Evaluating Diversity of Multiword Expressions in Annotated Text

Adam Lion-Bouton[1], Yağmur Öztürk[2],
Agata Savary[2], Jean-Yves Antoine[1]
University of Tours - LIFAT[1], Paris-Saclay University, CNRS - LISN[2],
{lion.adam.otman, yaamurozturk}@gmail.com
agata.savary@universite-paris-saclay.fr
jean-yves.antoine@univ-tours.fr

Abstract

Diversity can be decomposed into three distinct concepts, namely: variety, balance and disparity. This paper borrows from the extensive formalization and measures of diversity developed in ecology in order to evaluate the variety and balance of multiword expression annotation produced by automatic annotation systems. The measures of richness, normalized richness, and two variations of Hill's evenness are considered in this paper. We observe how these measures behave against increasingly smaller samples of gold annotations of multiword expressions and use their comportment to validate or invalidate their pertinence for multiword expressions in annotated texts. We apply the validated measures to annotations in 14 languages produced by systems during the PARSEME shared task on automatic identification of multiword expressions and on the gold versions of the corpora. We also explore the limits of such evaluation by studying the impact of lemmatization errors in the Turkish corpus used in the shared task.

1 Introduction

Diversity of naturally occurring phenomena and artefacts is a desirable property of many environments and systems. It has been modelled and measured in many domains, including linguistics but has rarely been formalized with respect to particular linguistic phenomena within one language. This paper addresses diversity of one particular phenomenon: *multiword expressions* (MWEs), which are combinations of words, such as ***out of the blue*** or ***pay*** *a* ***visit***, exhibiting idiosyncratic properties at lexical, morphological, syntactic and/or semantic level (Baldwin and Kim, 2010). Automatic annotation of multiword expression occurrences in texts – henceforth referred to as MWE *identification* following nomenclature from Constant et al. (2017) – has been the focus of many works, among which the PARSEME shared tasks (Savary et al., 2017; Ramisch et al., 2018, 2020). During these tasks, performances of participating systems were evaluated on the precision, recall, and F1-score of their annotations.

In order to get a better understanding of how participating MWE identifiers behave, performances were also measured for specific subtasks such as the annotation of light verb construction occurrences (***pay*** *a* ***visit***, ***give*** *a* ***lecture***), or the annotation of MWEs that were not seen during training. Such analysis of the resulting annotation stems from a need to make sure that the systems getting the best scores are not simply performing well on a few easier, more rewarding subtasks while ignoring others. Thus, these evaluation scenarios implicitly address some aspects of data and system diversity.

In this paper we follow the same objectives but we address diversity explicitly and formally. We suggest that diversity measures could later be used to put performance measures in perspective, by favoring NLP tools which cover diverse types and not only easy and repetitive cases.

The paper is organized as follows. We first present related work in estimating diversity in linguistics and NLP (Sec. 2). We present the data used in experiments (Sec. 3). We formalize diversity with respect to the MWE phenomenon and propose concrete measures (Sec. 4). Then we experiment with these measures to estimate diversity in MWE-annotated corpora and in annotations produced by systems participating in the PARSEME shared task 1.2 (Sec. 5) and we offer a discussion of the results (Sec. 6). Finally, we conclude and suggest perspectives for future work (Sec. 7).

2 Related Work in Linguistic Diversity

Measuring diversity, – often along its three dimensions: variety, balance and disparity (Sec. 4) – has been practiced in domains such as ecology, economy, public policy, information theory, social media, etc. (Morales et al., 2021).

Proceedings of the 29th International Conference on Computational Linguistics, pages 3285–3295
October 12–17, 2022.

Diversity is also a central notion in linguistic debates. Evans and Levinson (2009) oppose to the hypothesis of the existence of language universals (Greenberg, 1966) and suggest that linguistic research should rather use diversity as a starting point.

Quantifying linguistic diversity has been performed for decades. Greenberg (1956) measured the probability of monolingual members of a population to speak the same language. Nettle (1999), cited by Harmon and Loh (2010), modelled language diversity in terms of richness (the number of different languages in a given geographical area), phylogenetic diversity (the number of different lineages in the phylogenetic tree of languages) or structural diversity (variation among structures within languages). The Terralingua initiative[1] suggested that linguistic diversity should be regarded, from a holistic perspective, as part of biocultural diversity, and proposed indices to follow the number of world's active languages, the distribution of mother-tongue speakers among them and the rate of language extinction (Harmon and Loh, 2010). Also socio-linguistic diversity was measured in terms of the probability of using more than one common language in multilingual communication, as well as the degree of diversity of language policies (Gazzola et al., 2020).

All these measures are inter-linguistic. We are, conversely, interested in intra-linguistic measures which would represent diversity of linguistic phenomena within one language, and more precisely within NLP artefacts such as language resources and outcomes of NLP tools.

The need for diversity in training data and its impact of the performance of NLP tools has been stressed in parsing (Narayan and Cohen, 2015) or question answering (Yang et al., 2018). In these works, however, the notion of diversity was used loosely and in ad hoc manner (e.g. to describe adding noise to training data, or using multiple knowledge sources and topics) rather than formally defined.

