Programme Letonika – Fostering a Latvian and European Society project "Digital Resources and AI Technologies for the Sustainability of the Latvian Language (DigiLATE)" (VPP-IZM-LETONIKA-2025/1-0004) in synergy with the Latvian Council of Science project "Advancing Latvian computational lexical resources for natural language understanding and generation" (LZP2022/1-0443). We also thank our collegues Madara Stāde and Baiba Valkovska for valuable advice.

References

Hassan Al-Haj, Alon Itai, and Shuly Winter. 2014. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2):130–170.

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

[2] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/?page=home

Laimdots Ceplītis, editor. 1971–1996. *Latviešu literārās valodas vārdnīca*, volume 1–8. Zinātne, Rīga.

Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Bernhard Roll. 2022. Integration of Multi-Word Expressions into the Digital Dictionary of German Language (DWDS). Towards a Lexicographic Representation of Phraseological Variation. In *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, pages 851–860, Mannheim. IDS-Verlag.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. Mit Press.

Polona Gantar and Simon Krek. 2022. Creating the Lexicon of Multi-Word Expressions for Slovene. Methodology and Structure. In *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, pages 549–562, Mannheim. IDS-Verlag.

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.

Agute Klints, Mikus Grasmanis, Gunta Nešpore-Bērzkalne, Lauma Pretkalniņa, Madara Stāde, Normunds Grūzītis, Ilze Lokmane, Pēteris Paikens, Laura Rituma, and Andrejs Spektors. 2024. Tēzaurs as a digital multifunctional lexical resource. *Baltic Journal of Modern Computing*, 12(4):513–525.

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

Alīse Laua, Aija Ezeriņa, and Silvija Veinberga. 1996. *Latviešu frazeoloģijas vārdnīca*, volume 1–2. Avots, Rīga.

Jan Odijk and Martin Kroon. 2024. A canonical form for flexible multiword expressions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 91–101, Torino, Italia. ELRA and ICCL.

Laura Rituma, Gunta Nešpore-Bērzkalne, Agute Klints, Ilze Lokmane, Madara Stāde, and Pēteris Paikens. 2024. Classifying multi-word expressions in the latvian monolingual electronic dictionary tēzaurs.lv. In *6th International Conference Computational Linguistics in Bulgaria (CLIB)*, pages 113–118.

Baiba Saulīte, Roberts Darģis, Normunds Grūzītis, Ilze Auziņa, Kristīne Levāne-Petrova, Lauma Pretkalniņa, Laura Rituma, Pēteris Paikens, Artūrs Znotiņš, Lauma Strankale, Kristīne Pokratniece, Ilmārs Poikāns, Guntis Bārzdiņš, Inguna Skadiņa, Anda Baklāne, Valdis Saulespurēns, and Jānis Ziediņš. 2022. Latvian national corpora collection – korpuss.lv. In *13th Language Resources and Evaluation Conference (LREC)*, pages 5123–5129.

Agata Savary, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Păiș, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto A. Díaz Hernández, Victoria Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, Chaya Liebeskind, Irina Lobzhanidze, Aleksandra Marković, Gunta Nešpore-Bērzkalne, Adriana Pagano, Mehrnoush Shamsfard, Ranka Stanković, Vahide Tajalli, Carole Tiberius, and Aakanksha Padhye. 2026. Parseme 2.0 multilingual corpus of multiword expressions. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4819–4834.

Hana Skoumalová, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondřička, and Milena Hnátková. 2024. Lemur: A lexicon of czech multiword expressions. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword Expressions in Lexical Resources: Linguistic, Lexicographic, and Computational Perspectives*, pages 1–37. Language Science Press, Berlin.

Aline Villavicencio, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical encoding of MWEs. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 80–87, Barcelona, Spain. Association for Computational Linguistics.

Alenka Vrbinc and Marjeta Vrbinc. 2016. Canonical forms of idioms in online dictionaries. *GEMA Online Journal of Language Studies*, 16:1–15.