Adapting Definition Modeling for New Languages: A Case Study on Belarusian

Daniela Kazakouskaya **Timothee Mickus** **Janine Siewert**

University of Helsinki
firstname.lastname@helsinki.fi

Abstract

Definition modeling, the task of generating new definitions for words in context, holds great prospect as a means to assist the work of lexicographers in documenting a broader variety of lects and languages, yet much remains to be done in order to assess how we can leverage pre-existing models for as-of-yet unsupported languages. In this work, we focus on adapting existing models to Belarusian, for which we propose a novel dataset of 43,150 definitions. Our experiments demonstrate that adapting a definition modeling systems requires minimal amounts of data, but that there currently are gaps in what automatic metrics do capture.

1 Introduction

Dictionaries are invaluable resources. On a sociological level, it is fairly well documented that dictionaries are linked to cultural identity (Dollinger, 2016). From the point of view of the NLP scientist, lexicographic data has historically proven very useful for tasks ranging from word sense disambiguation (Lesk, 1986) to representation learning (Hill et al., 2016). On the other hand, lexicography is a complex enterprise: writing a dictionary from scratch is a time-consuming process, which often limits the number of languages, dialects and sociolects which can effectively be documented.

Definition modeling, the NLP task of generating definitions for words in context, is a promising direction to better support lexicographers in their work. Definition modeling has grown as a field since the seminal work of Noraset et al. (2017): we now have access to mature systems that can produce definitions automatically for English, Russian and other languages (Kutuzov et al., 2024). A direction that remains to be explored is whether these available pretrained definition modeling systems can be leveraged for as-of-yet unsupported languages. We take the Belarusian language as the object of our case study. Our main research question is to explore what is necessary to adapt a definition model to a new language — are large amounts of data necessary? Do we need base models trained for similar languages? To that end, we introduce a novel dataset of over 43,000 definitions for Belarusian, with which we demonstrate that a minimal amount of data is often sufficient to adapt to a novel language with reasonable performance.

This object of study also requires, as a complementary step, that we discuss how these systems should be evaluated. This has already been a point of inquiry in previous works — e.g., Bevilacqua et al. (2020) whereas Segonne and Mickus (2023) conducted manual evaluation. Here, we contrast measurements from automatic and manual evaluation, and underscore current limitations in the evaluation of definition modeling. We make our code and data available at github.com/kozochkadaniela/tsbm.

2 Related works

Definition modeling, initially introduced by Noraset et al. (2017), is the NLP task that consists in generating definitions (Gardner et al., 2022). If the original formulation of Noraset et al. involved static word embeddings as inputs, the field has since then shifted to contextualized definition modeling, where models are tasked to produce definitions for words in context (Gadetsky et al., 2018).

The most common use-case for a definition modeling system is to create tools that facilitate the understanding of rare or technical words (Balachandran et al., 2018; Huang et al., 2021; Jhirad et al., 2023; Huang et al., 2022b; Zielinski et al., 2025): the appearance of novel terminology, slang and neologisms outpaces often what lexicographers can handle manually. Another application is to automatize and support efforts for language documentation (Bear and Cook, 2021). As for this latter purpose, if efforts have been made towards studying definition

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 69–75
July 31, 2025

modeling in multilingual contexts (Mickus et al., 2022; Kutuzov et al., 2024, e.g.,), or for languages other than English (ranging from Portuguese, Dimas Furtado et al., 2024, to Japanese, Huang et al., 2022a), limited work has been devoted to cross-lingual transfer — a step necessary if we want to re-purpose systems to low-resource contexts where they are needed.

3 Experimental setting

Our overall approach is to (i) finetuning existing definition modeling systems for Belarusian, varying some key characteristics in their training, such as the amount of data they have access to and the base model we finetune; (ii) compare and contrast automatic metrics to the manual evaluation by a native Belarusian speaker, using a correlation analysis.

3.1 Dataset

We retrieve our data from the *Skarnik* online Russian-Belarusian dictionary,[1] originally based on the academic dictionary published by Kolas et al. (1984) and subsequently revised and regularly updated. The dataset was obtained directly from an open-access repository provided by its maintainers. To ensure the reliability and consistency of the data, additional preprocessing steps were applied. These included the removal of incorrect or misparsed entries, particularly words accompanied by unrelated example sentences. Words containing typographical errors or non-linguistic symbols were manually corrected. Additionally, several entries lacked explicit part-of-speech (POS) annotations or included only partial morphological information (e.g., gender, tense) without specifying the syntactic category. In such cases, full POS tags were added based on the available morphological information. Additionally, functional words (e.g., prepositions, conjunctions, determiners) were excluded from the dataset, and only content words were retained for analysis.