In other works more precise diversity measures do occur. This is especially the case in natural language generation (NLG), where the so-called quality-diversity tradeoff problem is observed (systems reduce the potential diversity of their generated outputs to better fit the reference). We notice, however, that the use of diversity in NLG is not standardized. Li et al. (2016) calculate the number of distinct unigrams and bigrams in generated text and Zhang et al. (2020) use Shannon's entropy, measures which relate to richness and balance, respectively. Agirre et al. (2016) define Word Embedding Similarity, i.e. the average cosine distance between utterance embeddings. Zhu et al. (2018) use SelfBLEU, i.e. the BLEU measure applied to generated utterances rather than to the reference. Palumbo et al. (2020) mix the 2 previous measures with Jaccard, i.e. the average word overlap across utterances. These are distance measures which might be used to model disparity (if items and types are properly defined). Some of those measures are also implemented in NLG toolboxes (Li et al., 2021). Thus, diversity estimation is becoming an inherent component of NLG models.

On the other hand, complexity, a notion somewhat similar to diversity has also been measured in NLP (Brunato et al., 2016), e.g. for the sake of language learning or text simplification.

We, conversely, are interested in diversity (a notion larger than complexity) and in its promotion in language resources and tools. In this paper, we mainly focus on two of the three aspects of diversity, variety and balance. Especially the latter seems to have rarely been formalised and measured in NLP.

3 Data

Before going into details on what our diversity measures will be, we take a look at the data and the ways we use them in our experiments.

The PARSEME shared task 1.2 on automatic identification of verbal MWEs (VMWEs) was conducted on 14 languages. The corpus of each language, annotated for morpho-syntax and VMWEs, was split into three corpora, TRAIN, DEV and TEST. The TRAINs, DEVs, and blind version of TESTs corpora (with annotation for VMWE hidden) were given to the participants of the shared task, which were tasked to annotate the blind TESTs with their systems.

We will apply our diversity measures to GOLD annotations (VMWEs manually annotated in the TEST corpora) and SYSTEM annotations (VMWEs automatically annotated by systems participating in the shared task 1.2). In addition, in section 5.1 we use the French corpus Sequoia (Candito et al., 2021) which is annotated in similar fashion,

[1] https://terralingua.org/what-we-do/the-loss-of-diversity/

not only for VMWEs, but for all syntactic types of MWEs.

Diversity measures for SYSTEM annotations make sense mainly for *correctly annotated MWE occurrences*. A MWE occurrence is considered correctly annotated in a given corpus if all its tokens annotated in the corresponding GOLD corpus, and only those tokens, have been annotated as part of the same MWE occurrence. Trivially, this definition also applies to MWEs from the GOLD corpus itself, which are all considered correctly annotated.

For example, if sentence 1 below is considered the gold annotation, where ***paid visit*** and ***out of the blue*** are respectively annotated as MWE A and B, then in sentence 2 only MWE D ***out of the blue*** is correctly annotated and in sentence 3 no MWE is correctly annotated.

1. I **paid**$_{\mathbf{A}}$ them a **visit**$_{\mathbf{A}}$ **out**$_{\mathbf{B}}$ **of**$_{\mathbf{B}}$ **the**$_{\mathbf{B}}$ **blue**$_{\mathbf{B}}$

2. I **paid**$_{\mathbf{C}}$ them **a**$_{\mathbf{C}}$ **visit**$_{\mathbf{C}}$ **out**$_{\mathbf{D}}$ **of**$_{\mathbf{D}}$ **the**$_{\mathbf{D}}$ **blue**$_{\mathbf{D}}$

3. I **paid**$_{\mathbf{E}}$ them a **visit**$_{\mathbf{F}}$ **out**$_{\mathbf{F}}$ **of**$_{\mathbf{F}}$ **the**$_{\mathbf{F}}$ **blue**$_{\mathbf{F}}$

In PARSEME shared task, identification systems were also tasked to assign a category (light-verb construction, verbal idiom, inherently reflexive verb, etc.) to each annotated MWE, however, we will not look at these annotations. A MWE can therefore be considered correctly annotated by a system even if its attributed category was wrong.

In this paper, we will be interested in the diversity both of GOLD annotation and of SYSTEM annotations. In the latter case, we will consider only correctly annotated MWEs.

4 Diversity Measures

The concept of diversity is usually divided into three distinct notions: *variety*, *balance* and *disparity* (Stirling, 1998). Each of these notions goes by other names and, adding to the confusion, is sometimes referred to as diversity.

The notion of diversity relies heavily on the concepts of *items* and *types*. In ecology, *items* usually refer to specimens/individuals, and *types* refer to the species these specimens are affiliated to.

In this paper, items will refer to the correctly annotated MWE occurrences and types will refer to what PARSEME calls MWE types, meaning the multisets of lemmas of the annotated MWE occurrences.

Formally, we define items and types as follows: Let I be a set of items, T a set of types, and $\tau : I \to T$ a mapping of each item to a type.

