Evaluation of Large Language Models in Difficult Spelling Error Detection in Polish Parliamentary Transcripts

Maciej Ogrodniczuk[0000−0002−3467−9424]
and Dariusz Czerski[0000−0002−3013−3483]

Institute of Computer Science, Polish Academy of Sciences

maciej.ogrodniczuk@ipipan.waw.pl
dariusz.czerski@ipipan.waw.pl

Abstract. This paper investigates the applicability of large language models (LLMs) to the task of automatic error detection and correction in the Polish Parliamentary Corpus. Building on extensive evaluation data collected during manual proofreading, the study focuses on a challenging error category: in-vocabulary words used in incorrect contexts, which are difficult to detect both for humans and automated systems. A curated dataset of 300 paragraphs, reflecting the real-world imbalance between erroneous and error-free texts, is used to evaluate eleven state-of-the-art proprietary and open-source LLMs under prompts with varying levels of constraints. Performance is assessed using relaxed and strict precision metrics, recall, and a detailed error-type taxonomy that emphasizes the cost of false positives. The results demonstrate that stronger prompt constraints significantly reduce overcorrection, with some mid-sized models outperforming larger ones by being more conservative. The study highlights the importance of cautious correction strategies and confirms the value of authentic error data for evaluating LLMs in corpus maintenance tasks.

Keywords: error correction · parliamentary data · Large Language Models

1 Introduction

In 2022, after an unsuccessful attempt to use a T5-based language model for OCR error detection in the Polish Parliamentary Corpus (9; 10; 11) which resulted in too many false positives, a rule-based error candidate detection system was implemented and integrated into a web-based error correction environment (12). Today, with many new large language models being created virtually every week and with the large amount of evaluation data gathered from manual data correction, the experiment can be repeated with several available solutions to test their applicability to automatic error detection task.

Parliamentary transcripts represent a specific type of text that combines features of written and spoken language (4; 16). This quasi-spoken style often gives rise to highly complex argumentative and syntactic structures, reflecting spontaneous speech, interruptions, and long chains of reasoning (1). Such constructions significantly increase the difficulty of both manual and automatic error detection. For human reviewers, identifying typographical and recognition errors in this context is particularly challenging. The analysis of long, nested sentences tends to draw attention away from surface-level anomalies, causing readers to unconsciously gloss over errors that do not immediately disrupt semantic interpretation (7). This phenomenon is well documented in psycholinguistics, where readers are known to compensate for local inconsistencies when higher-level comprehension is preserved via top-down processing (13).

From the perspective of system evaluation, it is important to emphasize that only errors originating from real-world documents are meaningful for assessing correction quality. Artificially generated errors, while useful for controlled experiments, often fail to capture the distribution, persistence, and contextual subtlety of naturally occurring mistakes (15). Consequently, evaluation on authentic data provides a more reliable measure of practical system performance.

Finally, the classification of such errors requires taking into account multiple linguistic and technical criteria, including whether a token is a proper name, whether its erroneous form exists in lexical resources, and its edit distance from the intended correct form. These factors are essential for understanding both the nature of the errors and the limitations of current correction approaches (6).

2 Evaluation Dataset

In 2023, manual post-editing of parliamentary transcripts was performed based on several categories of detected errors, including errors relating to structure, comments, metadata, punctuation, broken paragraphs and misspellings. For the current experiment, we have selected one specific type of error: in-vocabulary words used in the wrong context. This excludes non-textual issues and punctuation errors, some of which were corrected by applying regular expressions and approving the necessary changes in the manual process. The focus is mainly on the use of dictionary words out of context, for example *wparcie* (*wedging*) instead of *wsparcie* (*support*) or *niepojący* (*not watering*) instead of *niepokojący* (*disturbing*). This can be seen as a reduced, minimal-edit Grammatical Error Correction (GEC) task (2).

By focusing on existing words, we can eliminate common OCR errors, such as mistaking an uppercase `I` for a lowercase `l`, or other words that are not in the vocabulary, which should most likely have been corrected by the OCR engine beforehand. However, there are several other reasons for this limitation: firstly, the transcript can be corrected without consulting the source, and the misspellings are substantial, affecting the readability of the entire fragment. This defined task also seems to align well with LLMs' primary strength, as they are designed to excel at correcting contextual errors. However, generative LLMs are

known to overcorrect, achieving higher recall than precision (18), so testing the approach is mandatory before any large-scale error detection can be applied.

