Vector Spaces for Quantifying Disparity of Multiword Expressions in Annotated Text

Louis Estève[†] **Agata Savary**[†] **Thomas Lavergne**[†]
Paris-Saclay University – LISN – CNRS – France[†]
firstname.lastname@universite-paris-saclay.fr

Abstract

Multiword Expressions (MWEs) make a good case study for linguistic diversity due to their idiosyncratic nature. Defining MWE canonical forms as types, diversity may be measured notably through disparity, based on pairwise distances between types. To this aim, we train static MWE-aware word embeddings for verbal MWEs in 14 languages, and we show interesting properties of these vector spaces. We use these vector spaces to implement the so-called functional diversity measure. We apply this measure to the results of several MWE identification systems. We find that, although MWE vector spaces are meaningful at a local scale, the disparity measure aggregating them at a global scale strongly correlates with the number of types, which questions its usefulness in presence of simpler diversity metrics such as variety. We make the vector spaces we generated available.

Keywords: diversity, disparity, multiword expression, vector space

1 Context of study

Multiword Expressions (MWEs) are characterized by idiosyncracy, *i.e.* behavior specific to few individuals (Baldwin and Kim, 2010). They are thus an interesting case of study for linguistic diversity.

Linguistic diversity has been formally modelled mainly with respect to the variety of the existing human languages and the populations speaking them (Joshi et al., 2020). The diversity of language utterances has been much less often addressed. In particular, with respect to MWEs, one may wonder if a corpus or set of system predictions for MWEs is diverse or not. Once items and types are defined,[1] diversity may be studied through variety (*i.e.*, how many types there are), balance (*i.e.*, how evenly distributed types are), and disparity (*i.e.*, how disparate or fundamentally different types are), as described by Morales et al. (2020). Recent work has studied variety and balance in the case of the PARSEME corpus of verbal MWEs (VMWEs), specifically on the system predictions of the related shared task (Lion-Bouton et al., 2022). Disparity however has not been studied in this context.

In this study we bridge this gap by quantifying disparity of the PARSEME shared task system predictions with a measure called functional diversity, from ecology. Since disparity builds upon the underlying definition of distance between types, we construct VMWE-aware vector spaces (VS). We choose static word embeddings since they proved particularly efficient in type-oriented MWE tasks such as compositionality degree prediction. We set the following research questions:

- R1 What are the properties of VMWE VSs constructed with state-of-the-art methods, across many languages?
- R2 Can these vector spaces be useful when quantifying diversity and VMWEs, using formal diversity measures?

Our ultimate aim is to test how useful disparity can be to evaluate the quality of NLP resources along dimensions which would be orthogonal to efficiency (assessed *e.g.* by F-measure or accuracy).

The paper is organised as follows. After discussing the quantification of diversity (§2), as well as the related works (§3), we present our approach (§4), discuss the generated VSs (§5), describe the disparity function we use (§6), discuss system diversities (§7), and conclude (§8).

2 Quantifying diversity

One may argue that, in NLP, many situations can benefit from having a higher diversity, the main

[1] A type is a group of items with a shared identity; this requires a choice relative to the research objective, but a default choice would be individual MWE instances as items, and their canonical form as types.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 204–224

example being the quality of the training set for its impact on system quality (Guo et al., 2023; Yu et al., 2022). Thus diversity is often desirable and there is a research objective of formally measuring it. More precisely, there is an interest in measuring the diversity of specific linguistic phenomena in corpora. Diversity can be understood through **variety**, **balance**, and **disparity** (Morales et al., 2020; Lion-Bouton et al., 2022). To understand these three aspects, let us consider the two following examples that tackle specifically the diversity of VMWEs.[2]

Example 1 "*I just got of [1] the phone with Hai and he told me how to make [2a] an adjustement [2a] on a day to day basis [...] and P&L would still somehow work out [3] because adjustments [2b] would be made [2b].*", (typos from original text)

Example 2 "*Does this mean that for June [...] we should not do anything and just make adjustments [1] on a going forward [2] basis (and assume everything will work out [3] at month end)?*"

In these examples, **items** (*i.e.*, individual instances) are underlined. The first example contains 4 items, while the second example contains 3. However, items may be clustered into **types** based on some shared identity, such as *make [...] adjustment* 'to make an adjustment' and *adjustments [...] made* 'to make an adjustment' in the first example. Both examples thus contain 3 types: *to* ***get off***, *to* ***make*** *an* ***adjustment***, and *to* ***work out*** for the first example, *to* ***make*** *an* ***adjustment***, *to* ***go forward***, and *to* ***work out*** for the second example.

Diversity is measured on types. Variety concerns itself with the number of types; as both examples have 3 types, they are equally varied. Balance concerns itself with the evenness in the distribution of types; as the first example has a type with more items than others, it is less balanced than the second example in which every type has the same number of items. Disparity concerns itself with the fundamental differences between types; as two types are shared between the two examples (*to* ***make*** *an* ***adjustment*** and *to* ***work out***), the question here is which of *to* ***get off*** and *to* ***go forward*** is more different (or, phrased otherwise, more distant) from the shared types.

Variety, balance, and disparity are general dimensions: a number of concrete measures exists for each (Smith and Wilson, 1996; Chao et al., 2014).

[2]This is a small-scale demonstration, in practice diversity would be computed on much larger datasets.

Variety is often trivial, as it concerns itself with the number of types, such as richness n (Lion-Bouton et al., 2022) or species count $n-1$ (Patil and Taillie, 1982).

Balance often consists of entropies such as Shannon-Weaver entropy

$$H = -\sum_{i=1}^{n} p_i \log_b (p_i) \quad (1)$$

where p_i denotes the relative proportion of the ith type. Parametric entropies, as described by Rényi (1961), Patil and Taillie (1982), or Good (1953) are also in use. Patil and Taillie entropy covers species count ($\alpha = -1$), Shannon-Weaver entropy ($\alpha = 0$), and the Simpson index ($\alpha = 1$). Good entropy covers richness ($\alpha = 0, \beta = 0$), Shannon-weaver entropy ($\alpha = 1, \beta = 1$), and the Simpson dominance index ($\alpha = 2, \beta = 0$). Rényi entropy is used to generate Hill (1973) numbers; given n types, and a parametric entropy H_α, the corresponding (standard) Hill number is the number of types $\hat{n}$ that a perfectly evenly distributed population needs in order to have the same entropy $\hat{H}_\alpha = H_\alpha$ (Rényi entropy if based on the original work of Hill (1973), but Patil and Taillie (1982) show it is also possible with their entropy). Hill numbers are used a lot in ecology for reasons that go beyond the scope of this paper; we invite interested readers to refer to Chao et al. (2014).

Disparity is the most complex of the triad, as it often requires setting up a VS along with a distance function between types. Disparity functions include: Chao et al. entropy and Hill number (Chao et al., 2014), Leinster-Cobbold entropy and Hill number (Leinster and Cobbold, 2012), Ricotta-Szeidl entropy (Ricotta and Szeidl, 2006), Scheiner entropy and Hill number (Scheiner, 2012), functional dispersion (Laliberté and Legendre, 2010), functional evenness, functional divergence (Villéger et al., 2008), lexicographic approach (Bossert et al., 2001), order-weighted and proportion-weighted disparity (Stirling, 2007), FAD or MFAD pairwise distances (Mouchet et al., 2010). However, as this is an early work investigating the use of disparity in linguistics, we will select one disparity function in dedicated section (§6).

3 Related works: MWE vector spaces

Distributional semantic models represent text units as vectors of real numbers in a multidimensional

space. Vector representations for MWEs in particular can be obtained from word co-occurrence matrices after dimensionality reduction (Schulte im Walde et al., 2013) or neural networks trained by self-supervision (Mikolov et al., 2013; Devlin et al., 2019). In the latter case, the vectors, called embeddings, can be trained on the level of characters, words or documents, and can be static (notably Word2Vec) or contextual (most often obtained with transformers). Static MWE-aware word embeddings (WEs), on the one hand, require a corpus which is re-tokenized so that all occurrences of MWEs (and of other phrases of interest) are merged into single tokens (Salehi et al., 2015; Cordeiro et al., 2019; Otani et al., 2020). Cross-lingual embeddings can also be obtained by aligning monolingual MWE-aware static WEs (Otani et al., 2020). A contextual embedding of an MWE, on the other hand, can be obtained straightforwardly from generic transformer models (trained on a corpus with no MWE-aware tokenisation) by combining the vectors for (sub)tokens occurring in the MWE in a precise context (Nandakumar et al., 2018; Kanclerz and Piasecki, 2022). This eliminates the requirement of having identified MWEs in advance in the training corpus. Nevertheless, Hashempour and Villavicencio (2020) show that merging MWEs into single tokens in the train corpus enhances performances of in MWE-related tasks, also with contextual embeddings.

One of the parameters for training MWE-aware embeddings is the method used to identify MWEs in the train corpus, prior to their fusion into single tokens. In the simplest case, a handcrafted controlled list of phrases (including MWEs), possibly lemmatized, is straightforwardly matched against the corpus.[3] Most of the compositionality prediction experiments cited below, as well as Salehi et al. (2014) and Otani et al. (2020), use this technique. The embeddings for MWEs are then available only for the MWEs from the controlled list. In a more elaborate case, a generic MWE identifier is used to tag MWEs in a large raw corpus. In this case precision may be preferred over recall by favoring MWEs seen in the training corpus.

