When Precision Matters More Than Scale: A New Benchmark and a Translation System Surpassing Open-Source and Commercial Models for Polish–Silesian MT

Anonymous ACL submission

Abstract

Dialectal machine translation (MT) remains challenging due to limited data and strong linguistic variation not captured by standard benchmarks, which often assume standardized and well-edited text. We study Polish–Silesian MT using neural and rule-based systems, evaluating on SiLTT – a new Pol–Szl testset, alongside established BOUQuET and FLORES benchmarks. Results show our rule-based system is consistently strongest on SiLTT and BOUQuET datasets and that TranslateGemma fine-tuned on a curated dataset improves over strong neural baselines but does not surpass the rule-based system in dialectal settings. We release SiLTT and our best neural model to support further research.[1]

1 Introduction

Silesian (ISO 639-3: *szl*) is a West Slavic regional language of Poland with limited NLP resources and an ongoing debate about its linguistic status. According to the National Population and Housing Census 2021, 467,145 residents of Poland declared Silesian as a language used in home contacts (Statistics Poland, 2022). Despite growing interest in low-resource machine translation, Silesian remains largely absent from dedicated MT benchmarks and system comparisons.

A key challenge in developing Silesian language technologies is not only the scarcity of parallel data, but also its uneven quality. While Silesian text is present in online sources, much of it is inconsistent in orthography, heavily influenced by Polish, or shaped by ad-hoc writing conventions rather than stable linguistic norms. As a result, large-scale corpora derived from web data or generic multilingual resources often mix heterogeneous and noisy forms of Silesian, limiting their usefulness for training robust translation systems.

Existing multilingual benchmarks such as FLORES-101/200 (Guzmán et al., 2019; Goyal et al., 2022) include Silesian as part of broad evaluation coverage. However, they primarily reflect formal, edited text and do not capture the variability and informal usage found in real-world Polish–Silesian communication. This mismatch becomes more pronounced when models are trained on large but noisy corpora (e.g., OPUS and other web-mined datasets), where Silesian-like data often lacks linguistic consistency and contains substantial cross-lingual contamination.

In this paper, we investigate Polish–Silesian machine translation under realistic, data-constrained conditions. Rather than relying on large, noisy corpora, we construct a compact but curated parallel dataset of approximately 22k sentence pairs designed to better reflect usable Silesian. We compare neural and rule-based with commercial and other state of the art multilingual translation systems.

Our results show that systems trained on large, heterogeneous corpora do not necessarily generalize well to cleaner dialectal evaluation settings. In contrast, our rule-based system remains consistently strong, while neural models trained on the curated dataset outperform strong baselines but still fall short of rule-based performance in dialect-sensitive evaluation.

Our contributions are the following:

- SiLLT: a new Polish–Silesian translation evaluation dataset;
- a systematic comparison of neural and rule-based MT systems across three benchmarks;
- a publicly released evaluation dataset and fine-tuned neural model.

2 Related Work

Machine translation for low-resource and dialectal language varieties remains challenging due to lim-

[1] https://anonymous.4open.science/r/Pol2Szl-translator-FB6C/README.md

ited parallel data and high linguistic variation (Haddow et al., 2022; Ranathunga et al., 2023). Recent multilingual systems such as mT5 (Xue et al., 2021), NLLB (Team et al., 2022), and OPUS-MT (Tiedemann et al., 2024) showed that transfer learning and massively multilingual training can improve translation quality for under-resourced languages. Several studies further demonstrated that transfer between related languages is particularly effective in low-resource and zero-shot MT settings (Zoph et al., 2016; Dabre et al., 2017; Lakew et al., 2019).

Silesian is included in two multilingual evaluation benchmarks: FLORES-101/200 (Guzmán et al., 2019; Goyal et al., 2022) and BOUQuET (Team et al., 2025), enabling evaluation of multilingual MT systems on this language. However, FLORES has seen extensive and repeated use in multilingual MT evaluation, which can lead to benchmark saturation effects.

The closest work to ours is Signoroni (2023), who studied zero-shot Silesian–English translation using mT5 and ByT5 models fine-tuned on related Slavic languages. They found that cross-lingual transfer from related languages can improve performance, though the effect depends on model architecture and language similarity. Their evaluation relied on FLORES due to the lack of dedicated resources for Silesian.