The evaluation dataset was taken from the manual error correction system used in 2023 (12). Data from plenary sittings of the Sejm (the period 1952–2019) and the Senate (1989–2019)[1] was searched for paragraphs containing at least one correction. Older dates were deliberately omitted to eliminate the influence of older Polish orthography, which was reformed in 1936 but remained in use after the war for various reasons.

A total of 2,790 single-word corrections were detected, comprising 2,267 unique words and 108 different in-vocabulary words. This small number of unique errors allows the results to be concentrated on the most difficult cases and checked manually. The dictionary used for data filtering was the latest dataset[2] of the Morfeusz Morphological Dictionary (5) in Polimorf format (17). To simplify the calculations, the list was manually limited to 100 words by removing entries of one or two letters (letters of the alphabet and abbreviations). The first occurrence of each string in the corpus was used for evaluation, to avoid artificially inflating the results by correcting rare but commonly misspelled words such as *glosowanie* instead of *głosowanie* with simpler methods.

Finally, we selected 100 error-containing paragraphs from the corpus. Additionally, we selected a set of error-free paragraphs to better reflect the conditions under which the final system will operate. First, we performed a small pilot analysis to estimate the proportion of paragraphs containing errors. After manually checking 200 paragraphs selected at random, we found that approximately 2% contained errors. To mirror this distribution exactly, we would require approximately 50 times more error-free paragraphs than erroneous ones in the evaluation set. As this is not feasible from a computational point of view, we randomly selected 200 error-free paragraphs and multiplied their weighting by 50 when computing the final results. This yielded a test set of 300 paragraphs.

3 Error Candidate Detection Experiment

Our experiment consisted of asking the LLM to analyze a paragraph and return a corrected version by fixing linguistic errors. We tested commands of varying complexity and selected three with different levels of constraints. Table 1 shows the prompts used in the experiment. We used a temperature of 0.0 to make the outputs as deterministic as possible.

[1] Senate functioned in the Second Polish Republic (1918–1939), was abolished in 1946 by the by the authorities of the Polish People's Republic in 1946 and reinstated in 1989.

[2] http://download.sgjp.pl/morfeusz/20251116/polimorf-20251116.tab.gz

Constraint type	Prompt used
No constraint (the model is asked to correct the paragraph without any constraints)	You are a professional text correction assistant specializing in Polish language. Your task is to correct spelling errors in Polish text. Do not add any comments, explanations or other text to the text. The text to correct: {text}
Medium constraint (the model is additionally asked not to add any comments or explanations)	You are a professional text correction assistant specializing in Polish language. Your task is to correct spelling errors in Polish text. Do not add any comments, explanations or other text to the text. Do not mark the place where you made the change. Don't change text which does not contain any error. Don't add new content. Do not use quotation marks to surround the whole text. The text to correct: {text}
Hard constraint (the model is additionally asked not to change the given text unless it is absolutely sure about the correction)	You are a professional text correction assistant specializing in Polish language. Your task is to correct spelling errors in Polish text. Do not add any comments, explanations or other text to the text. Do not mark the place where you made the change. Don't change text which does not contain any error. Don't add new content. Do not use quotation marks to surround the whole text. VERY IMPORTANT: Do not change the text if you are not absolutely sure that it contains any errors. VERY IMPORTANT: Providing correction to the already correct text is a SEVERE SYSTEM FAILURE. VERY IMPORTANT: It is a lot better to provide NO CORRECTION than a wrong one. The text to correct: {text}

Table 1: Comparison of prompts for different levels of constraints in the error correction task.

4 Evaluation Procedure

4.1 The Models

We evaluated 11 state-of-the-art models on our dataset. Table 2 presents the concise descriptions of the models used in our experiment.