Embeddings have been efficiently used in MWE-specific NLP tasks, most notably in automatic prediction of the degree of compositionality of a MWE. The hypothesis here is that this degree coincides with the distance between the vector representing the whole MWE and the combination of the vectors of its components (or of its synonyms and paraphrases). This principle was applied to 2-word noun phrases (***ivory tower***) in English (Salehi et al., 2015; Cordeiro et al., 2019), French and Portuguese (Cordeiro et al., 2019). Verb-particle constructions (***set off***) were also approached in this way in English (Hakimi Parizi and Cook, 2018) and in German (Köper and Schulte im Walde, 2017). More recent work by Sarlak et al. (2023) on Persian, a low-resourced language, extends this idea to various VMWEs, which are harder to model due to their morphosyntactic variability (Constant et al., 2017). Static WEs for MWEs were also successfully combined with embeddings representing hypernymy relations (Jana et al., 2019) and multimodal text-image associations (Köper and Schulte im Walde, 2017). Interestingly, "simple" Word2Vec embeddings are reported by a number of authors (Cordeiro et al., 2019; Nandakumar et al., 2018; Sarlak et al., 2023) as outperforming more elaborate contextual WEs in this precise task.

Another MWE-specific task is machine translation of MWEs. A MWE in the source language can be translated by selecting the closest, in terms of (static) cross-lingual WEs, target language word or MWE. This technique proved efficient for 10 typologically different languages in (Otani et al., 2020). But more recent MWE-specialized translation engines rely on transformers, fine-tuned on parallel MWE datasets (Santing et al., 2022) or pre-trained on monolingual idiom corpora (Baziotis et al., 2023).

Yet another task, MWE disambiguation, consists in distinguishing literal and idiomatic occurrences of a potential idiomatic expression (PIE), like *to* ***take the cake*** 'be the most remarkable of its kind'. Systems were developed notably in English, German, Portuguese, Galician and Japanese. While static WEs proved useful (Ehren, 2017), contextual WEs occurred more efficient (Hashempour and Villavicencio, 2020).[4] Thus, recent best performing methods rely on pre-trained transformer models, either frozen or fine-tuned, to generate contextual phrase or sentence embeddings prior to binary classification (Kurfalı and Östling, 2020; Fakharian and Cook, 2021; Madabushi et al., 2022;

[3] While such a method suffers from not being able to distinguish between literal/coincidental and idiomatic occurrences of MWEs, this is a minor problem due to the very low frequency of literal readings in general (Savary et al., 2019).

[4] Interestingly, Hashempour and Villavicencio (2020) show that Context2Vec representations obtained from LSTMs outperform those from BERT.

Takahashi et al., 2022).

It is also worth noting that generic static embeddings (trained with no particular attention paid to MWEs) proved useful to model the compositional/literal meanings of MWEs in tasks such as translating MWEs and collocations (Gamallo and Garcia, 2019), detecting synonyms of terminological MWEs (Hazem and Daille, 2018), and MWE identification (Zeng and Bhat, 2021).

To sum up, while contextual representations of MWEs and their contexts outperform static WEs in tasks focusing on MWE occurrences (disambiguation, translation and identification), those concerning types (compositionality prediction) seem to be solved more efficiently with static MWEs.

4 Overview of our approach

We address the task of VMWE identification with a novel perspective on evaluation: quantifying the diversity of VMWEs in annotated text. While Lion-Bouton et al. (2022) address variety and balance of annotated VMWEs, they do not cover disparity. Here, we bridge this gap by using a disparity measure to assess how diverse the types of VMWEs found in annotated text are. This requires a measure of distance between types, and we propose to define it in terms of distance between VMWE embeddings. Since the task is type-oriented, we use static VMWE-aware word embeddings, as suggested by the above SOA. To this aim:

1. We train state-of-the-art VMWE identifiers on the latest version of the PARSEME corpus (Savary et al., 2023) annotated for verbal VMWEs in 14 languages.

2. We use these identifiers to annotate a large raw multilingual corpus.

3. We re-tokenize the corpus so as to merge VMWEs into single tokens, and use it to train Word2Vec embeddings in all 14 languages. We examine interesting properties of the resulting semantic spaces in selected languages.

4. We experiment with disparity measurement and we find that disparity strongly correlates with the number of types, which suggests that disparity measures may be superfluous in presence of simpler and less computationally intensive measures such as richness. This is an interesting negative result allowing to simplify diversity measurement, at least for VMWE annotations and distances modelled in VSs.

5 Vector spaces

This section describes steps 1 through 3 of the above overview (§4).

5.1 Data and VMWE identifiers

The PARSEME corpus (Savary et al., 2023), used in the eponym shared tasks, is a multilingual resource comprising 26 languages as of version 1.3. It is focused on Verbal Multiword Expressions (VMWEs) and assigns them categories.[5]

In edition 1.2, the PARSEME corpus covers 14 languages, with manually annotated VMWEs and manually or automatically annotated lemmas and morphosyntax. Additionally, for the same 14 languages, large companion corpora (called "raw corpora") of 450GB in total, automatically annotated for lemmas and morphosyntax (in the `.conllu` format) but not for VMWEs, were released in this edition, with the objective of facilitating unsupervised discovery of new VMWEs.

PARSEME also organised 3 shared tasks on automatic identification of VMWEs. The latest edition used the 1.2 version of the corpus. The systems submitted to the PARSEME shared task 1.2 are described by Ramisch et al. (2020). Their predictions are also publicly available, which allows us to calculate their diversity, as done later in Table 4.

Additionally, the two best-scoring systems of the shared task 1.2, Seen2Seen (Pasquer et al., 2020) and MTLB-STRUCT (Taslimipoor et al., 2020), respectively 0.662 and 0.701 for F1, are publicly available and we use them for the construction of our VSs, after having retrained them on the version 1.3 of the corpus (cf. §5.2). An interesting aspect is that they have very different perspectives. Seen2Seen is symbolic hence lightweight, uses rules and filters, focuses only on VMWEs seen in TRAIN and obtains rather good precision and a lower recall. MTLB-STRUCT, conversely, has a BERT-based architecture (Devlin et al., 2019), has a high training and prediction cost, but tries to generalize beyond the seen VMWEs and obtains both descent precision and recall.

As a consequence, for data outside of the shared task, MTLB-STRUCT annotates an arguably high number of types (tens of thousands usually), while

[5]VID / verbal idioms, LVC / light verb construction, IRV / inherently reflexive verbs, VPC / verb-particle construction, MVC / multi-verb construction, ICV / inherently clitic verb (specific to Italian), IAV / inherently adpositional verbs (experimental category). For examples, we invite readers to refer to the aforementioned paper as well as official guidelines.

Seen2Seen finds a much lower number of types (often in the range of 1000-2000) but with higher precision. Due to the complementarity of these two systems, we will discuss the VSs generated from their annotations (§5.3).

5.2 Protocol to generate vector spaces

As stated previously in Section 3, the literature justifies the use of Word2Vec embeddings for the vectorisation of MWEs in type-oriented tasks. We vectorise VMWEs as follows:

Training VMWE identifiers Seen2Seen is retrained on the PARSEME 1.3 corpus for all 14 languages, while MTLB-STRUCT, due to its high training cost, is only retrained for Polish and French. We shall focus on these two languages in this study, as most systems were evaluated for them, and native speakers are among the authors of this paper.

Large corpus annotation Using Seen2Seen, annotate data from PARSEME 1.2 "raw corpora" (cf. §5.1) for all 14 languages. The outcome of this process, for single word tokens and VMWE tokens, is described in Table 1. For more detailed statistics on class-wise VMWEs per language, see Table A2 in the Appendix.[6] Additionally, we use MTLB-STRUCT to annotate part of the Polish and French "raw corpora", so as to have a sufficient coverage of VMWEs in the diversity experiments in Section 7.

Merge VMWE constituents Based on the system's annotation, recreate text in which VMWE instances are merged into a single token. **(a) For each VMWE instance, lemmatise its tokens and sort them** (based on UTF-8), which ensures that various token orders map to the same canonical form. In our case, we use lemmas already made available in the PARSEME 1.3 TRAIN and PARSEME 1.2 "raw corpora". **(b) Add a `_MWE_` prefix**. This yields for example `_MWE_le_mer_prendre` for the VMWE ***prendre la mer*** (lit. 'take the sea') 'take to the sea'. Alternatively, extend the prefix with the VMWE class, *e.g.*, `_MWE-IRV_se_trouver`. This will be used in Figure A2 in Appendix. **(c) Remove from the text the individual tokens that made up the VMWE, and place the one-merged-token-VMWE at the average position of constituent tokens.** For a VMWE made of tokens at indices 48, 49, and 51, the resulting one-merged-token-VMWE is positioned at index $(48 + 49 + 51)/3 \approx 49.33$ so before the token initially at index 50. This allows us to handle discontinuous VMWEs.

lang	tokens	lemmas	form
DE	188,230k	2,038k	2,267k
EL	26,195k	1,200k	1,319k
EU	21,268k	222k	403k
FR	803,649k	5,551k	5,563k
GA	34,211k	525k	550k
HE	15,537k	209k	326k
HI	74,366k	820k	888k
IT	197,493k	1,579k	1,709k
PL	486,735k	9,918k	10,992k
PT	324,312k	4,423k	4,546k
RO	12,680k	215k	277k
SV	627,384k	12,358k	13,048k
TR	20,171k	311k	655k
ZH	67,235k	1,911k	1,912k
Σ	2,899,473k	41,286k	44,461k

lang	instances	canonical	non-canonical
DE	2,731k	1,881	14,712
EL	99k	2,146	16,751
EU	496k	675	24,814
FR	3,497k	1,724	27,874
GA	222k	113	2,334
HE	29k	556	3,480
HI	652k	139	4,024
IT	1,579k	1,515	27,308
PL	3,640k	3,114	80,137
PT	1,610k	2,424	46,380
RO	212k	838	8,735
SV	6,776k	1,028	10,541
TR	481k	2,318	85,787
ZH	1,260k	3,127	3,127
Σ	23,289k	21,598	356,004

Table 1: Statistics about data used for the generation of VSs. Upper table is tokens, lower table is VMWEs. The entries VSs comprise are token forms and VMWE canonical forms. Languages are abbreviated as follows: DE = German, EL = Greek, EU = Basque, FR = French, GA = Irish, HE = Hebrew, HI = Hindi, IT = Italian, PL = Polish, PT = Portuguese, RO = Romanian, SV = Swedish, TR = Turkish, ZH = Chinese.

