A Proposal for Evaluating the Linguistic Quality of Synthetic Spanish Corpora

Lucia Sevilla Requena
Universidad de Alicante / Alicante, Spain
lsr30@alu.ua.es

Abstract

Large language models (LLMs) rely heavily on high-quality training data, yet human-generated corpora face increasing scarcity due to legal and practical constraints. Synthetic data generated by LLMs is emerging as a scalable alternative; however, concerns remain about its linguistic quality and diversity. While previous research has identified potential degradation in English synthetic corpora, the effects in Spanish, a language with distinct grammatical characteristics, remain underexplored. This research proposal aims to conduct a systematic linguistic evaluation of synthetic Spanish corpora generated by state-of-the-art LLMs, comparing them with human-written texts. The study will analyse three key dimensions: lexical, syntactic, and semantic diversity, using established corpus linguistics metrics. Through this comparative framework, the proposal intends to identify potential linguistic simplifications and degradation patterns in synthetic Spanish data. Ultimately, the proposed outcome is expected to contribute valuable insights to support the creation of robust and reliable Natural Language Processing (NLP) models for Spanish.

1 Introduction

The development of Large Language Models (LLMs) has led to a paradigm shift in the field of Natural Language Processing (NLP), dramatically transforming the capabilities of current systems to understand and generate text (Touvron et al., 2023; van Noord et al., 2024). These models have achieved outstanding performance across a wide range of tasks, including machine translation, text generation, question answering, and semantic inference. However, their performance and robustness are critically dependent on the availability of high-quality, large-scale training data (Gandhi et al., 2024), yet obtaining such data has become a significant challenge (Villalobos et al., 2024; Chen et al., 2024).

The current training framework is heavily based on massive web-crawled corpora combined with curated datasets derived from books, scientific articles, and social media interactions (Penedo et al., 2023). Although this approach has been crucial in the evolution of LLMs, it faces significant structural limitations. On the one hand, scalability is constrained, as the amount of high-quality web data is finite and increasingly subject to legal, privacy, and copyright restrictions (Kurakin et al., 2024; Amin et al., 2025). On the other hand, much of the available crawled data suffers from quality issues, including noise, spam, misinformation, redundancy, toxic content, and increasingly low-quality machine-generated text (Trinh and Le, 2019; Kreutzer et al., 2022).

In response to growing data limitations, synthetic data generated by LLMs has emerged as a scalable and increasingly viable alternative (Long et al., 2024). Recent research demonstrates that current models can produce syntactically correct, semantically coherent, and stylistically diverse texts that are, in some cases, nearly indistinguishable from human-written content (Hartvigsen et al., 2022; Gao et al., 2023; Liu et al., 2024).

However, this approach introduces significant risks. A key concern is 'model collapse', which occurs when models are repeatedly trained on data generated by other models rather than on human-produced language (Gerstgrasser et al., 2024). This leads to a gradual degradation of linguistic quality (Shumailov et al., 2024), including loss of syntactic and semantic diversity, oversimplification of structures, increased redundancy, and a higher incidence of hallucinations, which are factually incorrect or incoherent outputs (Long et al., 2024). Over time, this severely undermines the model's ability to replicate the richness and complexity of natural

Proceedings of the Student Research Workshop associated with RANLP-2025,
pages 54–61, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/issn.2603-2821.2025_007

language (Bender et al., 2021; Penedo et al., 2023).

Despite recent studies exploring the benefits and risks of synthetic data (Liu et al., 2024; Gilardi et al., 2023), there is still a lack of methodological frameworks that rigorously assess the linguistic quality of synthetic data compared to real human data. This gap raises important concerns about whether synthetic data can truly support effective model training without introducing problems. Therefore, there is an urgent need for more rigorous and linguistic evaluation methods to assess whether synthetic corpora adequately reflect the qualities of human-produced text and can ensure the long-term reliability of NLP systems.

The present proposal seeks to address this gap by designing and implementing a systematic linguistic evaluation of synthetic Spanish data generated by state-of-the-art LLMs, focusing on three dimensions: lexical, syntactic and semantic diversity. While existing research has predominantly focused on English, the linguistic effects of synthetic data generation in other languages remain largely underexplored.