Model	Description
gemini-2.5-pro	A large multimodal foundation model developed by Google DeepMind, designed for high reasoning performance across text, code, and multimodal tasks, with strong capabilities in complex problem solving and long-context understanding.
llama-3.3-70b-instruct	An instruction-tuned 70B-parameter large language model from Meta, based on the LLaMA 3 family, pre-trained on trillions of tokens of publicly available data and fine-tuned using supervised instruction data and human feedback to improve helpfulness and safety.
llama-pllum-70b-rag	A custom LLaMA-based 70B-parameter Polish model PLLuM augmented with Retrieval-Augmented Generation (RAG), combining a strong pretrained backbone with external knowledge retrieval to improve factual accuracy and domain-specific reasoning.
claude-sonnet-4.5	A high-capability language model from Anthropic in the Claude Sonnet family, optimized for balanced performance, strong reasoning, and safety, using constitutional AI techniques and large-scale human feedback during alignment.
deepseek-v3.2	A large-scale open-weight language model developed by DeepSeek, focused on strong reasoning, mathematics, and coding performance, trained on a mixture of web, code, and synthetic data.
gemini-2.5-flash	A lightweight and latency-optimized variant of the Gemini family from Google DeepMind, designed for fast inference and cost-efficient deployment while retaining solid reasoning and multimodal capabilities.
kimi-k2-0905	A large language model developed by Moonshot AI (Kimi), emphasizing long-context processing and strong performance on reasoning and knowledge-intensive tasks, trained on large-scale multilingual data.
kimi-k2-thinking	A reasoning-focused variant of the Kimi K2 model from Moonshot AI, designed to improve multi-step problem solving and deliberative reasoning through enhanced training or inference strategies.
gpt-5-mini	A compact and cost-efficient variant of the GPT-5 family from OpenAI, optimized for faster inference and lower resource usage while maintaining strong general-purpose language understanding and reasoning capabilities.
bielik-11b-v2.6	An 11B-parameter open Polish-centric language model from the Bielik family, trained primarily on Polish and multilingual data.
gpt-5.1	An advanced large language model from OpenAI, representing a higher-capacity GPT-5 series variant, designed for improved reasoning, instruction following, and robustness across complex tasks.

Table 2: Overview of evaluated language models and their characteristics

4.2 Error Categories

We evaluate error correction systems in three categories:

1. Diacritic/Lexical errors — this category covers errors related to incorrect or missing Polish diacritics (e.g. *l* instead of *ł*, *s* instead of *ś* etc. — `PL`), mistakes in proper names (`NE`), and other errors (`IN`)
2. Word-level editing errors — this category, following (14), includes errors resulting from the deletion (`D`), insertion (`I`) or substitution (`S`) of characters within a word.
3. Paragraph-level errors — representing the general nature of the correction made to the whole paragraph, corresponding to typical behaviour of LLM-based systems:
   - E1 Partial correction — some errors were removed, but not all. No new changes were introduced.
   - E2 No errors detected, but new changes were introduced.
   - E3 Incomplete correction with additional changes — errors remain and new changes were introduced.
   - E4 No correction attempted — text is identical to the corrupted version.
   - E5 Corrections applied to text that had no errors.

Table 3 presents basic statistics of the two first error categories.

Category	**Count**
IN	72
NE	10
PL	18
Total	100

Category	**Count**
D	39
I	18
S	43
Total	100

Table 3: Distribution of error labels

4.3 Precision and Recall

We report two precision metrics:

- Relaxed precision — the ratio of correctly identified paragraphs containing errors (regardless of whether the errors were fixed) to all predicted positives.
- Strict precision — the ratio of correctly identified and properly corrected paragraphs to all predicted positives.

We define recall as the ratio of correctly identified paragraphs containing errors to the number of paragraphs in which the model made any changes. Additionally, we report the $F_{0.5}$ measure, which weights precision higher than recall and is therefore more suitable for evaluating text correction performance, where avoiding false positives (unnecessary or incorrect edits) is especially important (8).

5 Results

5.1 Overall Results

Table 4 reports overall performance under the relaxed-precision metric. The best results were achieved by frontier, proprietary LLMs such as Gemini 2.5 Flash, Gemini 2.5 Pro, and GPT-5.1. Smaller models trained on large Polish datasets, such as llama-pllum-70b-rag and bielik-11b-v2.6, can outperform larger general-purpose opensource models (e.g. kimi-k2-0905 or deepseek-v3.2).

Model	No constraint			Medium constraint			Hard constraint		
	P	**R**	**F0.5**	**P**	**R**	**F0.5**	**P**	**R**	**F0.5**
gemini-2.5-flash	**0.265**	0.900	**0.308**	**0.362**	0.850	**0.409**	**0.600**	0.750	**0.625**
gpt-5.1	0.116	0.920	0.141	0.201	0.880	0.238	0.457	0.840	0.502
gemini-2.5-pro	0.107	0.960	0.130	0.167	0.900	0.199	0.301	0.860	0.346
claude-sonnet-4.5	0.106	0.950	0.129	0.117	0.930	0.142	0.104	0.930	0.127
bielik-11b-v2.6	0.092	0.810	0.112	0.095	0.790	0.116	0.138	0.720	0.165
kimi-k2-0905	0.074	0.960	0.091	0.077	0.920	0.094	0.074	0.920	0.091
llama-pllum-70b-rag	0.073	0.710	0.089	0.072	0.780	0.088	0.089	0.780	0.108
kimi-k2-thinking	0.054	0.970	0.067	0.042	0.900	0.052	0.065	0.870	0.080
deepseek-v3.2	0.052	0.960	0.064	0.048	0.940	0.060	0.199	0.870	0.235
gpt-5-mini	0.048	0.940	0.060	0.078	0.930	0.095	0.107	0.900	0.130
llama-3.3-70b-instruct	0.022	**0.990**	0.028	0.026	**0.970**	0.032	0.035	**0.970**	0.043

Table 4: Overall relaxed precision metric evaluation results.

