Diversity quantification in natural language processing

Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik,
Agata Savary, Olha Kanishcheva

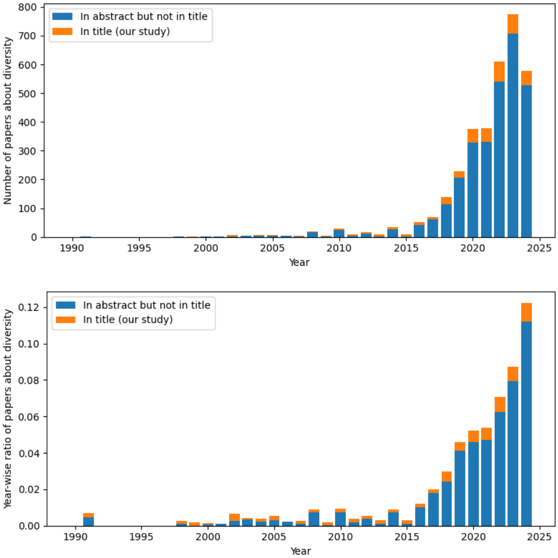
ZIL seminar, IPIPAN Warsaw, 16 June 2025

UniDive cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
Funded by the European Union
LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

A survey of diversity quantification in natural language processing: The why, what, where and how

- UniDive CA21167 COST action: Working Group 4 "Quantifying and promoting diversity"
- International collaboration:
  - Louis Estève, LISN, Université Paris-Saclay, France
  - Marie-Catherine de Marneffe, Université Catholique de Louven, Belgium
  - Nurit Melnik, The Open University, Israel
  - Agata Savary, LISN, Université Paris-Saclay, France
  - Olha Kanishcheva, Heidelberg University, Germany, SET University, Kiev, Ukraine
- TACL submission (under review)

Diversity: prevalence in NLP



Papers in the ACL Anthology from 1990 until 2024-07-26 with “diversity” or “diverse” in their title or abstract (3,653 in total)

Motivations and objectives

Diversity: lack of formalisation in NLP

- Disparate understanding of the notion of diversity
- Lack of a unified vocabulary and framework
- Limited attemps to systematize the notion of diversity
  [Tevet and Berant(2021), Ploeger et al.(2024)]
- NLP belongs to the “fields [. . .] where diversity is prominent in discussion, but remains undefined or analytically neglected”
  [Stirling(2007)]

Objectives

- Take steps towards a unified framework for **quantification** of diversity in NLP
- Take inspiration from fields where diversity has been theorized and systematically analysed, most prominently **ecology**

Contributions

- A review of the **308 papers** from the past 6 years containing "diverse" or "diversity" in their title, from the point of view of **diversity quantification**
- A **taxonomy** allowing to position various approaches:
  - motivations behind the quest for diversity (**why**),
  - objects on which diversity is quantified (**what**),
  - pipeline stages where diversity measures are applied (**where**)
  - types of diversity measures (**how**)

Theory of diversity in ecology (& Co.)

Mature topic

- Dozens of diversity **measures** defined [Smith and Wilson(1996)] and applied to various species and their habitats
- Measures borrowed from **information theory**: parameterized entropies [Patil and Taillie(1982)] and related transformations [Hill(1973)].
- **Distance** measures (underlying diversity) based on functional differences (body features, behavior, etc.) and positions in the pophylogenetic tree [Mouchet et al.(2010)].
- **Unified frameworks** [Leinster and Cobbold(2012), Scheiner(2012), Chao et al.(2014)].
- Debates on **properties** of diversity measures measures [Smith and Wilson(1996), Hoffmann and Hoffmann(2008), Jost(2009)].

Element/category dichotomy

- **Elements** (e.g. individuals) are apportioned into **categories** (e.g. species)

Dimensions of diversity (pl: *różnorodność*)

- **Variety** (pl: *rozmaitość*) – related to the number of categories
- **Balance** (pl: *równowaga/zrównoważenie*) – evenness of the distribution of elements in categories
- **Disparity** (pl: *zróżnicowanie*) – extent of the differences between categories

Why diversity is important in NLP

	Ethical	**Practical**
Goal	inclusiveness, equality, fairness	user expectation, naturalness
Means	deontology, best practices	performance, informativeness