3 Language Background

Silesian's orthography was standardized only in 2009 with the ślabikŏrz system, following centuries of ad hoc adaptations of Polish orthography or phonetic scripts (Bandtkie, 1821; Steuer, 1935). Its status as a distinct language or Polish dialect remains debated: some classify it as an Abstand language, isolated from Polish and neighboring dialects (Hentschel, 2018), while others argue for mutual intelligibility and shared features with Polish dialects (Dubisz, 2024; Siuciak, 2025). Lack of official recognition limits its use in education, and most speakers rely on Polish orthography, reinforced by the absence of default Silesian keyboard support. Silesian exhibits variation across orthography, regional morphology, and lexical borrowings from neighboring languages (Polish, Czech, German). These properties make direct sentence-level translation challenging for purely data-driven systems, especially in low-resource regimes For more details, see the Appendix A.

4 System Description

4.1 Rule-Based System

We implement a rule-based Polish–Silesian translator built on the `Apertium` platform (Forcada et al., 2011; Khanna et al., 2021). The system follows a shallow-transfer architecture and includes: (1) finite-state transducers for lexical transformations, (2) morphological analysis and generation, (3) Hidden Markov Models for POS tagging and ambiguity resolution, and (4) handcrafted transfer rules encoding systematic correspondences.

Our system contains 31,757 lexical translation pairs and 2,349 grammatical rules derived from dictionaries (Przywara, before 1904; Olesch; Zaręba, 1960; Pluta, 1973; Cząstka-Szymon et al., 2000), as well as theses and articles on Silesian vocabulary (e.g., Spectator (1908); Nikel (1908); Koszyk (1957); Iżykowska and Starczewska–Wojnar (2017)). We have included a detailed illustrated system pipeline in Appendix B.

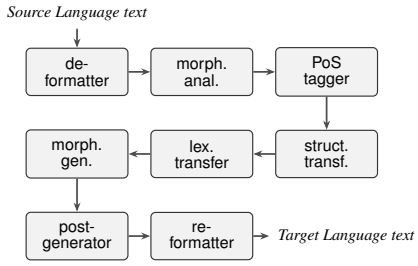

Figure 1: The eight modules that build the assembly line of the shallow-transfer machine translation system.

4.2 Neural models

To evaluate the performance of various neural machine translation architectures on the Silesian language, we selected a representative suite of models ranging from specialized multilingual systems to large-scale commercial language models (details in Appendix C):

- **PLLuM-12B-nc-chat**: Polish general-purpose LLM (Kocoń et al., 2025)
- **Bielik-PL-11B-v3.0-IT**: Polish-focused multilingual model with claimed Silesian support (Ociepa et al., 2025)
- **NLLB-200-3.3B, NLLB-54B MoE**: Silesian-capable MT models (Team et al., 2022)
- **MADLAD-400-10B-MT**: multilingual MT model based on the T5 architecture (Kudugunta et al., 2023)

- **GPT-5.4**: frontier zero-shot LLM (Singh et al., 2026)
- **Google Translate**: a production-grade MT system (Wu et al., 2016)
- **TranslateGemma-4B-IT (base, fine-tuned)**: instruction-tuned and custom finetuned MT variants (Finkelstein et al., 2026)

5 Datasets

5.1 Training Data

Our Polish-Silesian training data consists of three complementary sources. In particular, we combine manually curated and post-edited data with automatically generated resources obtained via pivoting through English and backtranslation. All data was filtered using a maximum length ratio threshold of 1.3.

- **22,783 sentences — Custom backtranslated dataset of Silesian sentences (CBT-22k):** We used articles provided us by the websites Wachtyrz.eu and Slazag.pl as well as several older dialect texts, the largest being the collection gathered by Malinowski (1899). The older texts were transliterated to the current alphabet. The Silesian sentences were translated into Polish using the rule-based system described in Section 4.1, and subsequently post-edited by a professional translator to ensure high-quality references.
- **50,000 sentences — NLLB corpora (pivoting via English):** Sampled from the OPUS corpora (Tiedemann and Nygaard, 2004), specifically the NLLB dataset (Schwenk et al., 2020; Fan et al., 2020), which contains approximately 1.7 million English–Silesian pairs. The English side was translated into Polish using a quantized version of EuroLLM-9B-Instruct (Martins et al., 2025) model.[2] We decided to sample only 50,000 sentences to avoid overflowing our custom high-quality dataset (CBT-22k).
- **28,107 sentences — HPLT backtranslation** Extracted from the monolingual Silesian HPLT v3.0 corpus (Oepen et al., 2026). Silesian sentences were translated into Polish using the NLLB-3.3B model to obtain a back-translated dataset.

