The Lock, Stock, and Barrel of Marathi Multiwords

Aakanksha Padhye, Ashwini Vaidya
Indian Institute of Technology Delhi
{aakanksha.p, avaidya} @hss.iitd.ac.in

Abstract

Multiword expressions are an important area of study in linguistics and natural language processing as they represent combination of words that function as a single unit, and display properties that cannot be predicated fully from their individual components. This paper describes annotated corpora of about 3000 multiword expressions across syntactic categories in Marathi. This is the first exhaustive resource for Marathi which includes both verbal and non-verbal multiwords. In order to develop the guidelines for annotation, we have used the existing literature on the identification and classification of these expressions. Following the PARSEME 2.0 guidelines, we discuss the categories of multiwords and their behaviour in the corpus. Throughout the annotation process, we encounter variability in compositionality and syntactic realization and discuss our design decisions during annotation. Such a dataset will further our understanding of how grammatical structure can be integrated with lexically stored multiword units in Marathi.

1 Introduction

Multiword expressions (MWEs) are a pervasive and heterogeneous class of linguistic units that are central to research in linguistics and natural language processing. Linguistically, these constructions challenge the notions of compositionality, argument structure, and the division of labor between syntax and lexicon. Efficient handling of MWEs would prove beneficial for natural language processing tasks like machine translation (Constant et al., 2017), semantic processing (Korkontzelos, 2010), information extraction, word sense disambiguation (Singh et al., 2016); and psycholinguistic studies like MWE representation and processing (Wittenberg and Piñango, 2011; Nenonen et al., 2002), etc.

Previous attempts to annotate Marathi MWEs restrict themselves to compound nouns and light verb constructions alone (Singh et al., 2016). In this paper, we describe our effort at creation of a more comprehensive database of MWEs in Marathi under the PARSEME project (Savary et al., submitted).[1] The annotation tags, annotation platform, and annotation schema strictly adhere to the guidelines of the PARSEME project (Savary et al., submitted). As a result, we do not revisit these details here. Instead, we primarily focus on reporting the methodological decisions adopted during the process of annotation, and the linguistic and empirical challenges encountered during the process.

2 Corpora and Annotation

The corpora needs to be representative and balanced (Pustejovsky and Stubbs, 2012) in order to capture the entire range of MWEs in Marathi. This is achieved by carefully determining the genre of the data. Ozarkar (2014) observes that Marathi light verb constructions may have originated in informal contexts. Keeping this in mind, we have chosen Marathi UD Treebank (Ravishankar, 2017), and Anuvaad (Tiedemann, 2012) corpora, primarily comprising stories from Wikisource and the lifestyle genre, respectively. Additionally, we have web-crawled children's stories that are randomly sampled from different sources. Table 1 reports the number of tokens in each type of corpus in this dataset.

Marathi UD Treebank (Ravishankar, 2017) is already annotated with gold standard POS tags, syntactic structures, and semantic relations in ConLL-U format. The remaining two corpora are raw. For the Anuvaad corpus and the children's stories, we used UDPipe for parsing and tagging the

[1] The data will be released under the PARSEME 2.0 (Savary et al., submitted) initiative.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 96–102
March 28, 2026 ©2026 Association for Computational Linguistics

Corpora	Tokens
Marathi UD Treebank	3849
Anuvaad Corpus	27956
Children’s Stories	4287

Table 1: Tokens in corpora chosen

raw Marathi text (Straka and Straková, 2017). We found that UDPipe (Straka and Straková, 2017) for Marathi is not very accurate, and we find errors for POS tagging and sentence segmentation. The Anuvaad (Tiedemann, 2012) corpus, and children’s stories contain only ‘silver standard’ tags and sentence segments. For this present work, we have prioritized the annotation of MWEs by not letting the errors influence the annotations.

The annotation task is carried out by a single annotator. Hence, we are unable to calculate inter-annotator agreement.

The upcoming sections discuss all possible MWEs in Marathi. Based on (Savary et al., submitted) guidelines, we broadly classify them into two categories: verbal MWEs and non-verbal MWEs. The latter is a broader class comprising nominal, adjectival, and adverbial MWEs.

3 Verbal MWEs

Structurally, verbal MWEs in Marathi are broadly classified into verb-verb and preverb-verb constructions in the literature. PARSEME 2.0 (Savary et al., submitted) refers to these constructions as multi-verb constructions, and light verb constructions respectively. This section presents the categories incorporated by these constructions, their identification along with the semantics they render.