Why diversity is important in NLP

Ethical reasons

- Diversity as a **goal**:
  - digital *inclusiveness* [Joshi et al.(2020)]
  - *equally* serving all users [Khanuja et al.(2023), Liu et al.(2024a)]
  - representing different *languages*, *language families* and *scripts* [Kodner et al.(2022), Goldman et al.(2023)]
  - mitigating the supremacy of English and *English-centric bias* [Pouran Ben Veyseh et al.(2022), Asai et al.(2022)]
  - *fair* account for diverse *cultures* [Yin et al.(2021), Mohamed et al.(2022), Keleg and Magdy(2023), Bhatia and Shwartz(2023), Liu et al.(2024a)], human *perspectives* [Parrish et al.(2024)] and *opinions* [Zhang et al.(2024)]
  - cover a large variety of *topics* in education [Hadifar et al.(2023)].
- Diversity as a **means** to achieve a goal
  - diverse prompt-response pairs ⇒ *less offensive* LLM answers [Song et al.(2024)]
  - diverse attention vectors ⇒ low sensitivity to adversarial attacks [Yang et al.(2024)]
  - diverse benchmark ⇒ reliable evaluation [Chen et al.(2023b)]
    - showing out-of-domain performance [Pradhan et al.(2022)].
    - highlighting the remaining challenges [Kim et al.(2023c)]
    - dataset's diversity more critical in evaluation than its size

Why diversity is important in NLP

Practical reasons

- Diversity as a **goal**:
  - inherent property of human language ⇒ *user expectation* towards machine-generated language
  - need for **one-to-many** scenarios: diverse spectrum of outputs rather than a single most optimal output [Kumar et al.(2019), Liu et al.(2020), Han et al.(2021), Shao et al.(2022), Puranik et al.(2023), E et al.(2023), Hwang et al.(2023)]
  - high diversity expectations in **dialog** [Lee et al.(2022)]: diverse system's reactions ⇒ highr user's engagement [Akasaki and Kaji(2019), Kim et al.(2023b)]
  - **naturalness**: diversity of human language ⇒ upper bound for systems [Schüz et al.(2021), Cegin et al.(2023), Liu et al.(2024b)]
- Diversity as a **means** to achieve a goal :
  - diverse training data ⇒ higher performance [Narayan and Cohen(2015), Liu and Zeldes(2023), Yang et al.(2018), Yadav et al.(2024), Tripodi et al.(2021), Shen et al.(2022), Li et al.(2016), Agirre et al.(2016), Zhu et al.(2018), Zhang et al.(2021), Thompson and Post(2020), Palumbo et al.(2020), Li et al.(2021)]
  - ensemble model with diverse submodels ⇒ better performances than a unique model [Song et al.(2021), Kobayashi et al.(2022)]
  - diverse keywords in class labels ⇒ more accurate classification [Yano et al.(2024)].
  - diverse generated text ⇒ less generic and more informative for users

Why diversity is important in NLP: Tendencies

Quest for diversity

- Most works advocate for an **increase of diversity**
- Few posit adjustment to the task: factual ⇒ low diversity, storytelling ⇒ high diversity
- Few see lower diversity of AI vs. human language as opportunity: bot detection, fack checking, protection of democracy

Quality/diversity trade-off

- Opposing objectives: quest for diversity vs. generative quality and consistence
  [Ma et al.(2024), Ippolito et al.(2019), Zhang et al.(2021), Shao et al.(2022), Chen et al.(2023a)]

Interest in theorizing diversity

- better understanding of the nature of **typological** diversity [Ploeger et al.(2024)]
- making **educated choices** of diversity measures
  [Tevet and Berant(2021), Lion-Bouton et al.(2022)]
- **comparative framework** in typological diversity for NLP [Poelman et al.(2024)]
- Our work

What diversity is measured on

In-text diversity

- Categories are inherent to a text: unique words, unique n-grams, sentences, syntactic trees
- Elements: word occurrences, n-gram occurrences, sentences, occurrences of syntactic trees

Meta-linguistic diversity

- Categories are metada of text: language, language family, branch in a phylogenetic tree, genre, domain, time period, racial identity or politycal opinion of the text author
- Elements: texts, language

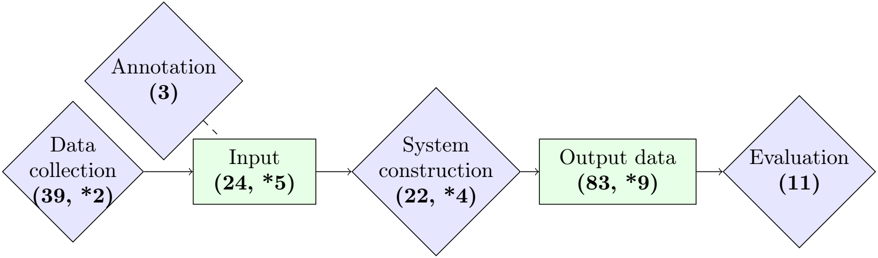
Diversity of processing

- Categories = elements: annotators, models (in an ensemble), NLP tasks, evaluation metrics, attention vectors
- diverse = several different

Where diversity is measured

NLP area	**#**
Generation	61
Corpus creation	22
Classification	17
Dialogue	15
Machine translation / Paraphrasing	10
Question answering / Summarization	9
Modeling	7
Recommendation / Parsing	5
Evaluation / Information extraction / Language Technology	4
Morphology / Vision / Inference	3
Speech / Survey or Opinion paper	2
Matching / Spellchecking	1

