Gathering Compositionality Ratings of Ambiguous Noun-Adjective Multiword Expressions in Galician

Laura Castro and **Marcos Garcia**
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{laura.castro,marcos.garcia.gonzalez}@usc.gal

Abstract

Multiword expressions pose numerous challenges to most NLP tasks, and so do their compositionality and semantic ambiguity. The need for resources that make it possible to explore such phenomena is rather pressing, even more so in the case of low-resource languages. In this paper, we present a dataset of noun-adjective compounds in Galician with compositionality scores at token level. These MWEs are ambiguous due to being potentially idiomatic expressions, as well as due to the ambiguity and productivity of their constituents. The dataset comprises 240 MWEs that amount to 322 senses, which are contextualized in two sets of sentences, manually created, and extracted from corpora, totaling 1,858 examples. For this dataset, we gathered human judgments on compositionality levels for compounds, heads, and modifiers. Furthermore, we obtained frequency, ambiguity, and productivity data for compounds and their constituents, and we explored potential correlations between mean compositionality scores and these three properties in terms of compounds, heads, and modifiers. This valuable resource helps evaluate language models on (non-)compositionality and ambiguity, key challenges in NLP, and is especially relevant for Galician, a low-resource variety lacking annotated datasets for such linguistic phenomena.

1 Introduction

Multiword expressions (MWE) are idiosyncratic word combinations that constitute both major challenges and interests in Natural Language Processing (Sag et al., 2002; Miletić and Walde, 2024). The reasons lie in their intricate nature, as MWEs can fall within a wide range of semantic compositionality, and both the expressions and their constituents may present different degrees of semantic ambiguity, among other challenging properties that complicate most NLP tasks (Constant et al., 2017). An example of the former is *dark horse*, which can be interpreted literally as a horse that is of a dark color or idiomatically as an *unexpected winner*, depending on the context. An example of the latter is *common sense*, which may be understood as the most frequent meaning of a word, expressions, etc., but can also be used to refer to a person's reasonable or good judgment, depending on the context.

In the last two decades, numerous datasets have been put forward to address the issues MWEs pose (Ramisch, 2023). Among them, we can find the dataset of English noun-compounds with compositionality ratings (Reddy et al., 2011), as well as its extensions for French and Portuguese (Cordeiro et al., 2019). For German, there exists a noun-noun compound dataset featuring compositionality ratings (Schulte im Walde et al., 2016b). Similarly, (Schulte im Walde, 2024) put forward a collection that comprises German compounds with compositionality ratings, where compound and constituent properties are also taken into account. Related datasets contain binary or three-way classification (literal/idiom/other) of MWEs and Potentially Idiomatic Expressions (PIE), such as the VNC-tokens dataset (Cook et al., 2008), comprising about 3,000 verb-noun combinations in English, or MAGPIE (Haagsma et al., 2020), featuring around 56k English PIEs in corpora-extracted sentences, also featuring the literal/idiomatic/other classification. Likewise, the SemEval-2022 Task 2 introduced binary classification datasets in English, Portuguese, and Galician (Tayyar Madabushi et al., 2022).

These datasets present highly valuable resources for NLP tasks. However, most of them are annotated at a type level (Reddy et al., 2011; Cordeiro et al., 2019). On the other hand, those that operate at a token level (Garcia et al., 2021) tend to comprise MWEs that convey the same meaning in all sentences compiled in the dataset. Therefore, such resources may not account for the wide variety of senses these idiosyncratic expressions can have.

Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), pages 32–40
May 4, 2025 ©2025 Association for Computational Linguistics

This situation is only more dire in the case of language varieties with few annotated resources, like Galician, for which few such works exist despite being essential to explore if language models can adequately process MWEs (Dankers et al., 2022; Miletić and Walde, 2024; He et al., 2025).

We address such shortcomings by presenting a collection of noun-adjective compounds in Galician. Their ambiguity is two-fold, since the dataset contains 1) potentially idiomatic MWEs with different degrees of compositionality, and 2) MWEs whose constituents present different degrees of ambiguity and productivity. These expressions are disambiguated, and their senses are contextualized and preliminarily classified in terms of compositionality. Overall, the dataset comprises 240 noun-adjective MWEs, and 322 senses. Each sense is contextualized in two manually-written and four corpora-extracted sentences, which account for a total of 1,858 contextualizing sentences.[1]

As a key contribution of this paper, this resource provides a set of human ratings on semantic compositionality levels for the 322 senses, along with additional linguistic information. In this regard, we enrich the dataset with frequency, ambiguity, and productivity data extracted from corpora and lexical resources for compounds and their constituents, used to explore potential correlations between linguistic features of the dataset. This publicly available dataset constitutes a valuable resource for evaluating language models on compositionality prediction and sense disambiguation tasks, among others.[2]

2 Creation of the Dataset

The goal was to create a dataset of potentially idiomatic, ambiguous MWEs in Galician, contextualized in validated sentences used to rate the expressions' senses in terms of compositionality.