3.1 Multi-Verb Constructions

PARSEME 2.0 (Savary et al., submitted) identifies multi-verb constructions (MVCs) as a sequence of two verbs functioning as a single predicate, having the same subject. referring to a single event, and denoting a single tense, aspect and polarity value. These characteristics are identified using Ozarkar (2014)’s classification of multi-verbs. The constructions below are annotated as MVCs in the corpus following her classification:

1. **Complex predicates (CPs)**: monoclausal and monoeventual sequences like basun rahɳe ‘sit stay’

2. **Factor verbs**: expressions stored in the mental lexicon as a single unit or a set formula. Example: nigʰun ʥaɳe | lit. ‘emerge go’ (‘depart’)

3. **Manner-adverbial CPs**: sequences like d̪ʰawət̪ jeɳe ‘run come’. The author notes that Marathi, unlike Hindi, does not give a serial reading for such constructions.

The MVCs usually occur in two forms in Marathi. Firstly, we have verb-verb sequences conjoined by the conjunctive particle (-un). Secondly, there are verb-verb sequences conjoined by an imperfective marker (-ət̪).

Ozarkar (2014) identifies a list of verbs conjoined by the conjunctive particle (-un). It includes ʥaɳe ‘go’, jeɳe ‘come’, d̪eɳe ‘give’, gʰeɳe ‘take’, ʈakɳe ‘throw’, ʈʰewɳe ‘keep’, bəsɳe ‘sit’, kaɖʰɳe ‘draw out’, and rahɳe ‘stay’ as light verbs. Pardeshi et al. (2006) add gʰalɳe ‘put’, pəɖɳe ‘fall’, and aɳɳe ‘bring’ to the list.

On the other hand, for the imperfective marker (-ət̪), Ozarkar (2014) observes that light verbs like bəsɳe ‘sit’, suʈɳe ‘be released’, ʦalɳe ‘walk’, ʥaɳe ‘go’, jeɳe ‘come’, and rahɳe ‘stay’ can be found. Kume (2011) mentions that certain perception verbs like pahɳe ‘see’ also function as light verbs.

We use the verb list for MVCs mentioned in the literature, and empirically investigate the occurrences of these verbs in the corpora. We have identified thirteen verbs functioning as light in the verb-verb sequence. Figure 1 shows the verbs with the highest frequencies in the corpora. These verbs are followed by d̪eɳe ‘give’, ʈakɳe ‘throw’, kaɖʰɳe ‘draw out’, bəsɳe ‘sit’, ʦalɳe ‘walk’ and pahɳe ‘see’. Light verbs with the lowest frequencies are suʈɳe ‘be released’ and aɳɳe ‘bring’.

Verbal reduplication is also attested in the corpus in a few rare examples. Instances like ʦalət̪ ʦalət̪ ‘walk walk’ are considered as MVCs as the entire verb is reduplicated to form a verb-verb sequence.

It should be noted that not all verbs mentioned above function as light verbs in all contexts when they appear as a second verb in the verb-verb sequence. Verbs like ʥaɳe ‘go’ and jeɳe ‘come’ can also function as passive markers. The passive constructions are not MVCs. Similarly, modals and auxiliaries do not constitute MVCs. Accordingly, passives, auxiliaries, modals, permissives,

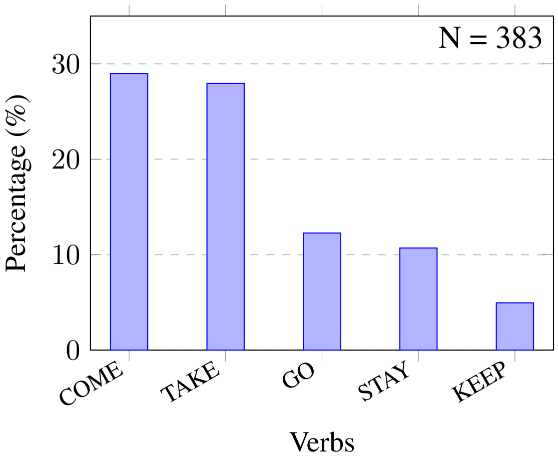

Figure 1: Distribution of verbs functioning as light in the MVC class. The figure illustrates top five light verbs with the highest frequencies amongst the thirteen light verbs identified.