Where diversity is measured



How diversity is measured

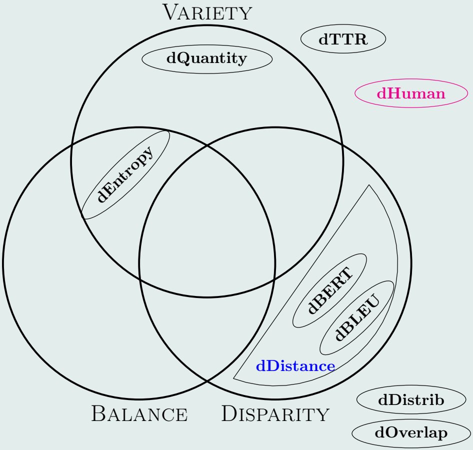
- 197 papers with actual quantifcation of diversity
- 150 different diversity measures ⇒ we group them into **3 approaches** and **9 families**
- 3 types of approaches:
  - **Absolute** quantification: a diversity score for the observed set independently of other sets
  - **Relative** quantification: a diversity of the observed set by comparison to a reference set
  - **Introspective** quantification: rank or score on a scale, by human judgement

How diversity is measured

Family of diversity measures	#
dQuantity: count categories	55
dBLEU: use BLEU for distances	41
dDistance: quantify differences between categories	37 + 8 + 41
dTTR: use the number of categories and normalize it by the number of elements	30
dEntropy: calculate unpredictability of categories	21
dOverlap: find the overlap between the categories in the observed set and in a reference set	9
dBERT: use BERT's contextual vector space for distances	8
dHuman: rely on a human evaluation	7
dDistrib: use the distance between observed and reference distributions	3
dOther: other measures	36

How diversity is measured

Diversity measures in NLP cast on 3 dimensions



Absolute quantification

Absolute quantification

Assigning a diversity score to the **observed set** independently of other sets.

Parameters

- n – number of (observed) categories
- m – number of (observed) elements
- $P = \langle p_1, ..., p_n \rangle$ – distribution of categories,
- $D = \langle \langle d_{1,1}, ..., d_{1,n} \rangle, ..., \langle d_{n,1}, ..., d_{n,n} \rangle \rangle$ – pairwise distances between the categories.

Absolute quantification

dQuantity $\in$ Variety

Variants of:

$$\texttt{richness}\,(n, m, P, D) = n \tag{1}$$

e.g. number of languages, language families, genres etc. in a dataset (meta-linguistic diversity).

dTTR $\notin$ {Variety, Balance, Disparity}

Variants of

$$\texttt{type-token-ratio}\,(n, m, P, D) = \frac{n}{m} \tag{2}$$

Frequently: `Distinct-n`, `Dist-n` or `Diverse-n`:

- ratio of **distinct n-grams** to the total number of tokens [Li et al.(2016)], $n \in [1, 4]$
- issues: not monotonic to n

Absolute quantification

dEntropy $\in$ {Variety, Balance}

Mostly [Shannon and Weaver(1949)]:

$$\texttt{entropy}\,(n, m, P, D) = \sum_{i=1}^{n} p_i \log_b \left(p_i^{-1}\right) \tag{3}$$

Monotonic with n. Maximum value $log_b(n)$ with uniform distribution.

dDistance $\in$ {Disparity}

- Aggregation and normalization of pairwise ditances between categories [Kim et al.(2024)], complexity $O\left(n^2\right)$:

$$\texttt{pairwise}\,(n, m, P, D) = \frac{2 * \sum_{i=1}^{n} \sum_{j=1}^{i-1} d_{i,j}}{n(n-1)} \tag{4}$$

- Volume of the geometry formed by vector vertices, e.g. convex hall [Yang et al.(2024)]
- Entropy of distances [Yu et al.(2022)]

Absolute quantification

dBLEU $\in$ dDistance $\in$ {Disparity}

Mostly average of BLEU between two texts [Zhu et al.(2018)], variant of `pairwise`:

$$\text{Self-BLEU}\left(n, m, P, D\right) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \text{BLEU}\left(\text{sent}_i, \text{sent}_j\right)}{n^2} \tag{5}$$

The higher BLEU, the larger the diversity.

dBERT $\in$ dDistance $\in$ {Disparity}

Mostly BERT score [Zhang* et al.(2020)]: F-measure between two texts X and Y:

$$R_{\text{BERT}} = \frac{1}{|X|}\sum_{x_i \in X} \max_{y_j \in Y} \vec{x_i}^\top \vec{y_j} \tag{6}$$

$$P_{\text{BERT}} = \frac{1}{|Y|}\sum_{y_j \in Y} \max_{x_i \in X} \vec{y_j}^\top \vec{x_i} \tag{7}$$

$$F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{8}$$

Relative quantification

Relative quantification

Assigning a diversity score to the **observed set** O in comparison to a **reference set** R. Two opposed variants:

