Diversity patterns run deep: Impact of diversity intake on multiword expression identification

Mathilde Deletombe, Manon Scholivet, Louis Estève, Thomas Lavergne, Agata Savary
Université Paris-Saclay, CNRS, LISN
first.last@lisn.fr

Abstract

Multiword expressions (MWEs) are good examples of a phenomenon where identification systems struggle with generalisation: MWE present in the test set but absent in the training set are rarely identified. This raises the question of the diversity of the test set, relative to that of the train set, and how this impacts performance. We set out to measure how much diversity of a train corpus increases when adding individual MWEs from the test corpus, and how this increase impacts MWE identification performance. We measure diversity across a three-dimension framework and find mostly consistent negative correlations with performance in 14 languages and 8 systems.

1 Introduction

Multiword expression (MWEs), such as *to **pay** a **visit***, *to **take off*** or *to **call it a day***, have been an object of interest and a major challenge in Natural Language Processing (NLP) for decades (Sag et al., 2002; Shwartz and Dagan, 2019), notably due to their prevalence in texts (Candito et al., 2021) and their semantic non-compositionality (Nandakumar et al., 2018; Cordeiro et al., 2019; Miletić and Schulte im Walde, 2025). MWE-related tasks defined by the NLP community include MWE identification in running text (Constant et al., 2017).

This task has received attention in the past decade, notably due to shared tasks such as DiMSUM for English (Schneider et al., 2016), and the PARSEME shared task on automatic identification of verbal MWEs (VMWEs) in up to 20 languages, with its three editions: 1.0 (Savary et al., 2017), 1.1 (Ramisch et al., 2018), and 1.2 (Ramisch et al., 2020). Edition 1.2, building on the findings from edition 1.1, introduced a focus on *unseen* VMWEs. A VMWE from the test corpus is considered seen if another VMWE with the same multiset of lemmas is annotated at least once in the train or the development corpus. Otherwise it is considered unseen. Ramisch et al. (2020) showed that the performances of VMWE identification systems more strongly (inversely) correlate with the number of unseen VMWEs than with the size of the train corpus. Savary et al. (2019) argued that this is due to the very nature of the MWE phenomenon and its distributional properties.

The number of MWEs seen in the test but not in the train can be interpreted as the lack of MWE diversity in the train, relative to the test. But diversity has many facets (Stirling, 1994, 2007; Ramaciotti Morales et al., 2021; Estève et al., 2025) and can refer not only to the number of categories (*variety*) but also to the evenness of their distribution (*balance*) and to their relative differences (*disparity*). Given that unseen data had such a predominant impact on system performance in PARSEME 1.2 shared task, we wish to examine how far these observations can be generalised to more widely understood diversity aspects. We introduce the notion of *train/test diversity intake*, or *diversity intake* for short, to denote the diversity that the test corpus adds to the train corpus.[1] In other words, we are interested in relative rather than absolute diversity quantification, as defined by Estève et al. (2025).

We address two research questions:

RQ1 How to estimate the train/test diversity intake?
RQ2 Does this intake correlate with performance in the MWE identification task?

To address RQ1, we take inspiration from interdisciplinary work on diversity, where this notion has been thoroughly conceptualised. We select three diversity indicators: richness delta, negated Zipfian curvature delta, and minimum tree edit distance. To tackle RQ2, we use the PARSEME 1.2 shared task corpora and system predictions. For the MWEs from a test corpus, we measure their

[1] For the sake of brevity, we consider that the development corpus (if any) is part of the train corpus.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 110–116

individual share in the diversity intake. We then calculate the correlation between this share and the fact of being correctly or wrongly predicted by a system. Our hypothesis is that diversity intake and performance are inversely correlated.

Data, codes and results of our experiments are openly available.[2]

2 PARSEME data

To examine how diversity intake correlates with performance, we use the open source corpora and system predictions from the PARSEME shared task 1.2.[3] The corpora cover 14 languages: Basque (EU), Chinese (ZH), French (FR), German (DE), Greek (EL), Hebrew (HE), Hindi (HI), Irish (GA), Italian (IT), Polish (PL), Brazilian Portuguese (PT), Romanian (RO), Swedish (SV), and Turkish (TR). Their sizes range from 35 thousand to over 1 million tokens per language, with 1 thousand to 9 thousand manually-annotated VMWEs.[4] They also include UD-style[5] morphosyntactic annotations.