We define items $i \in I$ as correctly annotated MWE occurrences (cf. Sec. 3), and types $t \in T$ as multisets of lemmas linked to items through the mapping τ defined bellow:

$$\tau(i) = \{\, \text{lemma}(w) \mid \forall w \in i \,\} \tag{1}$$

Here, an item i is seen as a sequence of the annotated wordforms of the MWE occurrence, and lemma(w) the function returning the lemma of a wordform.

Thus, under this definition, two correctly annotated MWE occurrences (items) are of the same type if their component words have the same lemmas, e.g. ***pay visit*** and ***visits paid*** are items of type $\{pay, visit\}$.

We note here that such a definition of a type, relying on the notion of a lemma, does not perfectly capture what we would instinctively refer to as MWE types. For instance, two senses of ***put down***: 'execute' and 'belittle' are assigned to the same type despite their different meanings. Sec. 5.3 brings to light other issues related to the data quality. Still, we consider our approximation of types good enough to be useful.

Given these definitions of items and types, we will see *variety* as a measure of the number of types in a set of items, that is to say, a measure of how many different MWE types are present in an annotation. The more MWE types an annotation has, the more varied it is. *Balance* is seen as a measure of the equilibrium of the distribution of items per type, meaning that it scores distribution based on how close the MWE types are from being equally represented. *Disparity* is seen as a measure of the distance among the types present in a set of items.

While some authors, such as Stirling (1998), advocate for a complete diversity measure – meaning that all three aspects of diversity are taken into account jointly – we take the opposite stance and aim to measure each aspect of diversity as independently of the others as possible. We consider that such an approach would be easier to interpret and sidesteps the issue of finding the right aggregation of our notions of diversity.

Henceforth, we will no longer address the notion of disparity in this paper and leave it for future work instead. Let us only mention two main challenges behind this concept:

1. Disparity usually relies on a notion of distance which can be computed by various means and by taking any of the properties of MWE types into consideration. The choice of the precise properties calls for insightful studies. 2. The disparity of a set of types is often described as an aggregation of the pairwise distances between the types (Stirling, 1998). Thus, the choice of the aggregation used imposes some defining properties on the disparity. One such property is the monotonicity in types (Weitzman, 1992), which states that the disparity of a set of types can only increase when a new type is added to the set. Two disparities based on the same notion of distance, one with this property, the other without, will work in completely different ways, showcasing how the question of the aggregation is central to disparity.

4.1 Variety

We measure the variety of a set of items by its richness (2). The richness is a simple count of the number of types actually represented in a set of items.

Formally, we can define richness as the following (with I being a set of items):

$$|\{\,\tau(i) \mid \forall i \in I\,\}| \tag{2}$$

By its simplicity, such a measure of variety can be quite effective when it is used to compare how two MWE identification systems produce item sets of different variety from the same corpus. It, however, does not allow for a comparison of variety of item sets generated from different corpora.

In an attempt to compare variety across differently sized corpora, we normalize the richness of a set of items by its number of items.

$$\frac{|\{\,\tau(i) \mid \forall i \in I\,\}|}{|I|} \tag{3}$$

We refer to this measure as the normalized richness. Similar concepts can be found in measures such as type-token ratio (TTR) (Richards, 1987).[2] Normalized richness is quite intuitive, note however that badly performing systems (e.g. a system with only 1 correctly annotated MWE occurrence) may have an optimal value for this measure.

4.2 Balance

Measuring balance is not as easy of a matter. A great number of balance measures have been proposed, and their properties described and compared to one another (Smith and Wilson, 1996; Tuomisto, 2012). Despite these efforts, none of the measures proved more appropriate than the others.

Balance measures are computed on a probability mass function p_T. In this paper we approximate p_T by the relative frequency f_T as shown in (4). [3]

$$p_T(t) \approx f_T(t) = \frac{|\{\,i \mid \forall i \in I, \tau(i) = t\,\}|}{|I|} \tag{4}$$

We will use one of the earliest evenness measures which was designed with the goal of separating the concept of variety and balance as much as possible. Hill (1973) defines the continuum of evenness measures (5) where x and y can take any real values (as long as $x > y$):

$$E_{x,y} = \frac{N_x}{N_y} \tag{5}$$

The Hill number (6) noted N_a, also known as 'true diversity', is defined as the exponentiation of a base b (usually 2, e or 10) to the power of H_a given in (7)[4].

$$N_a = b^{H_a} = \left(\sum_{t \in T} p_T(t)^a\right)^{\frac{1}{1-a}} \tag{6}$$

(7) was defined by Rényi (1961) as a class of entropy functions of order a, with the entropy of order 1 being Shannon's entropy (8) noted H.

$$H_a = \frac{1}{1-a} \log_b \sum_{t \in T} p_T(t)^a \tag{7}$$

$$\lim_{a \to 1} H_a = H_1 = H \tag{8}$$

While just about any values of x and y would form a valid evenness $E_{x,y}$, only $E_{1,0}$ and $E_{2,1}$ seem to have been of interest to the community. Many arguments have been advanced favoring one or the other (Alatalo, 1981; Gosselin, 2006; Tuomisto, 2012). We will, for now, only mention that $E_{2,1}$ is supposedly less sensitive to sampling bias than $E_{1,0}$ (Alatalo, 1981) and investigate further upon this in section 5.1.