- R is considered diverse, e.g. it is curated with diversity in mind, and O should be as close as possible to R [Samardzic et al.(2024)]
- O is expected to differ from R, e.g. generated utterances should be different from the training utterances [Murahari et al.(2019)]

dDistrib $\notin$ {Variety, Balance, Disparity}

Distributions P, Q of categories in R and O are compared, e.g.:

$$\texttt{cross-entropy}\left(Q, P\right) = \sum_{i=1}^{n} q_i \log_b \left(p_i^{-1}\right) \text{ (pl: } \textit{entropia krzyżowa}) \tag{9}$$

dOverlap $\notin$ {Variety, Balance, Disparity}

Categories in R and O are compared, e.g.:

$$\texttt{Jaccard}\left(n_R, n_O\right) = \frac{|n_R \cap n_O|}{|n_R \cup n_O|} \tag{10}$$

Introspective quantification

dHuman $\notin$ {Variety, Balance, Disparity}

Humans are asked to judge diversity by:

- ranking text samples for diversity [Liu et al.(2023)]
- scoring text samples along a diversity scale [Kim et al.(2023a)]

We cannot a priori know if humans rely on categories and elements for their judgment.

Discussion

- Diversity – prevalent concept in NLP
- . . . but used informally
  - 28% relevant papers judge diversity important without defining it
  - Further 20% only count the number of categories
  - Often coversing *a few* categories is already considered diverse
  - The choice of measures is rarely justified, their properties are not addressed
  - Some measures are unclear, even in their original definitions
- . . . used inconsistently across papers
  - no uniform terminology and methodology
  - 197 papers with diversity quantification ⇒ 150 different diversity measures
  - calling the same measure different names
  - using the same name for different measures
- Unawareness of the SOTA in other scientific domains
  - very few explicit links to longstanding theories of diversity from domains like ecology
  - hardly any references to variety, balance or disparity

Prototypical scenarios

Scenario 1: Corpus creation

- Where: *data collection*
- Why: ensuring inclusiveness and equality (*ethical goal*) and/or ensuring performance (*practical means*)
- What: *meta-linguistic* categories – text genres, languages, language genera, language families
- How: measures from *dQuantity* (variety)
- Example: highly multilingual morphological inflection [Vylomova et al.(2020)]

Scenario 2: Generation

- Where: *output data*
- Why: user expectation or naturalness (*practical goal*), e.g. enhance chatbot responses for diversity and relevance simultaneously*on-to-many* scenario
- What: *in-text* categories – text genres, languages, language genera, language families
- How: measures from *dTTR* or *dDistance*
- Example: enhance chatbot responses for diversity and relevance, by a summarizing latent variable inside an RNN [Liu et al.(2023)]

Diversity vs. naturalness

Correlation

- Scenario 1
  - Natural phenomenon: few languages are well-resourced and many others are not
  - Compensation by diversity-driven data selection
  - Diversity and naturalness are **opposed**
- Scenario 2
  - Diversity of human answers = upper bound for the systems' generations
  - Diversity and naturalness are **positively correlated**

Naturalness of categories

- (Meta-)linguistically meaningful (natural) categories: words, idiomatic expressions, sentences, syntactic trees, genres, language families, typological features, speakers, countries, ethnicities, NLP tasks, etc.
- Non-linguistic (artifical) categories: n-grams, BERT word pieces, word embeddings, attention vectors, points in a vector space, etc. (approximations of natural categories whose diversity might be too hard to compute)

Discreteness vs. continuousness

- 28% papers: diversity measures (dBLEU, dBERT) applied directly to elements
  - trivial disparity: elements = categories
  - variety = dataset size
  - balance is moot
- tension:
  - NLP – continuous representations
  - ecology – categorical modelling
  - reason for little popularity of the diversity theory in NLP ?

Future work

- Standardize diversity quantification in NLP
- Systematically incorporate diversity as an evaluation criterion in benchmarks
- Systematize and enhance methods for achieving diversity
  - corpus sampling strategies
  - model training techniques

Bibliography I

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (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), pp. 497–511, San Diego, California. Association for Computational Linguistics.

Akasaki, S. and Kaji, N. (2019).
Conversation initiation by diverse news contents introduction.
In J. Burstein, C. Doran, and T. Solorio, eds., 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), pp. 3988–3998, Minneapolis, Minnesota. Association for Computational Linguistics.