In this context, the proposed study takes a new perspective by examining whether the patterns of linguistic degradation observed in English synthetic data also manifest in Spanish, a language with fundamentally different grammatical properties. To this end, the study will develop a comparative framework, grounded in quantitative corpus-linguistic metrics, to systematically evaluate and contrast synthetic Spanish corpora with authentic human-written corpora of comparable size and genre. It is worth noting that this framework remains to be operationalised.

This comparative analysis aims to reveal whether risks such as linguistic simplification and loss of structural and semantic richness are universal phenomena or language-specific issues. This methodological approach aims to uncover whether said degradation previously observed in English also occurs in Spanish.

2 Background and Related Work

The increasing reliance on synthetic data used to overcome the limited availability of high-quality human-produced corpora has attracted growing attention in recent years. A substantial body of research has emerged examining the potential and limitations of synthetic datasets in the training of large language models (LLMs), particularly related to their linguistic properties and their implications for NLP systems. Hence, the present section reviews relevant literature on the risks associated with synthetic data, with particular emphasis on the loss of linguistic diversity in machine-generated texts. Situating this study within the broader context of these works provides the theoretical and empirical foundation for the proposed linguistic evaluation of synthetic Spanish corpora.

2.1 Risks in Synthetic Data

To commence, although synthetic data has been proposed as a scalable solution to the aforementioned problem of scarcity, ongoing research has identified several risks that can seriously affect the quality of models trained on this type of data (Marwala et al., 2023; Hao et al., 2024). These risks are diverse and impact not only the properties of the corpus itself but also the ability of models to perform well.

One of the most relevant issues is data bias, which occurs when synthetic data does not accurately reproduce the real characteristics of authentic data (Hao et al., 2024). This can lead models to learn inaccurate or unrealistic representations, reducing their reliability.

Closely related to this is the phenomenon of oversmoothing, where synthetic data tends to remove natural variation and rare patterns. As a result, the corpus becomes too homogeneous and simplified, lacking the complexity needed to train robust models (Hao et al., 2024). Such a loss of complexity contributes to the degradation of linguistic diversity in synthetic content.

Another common risk is incomplete or inaccurate information, as synthetic data does not always capture the full diversity of linguistic phenomena present in real texts. This is partly due to the limitations of generative models, which often suppress noise or contain algorithmic flaws (Marwala et al., 2023; Hao et al., 2024).

These risks are not just technical problems, but fundamental challenges that threaten the sustainability and reliability of natural language processing systems. As synthetic data becomes more widespread, understanding how it affects quality is key to designing strategies that can mitigate its negative impact.

2.2 Language Diversity Loss in Synthetic Data

Several recent studies have shown a growing interest in analysing how the use of LLMs affects

linguistic diversity, both in machine-generated text and in text produced by humans assisted by these models (Guo et al., 2024a). A common concern in this line of research is that, although LLMs have demonstrated remarkable capabilities in generating fluent and grammatically correct text, their use may lead to processes of linguistic homogenisation that reduce the richness and diversity of language. In particular, synthetic corpora often lack spelling mistakes and tend to underrepresent non-standard dialects, which further limits their applicability in real-world contexts.

Liang et al. (2024) identified a significant shift in lexical frequencies in academic writing, with an increase in the use of LLM-preferred words starting around five months after the release of ChatGPT in 2022. Similarly, Luo et al. (2024) demonstrated that machine translations exhibit lower morphosyntactic diversity and greater convergence compared to human translations. The authors attributed this outcome, in part, to the use of beam search, which biases outputs toward more frequent and less diverse patterns.

Finally, Padmakumar and He (2024) found that writing assisted by InstructGPT also reduces textual diversity compared to writing with GPT-3 or without model assistance. This effect is primarily driven by the model's output rather than by user behaviour. The authors warned that while reinforcement learning with human feedback (RLHF) improves the model's ability to follow instructions, it may also constrain personal expression. This highlights the need for user-centred evaluations and the development of more customisable models that preserve linguistic diversity.

In conclusion, systematic and language-specific evaluations of synthetic corpora are still scarce for languages such as Spanish. This study addresses said necessity through a comparative analysis of human and synthetic Spanish corpora across lexical, syntactic, and semantic levels.

