Formalising lexical and syntactic diversity for data sampling in French

Anonymous ACL submission

Abstract

Diversity is an important property of datasets and sampling data for diversity is useful in dataset creation. Finding the optimally diverse sample is expensive, we therefore present a heuristic significantly increasing diversity relative to random sampling. We also explore whether different kinds of diversity – lexical and syntactic – correlate, with the purpose of sampling for expensive syntactic diversity through inexpensive lexical diversity. We find that correlations fluctuate with different datasets and versions of diversity measures. This shows that an arbitrarily chosen measure may fall short of capturing diversity-related properties of datasets.

1 Introduction

Linguistic diversity is gaining a lot of attention in recent NLP research, for a lot of practical reasons. For instance, a desirable property of models is to perform comparatively well on diverse sets on languages (Liu et al., 2024a; Gueuwou et al., 2023). Outputs of generative models, notably in dialog systems, are expected to be not only accurate but also diverse (Gao et al., 2019; Alihosseini et al., 2019; Xiang et al., 2024; Kim et al., 2024). More diverse parameters (*e.g.*, attention vectors) make a model more relevant (Huang et al., 2019) and less sensitive to adversarial attacks (Yang et al., 2024). Diversity of human productions is seen as upper bound for the diversity of a model's output (Liu et al., 2024b). Diversity of output data being close to the one of input data is an indicator of fairness, representativeness or accurateness, *e.g.*, in summarisation (Zhang et al., 2024) and active learning (Xia et al., 2024). More diverse training data lead to better system performances *e.g.* in semantic parsing (Liu and Zeldes, 2023), dialog systems (Larson et al., 2019), or question answering (Yadav et al., 2024). To achieve such diverse datsets, data sampling based *e.g.* on embedding outlier detection (Larson et al., 2019) or uncertainty-and-diversity-based active learning (Kim, 2020) can be used.

In these works, the notion of diversity is used rather loosely, or is based on measures which are selected in an ad hoc way. We believe that this is largely due to the unawareness of long-standing formal approaches to diversity developed in other scientific domains, notably ecology (*cf.*, §2).

In this paper we are interested in applying ecology-inspired diversity measures, notably entropy, to the problem of data sampling. We aim at building a large corpus of French, automatically parsed for morphosyntax, with two constraints. Firstly, it should be large but manageable, *i.e.* its parsing, storage and maintenance cost should not be prohibitive. Secondly, it should still have sufficient lexical and syntactic diversity to cover long-tail phenomena.[1]

To this aim, we use very large raw corpora and we rely on formal lexical and syntactic diversity measures. We face two tractability problems: (i) while lexical diversity can easily be calculated for a raw text, finding the optimally diverse subset of a very large set of texts is intractable, (ii) syntactic diversity quantification requires the data to be parsed in advance, which is prohibitive with very large data. In this context, we propose a data sampling heuristic which is faster than an optimal method. We address the following research questions:

Q1 Can this method select a corpus whose diversity is significantly higher than at random?

Q2 Can lexical diversity help construct a syntactically diverse large corpus?

The experimental results allow us to give a positive answer to **Q1**. Concerning **Q2**, we show that it

[1] Long-tail phenomena are important *e.g.* in frame induction (QasemiZadeh et al., 2019), zero-shot identification of multiword expressions (Ramisch et al., 2020), probing language models for rare but interesting syntactic phenomena (Misra and Mahowald, 2024; Weissweiler et al., 2024), etc.

is necessary not to fix one's attention on one diversity measure (here Shannon-Weaver entropy) but to rather examine its generalisation (Rényi, 1961).

2 Diversity measures

Formal approaches to diversity (Morales et al., 2020; Chao et al., 2014) apportion **elements** into **categories**. For instance in ecology – a field with a long history of formal diversity –, species are categories and individuals are elements.

We apply this principle to NLP datasets. For **lexical diversity**, categories are unique token forms (henceforth: *forms*) and elements are their occurrences. For instance in Fig. 1, datasets ***LvHb*** (Low Variety High Balance) and ***HvLb*** (High Variety Low Balance) have $m = 10$ elements each, and $n = 8$ and $n = 9$ categories, respectively. For **syntactic diversity**, categories are complete syntactic subtrees containing only POS labels and dependency relations, and elements are all instances of these subtrees. For instance, in Fig. 1, ***LvHb*** and ***HvLb*** have 5 and 8 categories, respectively (*cf.*, Appendix 1), including the ones shown on the right, having two elements highlighted in blue and red.