5.2 Evaluation Data

We use three evaluation datasets:

BOUQuET: (Team et al., 2025) a multilingual, multicentric benchmark consisting of 854 Polish–Silesian sentence pairs translated by community contributors. We additionally conducted a manual evaluation with native Silesian and Polish speaker to verify overall quality of the translations.

SiLTT (**Si**lesian **L**iterary **T**ranslation **T**estset): 1,237 Polish–Silesian sentence pairs derived from human translations and automatically aligned using LaBSE (Feng et al., 2022). The dataset was compiled from literary and cultural corpora, including "*Cinderella*", "*Rapunzel*", "*Janko the Musician*" and "*A Christmas Carol*". We applied automatic alignment filtering based on cosine similarity of sentence embeddings, retaining only pairs with similarity above 0.6. Sentence pairs below this threshold were excluded from the final dataset. Dataset statistics are shown at Table 1. The resulting test set is released under the CC BY-NC 4.0 license.

Statistic	Polish	Silesian
Avg sentence length (chars)	93.1	79.4
Avg sentence length (words)	14.8	13.6
Max sentence length (chars)	542	445
Max sentence length (words)	82	73
Vocabulary size	7673	5393

Table 1: Dataset statistics of SiLTT.

FLORES-200: (Guzmán et al., 2019) A multilingual evaluation benchmark consisting of 1,012 high-quality sentence pairs for each language pair.

6 Experiments

We primarily focus on the Polish-to-Silesian translation direction, as it represents the more challenging task. Compared to Silesian, Polish is a high-resource language, making text generation substantially easier for modern models.

6.1 Fine-tuning the TranslateGemma Model

We fine-tuned the 4-billion-parameter instruction-tuned version of TranslateGemma[3] using QLoRA (Dettmers et al., 2023). We performed two fine-tuning experiments with different training data configurations: (1) using only the custom dataset of Silesian sentences (CBT-22k), and (2) using all corpora described in Section 5.1. This setup was motivated by concerns, based on manual inspection, that the Silesian side of the NLLB- and HPLT-derived data may contain noise and low-quality texts. Due to the low-resource nature

[2] https://huggingface.co/stelterlab/EuroLLM-9B-Instruct-AWQ

[3] https://huggingface.co/google/translategemma-4b-it

Example	Rule-based	TransGemma-FT (CBT-22k)	Google Translate
Chłopcy i dziewczynki byli bardzo szczęśliwi podczas zabawy. (PL)	Synki i dziouszki byli fest szczyńśliwi w czasie graczki.	Chłopcy i dziouchy byli bardzo szczyńśliwi podczas zabawy.	Syny i dziouchki byli barzo szczyńśliwi w czasie szpilu.
Boys and girls were very happy during play. (EN)	*Correct translation*	*Incorrectly left Polish "Chłopcy" and "zabawy" (instead: "Synki" and "graczki"; wrong diminutive in "dziouchy" (instead: "dziouszki".*	*Wrong morphology in "syny" (instead: "synki" and "dziouchki" (instead: "dziouszki"); incorrect translation of "zabawy" to "szpilu" (meaning: "game").*

Table 2: Qualitative comparison based on the BOUQuET dataset.

Model	SiLTT		BOUQuET	
	BLEU	chrF	BLEU	chrF
PLLuM-12B-nc-chat	1.7	22.3	3.4	28.4
Bielik-PL-11B-v3.0-IT	3.4	26.3	7.8	35.5
MADLAD-400-10B-MT	2.1	21.3	7.5	32.0
NLLB 3.3B	3.8	26.8	12.2	39.2
NLLB 54B MOE	3.2	26.0	9.9	38.3
TransGemma	2.5	23.6	7.4	30.9
GPT-5.4	7.4	31.7	20.0	48.6
Google Translate	4.2	29.3	21.8	49.4
ours				
Rule-Based	**10.3**	**34.5**	**36.7**	**61.0**
TransGemma-FT (CBT-22k)	8.0	31.9	26.3	52.8
TransGemma-FT (Mixed)	5.5	28.9	6.2	33.4

Table 3: Comparison of POL→SZL results on SiLTT and BOUQuET benchmarks. **Bold** indicates the best result, and underlining indicates the second-best result.

of Silesian, reliable automatic filtering methods are not available.