We then construct train, validation and test splits such that (i) headword types are only assigned to a single split, (ii) the proportion of Russian homographs is constant across splits and (iii) the train split contains at least 40K instances.

[1] `https://www.skarnik.by`

	Train	Val.	Test
N. items	40105	1486	1159
N. glosses	40073	1485	1558
N. headwords	28203	1060	1062
N. homographs	1879	70	71

Table 1: TSBM dataset statistics. N. items tracks the number of distinct instances (glosses and examples). N. homographs corresponds to the number of headwords with exact homographs in Russian.

3.2 Models

We finetune the Russian Definition Modeling system of Kutuzov et al. (2024), an `MT0-XL` model of 3.7B parameters fine-tuned on the CoDWoE dataset (Mickus et al., 2022). Taking inspiration from Kutuzov et al., inputs are formatted as in (1):

(1) `[EXAMPLE]` Что такое `[HEADWORD]`?

We use definition glosses as target outputs. Our models are all trained on the TSBM data (cf. above), using subsets of logarithmically-spaced sizes, namely $100^{0/4}\% = 1\%$, $100^{1/4}\% \approx 3.16\%$, $100^{2/4}\% = 10\%$, $100^{3/4}\% \approx 31.62\%$, and $100^{4/4}\% = 100\%$ of the available training data. We train three models for each subset with fixed random seeds. We furthermore report the performances of Kutuzov et al.'s (not re-trained) Russian Definition Modeling system as a baseline, which we refer to as training with 0% of the data. Lastly, to provide a better grasp as to the effects of language similarity on the performances we observe, we also duplicate our experiments using the two other `MT0-XL`–based models of Kutuzov et al., designed for Norwegian and English.

3.3 Automatic metrics

We report performances obtained with default metrics commonly used in NLG: BLEU (Papineni et al., 2002; Post, 2018), BERTScore (Zhang et al., 2020),[2] BLEURT (Sellam et al., 2020), and chrF++ (Popović, 2015; Post, 2018).

While BLEU assesses precision based on the number of exact matches in the candidate and the reference definition, BERTScore is more flexible as it does not compare the candidate and reference directly, but instead computes the similarity of their contextual embeddings. This makes it possible to recognize similar semantics despite different word use, which improves robustness against word swapping and leads to a higher overlap with human

[2] `bert-base-multilingual-cased` (Devlin et al., 2019)

judgments (Zhang et al., 2020). However, unlike BLEU, the usefulness of BERTScore depends on the quality of embeddings, which can be an issue in low-resource scenarios such as the one we are dealing with.

The other two metrics are less frequently used for definition modeling, but offer interesting perspectives worth investigating. The chrF++ metric of Popović assesses overlaps of character spans — which is useful to measure, given that generated definitions can rely on morphological relationships (Segonne and Mickus, 2023) and that character-level information can prove beneficial (Noraset et al., 2017). BLEURT, on the other hand, is a neural metric which is based on a small collection of variant models; the different existing models provide a tradeoff between computational costs and match with human assessments (Pu et al., 2021).

3.4 Manual evaluation

For the manual evaluation, we chose the criteria informativeness, fluency, and correct language and circularity.

Fluency. Fluency evaluates grammatical correctness, naturalness of phrasing and basic semantic coherence, i.e., whether the sentence makes sense even if it does not fully capture the intended meaning. Outputs rated 1 are fully natural, grammatically correct and fluent. A score of 0.5 is assigned to outputs with minor grammatical issues (e.g., an unexpected я–е alternation in the stem) or slightly unnatural phrasing. Outputs rated 0 exhibit clear grammatical errors, non-existent word forms, or constructions that are confusing or ungrammatical.

Informativeness. Informativeness assesses how well the output conveys the intended meaning of the gloss. Outputs rated with a score of 1 are clear and accurate. A score of 0.5 is assigned to definitions that are too broad, incomplete, or only partially informative. A score of 0 reflects outputs that are semantically uninterpretable, even if the general topic is somewhat correct, or cases where the model lists several synonyms and some of them are wrong.