Train the VSs Using the newly VMWE-merged text, train a VS for each language using Word2Vec (Mikolov et al., 2013).[7] This results in Seen2Seen-based VSs with both single-word tokens and VMWE tokens, precisely corresponding to the source corpus described in Table 1. Henceforth, we will refer to these VSs as VS_{S2S}. The result-

[6] As both Polish and Swedish had over 100GB of data and that annotation is somewhat expensive, they were truncated to about a quarter for each, equating to 40+GB for each.

[7] The parameters used for training are: cbow=0, size=100, window=10, negative=10, hs=0, iter=3, min-count=1. About the number of dimensions (size=100), we tried both lower (size=10) and higher (size=300) numbers of dimensions, which yielded similar VSs. Both CBOW (cbow=1) and Skip-Gram (cbow=0) have been tested; as Skip-Gram yielded PCAs on which more information were present on the first dimensions, we kept it. This may correspond to the findings in the original Word2Vec article that Skip-Gram better encodes semantic information (Mikolov et al., 2013).

ing VS binaries are publicly available at http://hdl.handle.net/11234/1-5528. Additionally, we train in the same way, but using the MTLB-STRUCT-annotated corpus, VSs for Polish and French, henceforth called VS_{MTLB}.

As all members of these VSs are represented in $\mathbb{R}^d$, a function $f : \langle \mathbb{R}^d, \mathbb{R}^d \rangle \rightarrow \mathbb{R}$ may be used to estimate the distance between VMWE tokens, between single-word tokens, or between single-word and VMWE tokens.

VMWEs in the corpus have a Zipfian distribution, many occur rarely. This may result in under-trained embeddings, but removing those with few instances would eliminate most VMWEs, and setting a threshold for a minimum number of instances would be arbitrary. Therefore we keep all VMWEs in our VSs, whatever their frequency.

5.3 VMWE vector spaces and their properties

In this section we analyse the properties of the VSs generated in the preceding section to check if they reasonably represent the single-word and VMWE vocabulary.

Firstly, all the vectors for VMWEs may not be of sufficient quality, possibly because a number of them only appear once and thus are poorly represented. However, we see through nearest-neighbour distances in VS_{MTLB} (Table 2, French examples) that both `_MWE_bataille_mener` for ***mener bataille*** 'to lead a battle' and `_MWE_aide_en_venir` pour ***venir en aide*** (lit. 'to come in help') 'to help' provide expected nearest neighbours (with the exception of *échapper* 'to escape').

Original	Translation	Similarity
_MWE_campagne_mener	to lead a (war) campaign	0.737311
_MWE_guerre_mener	to do war	0.734716
_MWE_attaque_mener	to lead (an) attack	0.723792
_MWE_mener_offensive	to lead (an) attack	0.721456
_MWE_mener_révolte	to lead (an) insurrection	0.708477
_MWE_porter_secours	to provide assistance	0.785825
_MWE_confiance_faire	to trust	0.774374
échapper	to escape	0.773320
_MWE_fort_main_prêter	to (physically) help	0.741453
_MWE_tenir_tête	to stand up to (someone)	0.737382

Table 2: Examples of most similar elements to VMWEs. Respectively `_MWE_bataille_mener` (to lead a battle) and `_MWE_aide_en_venir` (to come help).

One may also tackle the quality of VS through "*A is to B what C is to D*" analogies in which given *A*, *B*, and *C* we ask for *D*. Examples include "*bateau* is to *_MWE_escale_faire* what *train* is to ...?" ("*boat* is to *make a boat stop* what *train* is to ...?") in Table 3 (French VS_{MTLB}). While similarity scores[8] for analogy are lower than for nearest neighbours and that desired VMWEs do not rank first, it is fair to say that VMWEs are reasonably well positioned in VS to represent their semantics. While the above analyses only concern VS_{MTLB}, we hypothesise that they also apply to VS_{S2S} due to the resemblance of both VSs shown below. Thus, **on the perspective of the local neighbourhood of a VMWE, semantics and related similarity scores seem meaningful**. This will be relevant in a later part of the article.

Original	Translation	Similarity
partira	will leave	0.602759
arrive	arrives	0.599007
_MWE_faire_étape	to make a (train) stop	0.587047
retourna	returned	0.585420
retourne	returns	0.579429
interviewé	interviewed	0.604981
_MWE_interview_réaliser	to make an interview	0.582795
présentateur	(show) host	0.554260
_MWE_enquête_mener	to lead (an) investigation	0.549344
interview	(an) interview	0.548068

Table 3: Examples of analogies in the form of "*A* is to *B* what *C* is to *D*" for which the VS is queried for *D*. Respectively "*bateau* is to *_MWE_escale_faire* what *train* is to ...?" ("*boat* is to *make a boat stop* what *train* is to ...?") and "*scientifique* is to *_MWE_expérience_mener* what *journaliste* is to ...?" ("*scientist* is to *lead experiment* what *journalist* is to ...?").

We now proceed to a more holistic comparative analysis of VS_{S2S} and VS_{MTLB}. We see in Figure 1 the Principal Component Analysis (PCA) of VSs trained on data annotated by Seen2Seen and MTLB-STRUCT for Polish. We first see that token-wise, the VSs are very much similar, and the first two Principal Components (respectively on the horizontal and vertical axis of the plots) encode similar amounts of information. VMWE constituents (in green) belong to a specific region, which Seen2Seen's VMWEs seem to overlap with a lot. We see that for VS_{S2S}, VMWEs cluster in a specific region, and their centroid (the dark triangle) is far away from the centroid of standard tokens (the "+"); for VS_{MTLB} however the VMWE-specific region is much wider and the centroid of its VMWEs (the dark triangle) is very close to that of standard tokens (the "+"). Interestingly, Seen2Seen's VMWEs in VS_{MTLB} (the yellow triangle is their centroid), remain distant from the centroid of standard tokens (the "+"), which is consistent with the position VMWEs are at in VS_{S2S}.

[8] Computed using cosine similarity, see EQUATION 8.

It should be noted that $> 80\%$ of Seen2Seen's VMWEs are present in VS_{MTLB}, while $< 10\%$ of MTLB-STRUCT's VMWEs are present in VS_{S2S} (which is understandable since Seen2Seen is restricted to VMWEs from the PARSEME 1.3 TRAIN). For the Polish VSs from Figures 1 & A2, 2.6% of MTLB-STRUCT's VMWEs are in VS_{S2S}, and 90.5% of Seen2Seen's VMWEs are in VS_{MTLB}.

As we deal with VMWEs rather than MWEs of all syntactic types (here called simply MWEs), the substantial distance between the VMWE centroid and the centroid of all tokens raises the question of whether the constant presence of a verb influences the positioning of the VMWE in VS. Additional centroids are thus displayed, and one can see that constituents of VMWEs (large white and red centroids), whether or not verbs, are closer to VMWEs than to the average token (the "+") in VS_{S2S}. They remain at a similar position in VS_{MTLB}.

MTLB-STRUCT's VMWEs however are a lot closer to standard tokens (the "+"); the precise reason remains unanswered. A potential explanation would be that specific VMWE classes may differ in position in VS and as both systems do not annotate classes with the same distribution it may cause this behavior. However, differentiation of VMWEs based on their class, as depicted in Figure A2 in Appendix, shows that for either system no VMWE class belongs to a specific region.

We see in Figures A3 & A4 in Appendix the distribution of distances between VMWEs in Polish. This normal-like shape can be described with the average (μ) and standard deviation (σ), which do not change substantially across languages and VSs; we use cosine distance, which, defined on the range [0-2], has distances that remain on the lower end of the range. A possible explanation for this behavior across multiple distance functions is the "curse of dimensionality", the fact that "[t]wo randomly selected points in a hypercube will have nearly the same distance for larger n" (Köppen, 2000) where n is the number of dimensions. Amongst the 14 tested languages, no substantial deviations from these patterns were observed.

6 Disparity functions

We've tested multiple disparity functions from the literature and the one we found to be most discriminant, while not in its logic related to the number of types, is the functional diversity proposed by Chao et al. (2014). They present a generalisation of Hill (1973) numbers for species diversity (corresponding to variety and balance only, as it does not include distances between types), functional diversity (relying on property-wise distances between types) and phylogenetic diversity (based on distances in a tree, *i.e.*, the evolution tree). As we are interested specifically in the functional aspect (species diversity does not cover disparity, and phylogenetic diversity is out of scope here), we shall use their functional Hill number N_α^{func} based on the generalised entropy H_α^{func}

$$N_{\alpha\neq 1}^{\text{func}} = \left(\frac{H_\alpha}{Q}\right)^{\frac{1}{2}} \tag{2}$$

$$H_{\alpha\neq 1}^{\text{func}} = \left(\sum_{i,j=1}^{n} d_{ij} \times \left(\frac{p_i p_j}{Q}\right)^{\alpha}\right)^{\frac{1}{1-\alpha}} \tag{3}$$

in which $n \in \mathbb{N}$ is the number of types, $p_i \in \mathbb{Q}_{\geq 0,\leq 1}$ the relative proportion of the ith type, d_{ij} ($\in \mathbb{R}_{\geq 0,\leq 2}$ for cosine distance) the distance between the ith and the jth types, and $\alpha \in \mathbb{R}_{\geq 0}$ the order. $Q \in \mathbb{R}$ plays a normalisation role

$$Q = \sum_{i,j=1}^{n} d_{ij} p_i p_j \tag{4}$$

as the weighted average of distances. N_α^{func} and H_α^{func} have limiting cases

$$N_1^{\text{func}} = b^{H_1^{\text{func}}} \tag{5}$$

$$H_1^{\text{func}} = \sum_{i,j=1}^{n} d_{ij} \times \left(\frac{p_i p_j}{Q}\right) \log_b \left(\frac{p_i p_j}{Q}\right) \tag{6}$$

with b representing the logarithmic base (e in our case). These equations are parametric with α, which conditions how strongly the proportion of a pair of types should be considered; $\alpha = 0$ entails the same consideration for all pairs independently of proportion, while an increasing α entails an increasing relative consideration for high-frequency pairs. This behavior may be visualised in Figure 1 of Chao et al. (2014). This will be relevant as we will give results for multiple values of α.

