Quantifying intra-linguistic diversity: Case study of multiword expressions

Agata Savary **Yağmur Öztürk** **Adam Lion-Bouton**
Paris-Saclay University, CNRS - LISN,
Grenoble Alpes University, LIG,
University of Tours, LIFAT

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, such as as ecology, economy or information theory (Morales et al., 2021). In linguistics it was mainly addressed in the inter-lingual sense (Greenberg, 1956; Nettle, 1999; Harmon and Loh, 2010) but has rarely been formalized intra-lingually, i.e. with respect to particular linguistic phenomena within one language. In NLP, the need for intra-lingual diversity in training data and its impact on performances of NLP tools has been stressed in parsing (Narayan and Cohen, 2015), question answering (Yang et al., 2018) and natural language generation (Zhang et al., 2020; Agirre et al., 2016; Zhu et al., 2018; Palumbo et al., 2020; Li et al., 2021). However, in these works the notion of diversity was either understood informally or used in a rather restricted sense.

Our objective is to address intra-linguistic diversity more formally, on the topic of multiword expressions (MWEs), i.e. combinations of words, such as ***keep tabs*** *on sth* or ***pull strings***, exhibiting idiosyncratic properties at lexical, morphological, syntactic and/or semantic level (Baldwin and Kim, 2010). This abstract summarizes and extends our work published in (Lion-Bouton et al., 2022).

2 Dimensions of diversity

Formal definitions of diversity often rely on the notions of *items* and *types*. In ecology, *items* are specimens/individuals, while *types* refer to the species these specimens are affiliated to. For us, an item denotes an MWE occurrence in text and a type is a multiword lexeme.

Given a population of items classified into types, the concept of diversity is often defined along three distinct dimensions: *variety*, *balance* and *disparity* (Stirling, 1998). *Variety* is the quantity of types into which items can be classified. *Balance* is the extent to which the type-item distribution is uniform. *Disparity* is the degree to which types differ from each other, according to a distance metric defined on types.

Example: Consider two toy corpora (1) and (2). Both contain 4 MWE occurrences distributed into 2 types, i.e. they are of equal variety. The distribution of tokens in types is uniform in (2) but less so in (1), i.e. (2) is more balanced. If we measure the distance between MWE types on the grounds of meaning and/or the (canonical) syntactic structure, then ***keep tabs*** and ***pull strings*** are closer (verb-noun combinations relating to the meaning of control) than ***keep tabs*** and ***go with the wind***. Therefore, (2) shows a higher disparity than (1). The bottom line is then that, (2) is more diverse than (1).

(1) They ignore the [**strings**]$_3$ he used to [**pull**]$_3$ and the [**strings**]$_4$ he [**pulls**]$_4$ now. They might [**keep tabs**]$_1$ on the [**strings**]$_2$ he will [**pull**]$_2$ in the future.

(2) [**Tabs**]$_1$ are [**kept**]$_1$ on some things, some others [**go with the wind**]$_2$. Whether you [**keep tabs**]$_3$ on them or you let them [**go with the wind**]$_4$ is hard to decide.

3 Data

Our diversity estimations use 3 datasets, henceforth called GOLD, PRED and SEQ, respectively: (i) PARSEME multilingual corpus of verbal MWEs (VMWEs) edition 1.2[1], and more precisely the TEST subcorpus for all languages, (ii) predictions of VMWE identification systems from the PARSEME shared task 1.2 (Ramisch et al., 2020), these prediction were made on the blind versions of the same TEST subcorpora, (iii) French Sequoia corpus annotated for all (also non-verbal) MWEs and named entities (NEs) (Candito et al., 2021).

The theoretical notions of diversity obviously materialize in data in an approximate way. Namely, what we count as items are all MWE/NE annotations in GOLD and SEQ, and true positives in PRED. Types are approximated by multisets of

[1] http://hdl.handle.net/11234/1-3367

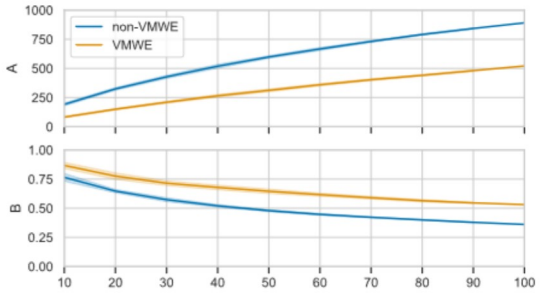

Figure 1: Richness (A) and Normalized Richness (B) in terms of SEQ sample size (in %).