[2] In TTR in NLP, though, types are often defined as different surface forms, not different lemmas. Thus, ***pay visit*** and ***visits paid*** would be considered different types for TTR, conversely to normalized richness.

[3] In accordance with established practice the parameter p_T of measures in equations (5), (6), (7), and (8) is omited for brevity sake.

[4] Where the base b is also used as logarithmic base in H_a.

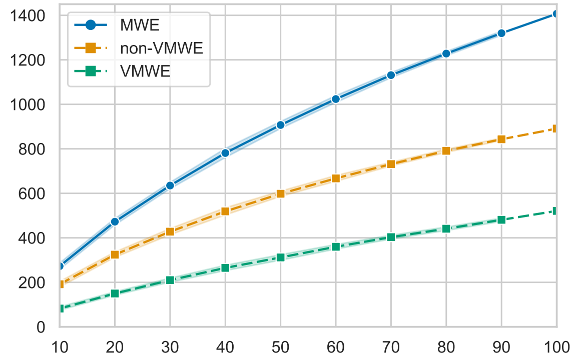

Figure 1: Richness in terms of Sequoia sample size (higher is more varied).

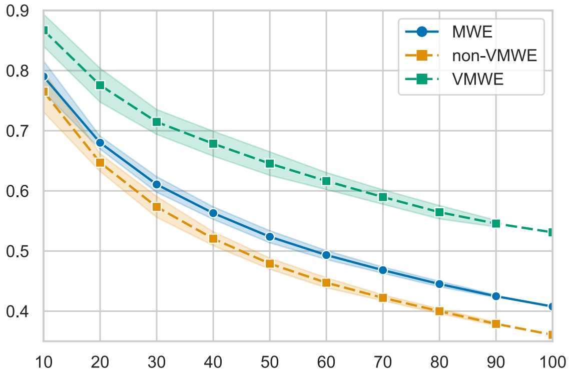

Figure 2: Normalized richness in terms of Sequoia sample size (higher is more varied).

5 Results

The previous section introduced formalization of measures for two diversity facets: variety and balance. In this section, we apply these measures on MWE-annotated corpora and MWE identifiers (Sec. 3) so as to assess their appropriateness to the MWE phenomenon (Sec. 5.1), use them as evaluation scenarios in the PARSEME shared task framework (Sec. 5.2), and show one of their limits related to data quality (Sec. 5.3).[5]

5.1 Diversity Measure Validation

We first focus on the French Sequoia corpus (Candito et al., 2021), which is one of the source corpora used for the PARSEME French corpus. It has the particularity of being annotated not only for verbal MWEs, but also for non-verbal ones (non-VMWEs).

In Figures 1, 2, 3, and 4 we apply richness (2), normalized richness (3), $E_{1,0}$ and $E_{2,1}$ (5) respectively, considering either all MWEs, only VMWEs, or non-verbal MWEs. We also take this occasion to consider how our indices behave when populations are randomly sampled. To this end, we randomly sample 10%, 20%, 30%, . . . , 100% of the sentences composing the Sequoia corpus, and compute our indices on these samples (12 repeats per sample size). Results are 12-sample averages plotted in function of the size of the sample used.

Quite unsurprisingly, in Figure 1, richness increases with the size of the sample. This growth however appears to be non-linear. Non-verbal MWEs are consistently richer than verbal ones. These results seem to be quite stable as the standard deviations (marked as bands around the lines) are barely visible on this plot.

Inversely, in Figure 2, normalized richness decreases with the size of the samples. As noted by Richards (1987), this is a consequence of the quasi linear relation between the number of sentences and number of items on the one hand, and the non-linearly slower growth of the number of types on the other hand (the latter is shown in Figure 1). This shows that, conversely to our previous intuitions, normalized richness does not allow us to reliably compare diversity of corpora of different sizes. A bigger corpus will most often be disadvantaged by this measure.

Regarding evenness (Figures 3 and 4), $E_{2,1}$, which we thought to be less sensitive to sampling bias, appears to be more volatile (i.e. having higher standard deviation) than its counterpart $E_{1,0}$. $E_{1,0}$ clearly considers the distributions of VMWEs to be more balanced than the distributions of all MWEs. On the other hand, $E_{2,1}$ finds the distributions of all MWEs slightly more balanced than those of VMWEs on large samples, but the opposite on small samples.

In an attempt to determine which evenness measure best fits the nature of the data, we first proceed to visually inspect the rank-frequency distributions of all MWEs for sample sizes 10% to 100% (Figure 5), which makes us believe that our rank-frequency distributions follow a Zipfian distribution. This would explain the non-linear relation between the number of sentences and richness. Given the apparent shape of our distributions and the relative ubiquity of Zipf distribution in linguistic phenomena (Ryland Williams et al., 2015), we will work

[5] Code and data of these experiments are available at https://github.com/AdamLionB/mwe_diversity_experiment_coling_2k22.