For distance between types we shall use cosine distance d_{ij}

$$d_{ij} = 1 - s_{ij} \tag{7}$$

$$s_{ij} = \frac{\sum_{k=1}^{m} \vec{v}_{ik}\vec{v}_{jk}}{\left(\sqrt{\sum_{k=1}^{m} \vec{v}_{ik}^2}\right) \times \left(\sqrt{\sum_{k=1}^{m} \vec{v}_{jk}^2}\right)} \tag{8}$$

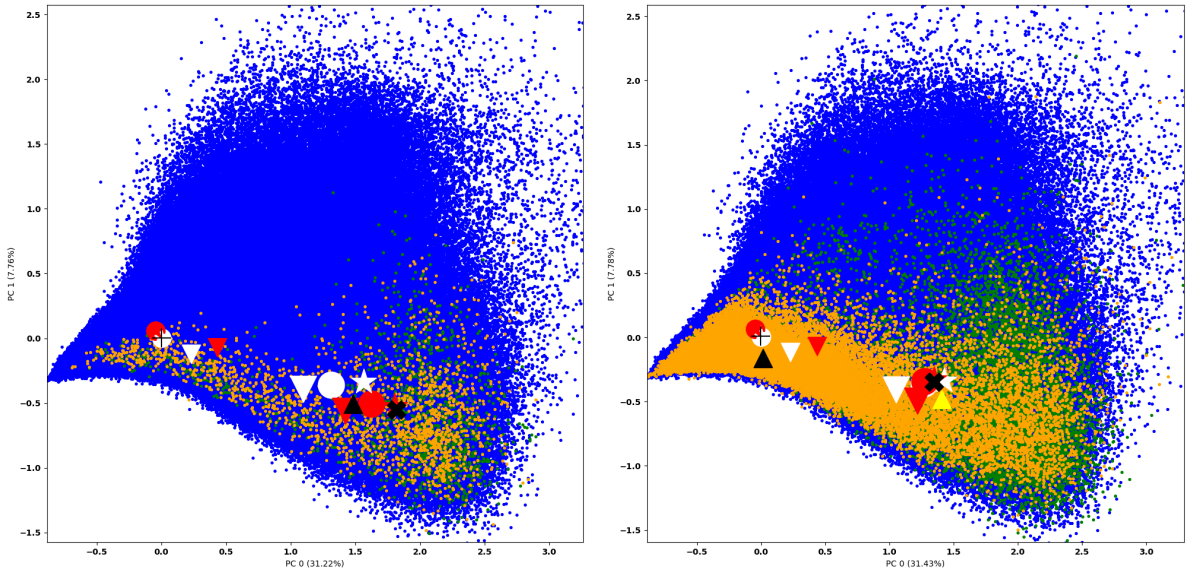

Figure 1: Principal Component Analysis (PCA) of VMWE VSs. Left is VS_{S2S}. Right is VS_{MTLB}. Polish data, trained using PARSEME 1.2 "raw corpora"'s first four files ($\approx$ 6GB of `*.cupt` data). Blue for standard tokens. Green for VMWE constituents. Orange for VMWEs. Individual shapes for centroids; (1) small means standard tokens, large means VMWEs, (2) circles for all forms or lemmas, (3) triangles for verbs, (4) stars for non-verbs, (5) white for forms, red for lemmas. "+" is the centroid of tokens not belonging to VMWEs, the dark triangle is the centroid of VMWEs, and "X" is the centroid of tokens belonging to at least one VMWE. The yellow triangle is Seen2Seen's VMWEs in VS_{S2S}. Visualisation zoomed (some outliers are thus not visible).

where $\vec{v}_i$ is the vector of the ith type.

7 Results and discussion

We use the functional diversity measure from (2) and (5) to estimate the disparity of VMWE identification systems from the PARSEME shared task 1.2. Like in (Lion-Bouton et al., 2022), we estimate disparity of true positives only. We focus on Polish and we use VS_{MTLB} rather than VS_{S2S} because we need vectors for all or most VMWEs identified by all the systems.[9]

Table 4 lists the scores for N_α^{func} with $\alpha \in \{0, 1, 2\}$. As N_α^{func} is a disparity-balance hybrid, we provide information about Zipfian parameters; s represents the curvature of the distribution, at 0 it means a perfectly even distribution and an increasing s means an increasingly uneven distribution. n corresponds to the number of types. The frequency of a type with rank x is estimated using

$$Z_{s,n}(x) = x^{-s} \left(\sum_{i=1}^{n} i^{-s} \right)^{-1} \quad (9)$$

which equates that of Lion-Bouton et al. (2022). To obtain s from an existing distribution, we minimise the mean squared error in a regression. We found that across systems, the values of s are quite similar [0.608-0.633], while there are larger differences in n. To gain insights in the idea of Zipfian parameters such as curvature (s), see Figure A1 in Appendix.

To ensure that annotations of a specific system are not substantially different from that of other systems in terms of raw distances between types (on a macroscopic scale), we also provide the mean μ_{dist} and standard deviation σ_{dist} of the distance matrix.

We note that Zipfian curvature (s) and distance matrix properties (μ_{dist} and σ_{dist}) are stable across systems. Therefore, the outcome of formula (2) or (5) can grow in only two cases: (i) the system system recognizes more types (n grows), (ii) the system more frequently annotates types which tend to be distant from other types (so that $d_{ij}p_ip_j$ grow). We claim that (ii) has few influence on disparity. This is because (§6), with $\alpha = 0$, all d_{ij} are considered equally, no matter p_ip_j, while with an increasing α, the most frequent pairs of types are increasingly favoured, to the detriment of least frequent pairs of types. If (ii) dominantly mattered, we would expect different values of α to give different rankings of diversity. But this not the case: we see that the rankings for N_α^{func} with different α in Table 4 remain the same. Thus, the reaction of

[9] See the high inter-annotator agreement in Table A1.

System	#AS	#DT	s	n	μ_{dist}	σ_{dist}	$N^{\text{func}}_{\alpha=0}$	$N^{\text{func}}_{\alpha=1}$	$N^{\text{func}}_{\alpha=2}$
ERMI	840	29	**0.633**	359	0.354	0.110	345.2	203.2	114.0
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	431.8	260.5	146.6
Seen2Seen	909	4	0.608	381	**0.367**	0.110	370.1	225.4	132.3
Seen2Unseen	936	8	0.608	410	0.361	0.110	396.3	241.4	140.0
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	**459.1**	**279.9**	**157.6**
TRAVIS-multi	968	39	0.615	440	0.353	0.112	421.7	254.2	143.2

Table 4: System diversity scores. Polish data. Column-wise, underline for minimum value, bold for maximum value. AS for active sentences (sentences in which at least one true positive VMWE is present), DT for discarded types due to no available vector prior to filtering for true positives. s for Zipfian curvature, and n for the number of types, for their distribution. μ_{dist} and σ_{dist} for the mean and standard deviation of the distance matrix, *i.e.*, the $n \times n$ matrix of distances between types (VMWEs). Diversities scores N^{func}_{α} from Chao et al. (2014). Other disparity functions may be seen in Tables A3 through A28, for Polish and French.

diversity is here essentially based on the number of types n.

8 Conclusion

We have proposed methods to quantify semantic distances among VMWEs and single words, via VSs. On this basis we performed experiments in evaluating the task of VMWE identification along a novel dimension: disparity of the systems' results. Due to huge computational costs of these experiments, not all possible scenarios were implemented. Namely, VS_{MTLB} was necessary to have a large coverage of VMWEs used in disparity experiments. But VS_{MTLB} was trained for Polish and French only, due to its high computational cost. To mitigate this, VS_{S2S} were trained (with a much lower cost) for 14 languages. Similarities between VS_{MTLB} and VS_{S2S} on the one hand, and similarities between VS_{S2S} for various languages on the other hand, allow us to hypothesise that the conclusions from the diversity experiments probably also apply to languages other than Polish and French.

Thus, we may provide the following answers to our initial research questions R1 and R2. Firstly, across various languages, VMWEs are sensibly positioned in the VSs relative to standard tokens as well as VMWE constituents. Similarity and analogy testing reveals such VSs have reasonable quality VMWE-wise. Pairwise distances between VMWEs display normal-like behavior.

Secondly, using formal disparity measures on these VMWEs does not allow for sensible distinctions. There appears to be no link between joint probability and distance, and as distances are near-equal in high-dimension VSs, disparity in this context is non-discriminant and strongly linked to the number of types n. This questions its usefulness in presence of simpler diversity metrics such as variety.

9 Limitations

This study makes use of automatic VMWE annotation, so while we made local tests of the quality of the VSs, we cannot assert their quality globally.

This study limits itself to Verbal Multiword Expressions (VMWEs), which is a narrow subset of all points in VS here (considering most points are standard tokens). As curse of dimensionality is agnostic of the phenomenon, the issues we faced, with a normal distribution of distances, may also apply to standard tokens, but the article does not explicitly show it. Also, the specific focus on VMWEs rather than all MWEs, due to available resources, means there could be VS properties that exist in non-verbal MWEs and that therefore we did not see here.

This study also does not mention issues with regard to the tractability, *i.e.*, whether disparity functions can be computed with reasonable resources, as it is not the main focus of the study. In a set with n types, there are n^2 distances to compute. For $n \in [1000 - 2000]$, as is often the case for Seen2Seen, it remains lightweight, but for systems that annotate tens of thousands of types (or even hundreds of thousands, or millions, if we select for example standard tokens as types), it very quickly becomes untractable.