Given n categories and m elements, diversity is measured along 3 dimensions: **variety**, **balance**, and **disparity** (Morales et al., 2020). We here only discuss variety and balance.

Variety tackles the number of categories: a (pure) variety function is monotonic to n and its most basic form is richness, *i.e.*, just n. Thus, a habitat with n species is more varied than with $n - 1$, and ***HvLb*** with 9 forms and 7 subtress is lexically and syntactically more varied than ***LvHb*** with 8 forms and 5 subtrees.

Balance tackles the distribution of elements in categories. A (pure) balance function reaches its maximum when all categories are equiprobable. Thus, a habitat with 50 octopuses and 50 squids is more balanced than one with 60 and 40. ***LvHb*** with 2 elements in each subtree is syntactically more balanced than ***HvLb*** with subtrees having more elements than others.

A large number of measures was proposed for variety and balance in the past (Hill, 1973; Smith and Wilson, 1996) and many are hybrids between the two. For conciseness, we restrict our study to entropies, as they encompass many diversity functions and have a strong background. Consider the non-zero probability distribution of categories $\Delta = \{p_1, ..., p_n\}$. Shannon and Weaver (1949) entropy H may thus be computed as (here $b = e$):

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

When $\forall p_i \in \Delta, p_i = \frac{1}{n}$, H reaches its maximum $\log_b(n)$. It is thus hybrid of variety and balance. (Rényi, 1961) generalizes entropy as H_α, which is pure variety when $\alpha = 0$. With growing α, H_α accounts more for balance and less for variety:

$$H_1 = H; H_{\alpha \neq 1}(\Delta) = \frac{1}{1-\alpha} \log_b \left(\sum_{i=1}^{n} p_i^\alpha \right) \quad (2)$$

Using data of our toy examples, we find:

	LvHb		*HvLb*	
Dataset	Lex.	Syn.	Lex.	Syn.
H_0	2.079	1.609	2.197	2.079
H_1	2.025	1.609	2.163	1.973
H_2	1.966	1.609	2.120	1.832

3 Source data

We wish to build a large syntactically parsed French corpus, with sufficient lexical and syntactic diversity to serve lexicon induction and morphosyntactically-based studies. We use the FTB-dep annotation schema (Candito et al., 2010) for its ability to finely represent French syntax.

Our first source, now called *BASE*, is BigScience (Laurençon et al., 2022), and more precisely its three subcorpora: Europarl, French part of the United Nations Corpus and Wikipedia[2], for a total of 833K documents and 1.5 billion tokens. Their advantages are large sizes, clearly identified sources and genres (parliamentary debates and encyclopedia entries), few multilingual texts, and clear licenses inspired by openness and fairness.

A major disadvantage of *BASE* is not to be sufficiently diverse as far as genres are concerned, which likely influences lexical and syntactic diversity. Therefore, we use HPLT (De Gibert et al., 2024), a massive multilingual dataset provided by Internet Archive and CommonCrawl. In the *cleaned* version of HPLT v1.2, texts were partly filtered for dubious sources (pornographic, racist, *etc*) and noisy paragraphs. They were also divided into languages by majority vote over a number of language predictors. We use the texts assigned to French, for a total of 99.59M documents and 122.88B tokens. Less clean than BigScience (due

[2]Wikisource was also considered, but had many issues.

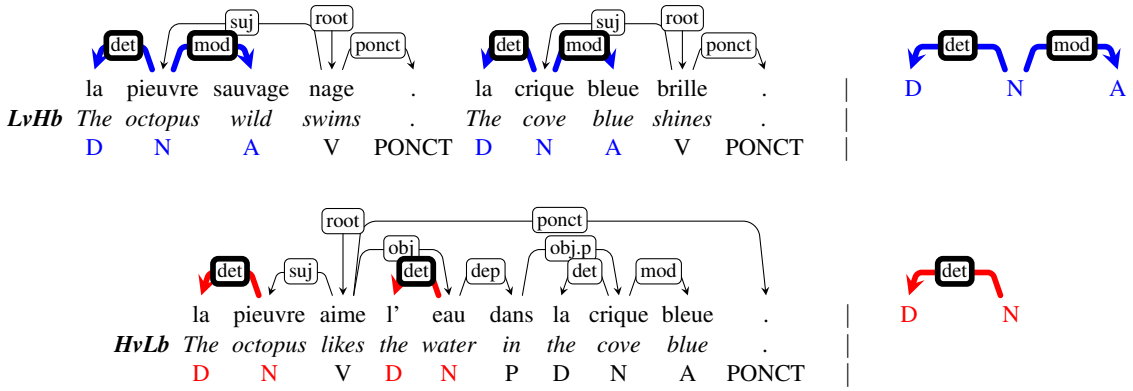


Figure 1: Two toy datasets with sample syntactic categories (on the right) and elements (inside the sentences).