Asai, A., Longpre, S., Kasai, J., Lee, C.-H., Zhang, R., Hu, J., Yamada, I., Clark, J. H., and Choi, E. (2022).
MIA 2022 shared task: Evaluating cross-lingual open-retrieval question answering for 16 diverse languages.
In A. Asai, E. Choi, J. H. Clark, J. Hu, C.-H. Lee, J. Kasai, S. Longpre, I. Yamada, and R. Zhang, eds., Proceedings of the Workshop on Multilingual Information Access (MIA), pp. 108–120, Seattle, USA. Association for Computational Linguistics.

Bhatia, M. and Shwartz, V. (2023).
GD-COMET: A geo-diverse commonsense inference model.
In H. Bouamor, J. Pino, and K. Bali, eds., Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 7993–8001, Singapore. Association for Computational Linguistics.

Bibliography II

Cegin, J., Simko, J., and Brusilovsky, P. (2023).
ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness.
In H. Bouamor, J. Pino, and K. Bali, eds., Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 1889–1905, Singapore. Association for Computational Linguistics.

Chao, A., Chiu, C.-H., and Jost, L. (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.

Chen, W.-L., Wu, C.-K., Chen, H.-H., and Chen, C.-C. (2023a).
Fidelity-enriched contrastive search: Reconciling the faithfulness-diversity trade-off in text generation.
In H. Bouamor, J. Pino, and K. Bali, eds., Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 843–851, Singapore. Association for Computational Linguistics.

Chen, Y., Liu, Y., Xu, R., Yang, Z., Zhu, C., Zeng, M., and Zhang, Y. (2023b).
UniSumm and SummZoo: Unified model and diverse benchmark for few-shot summarization.
In A. Rogers, J. Boyd-Graber, and N. Okazaki, eds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12833–12855, Toronto, Canada. Association for Computational Linguistics.

E, V., Maurya, K., Kumar, D., and Desarkar, M. S. (2023).
DivHSK: Diverse headline generation using self-attention based keyword selection.
In A. Rogers, J. Boyd-Graber, and N. Okazaki, eds., Findings of the Association for Computational Linguistics: ACL 2023, pp. 1879–1891, Toronto, Canada. Association for Computational Linguistics.

Bibliography III

Goldman, O., Batsuren, K., Khalifa, S., Arora, A., Nicolai, G., Tsarfaty, R., and Vylomova, E. (2023).
SIGMORPHON–UniMorph 2023 shared task 0: Typologically diverse morphological inflection.
In G. Nicolai, E. Chodroff, F. Mailhot, and Ç. Çöltekin, eds., Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 117–125, Toronto, Canada. Association for Computational Linguistics.

Hadifar, A., Bitew, S. K., Deleu, J., Hoste, V., Develder, C., and Demeester, T. (2023).
Diverse content selection for educational question generation.
In E. Bassignana, M. Lindemann, and A. Petit, eds., Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 123–133, Dubrovnik, Croatia. Association for Computational Linguistics.

Han, J., Beck, D., and Cohn, T. (2021).
Generating diverse descriptions from semantic graphs.
In A. Belz, A. Fan, E. Reiter, and Y. Sripada, eds., Proceedings of the 14th International Conference on Natural Language Generation, pp. 1–11, Aberdeen, Scotland, UK. Association for Computational Linguistics.

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

Hoffmann, S. and Hoffmann, A. (2008).
Is there a "true" diversity?
Ecological Economics, **65**(2), 213–215.

Bibliography IV

Hwang, E., Thost, V., Shwartz, V., and Ma, T. (2023).
Knowledge graph compression enhances diverse commonsense generation.
In H. Bouamor, J. Pino, and K. Bali, eds., Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 558–572, Singapore. Association for Computational Linguistics.

Ippolito, D., Kriz, R., Sedoc, J., Kustikova, M., and Callison-Burch, C. (2019).
Comparison of diverse decoding methods from conditional language models.
In A. Korhonen, D. Traum, and L. Màrquez, eds., Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3752–3762, Florence, Italy. Association for Computational Linguistics.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (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, pp. 6282–6293, Online. Association for Computational Linguistics.

Jost, L. (2009).
Mismeasuring biological diversity: Response to Hoffmann and Hoffmann (2008).
Ecological Economics, **68**(4), 925–928.

Keleg, A. and Magdy, W. (2023).
DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models.
In A. Rogers, J. Boyd-Graber, and N. Okazaki, eds., Findings of the Association for Computational Linguistics: ACL 2023, pp. 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Bibliography V

Khanuja, S., Ruder, S., and Talukdar, P. (2023).
Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages.
In A. Vlachos and I. Augenstein, eds., Findings of the Association for Computational Linguistics: EACL 2023, pp. 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

Kim, D., Ahn, Y., Lee, C., Kim, W., Lee, K.-H., Shin, D., and Lee, Y. (2023a).
Concept-based Persona Expansion for Improving Diversity of Persona-Grounded Dialogue.
In A. Vlachos and I. Augenstein, eds., Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 3471–3481, Dubrovnik, Croatia. Association for Computational Linguistics.