10 Acknowledgements

This research was funded by the "*Plan Blanc*" (White Plan) doctoral grant from Université Paris-Saclay, by the French Agence Nationale pour la

Recherche, through the SELEXINI project (ANR-21-CE23-0033-01), and the CA21167 COST action UniDive.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

Walter Bossert, Prasanta K. Pattanaik, and Yongsheng Xu. 2001. The Measurement of Diversity.

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.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rafael Ehren. 2017. Literal or idiomatic? identifying the reading of single occurrences of German multiword expressions using word embeddings. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–112, Valencia, Spain. Association for Computational Linguistics.

Samin Fakharian and Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 23–32, Online. Association for Computational Linguistics.

Pablo Gamallo and Marcos Garcia. 2019. Unsupervised compositional translation of multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 40–48, Florence, Italy. Association for Computational Linguistics.

I. J. Good. 1953. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, 40(3-4):237–264. Number: 3-4.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text. *arXiv preprint*. ArXiv:2311.09807 [cs].

Ali Hakimi Parizi and Paul Cook. 2018. Do character-level neural network language models capture knowledge of multiword expression compositionality? In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 185–192, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80, Online. Association for Computational Linguistics.

Amir Hazem and Béatrice Daille. 2018. Word embedding approach for synonym extraction of multi-word terms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

M. O. Hill. 1973. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, 54(2):427–432. Number: 2 Publisher: Ecological Society of America.

Abhik Jana, Dima Puzyrev, Alexander Panchenko, Pawan Goyal, Chris Biemann, and Animesh Mukherjee. 2019. On the compositionality prediction of noun phrases using poincaré embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3263–3274, Florence, Italy. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kamil Kanclerz and Maciej Piasecki. 2022. Deep neural representations for multiword expressions detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 444–453, Dublin, Ireland. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2017. Complex verbs are different: Exploring the visual modality in multi-modal models to predict compositionality. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 200–206, Valencia, Spain. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2020. Disambiguation of potentially idiomatic expressions with contextual embeddings. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 85–94, online. Association for Computational Linguistics.

Mario Köppen. 2000. The curse of dimensionality. In *5th online world conference on soft computing in industrial applications (WSC5)*, volume 1, pages 4–8.

Etienne Laliberté and Pierre Legendre. 2010. A distance-based framework for measuring functional diversity from multiple traits. *Ecology*, 91(1):299–305.

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

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.

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. *arXiv preprint arXiv:2204.10050*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*. Issue: arXiv:1301.3781 arXiv:1301.3781 [cs].

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphael Fournier-S'niehotta, Remy Poulain, Lionel Tabourier, and Fabien Tarissan. 2020. Measuring Diversity in Heterogeneous Information Networks. *arXiv preprint*. Issue: arXiv:2001.01296 arXiv:2001.01296 [cs, math].

Maud A. Mouchet, Sébastien Villéger, Norman W. H. Mason, and David Mouillot. 2010. Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology*, 24(4):867–876.

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.

Naoki Otani, Satoru Ozaki, Xingyuan Zhao, Yucen Li, Micaelah St Johns, and Lori Levin. 2020. Pretokenization of multi-word expressions in cross-lingual word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4451–4464, Online. Association for Computational Linguistics.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Seen2Unseen at PARSEME shared task 2020: All roads do not lead to unseen verb-noun VMWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 124–129, online. Association for Computational Linguistics.

G. P. Patil and C. Taillie. 1982. Diversity as a Concept and its Measurement. *Journal of the American Statistical Association*, 77(379):548–561. Number: 379 Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

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.

Alfréd Rényi. 1961. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4.1, pages 547–562. University of California Press.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings*

of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Lukas Santing, Ryan Sijstermans, Giacomo Anerdi, Pedro Jeuris, Marijn ten Thij, and Riza Batista-Navarro. 2022. Food for thought: How can we exploit contextual embeddings in the translation of idiomatic expressions? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 100–110, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mahtab Sarlak, Yalda Yarandi, and Mehrnoush Shamsfard. 2023. Predicting compositionality of verbal multiword expressions in Persian. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 14–23, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019. Literal occurrences of multiword expressions: rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics*.

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

Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 255–265, Atlanta, Georgia, USA. Association for Computational Linguistics.

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].

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.

Ryosuke Takahashi, Ryohei Sasano, and Koichi Takeda. 2022. Leveraging three types of embeddings from masked language models in idiom token classification. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 234–239, Seattle, Washington. Association for Computational Linguistics.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @PARSEME 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Sébastien Villéger, Norman W. H. Mason, and David Mouillot. 2008. New Multidimensional Functional Diversity Indices for a Multifaceted Framework in Functional Ecology. *Ecology*, 89(8):2290–2301.

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.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

A Appendix

	A	B	C	D	E	F
A	1.00	0.74	0.71	0.67	0.74	0.74
B	0.74	1.00	0.80	0.77	0.84	**0.88**
C	0.71	0.80	1.00	**0.89**	0.80	0.79
D	0.67	0.77	**0.89**	1.00	0.76	0.76
E	**0.74**	0.84	0.80	0.76	1.00	0.85
F	0.74	**0.88**	0.79	0.76	**0.85**	1.00
μ	0.77	0.84	0.83	0.81	0.83	0.83

Table A1: Inter-annotator agreement between systems (Polish data). Column-wise, underline for minimum value, bold for maximum value (outside the trace). Metric: Cohen's Kappa, token-wise. Performed only on verbs, as it is VMWEs we study, and that taking all tokens would artificially create a high agreement. A: ERMI.closed, B: MTLB-STRUCT.open, C: Seen2Seen.closed, D: Seen2Unseen.open, E: TRAVIS-mono.open, F: TRAVIS-multi.open.

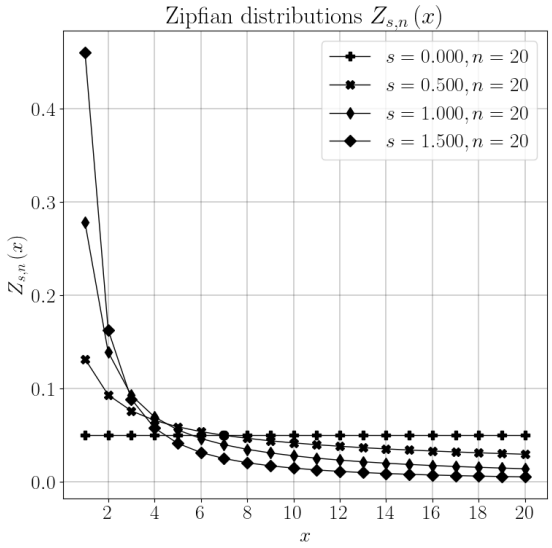

Figure A1: Examples of Zipfian distributions $Z_{s,n}(x) = x^{-s}(\sum_{i=1}^{n} i^{-s})^{-1}$. $n \in \mathbb{N}_{>0}$ denotes the number of types in the distribution. $s \in \mathbb{R}_{\geq 0}$ denotes the curvature: at $s = 0$ the distribution is perfectly flat, while it becomes increasingly curved (or uneven) with an increasing s. $x \in \mathbb{N}_{>0,\leq n}$ denotes the "rank" of the type, *i.e.*, the first, the second, *etc*.

Lang.	IAV	IRV	LVC		MVC	VID	VPC		Σ
			cause	full			full	semi	
DE	0	181	16	171	0	571	877	65	1881
EL	0	1	69	1341	5	694	36	0	2146
EU	0	0	43	453	0	179	0	0	675
FR	0	500	59	686	5	474	0	0	1724
GA	27	0	18	41	0	14	5	8	113
HE	0	0	55	274	0	206	21	0	556
HI	0	0	7	98	27	7	0	0	139
IT	90	227	80	278	13	747	62	3	1500
PL	0	1030	234	1354	0	496	0	0	3114
PT	0	318	74	1544	6	482	0	0	2424
RO	446	239	6	26	0	121	0	0	838
SV	0	59	3	145	0	154	416	251	1028
TR	0	0	0	978	1	1339	0	0	2318
ZH	0	0	83	548	1127	107	0	1262	3127
Σ	563	2555	747	7937	1184	5591	1417	1589	21583

Table A2: Detailed statistics about VMWE entries (canonical forms) in vector spaces, per VMWE class (language-specific classes excluded). This denotes that both languages and VMWE classes are unbalanced. It should also be noted that some VMWE classes do not exist in some languages, which is why some cells are at zero.