2.1 Multiword expressions and sentences

The Galician version of the Wikipedia, parsed with UDPipe (Straka, 2018), was used to extract noun-adjective compounds, which were ranked by number of occurrences. From them, a manual selection was carried out. The goal was to obtain MWEs with different degrees of frequency, compositionality, polysemy, and semantic ambiguity. For that matter, 240 compounds spanning different frequency ranges were selected. Then, for each of them, a manual definition of the potential senses the MWEs could take up depending on the context was carried out, totaling 322 senses. As a preliminary classification, senses were classified in terms of compositionality as *compositional*, *partial*, or *idiomatic*, depending on the transparency of their constituents or lack of thereof.

Thus, in those instances where the transparency of both constituents made it possible to infer the meaning of the compound as a whole, the expressions where classified as *compositional*. In cases where only the meaning of one of the constituents was transparent, expressions where ranked as *partial*. Lastly, when the meaning of the expressions could not be inferred from the semantics of their constituents, they were graded as *idiomatic*. Multiword expressions themselves were also classified in an identical manner, although a fourth label was used for those polysemic expressions whose different senses could take up more than one classification depending on the context. In these cases, expressions where classified as *Potentially Idiomatic Expressions* (PIE).

Compositional examples include *especie vexetal* ('plant species') and *enfermidade mental* ('mental illness'). Examples of partially idiomatic senses include *sentido común* ('common sense', meaning a person's 'sound judgment') and *tubo dixestivo* (which does not literally refer to a 'digestive tube', but to the 'digestive tract'). As for idiomatic expressions, the dataset includes *aire libre* (which does not literally refer to 'free air', but to the 'outdoors'), and *fillo predilecto* (which is not a 'favourite child', but a honorary title towns and cities give to remarkable citizens that were born within their jurisdiction). Potentially idiomatic MWEs include other noun-adjective compounds, such as *red line*, which can be used literally to talk about a line of a red color which is painted on a paper, for example, as well as idiomatically to talk about a personal *boundary* or limit that shall not be crossed.

Additionally, to contextualize the expressions comprised in the dataset, two sets of sentences were constructed. Firstly, a language expert created a set of two manually-written sentences per sense (644 in total). Secondly, the Wikipedia corpus and other textual resources were used to extract examples containing the MWEs, of which four were selected

[1] We build upon the expressions and sentences previously compiled in Castro et al. (2025), enriching them with additional information to enhance their scope and applicability in this work.

[2] The dataset can be found at: `https://github.com/Castro-L/MWE_dataset_gl`

	MWEs				Senses			Sentences	
	Comp.	**Part.**	**Idiom.**	**PIE**	**Comp.**	**Part.**	**Idiom.**	**Manual**	**Corpora**
Number	115	65	18	42	189	85	48	644	1,214
Total	**240**				**322**			**1,858**	

Table 1: Distribution and total number of multiword expressions, senses, and sentences contained in the dataset. The numbers for MWEs and senses correspond to the preliminary classification in Compositional, Partial, Idiomatic, and Potentially Idiomatic Expressions (PIEs).

per sense by the language expert (1,214 in total). The set of manual sentences was curated by two linguists that reviewed the expressions, senses, and examples. The set of extracted sentences was validated by five other linguists, who verified that at least the first of the sentences had the same meaning as one of the manually-written, curated examples. Table 1 summarizes the composition of the dataset, while Table 5 (Appendix C) contains a set of examples of multiword expressions and corpora-extracted sentences comprised in the dataset. A more detailed description of the creation process and the composition of the dataset can be found in Castro et al. (2025).

2.2 Annotation of compositionality levels

Once the dataset was completed, an annotation task was carried out to gather human judgments on the semantic compositionality of the expressions and their constituents.