to imperfect filtering and language prediction), this dataset covers a wide range of fields and will be useful to increase the diversity of $BASE$.

4 Diversity-driven data selection

As HPLT contains too much text, we select only a number of tokens similar to BigScience (*i.e.*, around 1.5 billion). To this aim, we use a data augmentation process driven by lexical diversity, which we define here as entropy (*cf.*, §2), where elements are tokens and categories are unique forms.

We start by randomly selecting around 6 billion documents from HPLT, to keep computation time reasonable. Henceforth, this subset is called HPLT_{small}. We then apply Algorithm 1. Its objective is to extend a dataset with a restricted number of new documents, while maximising lexical diversity. Precisely, given an initial dataset I (here $BASE$), a new dataset N (here HPLT_{small}), and a target size S (here 3.1 billion), we select documents from N which, added to I, increase I's size to S while maximising lexical diversity. Documents from N selected this way to be added to I are called $N_{diverse}$ (here $\text{HPLT}_{diverse}$), and the final corpus, containing I and $N_{diverse}$ is called $TOTAL$.

The optimal choice of $N_{diverse}$ would require a very high computational cost. Consider a dataset as a set of undividable blocks $\mathbb{A}$. (*e.g.*, a set of sentences or a set of texts). The question is: which subset $\mathbb{B} \subseteq \mathbb{A}$ maximises H_α? An exhaustive search of $\mathcal{P}(\mathbb{A})$ – the power set of $\mathbb{A}$, the set of all possible subsets of $\mathbb{A}$ – would take $O\left(2^{|\mathbb{A}|}\right)$, which is not tractable. Therefore, we use a heuristic to approximate it.

We start with the working corpus W equal to I (l. 1). In the internal loop (l. 5-15), we consider one candidate document n from N at a time. We filter and normalise it (l. 6) to avoid artificial increase in diversity.[3] We check if, added to W, n increases entropy (l. 7). If so, we check if this increase in entropy is higher than for a previously found document d (l. 9). If so, d becomes n (l. 10). Note that, the next optimal document d to add is not chosen from the whole corpus N but, for tractability, we stop when we found "enough" documents increasing entropy (l. 8 and 11). Then we pick the best of them d and add it to the working corpus W (l. 12).

Variable e, for exhaustivity of search, tells us how many documents we have to look at before we pick the optimal one to append to W. If e is too high, we might not reach the intended size S. Therefore, the array E gives several exhaustivity values in decreasing order. If S has not been reached, we become less exhaustive, *i.e.*, we go to the next e.

We return W (which then becomes $TOTAL$) as soon as it has exceeded S (l. 15), or when all exhaustivity levels have been considered.

5 Diversity evaluation

The initial lexical H of $BASE$ is 7.02 for 1,538,617,909 tokens. Applying Algorithm 1, $TOTAL$ obtains an entropy $H_{diverse}$ of 7.98. While 0.96 may appear as a small increase, it is in fact large due to the logarithmic nature of entropy. To verify the effectiveness of Algorithm 1 in selecting a diverse corpus, we will answer the two research questions defined in §1.

[3] (Telephone) numbers, HTML and XML tags, URLs, file paths, emoticons, series of punctuations, phonetic characters, series of alphanumerical tokens, and characters outside of the French range are represented by a unique token for each category, *e.g.*, `[NUMBER]`.

Algorithm 1 Algorithm to heuristically sample data while maximising entropy.