Kim, D., Ahn, Y., Kim, W., Lee, C., Lee, K., Lee, K.-H., Kim, J., Shin, D., and Lee, Y. (2023b).
Persona expansion with commonsense knowledge for diverse and consistent response generation.
In A. Vlachos and I. Augenstein, eds., Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1139–1149, Dubrovnik, Croatia. Association for Computational Linguistics.

Kim, J., Kim, Y., Baek, I., Bak, J., and Lee, J. (2023c).
It ain't over: A multi-aspect diverse math word problem dataset.
In H. Bouamor, J. Pino, and K. Bali, eds., Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 14984–15011, Singapore. Association for Computational Linguistics.

Bibliography VI

Kim, Y., Rome, S., Foley, K., Nankani, M., Melamed, R., Morales, J., Yadav, A. K., Peifer, M., Hamidian, S., and Huang, H. H. (2024).
Improving Content Recommendation: Knowledge Graph-Based Semantic Contrastive Learning for Diversity and Cold-Start Users.
In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, eds., Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 8743–8755, Torino, Italia. ELRA and ICCL.

Kobayashi, S., Kiyono, S., Suzuki, J., and Inui, K. (2022).
Diverse lottery tickets boost ensemble from a single pretrained model.
In A. Fan, S. Ilic, T. Wolf, and M. Gallé, eds., Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, pp. 42–50, virtual+Dublin. Association for Computational Linguistics.

Kodner, J., Khalifa, S., Batsuren, K., Dolatian, H., Cotterell, R., Akkus, F., Anastasopoulos, A., Andrushko, T., Arora, A., Atanalov, N., Bella, G., Budianskaya, E., Ghanggo Ate, Y., Goldman, O., Guriel, D., Guriel, S., Guriel-Agiashvili, S., Kieraś, W., Krizhanovsky, A., Krizhanovsky, N., Marchenko, I., Markowska, M., Mashkovtseva, P., Nepomniashchaya, M., Rodionova, D., Scheifer, K., Sorova, A., Yemelina, A., Young, J., and Vylomova, E. (2022).
SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection.
In G. Nicolai and E. Chodroff, eds., Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 176–203, Seattle, Washington. Association for Computational Linguistics.

Bibliography VII

Kumar, A., Bhattamishra, S., Bhandari, M., and Talukdar, P. (2019).
Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation.
In J. Burstein, C. Doran, and T. Solorio, eds., 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), pp. 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

Lee, J. Y., Lee, K. A., and Gan, W. S. (2022).
A randomized link transformer for diverse open-domain dialogue generation.
In B. Liu, A. Papangelis, S. Ultes, A. Rastogi, Y.-N. Chen, G. Spithourakis, E. Nouri, and W. Shi, eds., Proceedings of the 4th Workshop on NLP for Conversational AI, pp. 1–11, Dublin, Ireland. Association for Computational Linguistics.

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

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (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, pp. 110–119, San Diego, California. Association for Computational Linguistics.

Li, J., Tang, T., He, G., Jiang, J., Hu, X., Xie, P., Chen, Z., Yu, Z., Zhao, W. X., and Wen, J.-R. (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, pp. 30–39, Online. Association for Computational Linguistics.

Bibliography VIII

Lion-Bouton, A., Ozturk, Y., Savary, A., and Antoine, J.-Y. (2022).
Evaluating diversity of multiword expressions in annotated text.
In Proceedings of the 29th International Conference on Computational Linguistics, pp. 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Liu, C., Koto, F., Baldwin, T., and Gurevych, I. (2024a).
Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings.
In K. Duh, H. Gomez, and S. Bethard, eds., Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Liu, D., Gong, Y., Yan, Y., Fu, J., Shao, B., Jiang, D., Lv, J., and Duan, N. (2020).
Diverse, controllable, and keyphrase-aware: A corpus and method for news multi-headline generation.
In B. Webber, T. Cohn, Y. He, and Y. Liu, eds., Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6241–6250, Online. Association for Computational Linguistics.

Liu, G., Li, Y., Fei, Z., Fu, H., Luo, X., and Guo, Y. (2024b).
Prefix-diffusion: A lightweight diffusion model for diverse image captioning.
In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, eds., Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 12954–12965, Torino, Italia. ELRA and ICCL.

Bibliography IX

Liu, Y., Feng, S., Wang, D., Zhang, Y., and Schütze, H. (2023).
PVGRU: Generating diverse and relevant dialogue responses via pseudo-variational mechanism.
In A. Rogers, J. Boyd-Graber, and N. Okazaki, eds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3295–3310, Toronto, Canada. Association for Computational Linguistics.