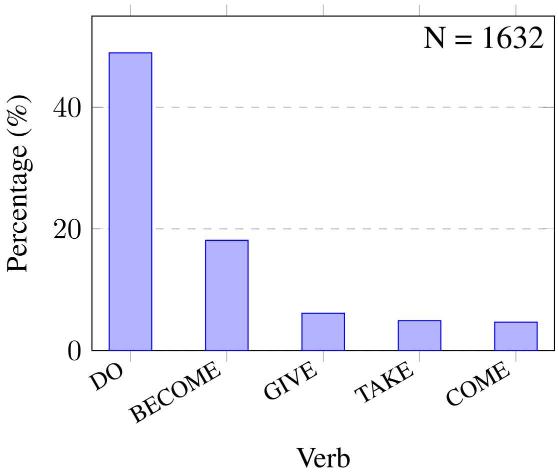

Figure 2: Distribution of verbs functioning as light in the LVC.full class. The figure illustrates top five light verbs with the highest frequencies amongst the twenty-three light verbs identified.

other such seemingly similar structures are carefully separated from the actual MVCs during the annotation process.

3.2 Light Verb Constructions

PARSEME 2.0 (Savary et al., submitted) defines light verb constructions (LVCs) as expressions formed by a verb, and a wide range of preverbs like nouns, adjectives, prepositions, etc (Family, 2014). Similar to Hindi, Marathi has nouns and adjectives as preverbs. There is also evidence of adverbs as preverbs in the database. In this subsection, we talk about the characteristics of these preverbs, and also discuss their identification strategies.

Literature on LVCs in Marathi is rather sparse, though Kulkarni (2019) and Hook and Pardeshi (2009) touch upon marɳe 'hit' and k^{h}aɳe 'eat' expressions briefly. Family (2014)'s identification of Persian light verbs under this category can be extended to Marathi for the purposes of annotation. These light verbs include kərɳe 'do', pəɖɳe 'fall', hoɳe or bənɳe 'become', miɭɳe 'get', aɳɳe 'bring', ʥaɳe 'go', and jeɳe 'come'. Certain perception verbs like pahɳe 'see' also function as light verbs (Kume, 2011). Accordingly, we have considered synonyms of 'see' like d̪isɳe 'see' as light verbs. Additionally, we have annotated verbs like lagɳe 'be attached', waʈɳe 'seem', along with some of the verbs recognized as light verbs in the MVC category like kaɖ^{h}ɳe 'draw out'.

We refer to this list to annotate the verbs, and to examine their empirical distribution. We have identified twenty-three such verbs. Figure 2 shows the verbs with the highest frequencies in the data. Verbs like k^{h}aɳe 'eat' and suʈɳe 'be released' have the lowest frequencies. We encounter certain verbs like suʈɳe 'be released' and soɖɳe 'be release-cause', bənɳe 'make' and bənəwɳe 'make-cause', wherein the second verb is in the causative form of the first verb. The causative light verb has been annotated as LVC.cause, following PARSEME 2.0 (Savary et al., submitted) guidelines and the non-causative form is LVC.full. LVC.cause are very few in number, and only attested with four verbs like soɖɳe 'be release-cause', bənəwɳe 'make-cause', miɭəwɳe 'get-cause', and d̪ak^{h}əwɳe 'see-cause'. Verbs like bənəwɳe 'make-cause' and miɭəwɳe 'get-cause' have the highest frequencies while soɖɳe 'be release-cause' and d̪ak^{h}əwɳe 'see-cause', the lowest.

Bonial (2021) notes that the event semantics of LVCs stems from the nouns (and other preverbs), rather than the verbs alone. These nouns (and other preverbs) also distinguish these verbs from their full verb and light verb usages. Verbs in their full verb usages denote their literal, canonical sense, while the light verbs constitute the nonliteral senses. When the nominal preverbs are abstract denoting events or states the verb is light, else full. The corpora show that while this holds true for most of the light verbs there are certain light verbs that have no such selectional restrictions. Light verbs like hoɳe 'become', ʥaɳe 'go', bənɳe 'make', kərɳe 'do', also as noted by Family (2014), select nominal preverbs that are not abstract.