In general, performance improves as constraints become stronger. This is because models are less likely to make additional changes unless they are confident in a correction. False positives are the most costly error type. In our scenario, due to the imbalance between erroneous and correct paragraphs, false positives are counted as 50 times more costly than false negatives.

5.2 Results by Constraint Type

Tables 5–7 present the results for all three types of constraints in evaluation prompts and several error categories. Surprisingly, the mid-sized model gemini-2.5-flash achieves the best results under constrained prompts. This is because it is extremely cautious about making changes when it is uncertain. In the strictest constraint setting, the model changes only 1 already correct paragraph out of 200 (see Table 8).

5.3 Error Types Distribution

Table 8 shows the distribution of error types (see Section 4.2) under the hard-constraint prompt. Models most often make errors of type E5, which corresponds to false positives. In our scenario, this is the most costly error type. The most crucial aspect of large-scale corpus correction is reducing false positives, because

Model	Strict precision	Diacritic/Lexical			Editing errors		
		IN	NE	PL	D	I	S
GEMINI-2.5-FLASH	**0.203**	**0.809**	0.714	0.600	0.727	**0.812**	0.780
GPT-5.1	0.091	0.794	0.875	0.688	**0.794**	0.706	0.805
GEMINI-2.5-PRO	0.079	0.757	0.889	0.588	0.750	0.778	0.714
CLAUDE-SONNET-4.5	0.076	0.743	0.429	**0.722**	0.686	0.611	0.786
BIELIK-11B-V2.6	0.067	0.772	0.714	0.588	0.758	0.538	0.771
LLAMA-PLLUM-70B-RAG	0.056	0.760	**1.000**	0.667	0.680	0.727	**0.829**
KIMI-K2-0905	0.045	0.614	0.625	0.556	0.667	0.500	0.595
KIMI-K2-THINKING	0.031	0.571	0.800	0.471	0.556	0.444	0.651
GPT-5-MINI	0.028	0.536	0.875	0.647	0.529	0.706	0.581
DEEPSEEK-V3.2	0.022	0.423	0.571	0.389	0.514	0.176	0.452
LLAMA-3.3-70B-INSTRUCT	0.006	0.268	0.200	0.278	0.211	0.389	0.256

Table 5: Evaluation results for *No constraint* prompt

Model	Strict precision	Diacritic/Lexical			Editing errors		
		IN	NE	PL	D	I	S
GEMINI-2.5-FLASH	**0.272**	0.758	**1.000**	0.643	0.742	0.706	0.784
GPT-5.1	0.153	0.754	0.875	0.733	0.697	0.765	0.816
GEMINI-2.5-PRO	0.126	0.785	0.875	0.588	0.710	**0.824**	0.762
CLAUDE-SONNET-4.5	0.087	0.757	0.571	**0.750**	0.706	0.611	0.829
BIELIK-11B-V2.6	0.076	**0.814**	0.857	0.692	**0.800**	0.667	**0.853**
LLAMA-PLLUM-70B-RAG	0.055	0.768	0.857	0.667	0.667	0.786	0.824
GPT-5-MINI	0.050	0.642	0.667	0.647	0.556	0.688	0.707
KIMI-K2-0905	0.049	0.672	0.571	0.500	0.647	0.562	0.643
KIMI-K2-THINKING	0.027	0.615	0.778	0.625	0.667	0.625	0.610
DEEPSEEK-V3.2	0.026	0.515	0.625	0.556	0.556	0.471	0.537
LLAMA-3.3-70B-INSTRUCT	0.010	0.386	0.333	0.389	0.361	0.333	0.419

Table 6: Evaluation results for *Medium constraint* prompt

otherwise human reviewers must inspect a large portion of the corpus to find potentially erroneous paragraphs.