6.2 Evaluation Metrics

We use sacreBLEU (Post, 2018) to calculate the BLEU[4] (Papineni et al., 2002) and chrF [5] (Popović, 2015) scores. We do not use neural-based evaluation metrics such as COMET (Rei et al., 2020) or LLM-as-a-Judge approaches, as the low-resource setting may lead to unreliable and potentially misleading results in this context.

6.3 Results

The evaluation results on SiLTT and the BOUQuET POL→SZL task are presented in Table 3 while results on FLORES and the reverse direction (SZL→POL) are provided in Appendix D.

The rule-based system achieves the best results, maintaining a clear advantage across both benchmarks. The base TranslateGemma model performs poorly on all evaluation sets, indicating limited zero-shot capability for this low-resource language pair. However, fine-tuning on the CBT-22k corpus yields substantial improvements, producing the strongest neural system and surpassing both GPT-5.4 and Google Translate. These results highlight the importance of high-quality parallel data for effective adaptation in low-resource MT. In contrast, augmenting the training data with additional NLLB and HPLT corpora consistently degrades performance compared to the CBT-22k-only setup, likely due to noise.

The Polish models Bielik and PLLuM also achieve relatively weak results across all evaluation settings. Although Bielik was pretrained on Silesian text, its performance remains limited, only slightly better than that of the base TranslateGemma model. PLLuM performs poorly, failing to generate usable translations. Overall, strong Polish language ability or incidental exposure to Silesian is insufficient for effective Polish–Silesian MT without dedicated bilingual supervision.

6.4 Qualitative Analysis

We conduct an qualitative analysis of selected best models on the BOUQuET dataset (see Table 2). Key observation: explicit linguistic rules provide robustness in low-resource dialect translation. The rule-based system provides the best results in the context of closely related languages. Google Translate seems to create plausible but hallucinated phrases. Fine-tuned TransGemma high-quality CBT-22k dataset provides promising results despite minor mistakes.

7 Conclusion

We presented a new benchmark and evaluation of Polish–Silesian machine translation systems. Our findings suggest that rule-based systems remain highly competitive in dialectal and low-resource settings, particularly when evaluated on newer datasets. We open-source the testset and the best performing neural model.

[4] BLEU signature: "nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.6.0

[5] chrF signature: "nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.6.0

Limitations

This study is limited by the size and scope of the available evaluation and training data. Although the CBT-22k corpus is carefully curated, it remains relatively small compared to large-scale multilingual MT datasets, which may constrain the generalization of the neural models.

Our evaluation focuses on Polish–Silesian translation and three benchmark datasets; while these cover both curated and existing benchmark settings, they do not fully capture the breadth of real-world Silesian usage across domains, speakers, and registers. In particular, spoken language and highly informal communication styles are underrepresented.

The rule-based system relies on manually designed correspondences and linguistic assumptions, which may not scale to broader coverage or other dialectal language pairs without significant adaptation effort. While it performs strongly in the evaluated setting, its portability and maintenance cost remain open questions.

References

Jerzy Samuel Bandtkie. 1821. Wiadomości o ięzyku polskim w szląsku i o polskich szlązakach. *Mrówka Poznańska*.

Bożena Cząstka-Szymon, Jerzy Ludwig, and Helena Synowiec. 2000. *Mały słownik gwary Górnego Śląska*. Leksykon.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia conference on language, information and computation*, pages 282–286.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Stanisław Dubisz. 2024. Lingwistycznie o godce ślunskij i języku regionalnym. *Poradnik Językowy*, 815(6):105–110.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. Translategemma technical report. *Preprint*, arXiv:2601.09012.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Government of Poland. 2020. Comments of the Government of Poland on the Fourth Opinion of the Advisory Committee on the implementation of the Framework Convention for the Protection of National Minorities by Poland. official memorandum.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6098–6111.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli-Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Gerd Hentschel. 2018. Śląski: Gwara – dialekt – język? spojrzenie z zewnątrz. *Polonistyka na początku XXI wieku : diagnozy, koncepcje, perspektywy. T. 4 : Pogranicza, mniejszości, regiony, etnolingwistyka*, 4:41–66.