Require: I, an initial dataset
Require: N, another dataset to select from
Require: E, a decreasing array of exhaustivity search parameters (positive integers)
Require: S, maximum size of resulting dataset
Ensure: I increased with fragments of N (while maximising entropy)

```
W ← I, the working dataset
for each e ∈ E do, for each exhaustivity level
    f ← 0, counter for current exhaustivity
    d ← ∅, document to append
    for each n ∈ N do, for each document
        n ← normalised (n)
        if H (W ∪ n) > H (W) then
            f ← f + 1
            if H (W ∪ n) > H (W ∪ d) then
                d ← n
            if f = e then
                W ← W ∪ d
                f ← 0
                d ← ∅
                if |W| > S then return W
return W
```

Q1 To test **Q1**, in addition to $N_{diverse}$, we construct N_{random} by randomly selecting sentences from $HPLT_{small}$ until reaching S. We repeat this with 20 different seeds. We test whether entropy for these 20 random samples follows a normal distribution $H_{random} \sim \mathcal{N}(x; \mu, \sigma)$. We consider the null hypothesis h_0 to be H_{random} *follows a normal distribution*, and the alternative hypothesis h_1 to be *it does not follow a normal distribution*.[4] As it is not significant ($p \approx 0.26$), it is likely we have a normal distribution.

For this normal distribution, $\mu \approx 7.656$ and $\sigma \approx 9.027\text{e}-4$, and the value of $H_{diverse}$ is away from μ by $\approx 347\sigma$. The p-value is ≈ 0.[5] Thus, $N_{diverse}$ is highly significantly more diverse than at random, which confirms the effectiveness of Algorithm 1.

Q2 Checking the correlation between lexical and syntactic diversity requires the corpus to be parsed, which is prohibitive even for HPLT_{small}. Since $BASE$ has already been parsed, we use it for the correlation estimation. We split it into 705 blocks of 100K sentences. Lexical and syntactic H_α with α ranging from 0 to 5 are calculated for each block. Spearman and Pearson correlations are calculated at each level of α.

We find that the Correlation scores between Lexical and Syntactic Diversities (CLSD) cannot easily be exploited (*cf.*, Fig. 2). Firstly, the choice of α highly impacts CLSD. The precise choice of α (*i.e.*, whether considering more variety or more balance) must thus be thought through. Secondly, across datasets and α values, CLSD is not constant. Approximating syntactic through lexical diversity would therefore require some amount of parsing to see if for the studied dataset CLSD are usable. Thirdly, within a specific genre, CLSD do not seem consistent: we see little resemblance between Europarl and UN corpus, despite both of them being parliamentary debates. Finally, the union of datasets (*cf.*, solid curves) shows special properties: while Pearson correlation is positive or slightly below 0 for subcorpora (Europarl, UN corpus, and Wikipedia), the union dips at -0.43 near $\alpha = 1$.

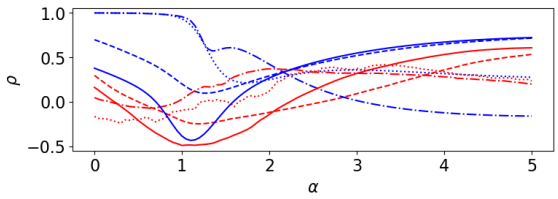


Figure 2: Correlation between lexical and syntactic H_α, according to α. Europarl (dotted), UN corpus (dashed), Wikipedia (dash-dotted), and the union of all three (solid). Blue for Pearson, red for Spearman.

These results might be interpreted by non-compositionalty of the diversity calculus. Given two datasets $D1$ and $D2$, the diversity of their union depends on their similarity. For instance, the variety of the union depends on if $D1$ and $D2$ have disjoint categories or not. The balance of the union can be high when $D1$ and $D2$ are unbalanced but share categories with inverse distribution patterns. More insight is needed to exploit these results.

These results do not provide a definitive conclusion for **Q2**. The use of lexical diversity did not increase syntactic diversity, but there are still research directions to be explored in this area.

6 Conclusions and future work

We proposed an algorithm for diversity-driven data sampling which is tractable compared to optimal solution and still significantly increases the lexical diversity of a dataset. We showed that correlation between lexical and syntactic diversities is not reliable enough for syntactic diversity to be approximated by lexical diversity. The algorithm may be use to sample for high syntactic diversity, but more insights are needed to reduce annotation cost.

[4] See Python's `scipy.stats.normaltest`.
[5] p is lower than the precision of a 64-bit double.

7 Limitations

The block size in the second experiment is arbitrary, as such we cannot ensure that another value would have yielded the same results. Our experiments are limited to French. The way the algorithm is coded favours sentences at the beginning of files, and longer documents. The syntax used to compute the syntactic diversity is predicted and as such contains errors. Some rare phenomena might especially be badly predicted, which may impact the diversity scores.