We use the predictions of 8 out of 9 systems participating in the shared task.[6] MTLB-STRUCT, TRAVIS-multi and TRAVIS-mono were based on BERT-finetuning; ERMI and MultiVitamin used simpler neural networks; HMSid and Seen2Unseen applied association measures; Seen2Seen and FipsCo were rule-based.

3 Diversity measures

For diversity quantification, we use the conceptual framework defined by Stirling (1994, 2007) to unify previous work in several scientific fields, most prominently ecology (Ricotta and Szeidl, 2006; Leinster and Cobbold, 2012; Scheiner, 2012; Chao et al., 2014; Chao and Ricotta, 2019). This framework has been recently applied in NLP (Estève et al., 2025), and to MWEs in particular (Lion-Bouton et al., 2022). It assumes that diversity is a property of sets whose *elements* can be aportioned into *categories*. Like Lion-Bouton et al. (2022), we define categories as VMWE canonical forms represented by multisets of lemmas of their components. For instance, for the MWE *to* ***call a spade a spade***, the multiset of lemmas is {*a*, *a*, *call*, *spade*, *spade*}. Elements are occurrences of these MWE canonical forms.[7]

Given the category/element dychotomy, diversity can be characterized along three dimensions (Stirling, 2007): *variety*, *balance*, and *disparity*. All other things being equal, the higher the variety the higher the diversity of a set. The same holds for balance and disparity.

Variety relates to the number of categories. A simple and widely used variety measure is *richness*, i.e. simply the number of categories, and we will use it for estimating variety intake.

Balance relates to the evenness of the distribution of the elements in categories. Balance reaches its optimum when the distribution is perfectly uniform. In fields like ecology, the distribution of categories (e.g. species) is often hard to estimate reliably, and then so-called non-parametric diversity measures, like Shannon evenness (Smith and Wilson, 1996), are used. But if a particular distribution can be assumed, so-called parametric measures apply (Magurran, 2004). In NLP, Zipfian distributions are frequently encountered and apply to MWEs (Ryland Williams et al., 2015). A Zipfian distribution is characterised by the probability mass function $Z_{s,n}(i) = i^{-s}\left(\sum_{j=1}^{n} j^{-s}\right)^{-1}$, where, in our case, n is the number of VMWE categories, i is the rank of the i's most frequent VMWE category, and s is the exponent characterizing the curvature of the distribution. When $s = 0$, the distribution is uniform, and the higher s, the more curved (more unbalanced) the distribution is. Therefore, the opposite of curvature, i.e. $-s$, can be considered a measure of balance (Zhang et al., 2023).

Disparity reflects the extent to which categories are different from each other, which calls for an appropriate distance measure between categories. Recently, various disparity measures using semantic vector spaces have been used in NLP (Yang et al., 2024; Yu et al., 2022; Puranik et al., 2023; E et al., 2023; Kim et al., 2023; Cao and Wan, 2020) but it was also shown that such measures strongly correlate with variety, in particular when MWE are concerned (Estève et al., 2024). Therefore, in-

[2] https://gitlab.lisn.upsaclay.fr/deletombe/repo

[3] https://gitlab.com/parseme/sharedtask-data

[4] The annotation follows unified guidelines with a VMWE taxonomy including verbal idioms (***go bananas***), light verb constructions (***pay a visit***, ***grants rights***), inherently reflexive verbs (***help oneself***), verb-particle construction (***do in***), multi-verb constructions (***let go***) and inherently adpositional verbs (***rely on***).

[5] https://universaldependencies.org/format.html

[6] The 9th system, MultiVitaminBooster, had per-language scores below 1% F-measure.

[7] Note that a non-idiomatic co-occurrence of a multiset of lemmas does not count as an element, e.g. in *she called this thing a spade but a spade is something else*.