Circularity. Circularity assesses the extent to which a model repeats the headword in its generated definition. A definition is considered fully circular if it includes the headword itself or one of its inflected forms. If the definition uses a derivational form of the headword, it is classified as partially circular. Definitions that do not contain the headword

Metric	Model	Data size					
		0%	1%	3%	10%	31%	100%
BERTscore	EN	63.04	69.64	70.52	70.95	71.49	72.66
	NO	62.16	70.02	70.87	71.13	71.81	72.82
	RU	63.28	69.72	70.61	71.01	71.67	72.87
BLEU	EN	4.04	8.26	10.14	11.60	12.58	14.20
	NO	1.83	8.31	10.51	11.72	13.09	14.31
	RU	4.66	8.43	10.55	11.69	12.65	14.22
BLEURT 20 D3	EN	8.61	26.91	28.55	29.48	31.06	33.26
	NO	6.55	26.75	28.70	29.60	31.35	33.35
	RU	11.74	25.62	28.60	29.56	31.13	33.63
BLEURT 20 D6	EN	8.13	25.51	27.75	28.87	30.41	32.44
	NO	7.60	25.49	27.91	29.21	30.76	32.68
	RU	13.18	24.85	27.93	29.00	30.38	32.81
BLEURT 20 D12	EN	9.26	23.43	25.57	26.99	28.35	30.79
	NO	9.04	23.45	25.95	27.27	28.83	31.02
	RU	13.40	23.51	25.71	26.81	28.35	31.00
BLEURT 20	EN	5.67	24.54	27.78	29.30	30.95	33.86
	NO	6.59	24.65	28.02	30.08	31.71	34.12
	RU	12.67	25.10	27.87	29.48	31.51	34.24
chrF++	EN	2.05	14.25	16.82	18.40	20.34	22.66
	NO	0.76	14.20	16.68	18.32	20.49	22.73
	RU	9.91	14.04	17.03	18.41	20.38	22.97

Table 2: Overview of automatic metrics (average of 3 runs; all metrics in a 0–100 range).

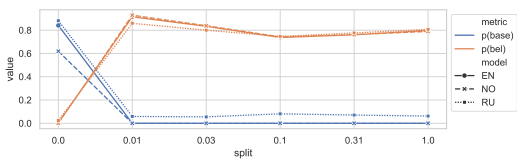

Figure 1: Language identification probability for Belarusian (p(bel)) and base model language (p(base))

or any of its inflectional or derivational variants are labeled as not circular. This categorization helps assess whether the model can produce semantically informative paraphrases without relying on forms morphologically related to the headword.

4 Results & discussion

Automatic metrics. Corresponding performances are shown in Table 2. As is apparent, we observe higher scores for larger datasets. The progress is usually highly similar across all metrics: the average across all datasets is usually obtained with 10% of the data; performances increase to +1 std. dev. above this average when using 100% of the data; even 1% of the data significantly mitigates the poor zero-shot performances of the base models. Difference between base models are rarely significant outside of zero-shot conditions.

We also consider whether our models' outputs are indeed in Belarusian, or whether the base model being trained on another language impacts the output. We assess this using `langid.py` (Lui and Baldwin, 2012), in Figure 1: any amount of training data immediately gears all three models toward produc-

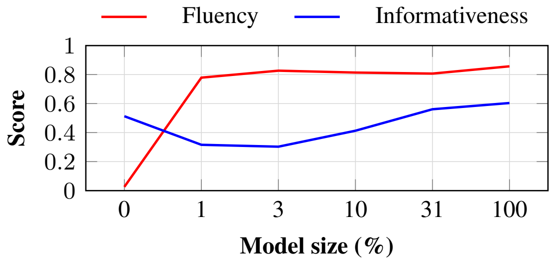

Figure 2: Fluency and informativeness across data size.

ing Belarusian, with a slight *decrease* when using more than 1% of the data as the model learn to produce more informative definitions.

It is worth remarking on the fact that metrics are surprisingly stable *regardless of the language of the base model*. Performances with a Russian model re-trained for Belarusian are on par with what we observe with the Norwegian or English baselines. This strongly suggests that adaptation does not depend on the similarity of the languages considered.

Manual analyses. For the manual analysis we examined 27 words with homographs in Russian and 50 without. We include examples of model productions for the criteria we annotate in Table 3.