3 Main Hypothesis and Objectives

The present research proposal is based on the hypothesis that synthetic data generated by large language models (LLMs) in Spanish may exhibit lower linguistic richness and diversity compared to human-produced data. If synthetic data is continuously used for model training, it could lead to a degradation of the linguistic quality of LLMs. Specifically, artificially generated texts are expected to show a more limited and repetitive vocabulary, simpler and less varied syntactic structures, and lower semantic coherence, resulting in discourse that is less connected, redundant, or even inconsistent (Guo et al., 2024b). Such linguistic deficiencies could negatively impact the ability of models trained with synthetic data to understand and produce natural language in real-world contexts, thereby compromising their performance on complex linguistic tasks.

From this perspective, the main objective of this research proposal is to perform a detailed linguistic evaluation of the synthetic Spanish corpora generated by LLMs. The evaluation will focus on three key dimensions: lexical, syntactic, and semantic. The purpose is to assess how the synthetic data reflects the natural variability and structural richness of the Spanish language. This will be done through a comparison between synthetic texts and human Spanish corpora of similar size and genre.

To achieve this general goal, the study proposes the following specific objectives:

- **O1:** To assess lexical diversity by applying established corpus linguistics metrics such as type-token ratio (TTR), lexical density, and vocabulary growth measures. These metrics will help determine whether synthetic texts maintain a wide and varied vocabulary comparable to that found in natural Spanish.

- **O2:** To examine syntactic complexity by analysing the presence and frequency of complex sentence constructions, including subordinate clauses, coordination, and sentence embedding. This will help determine whether synthetic data reproduces the grammatical sophistication of human language use.

- **O3:** To evaluate semantic diversity by measuring how much the synthetic texts cover different meanings and topics. This will be done using sentence embeddings to calculate semantic dispersion and topic modelling to assess the range and balance of themes. These metrics will assess if synthetic data reflects the richness and variability of natural Spanish.

- **O4:** To conduct a human evaluation aimed at identifying specific patterns of linguistic degradation in synthetic data through systematic comparison with natural corpora. Understanding these patterns will help guide the

creation of higher-quality synthetic datasets that better support the training of reliable and robust Spanish language models.

- **O5:**To compare the impact of synthetic data on Spanish with previously reported effects in English, thereby distinguishing universal patterns of linguistic simplification from phenomena specific to Spanish.

Through these objectives, the study seeks to provide a clearer picture of the current limitations of synthetic data in Spanish and contribute to the construction of higher-quality data.

4 Proposed Methodology

This study proposes a methodology for the evaluation of the linguistic quality of synthetic data generated by LLMs in Spanish, structured in different stages. The approach is grounded in the framework developed by Guo et al. (2024b) in "The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text", who demonstrated that synthetic data, while effective for improving task performance, systematically exhibit a decline across three key dimensions: lexical, syntactic and semantic diversity when compared to human-written texts. Their findings underscore the importance of incorporating fine-grained linguistic analysis into the evaluation of synthetic corpora, especially when these corpora are intended for use in training language models.

4.1 Data Gathering and Generation

The first stage of this proposal involves the careful selection and preparation of datasets. To carry out the study, two primary datasets will be established: (1) a natural corpus consisting of texts authored by humans, and (2) a synthetic corpus generated artificially by LLMs. The natural corpus will be an existing and compiled dataset, ensuring that the texts are available in open formats and preprocessed to guarantee comparability.

For the synthetic corpus, publicly available synthetic datasets will be collected, and additional texts will be generated using pretrained models like GPT-4, LLaMA 2, or Mistral, among others. Efforts will be made to produce a volume of text comparable to that of the natural corpus to ensure statistical validity. Generation prompts will be carefully crafted to yield texts with styles and thematic content closely matching the human-written corpus.

Finally, both corpora will undergo linguistic normalisation procedures to ensure that all subsequent comparisons are performed on consistent, noise-free data.

4.2 Linguistic Analysis of Corpora

In the second stage of the methodology, a thorough analysis will be carried out to assess the linguistic diversity present in the previously collected human and synthetic corpora. Following Tevet and Berant (2021), diversity can be understood in two main ways: content diversity, answering "What to say?", and form diversity, answering "How to say it?". In the words of Guo et al. (2024a), "lexical diversity and syntactic diversity are considered sub-aspects of form diversity, while semantic diversity reflects content diversity".