Based on this understanding, we come up with certain heuristics to distinguish the verbs into their light and full versions. The rephrasing test states that such constructions can be rephrased with one-word predicate. The Marathi corpora have examples like ut̪t̪ər d̪eɳe 'answer give' which can be paraphrased into corresponding single verb - ut̪t̪ərɳe 'answer'. But this test is not valid for most of the expressions as they cannot be mapped to their corresponding single verb forms. Thus, we come up with the following diagnostics beyond the rephrasing test to identify these monoclausal constructions:

1. **Omission of the preverb**: In the preverb-verb sequence, the preverb cannot be omitted. Example: ramne kʰoli swətʧʧʰə ʈʰewli | lit. 'Ram room clean keep' ('Ram kept the room clean') cannot be rewritten as *ramne kʰoli ʈʰewli 'Ram room kept'

2. **Co-ordination**: Event nouns as preverbs cannot be co-ordinated. Example: *t̪jane bʰeʈ aɳi məd̪ət̪ d̪ili 'he visit and help give'

3. **Limited compatibility with light verbs**: Certain nouns functioning as preverbs like bʰaʃəɳ 'speech' allow certain light verbs like kərɳe 'do' or d̪eɳe 'give'.

The LVC category is the most productive as compared to all other MWEs in Marathi across all the corpora that were examined (See Table 2).

3.3 Verbal Idioms

Verbal Idioms (VIDs) are a sequence whose meaning does not arise from the meaning of either of the component verbs. Example: aɖʰeweɖʰe gʰeɳe | lit. 'roundabout take' ('to make excuses'). These are relatively fewer in number as compared to MVCs and LVCs in the corpora.

4 Non-verbal MWEs

Non-verbal MWEs consist of a broad category of MWEs based on their syntactic role - nominal (NID), adjectival (AdjID), adverbial (AdvID), and other MWEs with other functional categories. Constant et al. (2017) assert that non-verbal MWEs can be grouped into the following schemes that are non-exhaustive and often overlapping. We follow the grouping to categorize the non-verbal MWEs identified in the Marathi data.

1. **Compounds**: can be further divided into two types: closed and open compounds. Closed compounds like kagad̪pət̪t̪rə 'document' are formed by two or more words functioning as a single token, and open compounds like mit̪t̪rə-məit̪riɳi 'friends' are formed from lexemes separated by spaces or hyphens.

2. **Mutiword term**: a multiword designation of a general concept in a specific subject field. Example: utʧʧə rəkt̪əd̪ab 'high blood pressure'

3. **Complex function word**: functional word formed by one or more lexeme. Example: ʥəwəɭpas 'nearby, almost'

4. **Idioms**: a group of lexemes whose meaning is established by convention. Example: ʤiw ki praɳ | lit. 'heart or spirit' ('immense love')

Constant et al. (2017) state that NIDs can also be classified into multiword named entities designating real-world entities like persons, organizations, locations, etc. The PARSEME 2.0 (Savary et al., submitted) guidelines do not identify these expressions as MWEs. Therefore, they are not annotated.

Reduplication is a morphophonological phenomenon found in several Indic languages. In this dataset, we look at them from the point of view of multi-word expressions. There are varieties of reduplication in the data. Total reduplication appears in examples like gərəm gərəm 'hot hot', and onomatopoeic expressions like fəɳ-fəɳ 'a kind of sound' while partial reduplication can be seen in expressions like awəɖi niwəɖi 'likes dislikes'. Moreover, semantic reduplication like t̪ʰəɳɖəgar | lit. 'cold cold' ('very cold') is found in the datasets. Pandharipande (1998) refers to such expressions as emphatic compounds, as this process intensifies the meaning of the first noun by the use of a synonym.

The semantic properties of the compound expressions are also taken into account during annotation. Pandharipande (1998) refers to expressions like hat̪-paj 'hand-feet' as a superordinate compound, as the two nouns belong to the same semantic class and there is no hierarchical head-embedding between the two. The expression overlaps with the class of a copulative compound. We have annotated these expressions, depending upon the class of the individual components. Accord-

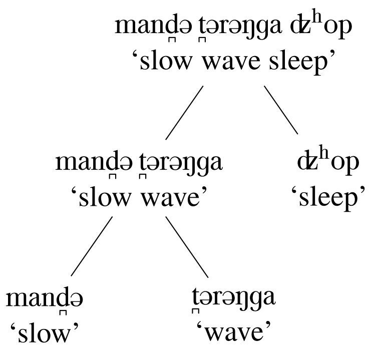

Figure 3: Stacked annotation observed in NID. The tree depicts one of the possible annotations of the NID appearing as a closed compound in the dataset. It suggests that NID has its own internal structure, and needs to be combined in the specific order to render desired meaning.

ingly, the example mentioned above is labeled as NID.