A more global picture for fluency and informativeness is presented in Figure 2. Fluency remains consistently high across all data sizes. With only 1% of the training data, the model already achieves a fluency score of 0.78, suggesting that it can produce natural and grammatically correct outputs even under low-resource conditions. Fluency slightly improves as more data become available, reaching 0.86 when the full dataset is used for fine-tuning. The proportion of Russian text in the retrained models doesn't exceed 2%, and it typically appeared as either a single Russian word or the letter и. In contrast, the informativeness shows a more significant improvement as the amount of training data increases. Starting from a modest score of 0.32 in 1%, informativeness increases to 0.60 when the entire dataset is used. This pattern highlights that, while fluency remains relatively stable even with limited training data, achieving accurate semantic alignment with the gloss requires larger datasets.

As shown in Table 4, full circularities decrease with model size, from 26% when using 1% of the data to 11% when using all available data, indicating that larger models are more effective at avoiding circular definitions. Partial circularities remain consistently common across models, suggesting that models frequently reuse morphological forms of the target word, a strategy also used in human-written glosses (Segonne and Mickus, 2023). However, some predicted glosses, even from larger models, rely on morphological patterns and ultimately produced semantically incorrect meanings. Non-circular outputs are most frequent in the largest model (53%), reflecting improved abstraction and lexical flexibility. Although we observe many non-circular outputs when using 1% of the data, a large portion of them were semantically inaccurate.

A closer analysis of the errors reveals several recurring patterns. The model sometimes struggles with sense disambiguation, especially when the senses are closely related. It may select the wrong meaning for polysemous or homographic entries. For example, убяліць means ‘to stain with something white’, but the model gives пакрыць белым слоем(‘to cover with a white layer’), which is not exactly correct in context. Another issue is the generation of circular definitions, both full and partial. Furthermore, the model tends to prioritize morphological similarity over semantic accuracy, which can lead to incorrect output. For example, банкаваць (‘to act as a banker in a card game‘) is misinterpreted as уступаць у банк (‘to join a bank‘).

To conclude, the model often captures important features, such as verb aspect, argument structure, and correct word class. However, it still tends to produce incorrect or overly surface-level definitions. High-scoring outputs typically contain semantically appropriate expressions, but they do not fully capture all the nuances unlike human-produced definitions.

Comparing manual and automatic assessments. In Table 5, we list coefficients of correlation between the automatic metrics and the manual annotation scores we detailed above. We can note several key points: Fluency is generally harder to capture than informativeness, with lower correlation scores; neural metrics such as BERTscore and BLEURT usually fare better than overlap metrics such as BLEU and chrF++. Commonplace metrics in NLG in general and definition modeling in particular, such as BLEU and BERTscore, are in fact not the most suitable for definition modeling, especially when it comes to informativeness: in fact, chrF++ proves to be remarkably fit. Lastly,

	Score	Word		Target	Hypothesis	Translation
Fluency	1	шанс	chance	Умова, якая можа забяспечыць удачу, поспех; магчымасць.	Шанц, магчымасць атрымаць што-н.	A chance, an opportunity to get something.
	0.5	звечарэць	to get dark	Пра надыход вечара, цемнаты.	Стаць вечарам.	To become evening
	0	скляваць	to peck	Клююючы, з'есці.	Сцерці клявым.	grammatically incorrect
Informativeness	1	берагчы	to take care of	Клапатліва ахоўваць каго-, штон., засцерагаць ад чаго-н.	Даглядаць, клапаціцца аб кім-, чым-н.	to look after, to take care of someone or something.
	0.5	банкетка	banquette (bench)	Тып нізкай табурэткі з мяккім сядзеннем прамавугольнай або круглай формы.	Упрыгожванне на кухні, канапа, каб сесці.	A decoration in the kitchen, a couch/sofa to sit on
	0	чаявыя	tips	Грошы, якія даюць за дробныя паслугі, абслугоўванне.	У Беларусі — штраф, аплачваецца чаем.	In Belarus, the fine is paid with tea
Circularity	N	палігон	military training area	Участак мясцовасці, спецыяльна абсталяваны для трэніровачнай стральбы і выпрабавання баявой тэхнікі.	Група вайсковых часцей, якая мае пэўныя мэты.	A formation of military units assigned to specific tasks
	P	дэбютаваць	to debut	Упершыню выступіць на сцэне.	Пачаць сваю дзейнасць, даць дэбют.	To start one's career and make a debut
	F	вокладка	book cover	Покрыўка кнігі, сшытка і пад.	Тое, што і вокладка.	Same as book cover