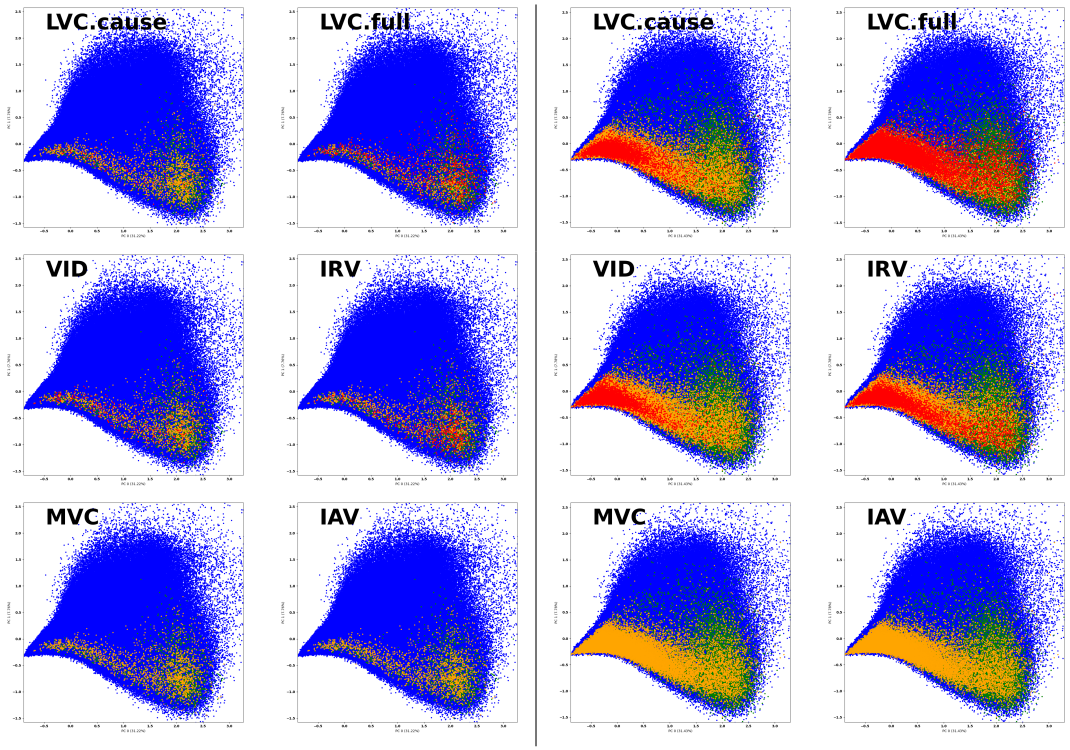

Figure A2: Vector space according to VMWE classes in Polish. Left: VS_{S2S}. Right: VS_{MTLB}. Red dots for to the specific VMWE type under study. Testing whether some VMWE classes have a special position in vector space is necessary as different systems may annotate VMWE classes with different proportions, and that this may influence disparity scores. We here see that no VMWE class has a clearly delimited region. Therefore, the tendency of systems to favour some VMWE classes is unlikely to have a substantial impact on disparity scores.

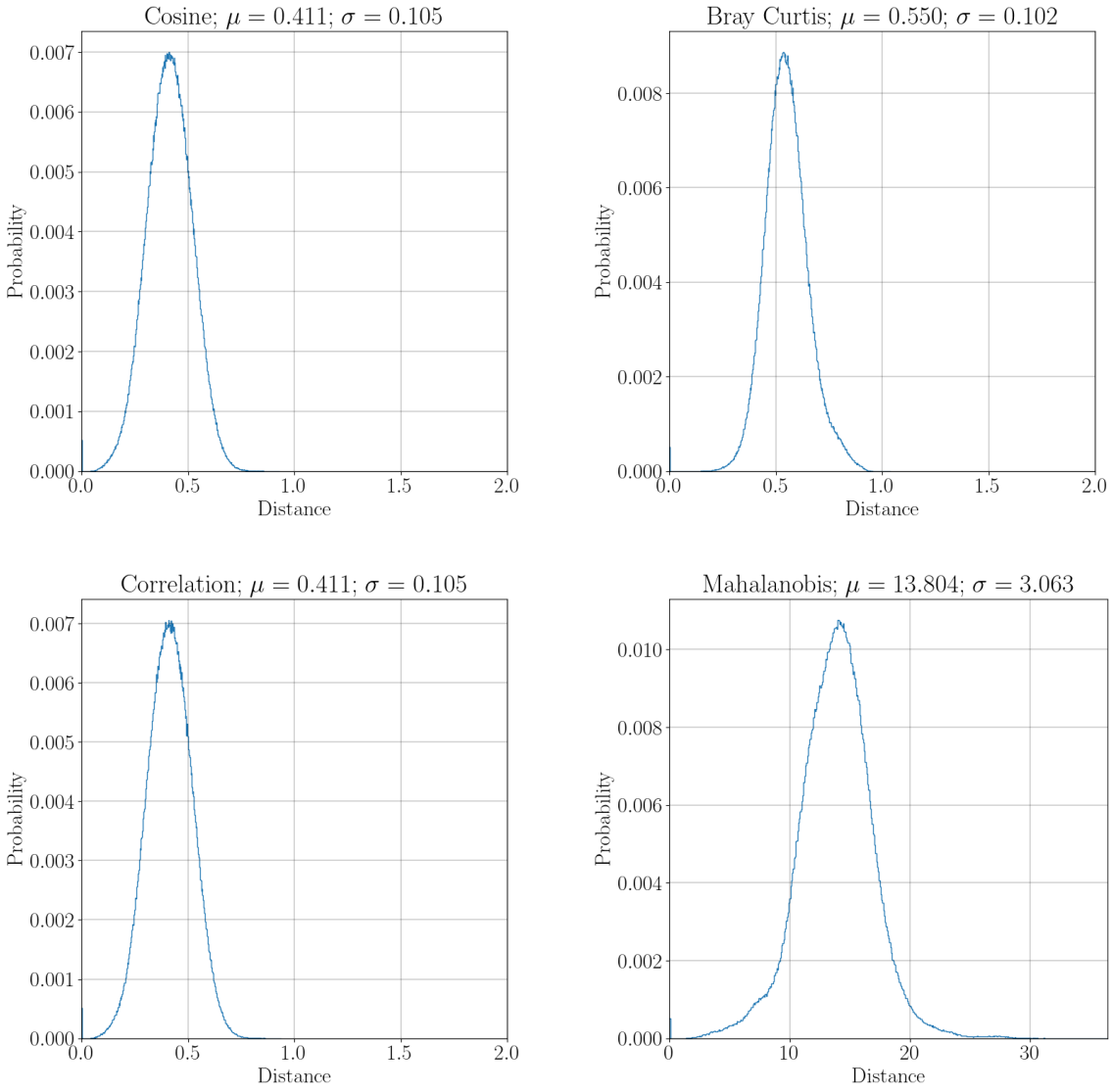

Figure A3: Distance functions and their distributions (first set). Distances between Polish VMWEs. μ for mean, σ for standard deviation. We see that functions have a near-normal distribution.

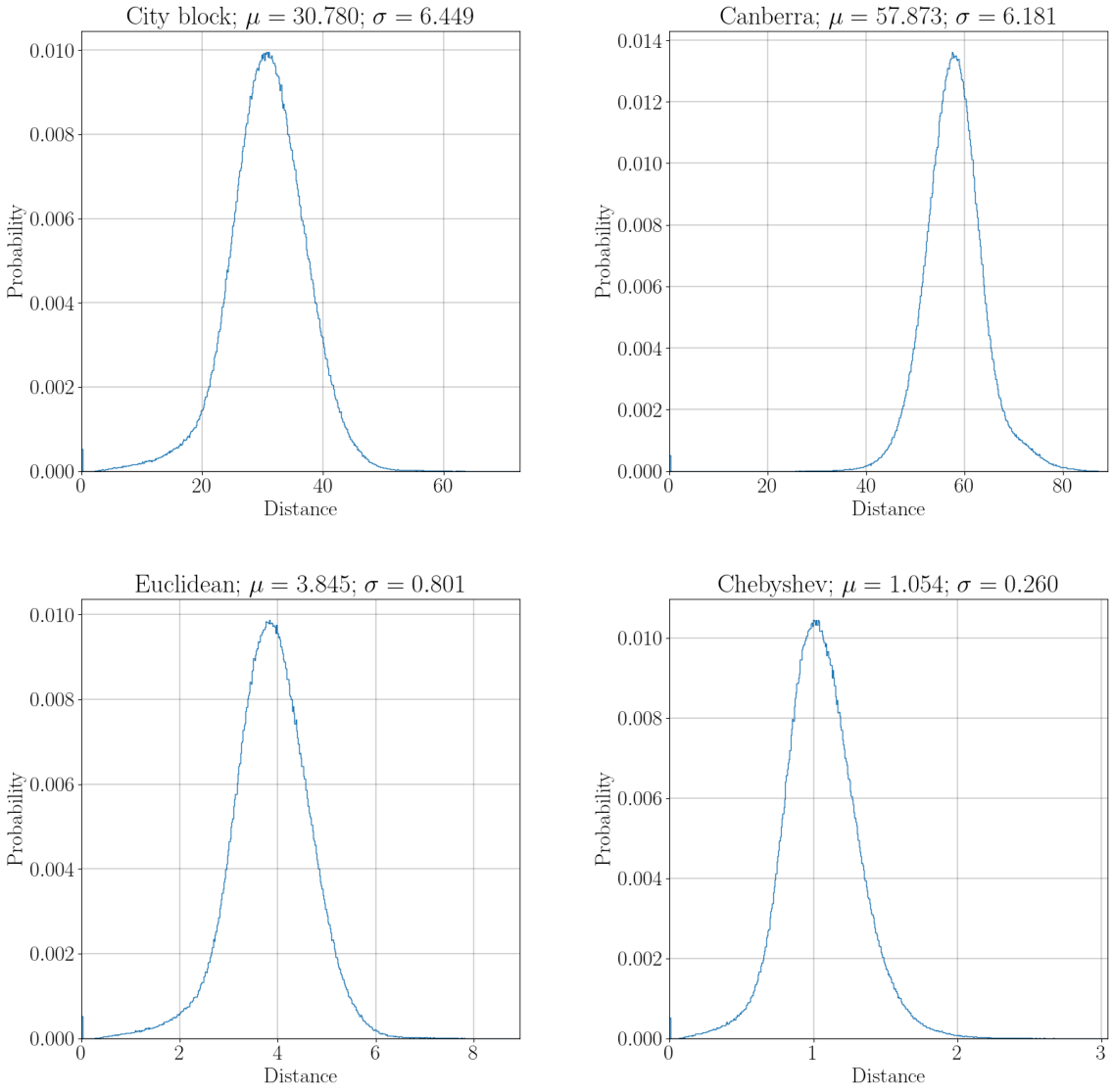

Figure A4: Distance functions and their distributions (second set). Distances between Polish VMWEs. μ for mean, σ for standard deviation. We see that functions have a near-normal distribution.