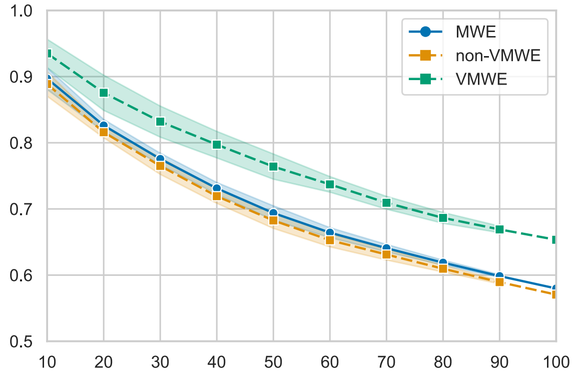

Figure 3: Evenness ($E_{1,0}$) in terms of Sequoia sample size (higher is more balanced).

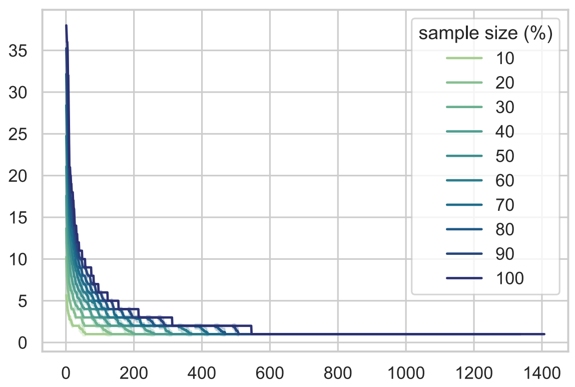

Figure 5: Rank-frequency distribution of types of all MWE according to sample size

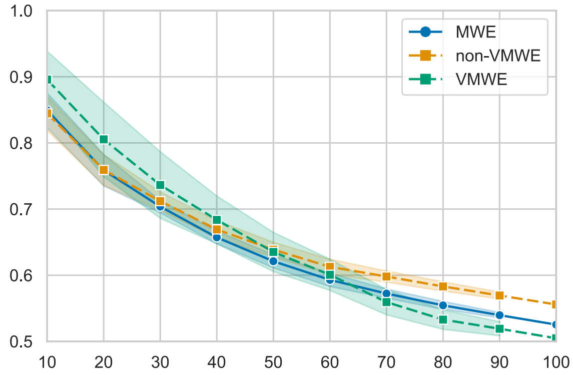

Figure 4: Evenness ($E_{2,1}$) in terms of Sequoia sample size (higher is more balanced).

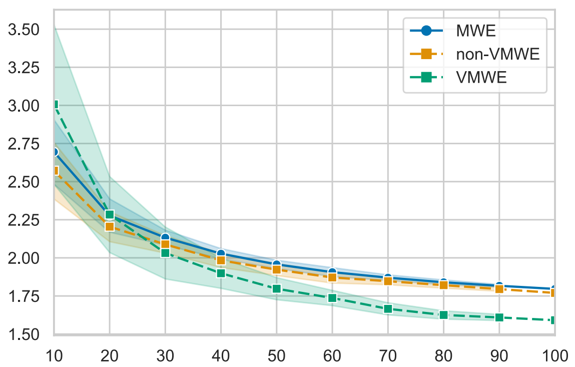

Figure 6: $\frac{1}{s}$ best fitting Sequoia samples in term of their size

under the hypothesis that the annotated MWEs follow a Zipfian distribution.

In the following, we argue that one of the parameters describing the Zipfian distribution can be used as a measure of balance. A random variable X following Zipfian distribution, noted $X \sim \text{Zipf}(s, N)$, is characterized by parameters s, N and the probability mass function (9).

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

Where N is the number of types in the distribution and s the exponent characterizing the curvature of the distribution. When N grows the distribution is slightly squeezed as follows: $Z_{s,N+1}(x) = Z_{s,N}(x) \cdot \left(1 + \frac{(N+1)^{-s}}{\sum_{n=1}^{N} n^{-s}}\right)^{-1}$ thus leaving room for the distribution to extend on its right while keeping its sum equal to 1. This hardly affects the shape of the distribution as the multiplication factor does not involve x. We therefore consider s (but not N) to be the parameter determining the shape of the distribution. Furthermore, the distribution is uniform when $s = 0$, it also becomes monotonically more and more skewed when s grows. We therefore argue that the value of s describing the Zipfian distribution which best fits an actual dataset constitutes an index of un-balance. When s is low the balance is high and vice versa. In other words, $\frac{1}{s}$ acts as a balance index.

Considering $\frac{1}{s}$ as a measure of balance, in Figure 6 we plot the values of $\frac{1}{s}$ found for our samples. s was optimised for the least square error disregarding overfitting, with N set to the number of types in the sample. By comparing Figure 6 to Figures 3 and 4 we find that $E_{2,1}$ and s both place all MWEs as more balanced than VMWEs for large samples and all MWEs as about as balanced as VMWEs for smaller samples. $E_{1,0}$ on the other hand always places VMWEs as more balanced than all MWEs. $E_{2,1}$ being in relative agreement with s in this study case, we will from now on use $E_{2,1}$ for the rest of this paper.