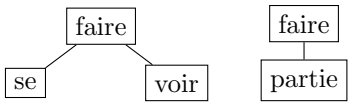

Figure 1: Syntactic trees of two French VMWEs: ***se faire voir*** (lit. 'make oneself see') 'show oneself in a fancy place' and ***faire partie*** (lit. 'make part') 'belong'.

spired by Guo et al. (2024), we endorse interest in syntactic (and partly semantic) diversity.

As the distance underlying our disparity measure, we use the *tree edit distance* by Zhang and Shasha (1989) between two VMWEs seen as simplified syntactic dependency subtrees, where only the lemmas and the head-dependent relations are represented, as in Figure 1. For such trees, we consider three elementary edit operations, allowed both for leaves and internal nodes, in any order: (i) deletion of a node, (ii) insertion of a node, (iii) replacement of the lemma of a node by another lemma. Operations (i) and (ii) each cost 1. The cost of (iii) is half of the cosine distance between the vectors representing the original lemma and the replacement lemma (i.e. its range is $[0, 1]$). We use the Word2Vec-style (Mikolov et al., 2013) vector spaces by Estève et al. (2024),[8] trained on the PARSEME corpus edition 1.3.[9] The tree edit distance is the cost of a minimal sequence of elementary edit operations transforming one tree into another. For instance, the edit distance between the two trees in Figure 1 is 1.282: 1 for the deletion of *se* 'oneself' and $0.564/2$ for the replacement of *voir* 'see' by *partie* 'part'.

4 Diversity intake

To address RQ1, we estimate diversity intake (DI), for each VMWE individually, along the three dimensions of diversity. For each diversity dimension, the DI of the test corpus is represented by a vector, comprising all individual DIs. For a given language L, let TRAIN and TEST be its train and test corpora. Let $E_{\text{TEST}} = (e_1, \ldots, e_n)$ be the list of the VMWE categories from TEST. Consider the toy example of TRAIN (1) and TEST (2) in French. Here, E_{TEST} =({*faire* 'do', *partie* 'part'}, {*faire* 'do', *se* 'oneself', *voir* 'see'}, {*faire* 'do', *sembler* 'seem'}).

(1) Il **s'agissait** de **faire partie** du show, donc il en **faisait partie**.
It was about taking part in the show, so he took part in it.

(2) Même s'il n'en **faisait** pas **partie**, il s'y **ferait voir** et **ferait semblant**.
Event if he didn't take part in it, we would show up and pretend.

We define the *variety intake* $DI_v(e_i)$ to be 1 if e_i is absent from TRAIN (i.e. it adds to TRAIN's variety), and 0 otherwise. Then the DI for the whole TEST is $DI_v = (DI_v(e_1), \ldots, DI_v(e_n))$. In (2) we have $DI_v = (0, 1, 1)$.

To calculate the *balance intake* $DI_b(e_i)$, we add e_i to the set of VMWEs from TRAIN and recalculate its Zipfian curvature s. The difference between $-s$ in TRAIN with and without e_i is the value of $DI_b(e_i)$. In (2), adding e_2 or e_3 to TRAIN flattens the curvature but adding e_1 increases the frequency of the most frequent category. Therefore, the DI_b vector has positive values at positions 2 and 3 and a negative one at position 1.

The *disparity intake* follows a slightly different logic than variety and balance intake. The idea is that a system might correctly identify e_i on the basis of a VMWE from TRAIN which is similar, even if not identical, to e_i. Therefore, $DI_d(e_i)$ is defined as the minimum edit distance between e_i and any VMWE in TRAIN. In (2), we have $DI_d \approx (0.00, 1.11, 0.16)$ respectively for **faire partie** / **faire partie** (identical means 0 distance), **s'agit** / **se faire voir** (0 distance between **se** and **se**, ≈ 0.11 between **agir** and **faire**, 1 to add **voir**), and **faire partie** / **faire semblant** (0 distance between **faire** and **faire**, ≈ 0.16 between **partie** and **semblant**).

To account for performance of system S on expression e_i, we define $Perf_S(e_i)$ to be 1 if S has correctly identified e_i and 0 otherwise. Then $Perf_S = (Perf_S(e_1), \ldots, Perf_S(e_n))$. In (2), if only the first two expressions are true positives, then $Perf_S = (1, 1, 0)$.