Liu, Y. J. and Zeldes, A. (2023).
Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity.
In A. Vlachos and I. Augenstein, eds., Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

Ma, Y., Chi, D., Li, J., Song, K., Zhuang, Y., and King, I. (2024).
VOLTA: Improving generative diversity by variational mutual information maximizing autoencoder.
In K. Duh, H. Gomez, and S. Bethard, eds., Findings of the Association for Computational Linguistics: NAACL 2024, pp. 364–378, Mexico City, Mexico. Association for Computational Linguistics.

Miao, S.-y., Liang, C.-C., and Su, K.-Y. (2020).
A diverse corpus for evaluating and developing English math word problem solvers.
In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds., Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 975–984, Online. Association for Computational Linguistics.

Mohamed, Y., Abdelfattah, M., Alhuwaider, S., Li, F., Zhang, X., Church, K., and Elhoseiny, M. (2022).
ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture.
In Y. Goldberg, Z. Kozareva, and Y. Zhang, eds., Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 8770–8785, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bibliography X

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

Murahari, V., Chattopadhyay, P., Batra, D., Parikh, D., and Das, A. (2019).
Improving generative visual dialog by answering diverse questions.
In K. Inui, J. Jiang, V. Ng, and X. Wan, eds., Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1449–1454, Hong Kong, China. Association for Computational Linguistics.

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

Palumbo, E., Mezzalira, A., Marco, C., Manzotti, A., and Amberti, D. (2020).
Semantic diversity for natural language understanding evaluation in dialog systems.
In Proceedings of the 28th International Conference on Computational Linguistics: Industry Track, pp. 44–49, Online. International Committee on Computational Linguistics.

Park, J.-H., Park, H., Kang, Y., Jeon, E., and Lee, S. (2023).
DIVE: Towards descriptive and diverse visual commonsense generation.
In H. Bouamor, J. Pino, and K. Bali, eds., Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 9677–9695, Singapore. Association for Computational Linguistics.

Bibliography XI

Parrish, A., Hao, S., Laszlo, S., and Aroyo, L. (2024).
Is a picture of a bird a bird? a mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models.
In G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, and S. Tonelli, eds., Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024, pp. 1–18, Torino, Italia. ELRA and ICCL.

Patil, G. P. and Taillie, C. (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.].

Ploeger, E., Poelman, W., de Lhoneux, M., and Bjerva, J. (2024).
What is "typological diversity" in NLP?
In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, eds., Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 5681–5700, Miami, Florida, USA. Association for Computational Linguistics.

Poelman, W., Ploeger, E., de Lhoneux, M., and Bjerva, J. (2024).
A call for consistency in reporting typological diversity.
In M. Hahn, A. Sorokin, R. Kumar, A. Shcherbakov, Y. Otmakhova, J. Yang, O. Serikov, P. Rani, E. M. Ponti, S. Muradoğlu, R. Gao, R. Cotterell, and E. Vylomova, eds., Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pp. 75–77, St. Julian's, Malta. Association for Computational Linguistics.

Bibliography XII

Pouran Ben Veyseh, A., Nguyen, M. V., Dernoncourt, F., and Nguyen, T. (2022).
MINION: a large-scale and diverse dataset for multilingual event detection.
In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, eds., Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2286–2299, Seattle, United States. Association for Computational Linguistics.

Pradhan, S., Bonn, J., Myers, S., Conger, K., O'gorman, T., Gung, J., Wright-bettner, K., and Palmer, M. (2022).
PropBank comes of Age—Larger, smarter, and more diverse.
In V. Nastase, E. Pavlick, M. T. Pilehvar, J. Camacho-Collados, and A. Raganato, eds., Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pp. 278–288, Seattle, Washington. Association for Computational Linguistics.

Puranik, V., Majumder, A., and Chaoji, V. (2023).
PROTEGE: Prompt-based diverse question generation from web articles.
In H. Bouamor, J. Pino, and K. Bali, eds., Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 5449–5463, Singapore. Association for Computational Linguistics.

Samardzic, T., Gutierrez, X., Bentz, C., Moran, S., and Pelloni, O. (2024).
A measure for transparent comparison of linguistic diversity in multilingual NLP data sets.
In K. Duh, H. Gomez, and S. Bethard, eds., Findings of the Association for Computational Linguistics: NAACL 2024, pp. 3367–3382, Mexico City, Mexico. Association for Computational Linguistics.

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

Bibliography XIII

Schüz, S., Han, T., and Zarrieß, S. (2021).
Diversity as a by-product: Goal-oriented language generation leads to linguistic variation.
In H. Li, G.-A. Levow, Z. Yu, C. Gupta, B. Sisman, S. Cai, D. Vandyke, N. Dethlefs, Y. Wu, and J. J. Li, eds., Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 411–422, Singapore and Online. Association for Computational Linguistics.

Shannon, C. E. and Weaver, W. (1949).
A Mathematical Theory of Communication. University of Illinois Press, Urbana.