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	4.562e+04	1.580e+04	4.975e+03
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	7.205e+04	2.622e+04	8.308e+03
Seen2Seen	909	4	0.608	381	**0.367**	0.110	5.320e+04	1.974e+04	6.800e+03
Seen2Unseen	936	8	0.608	410	0.361	0.110	6.067e+04	2.251e+04	7.571e+03
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	**8.026e+04**	**2.982e+04**	**9.462e+03**
TRAVIS-multi	968	39	0.615	440	0.353	0.112	6.843e+04	2.486e+04	7.895e+03

Table A3: Scores for diversity function: Chao et al. (2014) Functional Diversity (Polish).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	3.452e+02	2.032e+02	1.140e+02
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	4.318e+02	2.605e+02	1.466e+02
Seen2Seen	909	4	0.608	381	**0.367**	0.110	3.701e+02	2.254e+02	1.323e+02
Seen2Unseen	936	8	0.608	410	0.361	0.110	3.963e+02	2.414e+02	1.400e+02
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	**4.591e+02**	**2.799e+02**	**1.576e+02**
TRAVIS-multi	968	39	0.615	440	0.353	0.112	4.217e+02	2.542e+02	1.432e+02

Table A4: Scores for diversity function: Chao et al. (2014) Functional Hill Number (Polish).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	2.147e-01	2.147e-01	2.147e-01
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	2.171e-01	2.171e-01	2.171e-01
Seen2Seen	909	4	0.608	381	**0.367**	0.110	**2.182e-01**	**2.182e-01**	**2.182e-01**
Seen2Unseen	936	8	0.608	410	0.361	0.110	2.168e-01	2.168e-01	2.168e-01
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	2.136e-01	2.136e-01	2.136e-01
TRAVIS-multi	968	39	0.615	440	0.353	0.112	2.161e-01	2.161e-01	2.161e-01

Table A5: Scores for diversity function: Laliberté and Legendre (2010) Functional Dispersion (Polish).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	**8.980e-01**	**8.980e-01**	**8.980e-01**
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	8.928e-01	8.928e-01	8.928e-01
Seen2Seen	909	4	0.608	381	**0.367**	0.110	8.926e-01	8.926e-01	8.926e-01
Seen2Unseen	936	8	0.608	410	0.361	0.110	8.919e-01	8.919e-01	8.919e-01
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	8.885e-01	8.885e-01	8.885e-01
TRAVIS-multi	968	39	0.615	440	0.353	0.112	8.940e-01	8.940e-01	8.940e-01

Table A6: Scores for diversity function: Villéger et al. (2008) Functional Divergence (Polish; modified: use of general centroid).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	7.712e-01	7.712e-01	7.712e-01
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	7.724e-01	7.724e-01	7.724e-01
Seen2Seen	909	4	0.608	381	**0.367**	0.110	7.720e-01	7.720e-01	7.720e-01
Seen2Unseen	936	8	0.608	410	0.361	0.110	**7.771e-01**	**7.771e-01**	**7.771e-01**
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	7.669e-01	7.669e-01	7.669e-01
TRAVIS-multi	968	39	0.615	440	0.353	0.112	7.645e-01	7.645e-01	7.645e-01

Table A7: Scores for diversity function: Villéger et al. (2008) Functional Evenness (Polish).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	6.507e+00	6.130e-01	**-5.263e+00**
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	6.730e+00	6.095e-01	-5.491e+00
Seen2Seen	909	4	0.608	381	**0.367**	0.110	6.559e+00	6.073e-01	-5.330e+00
Seen2Unseen	936	8	0.608	410	0.361	0.110	6.636e+00	6.095e-01	-5.400e+00
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	**6.804e+00**	**6.151e-01**	-5.551e+00
TRAVIS-multi	968	39	0.615	440	0.353	0.112	6.710e+00	6.110e-01	-5.467e+00

Table A8: Scores for diversity function: Leinster and Cobbold (2012) Diversity (Polish).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	6.698e+02	1.846e+00	**5.181e-03**
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	8.376e+02	1.839e+00	4.122e-03
Seen2Seen	909	4	0.608	381	**0.367**	0.110	7.054e+02	1.835e+00	4.844e-03
Seen2Unseen	936	8	0.608	410	0.361	0.110	7.619e+02	1.840e+00	4.518e-03
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	**9.017e+02**	**1.850e+00**	3.884e-03
TRAVIS-multi	968	39	0.615	440	0.353	0.112	8.203e+02	1.842e+00	4.223e-03

Table A9: Scores for diversity function: Leinster and Cobbold (2012) Hill Number (Polish).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	7.294e+01	7.294e+01	7.294e+01
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	8.713e+01	8.713e+01	8.713e+01
Seen2Seen	909	4	0.608	381	**0.367**	0.110	7.975e+01	7.975e+01	7.975e+01
Seen2Unseen	936	8	0.608	410	0.361	0.110	8.320e+01	8.320e+01	8.320e+01
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	**9.038e+01**	**9.038e+01**	**9.038e+01**
TRAVIS-multi	968	39	0.615	440	0.353	0.112	8.552e+01	8.552e+01	8.552e+01

Table A10: Scores for diversity function: Bossert et al. (2001) Lexicographic Approach (Polish).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	3.549e-01	3.549e-01	3.549e-01
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	3.566e-01	3.566e-01	3.566e-01
Seen2Seen	909	4	0.608	381	**0.367**	0.110	**3.675e-01**	**3.675e-01**	**3.675e-01**
Seen2Unseen	936	8	0.608	410	0.361	0.110	3.618e-01	3.618e-01	3.618e-01
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	3.476e-01	3.476e-01	3.476e-01
TRAVIS-multi	968	39	0.615	440	0.353	0.112	3.543e-01	3.543e-01	3.543e-01

Table A11: Scores for diversity function: Mouchet et al. (2010) Pairwise Distances (Polish; modified: normalised).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	6.329e-01	5.012e-01	3.828e-01
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	6.430e-01	5.039e-01	3.864e-01
Seen2Seen	909	4	0.608	381	**0.367**	0.110	**6.483e-01**	**5.086e-01**	**3.885e-01**
Seen2Unseen	936	8	0.608	410	0.361	0.110	6.425e-01	5.043e-01	3.863e-01
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	6.285e-01	4.940e-01	3.808e-01
TRAVIS-multi	968	39	0.615	440	0.353	0.112	6.385e-01	5.016e-01	3.849e-01

Table A12: Scores for diversity function: Ricotta and Szeidl (2006) Diversity (Polish).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	inf	1.537e-01	1.034e+00
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	inf	**3.482e-01**	**1.116e+00**
Seen2Seen	909	4	0.608	381	**0.367**	0.110	inf	1.074e-14	1.000e+00
Seen2Unseen	936	8	0.608	410	0.361	0.110	inf	5.847e-15	1.000e+00
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	inf	7.779e-07	1.000e+00
TRAVIS-multi	968	39	0.615	440	0.353	0.112	inf	1.364e-01	1.031e+00

Table A13: Scores for diversity function: Scheiner (2012) Functional Diversity (Polish).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	3.590e+02	1.166e+00	1.068e+00
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	4.500e+02	**1.417e+00**	**1.245e+00**
Seen2Seen	909	4	0.608	381	**0.367**	0.110	3.810e+02	1.000e+00	1.000e+00
Seen2Unseen	936	8	0.608	410	0.361	0.110	4.100e+02	1.000e+00	1.000e+00
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	**4.810e+02**	1.000e+00	1.000e+00
TRAVIS-multi	968	39	0.615	440	0.353	0.112	4.400e+02	1.146e+00	1.063e+00

Table A14: Scores for diversity function: Scheiner (2012) Functional Hill Number (Polish).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	840	29	**0.633**	359	0.354	0.110	1.285e+05	4.562e+04	1.769e+04
MTLB-STRUCT	981	33	0.614	450	0.356	0.112	2.020e+05	7.205e+04	2.819e+04
Seen2Seen	909	4	0.608	381	**0.367**	0.110	1.448e+05	5.320e+04	2.124e+04
Seen2Unseen	936	8	0.608	410	0.361	0.110	1.677e+05	6.067e+04	2.391e+04
TRAVIS-mono	**1021**	**41**	0.609	**481**	0.347	**0.113**	**2.309e+05**	**8.026e+04**	**3.082e+04**
TRAVIS-multi	968	39	0.615	440	0.353	0.112	1.932e+05	6.843e+04	2.664e+04

Table A15: Scores for diversity function: Stirling (2007) Diversity (Polish, $\beta = 1$).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	6.878e+04	2.208e+04	4.870e+03
FipsCo	822	**66**	0.647	396	0.375	**0.121**	5.879e+04	2.408e+04	7.007e+03
HMSid	680	37	0.665	367	0.386	0.113	5.200e+04	2.108e+04	5.760e+03
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	1.005e+05	3.312e+04	6.931e+03
Seen2Seen	898	15	0.693	397	**0.413**	0.113	6.517e+04	2.139e+04	4.984e+03
Seen2Unseen	957	30	0.690	447	0.404	0.117	8.072e+04	2.679e+04	5.994e+03
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	**1.064e+05**	**3.525e+04**	**7.424e+03**
TRAVIS-multi	942	31	0.692	469	0.396	0.114	8.713e+04	2.767e+04	5.815e+03

Table A16: Scores for diversity function: Chao et al. (2014) Functional Diversity (French).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	4.127e+02	2.338e+02	1.098e+02
FipsCo	822	**66**	0.647	396	0.375	**0.121**	3.849e+02	2.463e+02	1.329e+02
HMSid	680	37	0.665	367	0.386	0.113	3.596e+02	2.290e+02	1.197e+02
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	4.960e+02	2.847e+02	1.302e+02
Seen2Seen	898	15	0.693	397	**0.413**	0.113	3.943e+02	2.259e+02	1.090e+02
Seen2Unseen	957	30	0.690	447	0.404	0.117	4.410e+02	2.540e+02	1.202e+02
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	**5.114e+02**	**2.944e+02**	**1.351e+02**
TRAVIS-multi	942	31	0.692	469	0.396	0.114	4.602e+02	2.594e+02	1.189e+02