One might wonder why then use $E_{2,1}$ to measure balance and not simply s. Using s made sense here since we saw the distribution of the data and assumed that it follows a Zipfian distribution. While this hypothesis might very well hold for other GOLD corpora given the Zipfian nature of the language, there is little reason to believe that this hypothesis holds for distribution of MWEs annotated by automatic identifiers. In the general case we will therefore favor a measure which is agnostic of the distribution followed by the data.

5.2 PARSEME Shared Task Use Case

In this section we apply our diversity measures, validated in the previous section, to the GOLD corpora of the PARSEME shared task and to the correct annotations (true positives) produced by the participating systems.

In Table 1 we see that MTLB-STRUCT produces the richest annotations on 8 out of 14 corpora, and one of the 3 richest on the other 6 corpora. Travis-mono and Travis-multi also produce quite rich annotations. Seeing MTLB-STRUCT, Travis-mono and Travis-multi as producing rich annotations is coherent with these systems' F-measure performances.[6][7] On the other hand, annotations produced by Seen2Seen have notably low richness despite its relatively high performances. This is consistent with the fact that Seen2Seen was not designed to identify MWEs unseen during training, and made the bulk of its score on seen MWEs, therefore limiting the number of types it could recognize and annotate. We note that the richness of the systems' annotations on Irish (GA), Hebrew (HE) and Hindi (HI) are notably low, compared to the richness of the GOLD annotations. Training sets for Irish and Hindi were very small, with very few MWE occurrences, which most likely explains these results. Finding reasons for Hebrew requires more insight and, likely, a native knowledge of the language.

Table 2 shows the evenness scores of the systems' annotations according to $E_{2,1}$. (Systems' evenness closest to GOLD underlined.) For each language except Hebrew, at least one system has more balanced correct annotations than GOLD. Moreover, for 7 languages (EU, FR, HI, PL, RO, SV, ZH) annotations of most systems are more balanced than of GOLD. We discuss the case of systems' annotations with higher eveness than the gold in Section 6. Seen2Seen produces the most balanced annotations on 7 out of 14 languages and a more balanced annotation than the GOLD on 11 out of 14 languages. In most cases, Seen2Unseen produces annotations very slightly less balanced than Seen2seen. Note also the particularly high scores of FipsCo and TRAVIS-mono on German and Hindi, respectively, with 36% and 2% of types correctly identified (cf. tab. 1). [8]

On the whole, MTLB-STRUCT and TRAVIS-mono stand out as the systems producing the most varied annotations and Seen2Seen and Seen2Unseen as those producing the most balanced annotations. While a better understanding of our diversity measures is still needed, we hope that this constitutes a good first step toward a more systematic evaluation of diversity in MWE identification, and possibly in NLP overall.

5.3 Limits of the Diversity Measures: Case Study of Turkish

The morphosyntactic annotation of the Turkish corpus of PARSEME[9] was realized automatically using UDPipe (Straka, 2018). Afterwards, VMWE annotation was made manually according to the unified annotation guidelines of PARSEME[10]. The automatic morphosyntactic annotation was then partly modified and enhanced for better identification of MWEs. Thus, we have access to two versions of the same PARSEME corpus for this language: one with automatically produced lemmas and one with manual corrections of some lemmatization errors (Öztürk et al., 2022). We decided to examine this corpus to gain a better understanding of our diversity measures, and more precisely to study how the quality of the lemmas, central to our definition of types, influences the estimation of MWE diversity.

As an agglutinative language, Turkish is highly inflectional and derivational, which results in high surface variability in word forms. This makes it

[6] Linear correlation between the richness of systems' annotation's and precision, recall and F1-score are 0.49, 0.83 and 0.78 respectively.

[7] http://multiword.sourceforge.net/sharedtaskresults2020

[8] Linear correlation between the $E_{2,1}$ of systems' annotation's and precision, recall and F1-score are -0.47, -0.62 and -0.61 respectively.

[9] https://gitlab.com/parseme/parseme_corpus_tr

[10] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/

	DE	EL	EU	FR	GA	HE	HI	IT	PL	PT	RO	SV	TR	ZH
GOLD	585	682	561	712	310	443	335	638	689	815	466	495	719	551
ERMI	230	352	327	392	37	95	183	165	355	453	312	241	390	272
FipsCo	212	191		392										
HMSid				454										
MTLB-STRUCT	**400**	**472**	**387**	505	**73**	**171**	**234**	330	465	**549**	365	**345**	455	365
Seen2Seen	274	358	280	406	22	116	67	315	384	489	250	208	389	249
Seen2Unseen	290	381	333	466	48	123	170	336	416	521	259	223	435	253
TRAVIS-mono	381	55		**531**			7	**341**	**487**		**369**	307	**472**	**398**
TRAVIS-multi	348	452	368	471	14	144	176	310	449					