Shao, C., Wu, X., and Feng, Y. (2022).
One reference is not enough: Diverse distillation with reference selection for non-autoregressive translation.
In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, eds., Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3779–3791, Seattle, United States. Association for Computational Linguistics.

Shen, Y., Liu, Q., Mao, Z., Wan, Z., Cheng, F., and Kurohashi, S. (2022).
Seeking diverse reasoning logic: Controlled equation expression generation for solving math word problems.
In Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, eds., Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 254–260, Online only. Association for Computational Linguistics.

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

Bibliography XIV

Song, B., Pan, C., Wang, S., and Luo, Z. (2021).
DeepBlueAI at SemEval-2021 task 7: Detecting and rating humor and offense with stacking diverse language model-based methods.
In A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, and X. Zhu, eds., Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 1130–1134, Online. Association for Computational Linguistics.

Song, F., Yu, B., Lang, H., Yu, H., Huang, F., Wang, H., and Li, Y. (2024).
Scaling data diversity for fine-tuning language models in human alignment.
In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, eds., Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 14358–14369, Torino, Italia. ELRA and ICCL.

Stirling, A. (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.

Tevet, G. and Berant, J. (2021).
Evaluating the evaluation of diversity in natural language generation.
In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 326–346, Online. Association for Computational Linguistics.

Bibliography XV

Thompson, B. and Post, M. (2020).
Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity.
In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, and M. Negri, eds., Proceedings of the Fifth Conference on Machine Translation, pp. 561–570, Online. Association for Computational Linguistics.

Tripodi, R., Conia, S., and Navigli, R. (2021).
UniteD-SRL: A unified dataset for span- and dependency-based multilingual and cross-lingual Semantic Role Labeling.
In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds., Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2293–2305, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vylomova, E., White, J., Salesky, E., Mielke, S. J., Wu, S., Ponti, E. M., Maudslay, R. H., Zmigrod, R., Valvoda, J., Toldova, S., Tyers, F., Klyachko, E., Yegorov, I., Krizhanovsky, N., Czarnowska, P., Nikkarinen, I., Krizhanovsky, A., Pimentel, T., Torroba Hennigen, L., Kirov, C., Nicolai, G., Williams, A., Anastasopoulos, A., Cruz, H., Chodroff, E., Cotterell, R., Silfverberg, M., and Hulden, M. (2020).
SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection.
In G. Nicolai, K. Gorman, and R. Cotterell, eds., Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 1–39, Online. Association for Computational Linguistics.

Bibliography XVI

Yadav, V., Kwon, H. j., Srinivasan, V., and Jin, H. (2024).
Explicit over implict: Explicit diversity conditions for effective question answer generation.
In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, eds., Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 6876–6882, Torino, Italia. ELRA and ICCL.

Yang, Y., Huang, P., Ma, F., Cao, J., and Li, J. (2024).
PAD: A robustness enhancement ensemble method via promoting attention diversity.
In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, eds., Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 12574–12584, Torino, Italia. ELRA and ICCL.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. (2018).
HotpotQA: A dataset for diverse, explainable multi-hop question answering.
In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yano, T., Takeoka, K., and Oyamada, M. (2024).
Relevance, diversity, and exclusivity: Designing keyword-augmentation strategy for zero-shot classifiers.
In D. Bollegala and V. Shwartz, eds., Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024), pp. 106–119, Mexico City, Mexico. Association for Computational Linguistics.

Bibliography XVII

Yin, D., Li, L. H., Hu, Z., Peng, N., and Chang, K.-W. (2021).
Broaden the vision: Geo-diverse visual commonsense reasoning.
In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds., Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2115–2129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yu, Y., Khadivi, S., and Xu, J. (2022).
Can data diversity enhance learning generalization?
In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, eds., Proceedings of the 29th International Conference on Computational Linguistics, pp. 4933–4945, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhang, H., Duckworth, D., Ippolito, D., and Neelakantan, A. (2021).
Trading off diversity and quality in natural language generation.
In A. Belz, S. Agarwal, Y. Graham, E. Reiter, and A. Shimorina, eds., Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pp. 25–33, Online. Association for Computational Linguistics.

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. (2020).
Bertscore: Evaluating text generation with bert.
In International Conference on Learning Representations.

Bibliography XVIII

Zhang, Y., Zhang, N., Liu, Y., Fabbri, A., Liu, J., Kamoi, R., Lu, X., Xiong, C., Zhao, J., Radev, D., McKeown, K., and Zhang, R. (2024).
Fair abstractive summarization of diverse perspectives.
In K. Duh, H. Gomez, and S. Bethard, eds., Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 3404–3426, Mexico City, Mexico. Association for Computational Linguistics.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. (2018).
Texygen: A benchmarking platform for text generation models.
In K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, and E. Yilmaz, eds., The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, pp. 1097–1100. ACM.