To address RQ2, in each language we calculate the diversity intake vectors, DI_v, DI_b and DI_d. We then measure the Pearson correlation between each of them and the performance vector $Perf_S$, for each system S. The results are described in the following section.

[8] This semantic space represents both single words and VMWEs. Here, we only use vectors for the former.

[9] Edition 1.3 contains consolidated versions of the VMWE-annotated corpora from 3 shared task editions in 26 languages.

Table 1: Pearson correlation measurement between variety/balance/disparity intake and performance

		DE	EL	EU	FR	GA	HE	HI	IT	PL	PT	RO	SV	TR	ZH
Variety	ERMI	−0.46	−0.47	−0.47	−0.43	−0.50	−0.61	−0.48	−0.46	−0.51	−0.38	−0.60	−0.53	−0.43	−0.44
	FipsCo	−0.18	−0.28		−0.33										
	HMSid				−0.26										
	MTLB-STRUCT	−0.53	−0.52	−0.56	−0.54	−0.48	−0.74	−0.42	−0.61	−0.58	−0.47	−0.63	−0.52	−0.44	−0.39
	Seen2Seen	**−0.89**	**−0.87**	**−0.86**	**−0.90**	**−0.74**	**−0.87**	**−0.75**	**−0.88**	**−0.94**	**−0.89**	**−0.65**	**−0.85**	**−0.88**	**−0.89**
	Seen2Unseen	−0.86	−0.82	−0.80	−0.78	−0.58	**−0.87**	−0.39	−0.84	−0.89	−0.81	−0.63	−0.81	−0.82	**−0.89**
	TRAVIS-mono	−0.41	−0.14		−0.49			−0.30	−0.48	−0.54		−0.52	−0.45	−0.38	−0.31
	TRAVIS-multi	−0.46	−0.54	−0.44	−0.55	−0.22	−0.67	−0.39	−0.53	−0.54		−0.47	−0.49	−0.44	−0.40
Balance	ERMI	**−0.31**	−0.31	**−0.35**	**−0.34**	−0.49	**−0.56**	−0.32	**−0.48**	**−0.39**	−0.35	−0.19	**−0.48**	**−0.36**	−0.32
	FipsCo	0.16	−0.07		−0.31										
	HMSid				−0.21										
	MTLB-STRUCT	−0.21	−0.23	−0.28	−0.25	−0.42	−0.43	−0.26	−0.35	−0.31	−0.29	−0.15	−0.40	−0.31	−0.23
	Seen2Seen	−0.27	−0.31	**−0.35**	−0.31	**−0.65**	−0.51	**−0.53**	−0.34	−0.36	**−0.36**	**−0.22**	**−0.48**	**−0.36**	**−0.33**
	Seen2Unseen	−0.27	−0.29	−0.32	−0.28	−0.54	−0.50	−0.36	−0.33	−0.34	−0.34	**−0.22**	−0.47	−0.34	**−0.33**
	TRAVIS-mono	−0.21	**−0.44**		−0.24			−0.50	−0.34	−0.28		−0.16	−0.40	−0.29	−0.21
	TRAVIS-multi	−0.23	−0.25	−0.31	−0.28	−0.34	−0.47	−0.33	−0.36	−0.32		**−0.22**	−0.42	−0.29	−0.25
Disparity	ERMI	−0.33	**−0.26**	**−0.33**	−0.26	−0.19	−0.23	−0.38	**−0.39**	−0.41	**−0.35**	−0.26	−0.27	**−0.27**	−0.02
	FipsCo	−0.34	−0.13		**−0.32**										
	HMSid				−0.29										
	MTLB-STRUCT	**−0.36**	−0.23	−0.32	−0.22	**−0.24**	−0.27	−0.44	**−0.39**	−0.41	−0.34	−0.29	−0.28	−0.24	**−0.09**
	Seen2Seen	**−0.36**	−0.21	−0.21	−0.21	−0.11	−0.23	−0.30	−0.35	−0.35	−0.27	−0.12	−0.25	−0.19	**−0.09**
	Seen2Unseen	**−0.36**	−0.22	−0.23	−0.24	−0.15	−0.24	−0.36	−0.36	−0.36	−0.29	−0.11	−0.25	−0.23	−0.08
	TRAVIS-mono	−0.32	−0.02		−0.21			−0.15	−0.37	−0.35		**−0.33**	−0.25	−0.22	−0.05
	TRAVIS-multi	−0.35	−0.21	**−0.33**	−0.23	−0.15	**−0.30**	**−0.46**	−0.36	**−0.42**		−0.28	**−0.32**	−0.26	−0.06