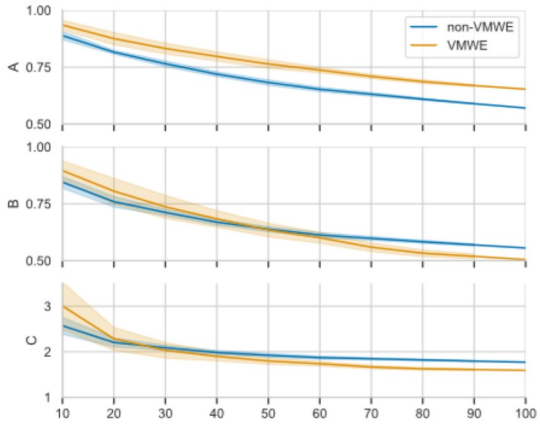

Figure 2: $E_{1,0}$ (A), $E_{2,1}$ (B) and Zipf balance (C) in terms of SEQ sample size (in %).

lemmas of tokens belonging to an MWE. For instance, the MWE types in (1) are represented as {*pull*,*string*} and {*keep*,*tab*}.

4 Experiments

The notions of variety, balance and disparity had various instantiations in past works and we checked the applicability of some of them to MWEs.

Variety is most often understood as richness (the number of types). It should increase with the size of the corpus, which we indeed observe on linearly growing samples of SEQ (Fig. 1). This growth is non-linear though and non-verbal MWEs consistently show higher variety than VMWEs.

We also measured normalized richness (richness divided by the number of items). We expected it to be more stable, but in fact it decreases with the corpus size. (Fig. 1). Thus, neither richness nor normalized richness are ideal to compare variety of differently sized corpora. They are useful though for evaluating systems on the same sample, which is the case in PRED. We observed that richness correlates with F-measure, e.g. the MTLB-STRUCT system (Taslimipoor et al., 2020) has the best global F-measure on PRED and also the highest richness in most languages. On the other hand, Seen2Seen (Pasquer et al., 2020), with the 2nd best global F-measure, has a weak richness since it focuses on identifying only previously seen VMWEs. Peculiarities of some systems can, thus, be highlighted if richness is used in evaluation.

For balance, a great number of measures have been proposed (Smith and Wilson, 1996; Tuomisto, 2012) but none proved universally optimal. We retained two evenness measures by Hill (1973) denoted $E_{1,0}$ and $E_{2,1}$, roughly based on a generalisation of Shannon's entropy. In Figure 2 we plotted them in function of linearly growing SEQ samples and compared them to a plot of the inverse of the estimated parameter s of the Zipfian distribution $Zipf(s, N)$ which we argue acts as the measure of balance of the Zipfian distribution. Since $E_{2,1}$ appeared more consistent with the Zipfian plot (and under the hypothesis that MWEs do exhibit Zipfian behavior), we prefer this version of evenness model balance in MWEs.

In experiments, our $E_{2,1}$ scores are mostly higher in PRED than in GOLD. They hardly coincide with global F-measure, which shows their complementarity with global performance measures, often biased by more frequent but less diverse phenomena.

Disparity also has many different instantiations, depending on how distance between types is defined and aggregated. One potential problem is the complexity of disparity calculation, e.g. if all items across all types have to be compared pairwise.[2] Therefore, as a first step, we propose to measure the distance between two MWEs as the difference between their meanings, approximated by static word embeddings (WEs). Note that, differently from contextual WEs, static WEs already aggregate items into types, since a single vector is an abstract representation of the meaning of a word independently of its context. Of course, due to the semantic non-compositionality of MWEs, their sense cannot be reliably represented by a simple combination of the WEs of their component

[2]For instance, the distance between two MWE types could be defined as the minimum distance between two items belonging to these two types.

words. Instead, we train custom MWE-aware WEs. First, a large raw corpus of French[3][4] extended with the French part of GOLD is automatically annotated for VMWEs with MTLB-STRUCT. Then the corpus is re-tokenized so that the components of discontinuous MWEs are rearranged and merged to become single tokens. The resulting corpus is used to train word2vec embeddings.

The distance between two MWE types t_1 and t_2 is derived from the cosine similarity of their WEs, i.e. $d(t_1, t_2) = (1 - cos(t_1, t2))/2$. Then, disparity is the average of the pairwise distances between all types observed in the set of items. Preliminary experiments tend to show that disparity decreases with the growing size of a corpus sample. Like for variety, disparity thus defined is of limited interest if MWEs in corpora of different sizes are compared. But, it can be usefully applied to PRED as yet another alternative quality measure. Performing these experiments with disparity-driven evaluation of VMWE identification is our ongoing work. In near future we would also like to experiment with other definitions of disparity e.g. based on morpho-syntactic features, lexical composition of MWEs, etc.

Our long-term objective is to integrate all three diversity measures in NLP evaluation and benchmarking for a variety of linguistic phenomena. We hope this can protect and promote intra-linguistic diversity in NLP resources and tools.

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.

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.

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.

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

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.

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.

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.

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.

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.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer

[3] https://gitlab.com/parseme/corpora/-/wikis/Raw-corpora-for-the-PARSEME-1.2-shared-task

[4] We chose French for these initial experiments because the quality of VMWE identification in this language is high and all co-authors are proficient enough in this language to discuss the results.

to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

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.

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.

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

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.