Małgorzata Iżykowska and Aleksandra Starczewska–Wojnar. 2017. *Przeczytano, przyjęto, podpisano: polskie protokoły sądów rozjemczych w powiecie strzeleckim w XIX wieku: edycja źródłowa ze zbiorów Archiwum Państwowego w Opolu*. Archiwum Państwowe w Opolu.

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hector Alos i Font. 2021. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, 35(4):475–502.

Jan Kocoń, Maciej Piasecki, Arkadiusz Janz, Teddy Ferdinan, Łukasz Radliński, Bartłomiej Koptyra, Marcin Oleksy, Stanisław Woźniak, Paweł Walkowiak, Konrad Wojtasik, Julia Moska, Tomasz Naskręt, Bartosz Walkowiak, Mateusz Gniewkowski, Kamil Szyc, Dawid Motyka, Dawid Banach, Jonatan Dalasiński, Ewa Rudnicka, and 80 others. 2025. Pllum: A family of polish large language models. *Preprint*, arXiv:2511.03823.

Szymon Koszyk. 1957. Z gwary „wojouków" opolskich. *Kwartalnik opolski*, 2:168–184.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Surafel M Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. 2019. Adapting multilingual neural machine translation to unseen languages. In *Proceedings of the 16th International Conference on Spoken Language Translation*.

Lucjan Malinowski. 1899. *Powieści ludu polskiego na Śląsku*. Akademia Umiejętności.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report. *Preprint*, arXiv:2506.04079.

Emmanuel Nikel. 1908. *Die polnische Mundart des oberschlesischen Industriebezirks*. Ph.D. thesis, Friedrich-Wilhelms-Universität zu Berlin.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, and Adrian Gwoździej. 2025. Bielik 11b v3: Multilingual large language model for european languages. *Preprint*, arXiv:2601.11579.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2026. Hplt 3.0: Very large-scale multilingual resources for llms and mt. mono- and bi-lingual data, multilingual evaluation, and pre-trained models. *Preprint*, arXiv:2511.01066.

Reinhold Olesch. *Der Wortschatz der polnischen Mundart von Sankt Annaberg*, volume 1. Osteuropa-Institut.

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.

Feliks Pluta. 1973. Słownictwo Dzierżysławic w powiecie prudnickim. *Zeszyty naukowe Wyższej Szkoły Pedagogicznej im. Powstańców Śląskich w Opolu / Seria B: Studia i monografie*.

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. *Preprint*, arXiv:1804.08771.

Michał Przywara. before 1904. Narzecza śląskie. Unpublished manuscript.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2020. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *Preprint*, arXiv:1911.04944.

Edoardo Signoroni. 2023. Fine-Grained Language Relatedness for Zero-Shot Silesian-English Translation. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023*, pages 145–156.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Mirosława Siuciak. 2025. „Język śląski" – czyli o mieszaniu porządków. *Poznańskie Studia Polonistyczne. Seria Językoznawcza*, 32(2):329–346.

Spectator. 1908. *Sprachmischung in Oberschlesien. Ein Beitrag zur oberslesischen Volkskunde*. Verlag von Gebrüder Böhm.

Statistics Poland. 2022. Preliminary results of the national population and housing census 2021 in the scope of national-ethnic structure and language used in home contacts. Accessed: 2026-05-25.

Feliks Steuer. 1935. *Ostatńi gwojźdźaurz*. Księgarnia i Drukarnia Katolicka S. A. Katowice.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

The Omnilingual MT Team, Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Cynthia Gao, Jean Maillard, Alex Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. Bouquet: dataset, benchmark and open initiative for universal quality evaluation in translation. *Preprint*, arXiv:2502.04314.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing neural machine translation with opus-mt. *Language Resources and Evaluation*, 58(2):713–755.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and 12 others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498.

Alfred Zaręba. 1960. Słownik Starych Siołkowic w powiecie opolskim. *Zeszyty naukowe Uniwersytetu Jagiellońskiego / Rozprawy i studia*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1568–1575.