5 Results

The results are shown in Table 1, with the strongest correlation for each language highlighted in bold. All results, except for TRAVIS-multi in ZH, ERMI for the same langage and TRAVIS-mono for EL, are statistically significant with threshold 0.05. We observe mostly negative correlation (with only three exceptions), which suggests that higher train/test diversity intake is associated with weaker system performance, and vice versa, which corroborates our hypothesis. Negative correlation is considered (i) strong, (ii) moderate and (iii) weak, if the scores fall (i) below -0.7 or above 0.7, (ii) from -0.7 to -0.3 or from 0.3 to 0.7 and (iii) from -0.3 to 0.3.

Regarding variety, a strong negative correlation with system performance is noticeable. Out of the 83 scores, the majority consists of moderate (53) and strong (25) correlations. The Seen2Seen and Seen2Unseen systems display the strongest negative correlation, reaching -0.94 for Seen2Seen in PL and -0.89 for Seen2Unseen in ZH and PL. This is expected, given that these systems focus on the MWEs seen in train. Conversely, FipsCo relies of external MWE lexicons, which is consistent with its relatively weak correlation with variety intake.

For balance intake we observe a weaker but still non-negligible negative correlation, with 52 moderate and 31 weak scores. The highest scores are shown for Seen2Seen (reaching -0.65 in GA) and ERMI (reaching -0.56 in HE). This might be partly due to the fact that unseen VMWEs, when added to TRAIN systematically increase both its variety and balance. This might strongly influence the systems like Seen2Seen and ERMI which use no external data (e.g. lexicons, pre-trained models).

As to disparity intake, a majority of correlations (50) are weak, which indicates that disparity intake has little or no impact on performance. This might mean that systems hardly capture syntactic and semantic similarities between VMWEs. There are only 33 moderate correlations, notably for TRAVIS-multi in HI and PL (-0.46 and -0.42), ERMI in PL (-0.41), and MTLB-STRUCT in HI and PL (-0.44 and -0.41).

6 Conclusions and future work

In this paper, we investigated the correlation between train/test diversity intake and the performance of VMWE identification systems. We confirmed prior findings from the PARSEME shared tasks showing a strong relationship between system performance and the rate of unseen VMWEs, which we re-interpreted as variety intake.

We extended previous findings to the two other dimensions of diversity — balance and disparity — by studying whether similar correlations could be observed. Balance intake, quantified through the curvature of the Zipfian distribution, was found to exhibit a moderate correlation with system performance. In contrast, disparity intake, modelled using a novel approach based on tree edit distance,

showed only weak correlation with system performance in our experiments.

In the future, we intend to apply the methods presented here to the corpora and system predictions from the latest edition 2.0 of the PARSEME shared task (Scholivet et al., 2026). We also wish to study alternative ways to measure diversity that may better correlate with performance. Such measures may help evaluate the systems but also improve their training and contribute to the creation of more balanced datasets.

7 Acknowledgements

This work received support from: (i) COST (European Cooperation in Science and Technology) through the CA21167 COST action UniDive, (ii) the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01), (iii) the "*Plan Blanc*" (White Plan) doctoral funding from Université Paris-Saclay (France).

8 Limitations

This work uses vector spaces to quantify dissimilarity between words, which, as any vector space trained on real-world data, cannot account for all possible structures and variations. The quality of the vector spaces, and subsequent experiments relying on them, while reasonable, cannot be a complete and perfect representation of these phenomena at work.

Diversity quantification has a very rich bibliography and many other diversity measures exist which could be applied in our context. More thorough criteria for selecting the most accurate measures are needed.

References

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2021. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2).