2.2.1 Annotation task

The annotation task featured the total of 322 senses. To properly contextualize them, one of the two manually-written sentences was randomly selected for each sense. Given that such sentences had been curated, they allowed us to ensure that examples were not ambiguous, represented each sense correctly, and provided enough context for annotators to make meaningful judgments. Due to time and personnel constrains, only one of the two sentences could be annotated. The procedure and subtasks were inspired by other relevant works where compositionality scores were gathered for MWEs (Reddy et al., 2011; Schulte im Walde, 2024).

Instructions: To instruct annotators on how to carry out the task, guidelines were provided. The goal was set to *reflect on each expression and the elements they are made up of, in terms of how literal or not they may be, based on the example sentences*, and annotators were asked to answer the questions in strict order.

Compound: Firstly, annotators were asked to consider the expressions out of context. The aim was to prompt linguistic reflection on each compound as a whole, both in terms of semantics and compositionality levels.

Example sentence: Subsequently, annotators were asked to read an example sentence. Given the length of some examples, and the fact that certain expressions allow for other linguistic elements to appear in-between constituents, both elements were highlighted in bold for readability's sake.

Compositionality of the compound: Next, annotators had to consider the meaning of the compound within the example, and to provide a score for it. To further prompt linguistic reflection, questions were posed as follows: *In the sentence, and on a scale from 0 (not literal) to 5 (literal), is [MWE] literally a [noun] that is [adjective]?*

Compositionality of the constituents: Then, annotators had to consider how literal or not literal the head and the modifier were based on the example, and to provide a score for it: *In the sentence, and on a scale from 0 (not literal) to 5 (literal), how literal or not literal is [noun/adjective]?*

2.2.2 Annotation process

Two sets of annotations were obtained. One of them was completed by the main language expert. The second annotation was carried out by six external annotators, all of them native speakers of Galician with background in Linguistics. Both the expert and the annotators were given the same instructions, and an identical annotation task to complete. In the case of the external annotators, given its size, the task was equally and randomly divided into six annotation sheets, so that each annotator would rate the same number of instances, up to completing a full annotation. As a result, we put forward two sets of annotations, as well as the mean values of both sets, for compounds, heads, and modifiers of all MWEs and senses featured in the original dataset.

2.3 Results

Compositionality scores: Mean values were determined for the compounds, heads, and modifiers of the senses that had been preliminarily classified as *compositional*, *partial*, and *idiomatic*. Table 2 shows the mean compositionality scores of the MWEs and their constituents in each of the three classes. In general, the scores per class for the MWEs and the constituents follow the same tendencies as in similar datasets for other languages, only diverging in the compositionality score of the partially idiomatic compounds (1.87). A more detailed distribution of compositionality ratings per category can be found in the bloxplots in Appendix A.

Element	Idiom	Part	Comp
Compound	1.00	1.87	3.60
Head	1.25	2.57	3.88
Modifier	1.28	2.26	3.83

Table 2: Mean compositionality scores for compounds, heads, and modifiers belonging to senses classified as Idiomatic, Partial, and Compositional.

Similarly, annotation scores allowed us to obtain threshold values for the three categories. Thus, values ranging from 0 to 1.44 would be considered idiomatic; values ranging from 2.73 to 5 would be labeled as compositional, and in-between values would correspond to partially idiomatic compounds. Following such thresholds, 167 senses scored compositional values, while 155 were rated as non-compositional — from those, 93 senses were partially idiomatic, and 62 senses were considered idiomatic. In comparison with the preliminary classification, there are 100 senses that correspond to a different category. Of those hundred cases, 67% of instances obtained higher scores in human ratings than the threshold values of the preliminary category they had been given, thus indicating that the dataset may be more non-idiomatic than it was previously classified as.

Inter-annotator agreement: We determined 1) Krippendorff's α (Krippendorff, 2011) for the whole dataset using the scores provided in both sets of annotations, and 2) weighted Cohen's κ (Cohen, 1960) for the values of each annotator in their corresponding subsets. Krippendorff's α is 0.70 for compounds, 0.66 for heads, and 0.58 for modifiers. κ values for subsets range from 0.34 to 0.70. The complete inter-annotator agreement scores can be seen in Appendix B.