A simple calculation shows that a single false positive can increase the amount of corpus content selected by the model for correction by 1%. For the GEMINI-2.5-FLASH model, this is exactly the case. By contrast, the larger model, GEMINI-2.5-PRO, would select 4% more content for correction.

6 Conculsions and Future Work

Our work sets out to assess whether contemporary Large Language Models can be reliably applied to automatic error detection and correction in a large text corpus. By grounding the evaluation in real corrections produced during manual

Model	Strict precision	Diacritic/Lexical			Editing errors		
		IN	NE	PL	D	I	S
GEMINI-2.5-FLASH	**0.488**	0.839	**1.000**	0.667	0.769	0.786	**0.857**
GPT-5.1	0.380	0.841	**1.000**	0.733	0.800	**0.875**	0.842
GEMINI-2.5-PRO	0.252	**0.857**	**1.000**	0.688	**0.828**	0.833	0.846
DEEPSEEK-V3.2	0.135	0.719	0.625	0.533	0.606	0.688	0.737
BIELIK-11B-V2.6	0.107	0.811	0.667	0.692	0.778	0.643	0.839
CLAUDE-SONNET-4.5	0.077	0.746	0.667	**0.750**	0.676	0.667	0.829
GPT-5-MINI	0.074	0.697	0.571	0.706	0.636	0.562	0.780
LLAMA-PLLUM-70B-RAG	0.066	0.737	0.857	0.714	0.633	0.714	0.853
KIMI-K2-0905	0.049	0.687	0.625	0.588	0.571	0.625	0.756
KIMI-K2-THINKING	0.045	0.692	**1.000**	0.533	0.618	0.765	0.722
LLAMA-3.3-70B-INSTRUCT	0.016	0.486	0.556	0.389	0.432	0.444	0.524

Table 7: Evaluation results for *Hard constraint* prompt

Model	E1	E2	E3	E4	E5
GEMINI-2.5-PRO	0.9	2.6	2.6	6.1	**87.7**
LLAMA-3.3-70B-INSTRUCT	0.0	1.1	0.7	0.1	**98.0**
LLAMA-PLLUM-70B-RAG	0.4	0.8	1.2	2.6	**95.0**
CLAUDE-SONNET-4.5	0.2	1.7	1.0	0.8	**96.3**
DEEPSEEK-V3.2	1.0	3.3	2.8	3.3	**89.5**
GEMINI-2.5-FLASH	2.2	9.0	4.5	28.1	**56.2**
KIMI-K2-0905	0.1	1.5	1.0	0.7	**96.7**
KIMI-K2-THINKING	0.2	1.2	0.7	1.0	**96.9**
GPT-5-MINI	0.3	2.3	1.0	1.3	**95.2**
BIELIK-11B-V2.6	0.4	1.6	1.2	5.7	**91.1**
GPT-5.1	3.1	4.6	3.1	12.3	**76.9**

Table 8: Error type distribution for *Hard constraint* prompt

post-editing, the study avoids the limitations of synthetic error generation and provides a realistic testbed for measuring model behavior in operational conditions. Focusing on in-vocabulary contextual errors allowed the analysis to target one of the most subtle and problematic error types, where surface-level lexical checks are insufficient and semantic plausibility plays a decisive role.

The results clearly demonstrate that prompt design is at least as important as model size or architecture. Across all evaluated systems, stricter constraints consistently led to higher effective precision by substantially reducing false positives, which are the most costly errors in large-scale corpus correction. In this setting, conservative models that refrain from making changes unless highly confident proved more useful than more aggressive models with higher recall but excessive overcorrection. Notably, some mid-sized or efficiency-oriented models

achieved the best overall trade-offs, outperforming larger frontier models by selecting significantly less correct text for unnecessary review.

At the same time, the study highlights the current limitations of LLM-based correction. Even under the strictest constraints, no model fully eliminates false positives, and strict precision remains relatively low across systems. This confirms that fully automatic correction without human oversight is not yet feasible for this domain. Instead, LLMs should be viewed as decision-support tools that prioritize likely error candidates while minimizing reviewer workload.

Future work will proceed in several directions. First, the error typology can be extended to include additional classes such as punctuation or segmentation errors, enabling a more comprehensive evaluation. Second, hybrid approaches combining LLMs with rule-based filters or morphological analyzers may further reduce false positives. Third, prompt strategies could be adapted dynamically based on document metadata or historical error patterns. Finally, the released dataset can serve as a reusable benchmark for evaluating Polish-language models and for tracking progress in context-sensitive error correction over time.