Although other sub-aspects of linguistic diversity exist, such as style or register, these tend to be more ambiguous, harder to measure, and often overlap with the three main dimensions. For these reasons, this study will focus specifically on the three clearly defined and quantifiable dimensions mentioned above (Guo et al., 2024b), which offer a solid foundation for comparative analysis.

To fulfil the goal of evaluating and comparing synthetic and human corpora, the analysis is organised around the following dimensions:

4.2.1 Lexical Diversity

Lexical diversity generally refers to the proportion of unique word types within a standardised text sample, such as the total number of tokens (Zheng, 2025). Laufer and Nation (1995) defined measures of lexical richness as attempts to "quantify the degree to which a writer is using a varied and large vocabulary." Consequently, lexical diversity is widely recognised as one of the most direct indicators of lexical richness (Vermeer, 2004).

Lexical diversity metrics quantify the range of vocabulary used in a text, which can reflect both the richness of a language model and its ability to generate varied language (Zheng, 2025). Following the hypothesis presented by Guo et al. (2024a), models trained on synthetic data tend to exhibit a more limited lexical repertoire, often resulting in repetitive and predictable language generation.

In the context of Spanish, the evaluation of lexical diversity presents additional challenges due to the rich inflectional morphology of the language.

In addition, variability caused by verb conjugations, along with gender and number agreements, can artificially inflate surface-level type counts. As a result, accurately assessing lexical variation becomes more complex.

To assess these challenges in Spanish, this study will adopt a set of lexical diversity metrics from corpus linguistics to ensure a comprehensive evaluation:

- **Type-Token Ratio (TTR)** (Johnson, 1944): The ratio between the number of lexical types (unique words) and the total number of tokens in a text. Due to its well-known sensitivity to text length, this metric is applied to texts truncated to a fixed length, following the approach proposed by Guo et al. (2024a).
- **Distinct-n** (Li et al., 2016): Computes the proportion of unique n-grams over the total number of n-grams. This study uses $n = 1$ (equivalent to TTR), $n = 2$, and $n = 3$, as this indicator is particularly informative to evaluate diversity in longer lexical sequences.
- **Self-BLEU** (Zhu et al., 2018): A metric originally developed for generative models that measures the similarity between generated sentences within the same data set. Lower Self-BLEU indicates higher diversity.

These metrics collectively provide a robust view of lexical diversity, accounting for both the superficial variety of word forms and the deeper variability of lexical patterns.

4.2.2 Syntactic Diversity

Syntactic diversity refers to the variety and complexity of sentence structures present in a text or corpus. It shows how flexibly different grammatical parts are used, such as phrases, clauses, and sentence types (Guo et al., 2024b).

According to Bastiaanse and Edwards (1998), higher syntactic diversity makes the text more expressive and adds subtle meaning, affecting its style and tone. Texts with high syntactic diversity have many different sentence forms, while texts with low diversity tend to use repetitive or simple sentences. Additionally, exposure to different syntactic structures is essential for language models to develop a deeper and more complex understanding of language (Aggarwal et al., 2022).

Despite its importance, syntactic diversity has been a relatively underexplored aspect in linguistic analyses (Guo et al., 2024b). This phenomenon is especially significant in Spanish, a language characterised by flexible word order, frequent subject ellipsis, and abundant use of subordinate clauses.

To evaluate this diversity, the present study will employ traditional syntactic complexity metrics commonly used in linguistic research. These metrics are as follows:

- **Syntactic Complexity Index (SCI)** (Lu, 2009): which integrates characteristics such as the average depth of dependency trees, the proportion of subordinate clauses and the mean sentence length.
- **Subordination Ratio** (Hunt, 1965): defined as the proportion of subordinate clauses relative to the total number of clauses, is a widely used metric in the research of syntactic complexity in Spanish.

Together, these metrics capture both the structural diversity and the richness in the syntactic configurations generated by the models.

4.2.3 Semantic Diversity

Semantic diversity refers to the range and variability of meanings, concepts, and topics expressed within a text or across a collection of texts. To capture this dimension, the present study will adopt a dual approach that combines embedding-based and network-based methods, which together provide a robust assessment of semantic variation.