The author states that in adjective-noun compounds like uʧʧə rəkt̪əd̪ab ‘high blood pressure’, noun is the semantic head, and adjective modifies the noun. The resulting expression is a noun. We have classified such examples as NID. However, noun-adjective expressions like praɳəgʰat̪ək ‘life-threatening’ are annotated as AdjIDs because the resulting expression is an adjective. Expressions like lakuɖt̪oɖ । lit. ‘wood break’ (‘the act of breaking a log of wood’) are noun-verb compounds wherein the derived compound functions as a noun. They are rarely found in the data, and following Pandharipande (1998), they are tagged as NID.

The corpora have certain expressions that have an internal structure, and span over multiple tokens. They result in nested annotations as seen in Figure 3. The current guidelines for annotation do not permit nested annotations for closed compounds. Therefore, we have annotated them as a flat structure.

5 Properties of Marathi MWEs

Marathi MWEs, as observed in the data, possess certain properties that are challenging for their annotation and representation. In this section, we briefly present an overview of their characteristics that provide the rationale underlying annotation decisions.

- **Heterogeneity**: Section 3 and Section 4 in the paper show that the annotated MWEs are not confined to any specific syntactic construction. They are linguistically diverse, and cannot be restricted to only compound nouns and light verbs.

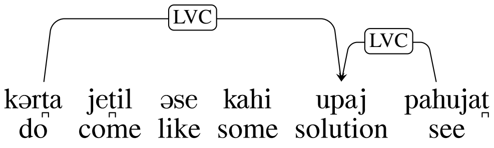
LVC
LVC
kərta jet̪il əse kahi upaj pahujat̪
do come like some solution see

Figure 4: The sentence can be roughly glossed as - Let us look at some solutions that can be done. The figure indicates that noun ‘solution’ is shared by the two verbs - ‘do’ and ‘see’.

- **Non-compositionality**: Within the entire class of Marathi MWEs, idioms are highly non-compositional. The rest of the categories fall on a continuum between compositionality and non-compositionality.

- **Overlap** (Schneider, 2014): There are some overlapping MWE instances. Figure 4 shows that the noun *solution* overlaps with two distinct verbs, acting as a preverb for both light verbs.

- **Gappy grouping** There are intervening elements between the components of MWEs, making them discontinuous (Constant et al., 2017). Schneider (2014) classifies the ‘gap’ as the argument gap formed by an argument of the predicate, and the modifier gap created due to the intervening adjective, adverb, or determiner.

(1) ha prajog < aʈʰəwɖjat̪un ek weɭa > kəra
this experiment < in a week one time > do
Perform this experiment once a week.

In (1), the LVC.full in blue is discontinuous, separated by an adverbial modifier.

Most of the constructions exhibiting this property belong to the LVC class. Marathi corpora reinforce the fact that MVCs are tightly integrated verbal units with restricted internal syntax, while LVCs permit intervening linguistic material (Butt, 1995).

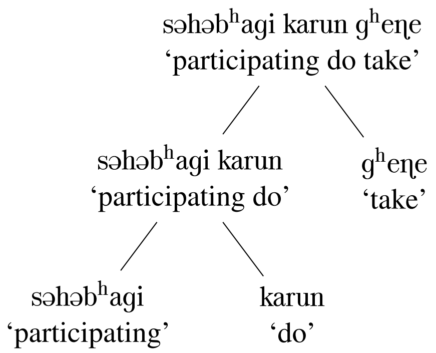

Figure 5: Stacked annotation observed in a verbal MWE. The root of the tree is an MVC. The terminal nodes together form an LVC. This construction is nested within a larger MVC with the light verb ‘take’.

- **Stacked annotation**: When an MWE contains another MWE it leads to a hierarchical structure. While this phenomenon is observed in Hindi (Jain and Vaidya, 2024), it is also found in both the verbal and non-verbal MWEs of Marathi. Figure 5 illustrates a verbal MWE within a verbal MWE. Whenever it is possible to preserve the embedded structure, that representation is preferred. However, closed compounds as discussed in Figure 3 are annotated as a flat structure.

6 Summary and Conclusion

We develop an exhaustive knowledge base of Marathi MWEs of all syntactic types - verbal, nominal, adjectival, adverbial, and other functional types. The consistency checks have been performed as per PARSEME 2.0 (Savary et al., submitted) guidelines. The Table 2 mentions the distributional patterns of MWEs in Marathi, revealing both frequent patterns and exceptional cases in the language.