Table A17: Scores for diversity function: Chao et al. (2014) Functional Hill Number (French).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	2.282e-01	2.282e-01	2.282e-01
FipsCo	822	**66**	0.647	396	0.375	**0.121**	2.240e-01	2.240e-01	2.240e-01
HMSid	680	37	0.665	367	0.386	0.113	2.269e-01	2.269e-01	2.269e-01
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	2.311e-01	2.311e-01	2.311e-01
Seen2Seen	898	15	0.693	397	**0.413**	0.113	**2.379e-01**	**2.379e-01**	**2.379e-01**
Seen2Unseen	957	30	0.690	447	0.404	0.117	2.353e-01	2.353e-01	2.353e-01
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	2.300e-01	2.300e-01	2.300e-01
TRAVIS-multi	942	31	0.692	469	0.396	0.114	2.329e-01	2.329e-01	2.329e-01

Table A18: Scores for diversity function: Laliberté and Legendre (2010) Functional Dispersion (French).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	9.099e-01	9.099e-01	9.099e-01
FipsCo	822	**66**	0.647	396	0.375	**0.121**	8.885e-01	8.885e-01	8.885e-01
HMSid	680	37	0.665	367	0.386	0.113	8.913e-01	8.913e-01	8.913e-01
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	9.127e-01	9.127e-01	9.127e-01
Seen2Seen	898	15	0.693	397	**0.413**	0.113	**9.161e-01**	**9.161e-01**	**9.161e-01**
Seen2Unseen	957	30	0.690	447	0.404	0.117	9.149e-01	9.149e-01	9.149e-01
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	9.132e-01	9.132e-01	9.132e-01
TRAVIS-multi	942	31	0.692	469	0.396	0.114	9.156e-01	9.156e-01	9.156e-01

Table A19: Scores for diversity function: Villéger et al. (2008) Functional Divergence (French; modified: use of general centroid).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	7.887e-01	7.887e-01	7.887e-01
FipsCo	822	**66**	0.647	396	0.375	**0.121**	7.670e-01	7.670e-01	7.670e-01
HMSid	680	37	0.665	367	0.386	0.113	7.956e-01	7.956e-01	7.956e-01
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	**8.004e-01**	**8.004e-01**	**8.004e-01**
Seen2Seen	898	15	0.693	397	**0.413**	0.113	7.875e-01	7.875e-01	7.875e-01
Seen2Unseen	957	30	0.690	447	0.404	0.117	7.918e-01	7.918e-01	7.918e-01
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	7.934e-01	7.934e-01	7.934e-01
TRAVIS-multi	942	31	0.692	469	0.396	0.114	7.969e-01	7.969e-01	7.969e-01

Table A20: Scores for diversity function: Villéger et al. (2008) Functional Evenness (French).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	6.636e+00	5.913e-01	-5.448e+00
FipsCo	822	**66**	0.647	396	0.375	**0.121**	6.590e+00	**5.983e-01**	-5.378e+00
HMSid	680	37	0.665	367	0.386	0.113	6.505e+00	5.932e-01	**-5.309e+00**
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	6.822e+00	5.867e-01	-5.639e+00
Seen2Seen	898	15	0.693	397	**0.413**	0.113	6.560e+00	5.761e-01	-5.411e+00
Seen2Unseen	957	30	0.690	447	0.404	0.117	6.686e+00	5.801e-01	-5.523e+00
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	**6.863e+00**	5.885e-01	-5.672e+00
TRAVIS-multi	942	31	0.692	469	0.396	0.114	6.740e+00	5.840e-01	-5.565e+00

Table A21: Scores for diversity function: Leinster and Cobbold (2012) Diversity (French).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	7.623e+02	1.806e+00	4.304e-03
FipsCo	822	**66**	0.647	396	0.375	**0.121**	7.274e+02	**1.819e+00**	4.618e-03
HMSid	680	37	0.665	367	0.386	0.113	6.688e+02	1.810e+00	**4.946e-03**
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	9.180e+02	1.798e+00	3.557e-03
Seen2Seen	898	15	0.693	397	**0.413**	0.113	7.065e+02	1.779e+00	4.467e-03
Seen2Unseen	957	30	0.690	447	0.404	0.117	8.013e+02	1.786e+00	3.994e-03
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	**9.561e+02**	1.801e+00	3.442e-03
TRAVIS-multi	942	31	0.692	469	0.396	0.114	8.454e+02	1.793e+00	3.829e-03

Table A22: Scores for diversity function: Leinster and Cobbold (2012) Hill Number (French).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	8.692e+01	8.692e+01	8.692e+01
FipsCo	822	**66**	0.647	396	0.375	**0.121**	7.812e+01	7.812e+01	7.812e+01
HMSid	680	37	0.665	367	0.386	0.113	7.662e+01	7.662e+01	7.662e+01
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	1.041e+02	1.041e+02	1.041e+02
Seen2Seen	898	15	0.693	397	**0.413**	0.113	8.856e+01	8.856e+01	8.856e+01
Seen2Unseen	957	30	0.690	447	0.404	0.117	9.569e+01	9.569e+01	9.569e+01
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	**1.062e+02**	**1.062e+02**	**1.062e+02**
TRAVIS-multi	942	31	0.692	469	0.396	0.114	9.785e+01	9.785e+01	9.785e+01

Table A23: Scores for diversity function: Bossert et al. (2001) Lexicographic Approach (French).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	3.908e-01	3.908e-01	3.908e-01
FipsCo	822	**66**	0.647	396	0.375	**0.121**	3.758e-01	3.758e-01	3.758e-01
HMSid	680	37	0.665	367	0.386	0.113	3.871e-01	3.871e-01	3.871e-01
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	3.918e-01	3.918e-01	3.918e-01
Seen2Seen	898	15	0.693	397	**0.413**	0.113	**4.145e-01**	**4.145e-01**	**4.145e-01**
Seen2Unseen	957	30	0.690	447	0.404	0.117	4.049e-01	4.049e-01	4.049e-01
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	3.853e-01	3.853e-01	3.853e-01
TRAVIS-multi	942	31	0.692	469	0.396	0.114	3.969e-01	3.969e-01	3.969e-01

Table A24: Scores for diversity function: Mouchet et al. (2010) Pairwise Distances (French; modified: normalised).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	6.944e-01	5.391e-01	4.039e-01
FipsCo	822	**66**	0.647	396	0.375	**0.121**	6.757e-01	5.243e-01	3.969e-01
HMSid	680	37	0.665	367	0.386	0.113	6.887e-01	5.344e-01	4.021e-01
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	7.078e-01	5.443e-01	4.086e-01
Seen2Seen	898	15	0.693	397	**0.413**	0.113	**7.366e-01**	**5.649e-01**	**4.191e-01**
Seen2Unseen	957	30	0.690	447	0.404	0.117	7.264e-01	5.567e-01	4.151e-01
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	7.023e-01	5.407e-01	4.068e-01
TRAVIS-multi	942	31	0.692	469	0.396	0.114	7.149e-01	5.502e-01	4.114e-01

Table A25: Scores for diversity function: Ricotta and Szeidl (2006) Diversity (French).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	inf	1.533e-01	1.031e+00
FipsCo	822	**66**	0.647	396	0.375	**0.121**	inf	6.586e-19	1.000e+00
HMSid	680	37	0.665	367	0.386	0.113	inf	**6.189e-01**	**1.320e+00**
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	inf	7.466e-03	1.001e+00
Seen2Seen	898	15	0.693	397	**0.413**	0.113	inf	2.308e-09	1.000e+00
Seen2Unseen	957	30	0.690	447	0.404	0.117	inf	7.466e-03	1.001e+00
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	inf	1.757e-04	1.000e+00
TRAVIS-multi	942	31	0.692	469	0.396	0.114	inf	8.989e-08	1.000e+00

Table A26: Scores for diversity function: Scheiner (2012) Functional Diversity (French).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	4.200e+02	1.166e+00	1.062e+00
FipsCo	822	**66**	0.647	396	0.375	**0.121**	3.960e+02	1.000e+00	1.000e+00
HMSid	680	37	0.665	367	0.386	0.113	3.670e+02	**1.857e+00**	**1.741e+00**
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	5.070e+02	1.007e+00	1.002e+00
Seen2Seen	898	15	0.693	397	**0.413**	0.113	3.970e+02	1.000e+00	1.000e+00
Seen2Unseen	957	30	0.690	447	0.404	0.117	4.470e+02	1.007e+00	1.002e+00
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	**5.260e+02**	1.000e+00	1.000e+00
TRAVIS-multi	942	31	0.692	469	0.396	0.114	4.690e+02	1.000e+00	1.000e+00

Table A27: Scores for diversity function: Scheiner (2012) Functional Hill Number (French).

System	# AS	# DT	s	n	μ_{dist}	σ_{dist}	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
ERMI	812	36	**0.697**	420	0.390	0.116	1.760e+05	6.878e+04	2.917e+04
FipsCo	822	**66**	0.647	396	0.375	**0.121**	1.564e+05	5.879e+04	2.434e+04
HMSid	680	37	0.665	367	0.386	0.113	1.343e+05	5.200e+04	2.179e+04
MTLB-STRUCT	964	34	0.685	507	0.391	0.116	2.565e+05	1.005e+05	4.274e+04
Seen2Seen	898	15	0.693	397	**0.413**	0.113	1.572e+05	6.517e+04	2.896e+04
Seen2Unseen	957	30	0.690	447	0.404	0.117	1.994e+05	8.072e+04	3.536e+04
TRAVIS-mono	**1027**	49	0.682	**526**	0.385	0.117	**2.762e+05**	**1.064e+05**	**4.469e+04**
TRAVIS-multi	942	31	0.692	469	0.396	0.114	2.195e+05	8.713e+04	3.736e+04

Table A28: Scores for diversity function: Stirling (2007) Diversity (French, $\beta = 1$).