On the one hand, semantic dispersion (Div_sem) is calculated by representing each sentence as a dense vector that captures its meaning within a multilingual semantic space, using SBERT (Reimers and Gurevych, 2019). Then, the average cosine distance between all pairs of sentence vectors is measured to estimate how far the document spreads across semantic space. A higher dispersion value reflects greater variety in the concepts covered.

On the other hand, topic diversity is measured using BERTopic (Grootendorst, 2022), which groups together semantically similar sentence vectors to identify underlying topics in the text. Diversity is then quantified by (a) counting the number of distinct topics found and (b) calculating topic entropy, which reflects how rich and evenly distributed the thematic content is across the document.

Lastly, this combined approach enables a detailed comparison of semantic diversity between human-authored and synthetic texts.

5 Expected results

Based on the proposed methodology, preliminary assumptions suggest that synthetic corpora in Spanish may display lower linguistic diversity compared to human-authored texts. For instance, synthetic texts are expected to exhibit reduced lexical richness, with comparatively lower type-token ratios (TTR), smaller distinct-n values, and higher Self-BLEU scores, indicating a tendency toward repetitive and homogeneous vocabulary. At the syntactic level, a decrease in syntactic complexity is anticipated, reflected in shallower dependency trees, shorter average sentence lengths, and lower subordination ratios, suggesting a preference for simpler and more uniform sentence structures. Finally, in the semantic dimension, synthetic corpora might cover a narrower range of topics and exhibit lower semantic dispersion, which would signal limited conceptual variability.

In conclusion, it is hypothesised that these results may align with previous findings in English. Moreover, given Spanish's greater morphological complexity and comparatively lower online representation, the negative impact of synthetic data is expected to be more pronounced. Nevertheless, these expectations remain tentative and will only be confirmed once the proposed evaluation framework is applied.

References

Arshiya Aggarwal, Jiao Sun, and Nanyun Peng. 2022. Towards robust NLG bias evaluation with syntactically-diverse prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6022–6032, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kareem Amin, Sara Babakniya, Alex Bie, Weiwei Kong, Umar Syed, and Sergei Vassilvitskii. 2025. Escaping collapse: The strength of weak data for large language model training.

Y.R.M. Bastiaanse and S. Edwards. 1998. Diversity in the lexical and syntactic abilities of fluent aphasic speakers. *Aphasiology*, 12(2):99 – 117.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. 2024. On the diversity of synthetic data and its impact on training large language models.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better synthetic data by retrieving and transforming existing datasets.

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024a. Benchmarking linguistic diversity of large language models.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024b. The curious decline of linguistic diversity: Training language models on synthetic text.

Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. 2024. Synthetic data in ai: Challenges, applications, and ethical implications.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Kellogg W Hunt. 1965. *Grammatical structures written at three grade levels*. 8. National Council of Teachers of English.

Wendell Johnson. 1944. Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol

Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. 2024. Harnessing large-language models to generate private synthetic text.

Batia Laufer and Paul Nation. 1995. Vocabulary size and use: Lexical richness in l2 written production. *Applied linguistics*, 16(3):307–322.

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.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the increasing use of LLMs in scientific papers. In *First Conference on Language Modeling*.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Xiaofei Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

Jiaming Luo, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation.

Tshilidzi Marwala, Eleonore Fournier-Tombs, and Serge Stinckwich. 2023. The use of synthetic data to train ai models: Opportunities and risks for sustainable development.

Rik van Noord, Taja Kuzman, Peter Rupnik, Nikola Ljubešić, Miquel Esplà-Gomis, Gema Ramírez-Sánchez, and Antonio Toral. 2024. Do language models care about text quality? evaluating web-crawled corpora across 11 languages.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity?

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross J. Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nat.*, 631(8022):755–759.

Guy Tevet and Jonathan Berant. 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*, pages 326–346, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Trieu H. Trinh and Quoc V. Le. 2019. A simple method for commonsense reasoning.

A. Vermeer. 2004. *The relation between lexical richness and vocabulary size in Dutch L1 and L2 children*, number 10 in Language Learning and Language

Teaching, pages 173–189. John Benjamins Publishing Company, Netherlands. Pagination: 17.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Will we run out of data? limits of llm scaling based on human-generated data.

Wanwan Zheng. 2025. Lexical richness viewed through lexical diversity, density, and sophistication. *Digital Scholarship in the Humanities*, 40(2):692–708.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models.