Table 1: Richness of GOLD and SYSTEMs annotations of PARSEME corpus

	DE	EL	EU	FR	GA	HE	HI	IT	PL	PT	RO	SV	TR	ZH
GOLD	43.40	63.95	30.30	40.42	63.88	91.08	42.48	57.05	51.38	71.67	30.89	69.28	58.59	38.12
ERMI	42.53	59.32	34.61	42.42	65.76	**90.75**	50.55	52.49	54.18	69.61	34.20	69.43	57.58	43.85
FipsCo	**85.67**	55.44		**47.98**										
HMSid				45.46										
MTLB-STRUCT	43.00	62.34	33.57	41.97	56.13	89.33	41.32	56.06	53.87	71.79	33.16	69.63	58.41	42.62
Seen2Seen	41.72	62.22	**39.05**	44.73	70.14	88.30	50.31	**58.34**	**56.73**	**72.65**	**35.54**	**74.38**	**60.06**	**45.66**
Seen2Unseen	41.68	62.14	36.94	43.52	60.90	88.20	40.89	58.03	55.84	72.02	35.24	73.81	59.61	45.59
TRAVIS-mono	42.09	**75.34**		41.76			**89.33**	56.18	54.15		33.20	69.15	56.97	42.25
TRAVIS-multi	41.52	62.46	34.28	41.91	**80.37**	90.34	64.68	55.34	54.09		35.46	72.94	57.93	45.04

Table 2: $E_{2,1}$ (%) evenness of GOLD and SYSTEMs annotations of PARSEME corpus

a good case study for diversity. The surface variability in the MWE occurrences can be observed in examples (10)–(12). All three examples contain the same VMWE with different surface forms. Next to the gloss (2nd line in each example) we report in parentheses the automatic lemmatization of the verb. In example (10) we can see the correct lemmatization ***(dava) aç***. Inadequate lemmatization can be observed in (11) and (12).

(10) **dava aç-tı**
dava aç-PAST (lemma: *aç*)
lawsuit open-PAST
'(someone) commenced lawsuit'

(11) **dava aç-ıl-abil-ir**
dava aç-PASS-POT-HAB (lemma: **açılab*)
lawsuit open-PASS-POT-HAB
'lawsuit could be commenced'

(12) **dava aç-ıl-acak**
dava aç-PASS-FUT (lemma: **açıla*)
lawsuit open-PASS-FUT
'lawsuit will be commenced'

All VMWE annotations were manually inspected in this corpus, to check the lemmas of all components and correct them if needed. Thus, the verb obtained the correct lemma *aç* in (11) and (12). After these corrections, the enhanced corpus was re-evaluated for richness and evenness. It was also used to retrain and re-evaluate Seen2Seen, one of the leading systems of the PARSEME shared task.

The results are presented in Table 3. In the corpus, the $E_{2,1}$ evenness becomes slightly lower, while the richness drops significantly, which signals that the number of types has decreased. This makes the Turkish corpus go down from the second to the fifth richest across all 14 languages in table 1. This outcome was expected since with correct lemmatization, items which were wrongly assigned to different types, like (10), (11) and (12), are now correctly assigned to one type.

As far as the MWEs correctly identified by Seen2Seen are concerned, evenness is almost stable but richness increases. This might be due to the fact that this system relies heavily on lemmas. It extracts the MWEs annotated in the training data and looks for co-occurrences of the same multisets of lemmas in the test corpus. If lemmas have better quality, this process is more efficient.

In brief, the main issue caused by inadequate lemmatization was that MWE items of the same type could be found in different clusters. The inadequate stripping of the suffixes resulted in more clusters (types) than there should be. These examples show the limits of diversity estimation on automatically annotated corpora. Namely, it heavily depends on the quality of the data. Here, when lemmatization is unreliable, both the richness and the evenness artificially go up. In other words, paradoxically, bad data quality "favors" diversity in this case. This experiment also shows that high diversity of a language at a certain level (here: morphological) might make the estimation of diversity at other levels (here: MWE level) less reliable.

		GOLD	Seen2Seen
Richness	TR	**719**	389
	TR'	660	**401**
$E_{2,1}$ (%)	TR	**58.59**	60.06
	TR'	58.14	**60.08**

Table 3: Diversity measures of GOLD and Seen2Seen on the Turkish corpus, before (TR) and after (TR') lemma corrections.

6 Discussion

In Section 4 we define items as correctly annotated MWE occurrences, meaning that incorrect annotations are ignored during diversity measurements. This was motivated by the fact that variety would otherwise be artificially increased by wrong MWE types and balance affected by the likely abundance of wrong MWE types with very few occurrences. As a consequence the variety of systems' annotations cannot be higher than that of the gold.

This is however not the case for balance (as can be seen in Figure 2). This raises the question of whether it is preferable for a system annotation to be more balanced than its target (gold) annotation or to be as close to its gold annotation as possible. When training identification systems the aim is usually to approach gold annotations as close as possible. This is however already measured through scores such as precision and balance. When used in conjunction with performance measures we believe that balance should be simply interpreted as "higher is better".