3 Frequency, ambiguity, and productivity

Following previous works on datasets of similar nature, frequency, ambiguity, and productivity data were obtained for compounds, heads, and modifiers of all senses, aimed at studying the relationships between these properties regarding their compositionality degrees (Schulte im Walde et al., 2016a; Schulte im Walde, 2024). For frequency and productivity, the original corpus of MWE extraction was used, while ambiguity data was extracted from Galnet (Gómez Guinovart, 2011).[3]

3.1 Frequency data

Regarding frequency, we enriched the dataset with the normalized frequencies of the compounds and their constituents: 1) **Compound frequency**, which represents the normalized frequency of each MWE within the original corpus; 2) **head frequency**, calculated using the total number of times it appears as a head in any noun-adjective compound, and 3) **modifier frequency**, computed also counting the total number of times it appears as a modifier in any noun-adjective compound within the corpus.

3.2 Ambiguity data

In this case, we have gathered: 1) **Head ambiguity**, where, for each syntactic head, the total number of synsets available in Galnet were extracted. In this case, we compiled two types of ambiguity data: 1.a) **overall head ambiguity** data, that represents the total number of synsets, regardless of its grammatical category, and 1.b) **category head ambiguity** data, where only those synsets corresponding to the *noun* category are accounted for. Besides, we obtained 2) **modifier ambiguity**, using the number of synsets available in Galnet for each modifier. As with heads, there are two types of data: 2.a) **overall modifier ambiguity** data, for the total number of synsets, and 2.b) **category modifier ambiguity** data, where only those synsets corresponding to the *adjective* category were taken into account.

3.3 Productivity data

Finally, we used the Wikipedia corpus to compile: 1) **Head productivity** data, where the total number of unique combinations within the extraction

[3] It is worth noting that Galnet is a relatively smaller lexical resource, containing approximately 36% of the synsets and 31% of the words found in the English WordNet (Guinovart et al., 2021).

	Frequency			Ambiguity				Productivity	
	Comp.	**Head**	**Modif.**	**Head-a**	**Modif-a**	**Head-c**	**Modif-c**	**Head**	**Modif.**
Comp.	0.063	0.081	0.136	-0.026	0.033	-0.022	0.023	0.082	0.143
Head	0.030	0.039	0.090	*-0.154*	0.141	*-0.152*	0.123	0.017	0.126
Modif.	0.103	0.093	0.130	0.085	-0.142	0.093	-0.138	0.105	0.083

Table 3: Spearman ρ correlations between the compositionality of the compounds, heads, and modifiers (rows) and frequency, ambiguity, and productivity (columns). Ambiguity includes overall (-*a*) and category-based (-*c*) results. Italics indicate p-values < 0.01, while underlining denotes p-values between ≥ 0.01 and 0.05. Results with p-values ≥ 0.05 remain unformatted.

corpus was determined for each head of the compounds present in the dataset. In this case, the normalized values are relative to the number of unique MWE combinations in the dataset; and 2) **modifier productivity**, where for each modifier in the dataset, the total number of unique combinations within the original corpus was determined. In the two cases, both raw and normalized values are provided.

3.4 Correlations

We computed Spearman's ρ correlation between the mean compositionality scores for compounds, heads, and modifiers and frequency data (compound, head, and modifier), ambiguity data (head and modifier, both overall and category-wise), and productivity data (head and modifier).

As it can be seen in Table 3, the correlations were overall weak, and mostly not significant. These results are in line with the findings of other related works, such as the German noun-noun compound dataset (Schulte im Walde et al., 2016b), where ρ between compositionality and productivity was -0.204 for heads and -0.023 for modifiers. Similarly, in a recent analysis of various datasets, Schulte im Walde (2024) found no correlations between compositionality scores and frequency, productivity, and ambiguity data across several English and German datasets, with the exception of the English NN-compounds dataset (Reddy et al., 2011), where moderate correlations were observed with frequency and productivity data. While our task was of a relative different nature, as ours operated at a token, not a type level, it is still worth noting that it follows the same general trend found in other works. However, since our dataset does account for the different senses MWEs can take up depending on the context, more exploration is needed, especially in relation to potential differences between monosemic and polysemic expressions.

4 Conclusions and Further work

In this work, we have introduced a dataset comprised of 240 noun-adjective MWEs in Galician that account for 322 senses, which present varying degrees of compositionality as well as semantic ambiguity. We have put forward human judgments on compositionality scores, which served to ascertain where MWEs fall within the spectrum of idiomaticity, and also provided frequency, ambiguity and productivity data. Using this information, we only found very weak and non-significant correlations with compositionality scores. The dataset, which comprises manually created sentences and examples extracted from corpora, fills a gap in annotated resources for Galician. The dataset will be freely released, except for the manually annotated sentences, which will be kept for evaluation purposes only to prevent data contamination in language models.