A Silesian Language Description

Despite the efforts of documenting the Silesian language reaching the first half of the 19th century (Bandtkie, 1821), the proposals for a single orthography had not reached a wider consensus until 2009 when the *ślabikŏrz* (meaning: "primer") orthography was agreed. Until then the few Silesian books and other publications used an ad hoc modified Polish orthography or – as in the case of a linguist Feliks Steuer's works – a special phonetic orthography (Steuer, 1935).

As a signatory of the European Charter for Regional or Minority Languages, Poland recognises languages of its national and ethnic minorities as well as Kashubian as a regional language. There have been several attempts to recognise Silesian as the second regional language, however all of them failed. The opponents of this solution argue that the term *regional language* does not cover dialects of the official state language (see eg. Government of Poland, 2020, page 10). This led to an intense discourse about the classification of Silesian as a separate language and not a dialect of Polish.

As proposed by Hentschel (2018), Silesian could be described as Abstand language or Abstand lect. Silesian has isolated itself from neighbouring dialects and Polish standard language. Since Polish dialects around Silesia have died out, there is no basis to consider Silesian a part of Polish dialect continuum.

On the other hand the supporters of preserving the status quo argue that Silesian is mutually intelligible with the Polish standard language and that the features of the Silesian language can be also found in other Polish dialects (Dubisz, 2024; Siuciak, 2025).

The lack of recognition barres the language from being taught at schools. Because of that a relatively low number of Silesians have mastered the new orthography. Although Silesian cultural organisations use it in their public communication, most of Silesian speakers still tend to phonetically write using Polish orthography. This is further reinforced by a caveat that could be named a sociotechnical one: there are no electronical devices offering Silesian keyboard by default or even as a preinstalled option. Silesian speakers have to deliberately download and install it. This usually leads to resorting to the Polish keyboard.

Silesian exhibits variation across orthography (non-standardized spelling conventions), regional

morphology, and lexical borrowings from neighbouring languages. These properties make direct sentence-level translation challenging for purely data-driven systems, especially in low-resource regimes.

B Rule-based system pipeline

To illustrate the rule-based system pipeline, consider the phrase:

system przekładający maszynowo

"machine-translating system"

The system first analyses the phrase using Constraint Grammar. The first of the three tokens "system" can have two morphological intepretations, the second "przekładający" can have nine, and finally "maszynowo" can have two. Constraint Grammar eliminates these ambiguities and leaves the three correct ones:

`"system" n mi sg nom`

`"przekładać" vblex impf pprs mi sg nom`

`"maszynowo" adv`

The disambiguated representation is then sent to the lexical transfer module which translates the result using the bilingual dictionary:

systym przekłŏdajōncy maszinowo

"machine-translating system"

The phrase would be understandable to a Silesian speaker, however this is a seldom used form. The structural transfer module detects the pattern and applies a rule for a more natural sounding output:

noun + present participle used adjectivally → noun + comma + "co" + present participle retagged as third-person present

The result is:

systym, co przekłŏdŏ maszinowo

"system that machine-translates"

Despite its simplicity, the system provides strong performance in low-resource settings due to explicit linguistic priors.

C Neural models

C.1 NLLB: No Language Left Behind

NLLB-200-3.3B and **NLLB 54B MOE** this MT models are included due to its explicit support for Silesian in its training data.

C.2 GPT-5.4

We include this model as state-of-the-art commercial benchmark. While its scale differs significantly from our models, it provides a vital reference for how frontier LLM handle low-resource tasks in a zero-shot setting. We obtained these translations through the OpenAI API.

The model was evaluated with default settings, without reasoning enabled, and with temperature set to 1.0. The prompt used for translation is shown in Figure 2. The prompt was written in Polish, as it was the source language of the translation task and provided the most natural interaction setting for the evaluated model.

```
Przetłumacz poniższe zdanie na
język śląski.
Zwróć wyłącznie tłumaczenie
po śląsku (bez komentarzy,
wyjaśnień ani dodatkowego
tekstu).

Tekst: {text}
```

Figure 2: Prompt used for Silesian translation with GPT-5.4, Bielik and PLLum models. English translation: "Translate the following sentence into the Silesian language. Return only the translation in Silesian (without comments, explanations, or additional text). Text: {text}."