Yue Cao and Xiaojun Wan. 2020. DivGAN: Towards diverse paraphrase generation via diversified generative adversarial network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2411–2421, Online. Association for Computational Linguistics.

Anne Chao, Chun-Huo Chiu, and Lou Jost. 2014. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45:297–324. Publisher: Annual Reviews.

Anne Chao and Carlo Ricotta. 2019. Quantifying evenness and linking it to diversity, beta diversity, and similarity. *Ecology*, 100(12):e02852. Number: 12.

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.

Silvio Ricardo Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57. Impact Factor: 1.319. http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00341.

Venkatesh E, Kaushal Maurya, Deepak Kumar, and Maunendra Sankar Desarkar. 2023. DivHSK: Diverse headline generation using self-attention based keyword selection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1879–1891, Toronto, Canada. Association for Computational Linguistics.

Louis Estève, Agata Savary, and Thomas Lavergne. 2024. Vector spaces for quantifying disparity of multiword expressions in annotated text. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 110–130, Bangkok, Thailand. Association for Computational Linguistics.

Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik, Agata Savary, and Olha Kanishcheva. 2025. A survey of diversity quantification in natural language processing: The why, what, where and how. *Preprint*, arXiv:2507.20858.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-Ho Lee, Jeonguk Kim, Donghoon Shin, and Yeonsoo Lee. 2023. Persona expansion with commonsense knowledge for diverse and consistent response generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1139–1149, Dubrovnik, Croatia. Association for Computational Linguistics.

Tom Leinster and Christina A. Cobbold. 2012. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489. Number: 3.

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne E. Magurran. 2004. *Measuring biological diversity*. Oxford: Blackwell Publishing Company , 2004, Oxford.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Filip Miletić and Sabine Schulte im Walde. 2025. Modeling the evolution of English noun compounds with feature-rich diachronic compositionality prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20071–20092, Vienna, Austria. Association for Computational Linguistics.

Navnita Nandakumar, Bahar Salehi, and Timothy Baldwin. 2018. A comparative study of embedding models in predicting the compositionality of multiword expressions. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 71–76, Dunedin, New Zealand.

Vinayak Puranik, Anirban Majumder, and Vineet Chaoji. 2023. PROTEGE: Prompt-based diverse question generation from web articles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5449–5463, Singapore. Association for Computational Linguistics.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphaël Fournier-S'Niehotta, Rémy Poulain, Lionel Tabourier, and Fabien Tarissan. 2021. Measuring diversity in heterogeneous information networks. *Theoretical Computer Science*, 859:80–115. Publisher: Elsevier.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, and 6 others. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Carlo Ricotta and Laszlo Szeidl. 2006. Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology*, 70(3):237–243. Number: 3.

Jake Ryland Williams, Paul R. Lessard, Suma Desu, Eric M. Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Zipf's law holds for phrases, not words. *Scientific Reports*, 5.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expresions: A Pain in the Neck for NLP. In *Proceedings of CICLING'02*. Springer.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Samuel M. Scheiner. 2012. A metric of biodiversity that integrates abundance, phylogeny, and function. *Oikos*, 121(8):1191–1202. Number: 8.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Păiș. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Benjamin Smith and J. Bastow Wilson. 1996. A Consumer's Guide to Evenness Indices. *Oikos*, 76(1):70–82. Number: 1 Publisher: [Nordic Society Oikos, Wiley].

Andrew Stirling. 1994. Diversity and ignorance in electricity supply investment. *Energy Policy*, 22(3):195–216.

Andy Stirling. 2007. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719. Number: 15 Publisher: Royal Society.

Yuting Yang, Pei Huang, Feifei Ma, Juan Cao, and Jintao Li. 2024. PAD: A robustness enhancement ensemble method via promoting attention diversity. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12574–12584, Torino, Italia. ELRA and ICCL.

Yu Yu, Shahram Khadivi, and Jia Xu. 2022. Can data diversity enhance learning generalization? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4933–4945, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18(6):1245–1262.

Xinran Zhang, Maosong Sun, Jiafeng Liu, and Xiaobing Li. 2023. Lingxi: A diversity-aware Chinese modern poetry generation system. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 63–75, Toronto, Canada. Association for Computational Linguistics.