For future work, we aim to collect additional human ratings to strengthen the annotation of the dataset presented in this paper, ensuring greater reliability and consistency. Furthermore, we plan to apply the same methodology to construct similar datasets for other types of linguistic expressions, such as verb-object combinations. Additionally, it would be valuable to explore other linguistic properties and contextual cues that may influence human perception of semantic compositionality, providing deeper insights into the factors that shape meaning construction.

Limitations

Our dataset comprises a compilation of MWEs, senses, and contextualizing sentences. Additionally, it provides compositionality scores. They were the result of a meticulously crafted annotation task that contextualizes compounds in curated examples to ensure an adequate representation of senses. However, our main limitation is the number of human annotations obtained per sense. Our work was limited to seven annotators, which put forward two sets of ratings. Although insightful, the data provided could be greatly enriched by a higher number of ratings that fully represent the degrees of compositionality of the MWE senses. Additionally, Galnet made it possible to obtain four types of ambiguity data with which to explore the relationships between linguistic phenomena. However, it shall be pointed out that Galnet is a limited resource size-wise. Thus, further work is needed to gather more human judgments, as well as to further expand Galnet's number of synsets to allow for a finer representation of constituents' ambiguity.

Acknowledgments

This work was funded by MCIN/AEI/10.13039/501100011033 (grants with references PID 2021-128811OA-I00, and TED 2021-130295B-C33, the latter also funded by "European Union Next Generation EU/PRTR"), by the Galician Government ("Centro de investigación de Galicia" accreditation 2024-2027 ED431G-2023/04, and by the project with reference ED431F 2021/01) and the European Union (European Regional Development Fund - ERDF), and by a Ramón y Cajal grant (RYC2019-028473-I). This publication was also produced within the framework of the Nós Project, which is funded by the Spanish Ministry of Economic Affairs and Digital Transformation and by the Recovery, Transformation, and Resilience Plan, Funded by the European Union - NextGenerationEU (reference 2022/TL22/00215336), and by the Xunta de Galicia through the collaboration agreements signed in with the University of Santiago de Compostela.

References

Laura Castro, Anna Temerko, and Marcos Garcia. 2025. Compositionality and Ambiguity in Multiword Expressions: A Dataset for the Evaluation of Language Models in Galician. In *Progress in Artificial Intelligence*, volume 14969 of *Lecture Notes in Computer Science*, pages 228–240, Cham. Springer Nature Switzerland.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Xavier Gómez Guinovart, Itziar Gonzalez-Dios, Antoni Oliver, and German Rigau. 2021. Multilingual Central Repository: a Cross-lingual Framework for Developing Wordnets. *arXiv preprint arXiv:2107.00333*.

Xavier Gómez Guinovart. 2011. Galnet: WordNet 3.0 do Galego. *Linguamática*, 3(1):61–67.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Investigating Idiomaticity in Word Representations. *Computational Linguistics*.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. In *Departmental Paper (ASC)*, volume 43.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.

Carlos Ramisch. 2023. *Multiword expressions in computational linguistics: Down the rabbit hole and through the looking glass*. Aix Marseille Université.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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.

Sabine Schulte im Walde. 2024. Collecting and investigating features of compositionality ratings. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword Expressions in Lexical Resources. Linguistic, Lexicographic and Computational Perspectives*, chapter 8, pages 269–308. Language Science Press.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016a. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany. Association for Computational Linguistics.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016b. GhoSt-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2285–2292, Portorož, Slovenia. European Language Resources Association (ELRA).

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

A Appendix: Distribution of compositionality scores

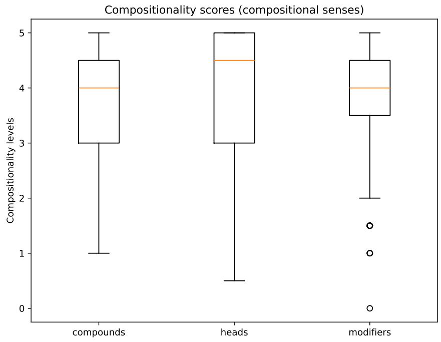

Figure 1: Scores for compounds, heads, and modifiers of expressions classified as compositional.