C.3 Google Translate

This system is included as a representative industry standard for production-grade translation quality. To ensure reproducibility, we note that the translations were processed in April and May 2026, through the Google Cloud Translation API (Wu et al., 2016),

C.4 TranslateGemma-4B-IT (Base and Fine-tuned)

We evaluate both base and custom fine-tuned variants, the details of which are provided in section 6. (Finkelstein et al., 2026)

We selected this model because it was trained on the English–Silesian translation direction, which we expected could transfer well to the Polish–Silesian translation task. However, the model's chat template did not include support for the Silesian language. To address this limitation, we manually extended the chat template by adding the Silesian language support as "szl": "Silesian (Śląski)" and then performed inference using the modified chat template.

C.5 MADLAD-400-10B-MT

MADLAD-400-10B-MT is a multilingual machine translation model based on the T5 architecture. It was trained on 250 billion tokens of publicly available data covering more than 450 languages, including Silesian.

C.6 Bielik-PL-11B-v3.0-IT

This is a multilingual model with a strong emphasis on Polish. Notably, it was also pretrained on Silesian-language data. We used the system prompt shown in Figure 3 and the user prompt shown in Figure 2.

```
Jesteś ekspertem od tłumaczeń polsko-śląskich.
```

Figure 3: System prompt used for Silesian translation with Bielik model. English translation: "You are an expert in Polish-Silesian translations."

C.7 PLLuM-12B-nc-chat

This model was optimized for Polish and designed for general-purpose chat, question answering, and text generation tasks. We used the user prompt shown in Figure 2.

C.8 Prompt Configuration

To ensure a fair comparison across models, we kept the prompting configuration as close as possible to the default settings provided by each model interface. The goal of this study was to compare the translation capabilities of the evaluated models rather than optimize their performance through extensive prompt engineering or parameter tuning. For this reason, we avoided additional system-level modifications, handcrafted instructions, or decoding adjustments beyond the prompts presented in Figures 2 and 3.

D Expanded results

D.1 Silesian to Polish direction

Although our primary focus is the Polish-to-Silesian direction, we additionally evaluate on the reverse Silesian-to-Polish translation task using the BOUQuET benchmark. Since Polish is a high-resource language, this direction is generally easier for translation systems and generative models.

Table 4 presents the evaluation results. GPT-5.4 achieves the best overall performance with a BLEU score of 41.1 and chrF of 63.2, slightly outperforming Google Translate. The rule-based system also achieves competitive results, reaching 31.7 BLEU and 57.4 chrF, outperforming both TranslateGemma and NLLB 3.3B. This suggests that manually designed linguistic rules and dictionaries remain effective for closely related language pairs such as Silesian and Polish, where systematic lexical and orthographic correspondences can be captured explicitly. TranslateGemma obtains moderate results, while NLLB 3.3B performs substantially worse than the strongest proprietary systems. In many cases, NLLB 3.3B failed to translate Silesian-specific words and expressions, leaving them unchanged in the Polish output.

Model	BLEU	chrF
NLLB 3.3B	18.0	32.5
TransGemma	24.8	48.4
GPT-5.4	**41.1**	**63.2**
Google Translate	38.3	61.7
ours		
Rule-Based	31.7	57.4

Table 4: Evaluation results on the BOUQuET SZL→POL translation task.

D.2 Flores

Table 5 show that Google Translate achieves the best performance on FLORES system. The unusually strong performance may indicate that the model was exposed during training to FLORES sentences, especially since this benchmark has been publicly available for a long time. However, this advantage does not transfer to the SiLTT and BOUQuET datasets.

Model	**BLEU**	**chrF**
PLLuM-12B-nc-chat	0.1	12.1
Bielik-PL-11B-v3.0-IT	5.7	35.1
MADLAD-400-10B-MT	9.8	37.2
NLLB 3.3B	7.4	34.8
NLLB 54B MOE	7.4	37.5
TransGemma	5.1	32.1
GPT-5.4	11.9	44.2
Google Translate	**17.7**	**48.6**
ours		
Rule-Based	14.9	48.3
TransGemma-FT (CBT-22k)	12.9	45.8
TransGemma-FT (Mixed)	10.4	41.4

Table 5: Results on the FLORES benchmark for POL→SZL translation. Google Translate achieves the best performance overall, followed by the rule-based system. TranslateGemma fine-tuned on CBT-22k improves over the base model but remains below the strongest systems.