In Section 4 we choose to define our types through a lexical approach of MWEs. This decision was motivated by multiple reasons: (i) it is easy to implement, (ii) non-parametric, (iii) it results in clear-cut types (non-fuzzy), (iv) it offers a good granularity (for balance measures), (v) it is very similar to the PARSEME notion of MWE types and (vi) it is a quite adequate approximation of what we would consider linguistically motivated MWE types (MWEs sharing lexemes and meaning). In future work we might use alternative definitions of types, e.g. cluster MWEs having the same syntactic structure (verb-object, subject-verb, adjective-noun), the same semantics (*to* ***kick the bucket***, *to* ***bite the dust***), or the same MWE categories (light verb construction, verbal idiom).

7 Conclusions and Future Works

In this paper, we borrowed the formalization of the notion of diversity from the literature in ecology. We focused on two out of the three main aspects of diversity, namely variety and balance. Our contribution is to apply these measures to assess intralinguistic diversity, focusing on the particular phenomenon of multiword expressions. We not only formalize variety and balance measures in this context but we also put forward methods for selecting those variants of these measure which fit the nature of the MWE phenomenon. This validation methodology is based on corpus sampling with variable sample size. As a result, we retain richness and the $E_{2,1}$ evenness as the optimal variety and balance measures for MWEs (among those studied by us). We apply these measures to the corpora and system results in the PARSEME shared task on automatic identification of MWEs. The results show that richness of the correct annotations produced by the systems is roughly consistent with their F-measure performances. However, their balance is much less correlated with more traditional measures. We also display the limits of the richness and balance measures, when calculated on automatically annotated data, due to incorrect approximation of types under improper lemmatization in a morphologically rich language.

Further investigation, particularly on the evenness, is needed as an impressive amount of evenness measures have been proposed throughout the years. Furthermore, the notion of disparity was only briefly touched upon in this paper. Disparity measures might be based on lexical overlap between types, similarity of syntactic structures or distributional semantics. These directions constitute future work.

Acknowledgements

We are grateful to Emmanuel Schang for his remarks on previous versions of this paper. We also thank anonymous reviewers for their insightful reviews.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Rauno V Alatalo. 1981. Problems in the measurement of evenness in ecology. *Oikos*, pages 199–204.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second edition*, pages 267–292. CRC Press, Boca Raton.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Thomas François, and Philippe Blache, editors. 2016. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. The COLING 2016 Organizing Committee, Osaka, Japan.

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):415–479.

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.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448.

Michele Gazzola, Torsten Templin, and Lisa J. McEntee-Atalianis. 2020. Measuring diversity in multilingual communication. *Social Indicators Research*, 147:545–566.

Frédéric Gosselin. 2006. An assessment of the dependence of evenness indices on species richness. *Journal of theoretical biology*, 242(3):591–597.

Joseph H. Greenberg. 1956. The measurement of linguistic diversity. *Language*, 32(1):109–115.

Joseph H. Greenberg, editor. 1966. *Universals of language*, 2nd edition. MIT Press, Cambridge, MA.

David Harmon and Jonathan Loh. 2010. The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages. *Language Documentation and Conservation*, 4.

Mark O Hill. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Gaole He, Jinhao Jiang, Xiaoxuan Hu, Puzhao Xie, Zhipeng Chen, Zhuohao Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2021. TextBox: A unified, modularized, and extensible framework for text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 30–39, Online. 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.

Shashi Narayan and Shay B. Cohen. 2015. Diversity in spectral learning for natural language parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1868–1878, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Nettle. 1999. *Linguistic diversity*. Oxford University Press, Oxford.

Yağmur Öztürk, Najet Hadj Mohamed, Adam Lion-Bouton, and Agata Savary. 2022. Enhancing the parseme turkish corpus of verbal multiword expressions. In *Proceedings of The 18th Workshop on Multiword Expressions @LREC2022*, pages 100–104, Marseille, France. European Language Resources Association.

Enrico Palumbo, Andrea Mezzalira, Cristina Marco, Alessandro Manzotti, and Daniele Amberti. 2020. Semantic diversity for natural language understanding evaluation in dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 44–49, Online. International Committee on Computational Linguistics.

Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, et al. 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.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, et al. 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.

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*, pages 547–561. University of California Press.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

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(1):1–7.

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemi Zadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, et al. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *The 13th Workshop on Multiword Expression at EACL*, pages 31–47.

Benjamin Smith and J Bastow Wilson. 1996. A consumer's guide to evenness indices. *Oikos*, pages 70–82.

Andrew Stirling. 1998. On the economics and analysis of diversity. *Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper*, 28:1–156.

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

Hanna Tuomisto. 2012. An updated consumer's guide to evenness and related indices. *Oikos*, 121(8):1203–1218.

Martin L Weitzman. 1992. On diversity. *The Quarterly Journal of Economics*, 107(2):363–405.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2020. Trading off diversity and quality in natural language generation. *arXiv preprint arXiv:2004.10450*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.