Table 3: Examples illustrating annotation scores

	1%	3%	10%	31%	100%
No %	52.21	35.24	32.52	49.85	53.41
Part %	22.02	32.19	36.27	35.55	35.33
Full %	25.77	32.57	31.21	14.60	11.26

Table 4: Proportion of circular definitions

	BERT-score	BLEU	BLEURT D3	BLEURT D6	BLEURT D12	BLEURT 20	chrF ++
Fluent	11.56	6.60	11.89	13.63	12.57	10.91	6.64
Informative	25.53	13.07	34.26	34.46	39.79	36.17	40.38

Table 5: Comparison of manual and automatic assessment using Spearman's ρ ($\times 100$).

what works for other NLG subfields need not apply in definition modeling contexts: while Pu et al. (2021) find BLEURT 20 to be a better model of human preferences than all of its distilled variants, here, BLEURT 20 D12 captures informativeness more appropriately, while BLEURT D6 is more appropriate as a model of fluency.

5 Conclusions

In this paper, we have studied how to adapt existing definition modeling systems to Belarusian.

To that end, we introduce a large dataset of Belarusian definitions and conduct extensive experimentation. Small datasets can already achieve some success: even 1% of the data collected was sufficient to ensure the generated definitions would be in Belarusian with a reasonably high degree of fluency. Other characteristics often benefit from more data — e.g., informative, non-circular definitions are more frequent in models trained on larger datasets.

Lastly, further research is necessary in order to properly automatize the assessment the quality of generated definitions: metric rankings from previous work do not translate to definition modeling in Belarusian; none of the metrics we tested capture fluency; and metrics can very greatly in their ability to describe informativeness.

Acknowledgments

This work is supported by the Research Council of Finland through projects No. 342859 "CorCoDial – Corpus-based computational dialectology" and No 353164. "Green NLP – controlling the carbon footprint in sustainable language technology."

References

Vidhisha Balachandran, Dheeraj Rajagopal, Rose Catherine Kanjirathinkal, and William Cohen. 2018. Learning to define terms in the software domain. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 164–172, Brussels, Belgium. Association for Computational Linguistics.

Diego Bear and Paul Cook. 2021. Cross-lingual wolastoqey-English definition modelling. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 138–146, Held Online. INCOMA Ltd.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anna Beatriz Dimas Furtado, Tharindu Ranasinghe, Frederic Blain, and Ruslan Mitkov. 2024. DORE: A dataset for Portuguese definition generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5315–5322, Torino, Italia. ELRA and ICCL.

Stefan Dollinger. 2016. National dictionaries and cultural identity: insights from austrian, german, and canadian english. In Philip Durkin, editor, *The Oxford Handbook of Lexicography*, chapter 37, pages 577–589. Oxford University press.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Noah Gardner, Hafiz Khan, and Chih-Cheng Hung. 2022. Definition modeling: Literature review and dataset analysis. *Applied Computing and Intelligence*, 2(1):83–98.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2022a. JADE: Corpus for Japanese definition modelling. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6884–6888, Marseille, France. European Language Resources Association.

Jie Huang, Hanyin Shao, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022b. Understanding jargon: Combining extraction and generation for definition modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4004, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Jhirad, Edison Marrese-Taylor, and Yutaka Matsuo. 2023. Evaluating large language models' understanding of financial terminology via definition modeling. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 93–100, Nusa Dua, Bali. Association for Computational Linguistics.

Y Kolas, K Krapiva, and P Hlebka. 1984. Тлумачальны слоўнік беларускай мовы *(Explanatory Dictionary of the Belarusian Language)*, volume 1–5.

Andrey Kutuzov, Mariia Fedorova, Dominik Schlechtweg, and Nikolay Arefyev. 2024. Enriching word usage graphs with cluster definitions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6189–6198, Torino, Italia. ELRA and ICCL.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26, New York, NY, USA. Association for Computing Machinery.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Thanapon Noraset, Chen Liang, Lawrence Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *AAAI*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vincent Segonne and Timothee Mickus. 2023. Definition modeling : To model definitions. generating definitions with little to no semantics. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 258–266, Nancy, France. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Andrea Zielinski, Simon Hirzel, and Sonja Arnold-Keifer. 2025. *Enhancing Digital Libraries with Automated Definition Generation*. Association for Computing Machinery, New York, NY, USA.