Wherever the precise MWE identificational criteria are not studied, we attempt to propose them based on the empirical evidence from the corpora. However, determining the MWE-hood status of these expressions remains challenging, owing to their structural and semantic properties.

7 Limitations

There are certain limitations affecting the applicability of the resource. First, the annotations are performed by a single annotator. Therefore, inter-annotator agreement cannot be reported. Secondly, the resource depends on the automatic preprocessing done using UDPipe (Straka and Straková, 2017). This has led to errors in tokenization, POS tagging. Though we have prioritized the MWE annotations, we plan to manually review and correct these automatically generated annotations for future release.

References

Claire Bonial. 2021. Précis of take a look at this! form, function, and productivity of english light verb constructions. *Colorado Research in Linguistics*, 25.

Miriam Butt. 1995. *The structure of complex predicates in Urdu*. Center for the Study of Language (CSLI).

Mathieu Constant, Gülen Eryiit, 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.

Neiloufar Family. 2014. *Semantic spaces of Persian light verbs: A constructionist account*, volume 6. Brill.

Peter Hook and Prashant Pardeshi. 2009. A taxonomy of eat expressions in marathi. *Annual Review of South Asian Languages and Linguistics*, pages 41–63.

Kanishka Jain and Ashwini Vaidya. 2024. Revisiting VMWEs in Hindi: Annotating layers of predication. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 98–105, Torino, Italia. ELRA and ICCL.

Ioannis Korkontzelos. 2010. Unsupervised learning of multiword expressions. Unpublished.

Aaditya Kulkarni. 2019. Semantics of hit expressions in marathi.

Yusuke Kume. 2011. On the complement structures and grammaticalization of see as a light verb. *The Electronic Library*, 28:206–221.

Marja Nenonen, Jussi Niemi, and Matti Laine. 2002. Representation and processing of idioms: Evidence from aphasia. *Journal of Neurolinguistics*, 15(1):43–58.

Renuka Ozarkar. 2014. *Structures of Marathi verbs*. Ph.D. thesis, Doctoral dissertation, University of Mumbai.

Rajeshwari V Pandharipande. 1998. *Marathi*. Routledge.

Corpora	MVC	LVC.full	LVC.cause	VID	NID	AdjID	AdvID
UD Treebank (%)	46 (1.19)	120 (3.11)	1 (0.02)	17 (0.44)	107 (2.77)	12 (0.31)	15 (0.38)
Anuvaad Corpus (%)	243 (0.86)	1279 (4.57)	2 (0.007)	2 (0.007)	868 (3.10)	55 (0.19)	62 (0.22)
Children's Stories (%)	94 (2.19)	233 (5.43)	5 (0.11)	3 (0.02)	24 (0.55)	11 (0.25)	13 (0.30)
Total (%)	**383**(1.06)	**1632** (4.52)	**8** (0.02)	**22** (0.06)	**999** (2.76)	**78** (0.21)	**90** (0.24)

Table 2: Distribution of MWEs in Marathi. The table presents the entire landscape of MWEs identified and annotated. PARSEME 2.0 (Savary et al., submitted) identifies other constructions like inherently reflexive verbs (IRV), inherently adpositional verbs (IAV), etc. They are not attested in Marathi data.

Prashant Pardeshi, Peter E. Hook, and Sung-Yeo Chung. 2006. In search of the origins of compound verbs in marathi. Handout of the presentation made at SALA 26, CIIL, Mysore.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* " O'Reilly Media, Inc.".

Vinit Ravishankar. 2017. A universal dependencies treebank for marathi. In *Proceedings of the 16th international workshop on treebanks and linguistic theories*, pages 190–200.

Agata Savary, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Păiş, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto A. Díaz Hernández, Victoria Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, and 9 others. submitted. PARSEME 2.0: Multilingual corpus of multiword expressions. Submitted to LREC 2026.

Nathan Schneider. 2014. *Lexical Semantic Analysis in Natural Language Text*. Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Dhirendra Singh, Sudha Bhingardive, and Pushpak Bhattacharyya. 2016. Multiword expressions dataset for indian languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2331–2335.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pages 88–99.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Eva Wittenberg and Maria Piñango. 2011. Processing light verb constructions. *The Mental Lexicon*, 6.