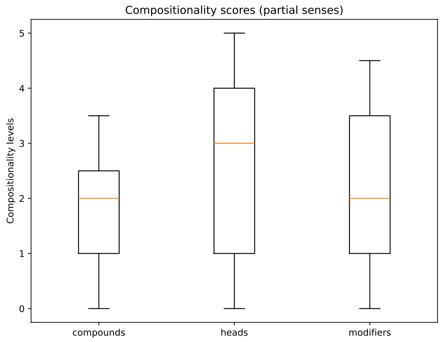

Figure 2: Scores for compounds, heads, and modifiers of expressions classified as partially idiomatic.

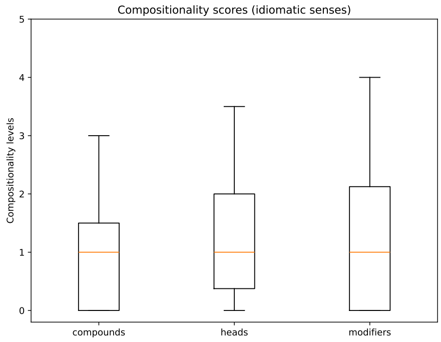

Figure 3: Scores for compounds, heads, and modifiers of expressions classified as idiomatic.

B Appendix: Inter-annotator agreement scores

Annot.	**Compound**	**Head**	**Modifier**
All (α)	0.705	0.663	0.584
Set-1 (κ)	0.549	0.518	0.489
Set-2 (κ)	0.520	0.565	0.434
Set-3 (κ)	0.473	0.463	0.345
Set-4 (κ)	0.528	0.526	0.525
Set-5 (κ)	0.640	0.708	0.540
Set-6 (κ)	0.556	0.558	0.577

Table 4: Agreement for compounds, heads, and modifiers per annotators' subsets. Top row are Krippendorff's α values for the whole dataset, while bottom rows refer to the weighted Cohen's κ of individual sets of MWEs.

C Appendix: Examples of multiword expressions and contextualizing sentences

Category	MWE	Galician Sentence	English Translation
Comp.	**bebida alcohólica** ('alcoholic drink')	A cervexa e todas as *bebidas alcohólicas* feitas a partir da fermentación tamén son produtos fúnguicos.	Beer and all *alcoholic drinks* made from fermentation are also fungal products.
	incendio forestal ('forest fire')	Este fenómeno aumenta considerablemente o perigo de *incendios forestais* nos outeiros e montañas.	This phenomenon considerably increases the risk of *forest fires* in hills and mountains.
	bandeira vermella ('red flag')	O 22 de setembro, a *bandeira vermella* reapareceu e pouco tempo despois a bandeira tricolor estoniana foi retirada.	On September 22nd, the *red flag* reappeared and shortly afterwards the Estonian tricolor flag was withdrawn.
Part.	**partido amigable** ('friendly game')	Por tal motivo a selección brasileira xoga os seus *partidos amigables* e clasificatorios en diferentes escenarios.	For this reason, the Brazilian national team plays its *friendly* and qualifying *matches* in different settings.
	paraíso fiscal ('fiscal paradise')	Moitos países teñen tratados fiscais bilaterais que evitan ao seus residentes pagar impostos dobres, pero poucos teñen tratados cos *paraísos fiscais*.	Many countries have bilateral tax treaties that prevent their residents from paying double taxes, but few have treaties with *tax havens*.
	bandeira vermella ('red flag')	Massa provocou unha *bandeira vermella* logo de chocar contra as barreiras na curva 3.	Massa caused a *red flag* after crashing into the barriers at Turn 3.
Idiom.	**sangue frío** ('cold blood')	A miña tía María recuperou o seu *sangue frío* e contestoulle con certa sequidade.	My aunt María regained *her composure* and answered him with certain dryness.
	vida útil ('useful life')	Aínda así, un uso prolongado do óxido nitroso pode acabar danando motor e acurtando a súa *vida útil*.	Even then, a prolonged use of nitrous oxide can end up damaging the engine and shortening its *service life*.
	bandeira vermella ('red flag')	Na lista de verificación de relacións emocionalmente abusivas, a manipulación é unha das *bandeiras vermellas* destacadas.	On the checklist of emotionally abusive relationships, manipulation is one of the prominent *red flags*.

Table 5: Examples of Compositional, Partial, and Idiomatic multiword expressions and corpora-extracted sentences contained in the dataset. Note that some of them, e.g., *bandeira vermella* are Potentially Idiomatic Expressions, with different compositionality scores depending on the context.