8 Ethical statement

The algorithm we presented in this article is agnostic of the data. As such, when given a "clean" *BASE* (*i.e.*, that does not contain undesirable content), it tends to select previously unseen data, since such data often increase diversity scores. The selected data may have inappropriate content (*e.g.*, pornographic or racist data), even after filtering for data sources, as filters are often imperfect.

The authors have no known conflict of interest with the authors of source data.

References

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical French dependency parsing: Treebank conversion and first results. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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.

Ona De Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer Van Der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, et al. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. 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 1229–1238, Minneapolis, Minnesota. Association for Computational Linguistics.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. 2023. JWSign: A highly multilingual corpus of Bible translations for more diversity in sign language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9907–9927, Singapore. Association for Computational Linguistics.

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.

Po-Yao Huang, Xiaojun Chang, and Alexander Hauptmann. 2019. Multi-head attention with diversity for learning grounded multilingual multimodal representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1461–1467, Hong Kong, China. Association for Computational Linguistics.

Yejin Kim, Scott Rome, Kevin Foley, Mayur Nankani, Rimon Melamed, Javier Morales, Abhay K. Yadav, Maria Peifer, Sardar Hamidian, and H. Howie Huang. 2024. Improving content recommendation: Knowledge graph-based semantic contrastive learning for diversity and cold-start users. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8743–8755, Torino, Italia. ELRA and ICCL.

Yekyung Kim. 2020. Deep active learning for sequence labeling based on diversity and uncertainty in gradient. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 1–8, Suzhou, China. Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019. Outlier detection for improved data quality and diversity in dialog systems. 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 517–527, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral,

Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2022. The bigscience roots corpus: A 1.6tb composite multilingual dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024a. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Guisheng Liu, Yi Li, Zhengcong Fei, Haiyan Fu, Xiangyang Luo, and Yanqing Guo. 2024b. Prefix-diffusion: A lightweight diffusion model for diverse image captioning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12954–12965, Torino, Italia. ELRA and ICCL.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. *Preprint*, arXiv:2403.19827.

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].

Behrang QasemiZadeh, Miriam R. L. Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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

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.

Claude Elwood Shannon and Warren Weaver. 1949. *A Mathematical Theory of Communication*. University of Illinois Press, Urbana.

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].

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2024. Hybrid human-llm corpus construction and llm evaluation for rare linguistic phenomena. *Preprint*, arXiv:2403.06965.

Yu Xia, Xu Liu, Tong Yu, Sungchul Kim, Ryan Rossi, Anup Rao, Tung Mai, and Shuai Li. 2024. Hallucination diversity-aware active learning for text summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8665–8677, Mexico City, Mexico. Association for Computational Linguistics.

Jianxiang Xiang, Zhenhua Liu, Haodong Liu, Yin Bai, Jia Cheng, and Wenliang Chen. 2024. DiffusionDialog: A diffusion model for diverse dialog generation with latent space. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4912–4921, Torino, Italia. ELRA and ICCL.

Vikas Yadav, Hyuk joon Kwon, Vijay Srinivasan, and Hongxia Jin. 2024. Explicit over implict: Explicit diversity conditions for effective question answer generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6876–6882, Torino, Italia. ELRA and ICCL.

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

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2024. Fair abstractive summarization of diverse perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3404–3426, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

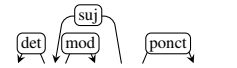
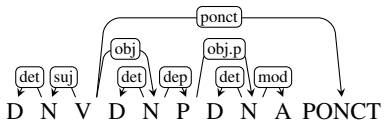
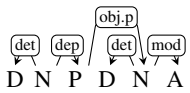
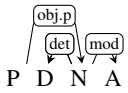
Dataset	Lexical		Syntactic	
	Category	Nb Elem	Category	Nb Elem
LvHb	.	2		2
	la	2		2
	bleue	1	D	2
	brille	1	A	2
	crique	1	PONCT	2
	nage	1		
	pieuvre	1		
	sauvage	1		
HvLb	la	2		1
	.	1		1
	aime	1		1
	bleue	1		1
	crique	1		1
	dans	1	D	3
	eau	1	A	1
	l’	1	PONCT	1
	pieuvre	1		

Table 1: Extracted categories per dataset in Figure 1.