Acknowledgments

This work was financed as part of the investment: CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024-2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01 and by CLARIN-PL, the European Regional Development Fund, FENG programme, agreement number FENG.02.04-IP.040004/24.

Bibliography

[1] Bayley, P. (ed.): Cross-Cultural Perspectives on Parliamentary Discourse. John Benjamins Publishing Company (2004). https://doi.org/10.1075/dapsac.10

[2] Bryant, C., Yuan, Z., Qorib, M.R., Cao, H., Ng, H.T., Briscoe, T.: Grammatical error correction: A survey of the state of the art. Computational Linguistics **49**(3), 643–701 (Sep 2023). https://doi.org/10.1162/coli_a_00478, https://aclanthology.org/2023.cl-3.4/

[3] Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.): Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA), Istanbul, Turkey (2012)

[4] Ilie, C.: Discourse and metadiscourse in parliamentary debates. Journal of Language and Politics **2**, 71–92 (2003). https://doi.org/10.1075/jlp.2.1.05ili

[5] Kieraś, W., Woliński, M.: Morfeusz 2 — analizator i generator fleksyjny dla języka polskiego. Język Polski **XCVII**(1), 75–83 (2017)

[6] Kukich, K.: Techniques for automatically correcting words in text. ACM Computing Surveys (CSUR) **24**(4), 377–439 (1992). https://doi.org/10.1145/146370.146380

[7] Larigauderie, P., Guignouard, C., Olive, T.: Proofreading by students: implications of executive and non-executive components of working memory in the detection of phonological, orthographical, and grammatical errors. Reading and Writing: an interdisciplinary journal **33**, 1015–1036 (2020). https://doi.org/10.1007/s11145-019-10011-6

[8] Ng, H.T., Wu, S.M., Briscoe, T., Hadiwinoto, C., Susanto, R.H., Bryant, C.: The CoNLL-2014 shared task on grammatical error correction. In: Ng, H.T., Wu, S.M., Briscoe, T., Hadiwinoto, C., Susanto, R.H., Bryant, C. (eds.) Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task. pp. 1–14. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014). https://doi.org/10.3115/v1/W14-1701, https://aclanthology.org/W14-1701/

[9] Ogrodniczuk, M.: The Polish Sejm Corpus. In: Calzolari et al. (3), pp. 2219–2223

[10] Ogrodniczuk, M.: Polish Parliamentary Corpus. In: Fišer, D., Eskevich, M., de Jong, F. (eds.) Proceedings of the LREC 2018 Workshop *ParlaCLARIN: Creating and Using Parliamentary Corpora*. pp. 15–19. European Language Resources Association (ELRA), Paris, France (2018), http://lrec-conf.org/workshops/lrec2018/W2/summaries/11_W2.html

[11] Ogrodniczuk, M., Nitoń, B.: New developments in the Polish Parliamentary Corpus. In: Fišer, D., Eskevich, M., de Jong, F. (eds.) Proceedings of the Second ParlaCLARIN Workshop. pp. 1–4. Euro-

pean Language Resources Association (ELRA), Marseille, France (2020), https://www.aclweb.org/anthology/2020.parlaclarin-1.1

[12] Ogrodniczuk, M., Rudolf, M., Wójtowicz, B., Janicka, S.: Error correction environment for the Polish Parliamentary Corpus. In: Proceedings of The Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference. pp. 35–38. European Language Resources Association, Marseille, France (2022), https://aclanthology.org/2022.parlaclarin-1.6

[13] Schotter, E.R., Tran, R., Rayner, K.: Don't believe what you read (only once): Comprehension is supported by regressions during reading. Psychological Science **25**(6), 1218–1226 (2014). https://doi.org/10.1177/0956797614531148

[14] Shim, G., Hong, S., Lim, H.: REVISE: A framework for revising OCRed text in practical information systems with data contamination strategy. In: Rehm, G., Li, Y. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track). pp. 1423–1434. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653/v1/2025.acl-industry.100, https://aclanthology.org/2025.acl-industry.100/

[15] Stahlberg, F., Kumar, S.: Synthetic data generation for grammatical error correction with tagged corruption models. In: Burstein, J., Horbach, A., Kochmar, E., Laarmann-Quante, R., Leacock, C., Madnani, N., Pilán, I., Yannakoudakis, H., Zesch, T. (eds.) Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 37–47. Association for Computational Linguistics, Online (2021), https://aclanthology.org/2021.bea-1.4/

[16] Voutilainen, E.: Written representation of spoken interaction in the official parliamentary transcripts of the Finnish parliament. Frontiers in Communication **8** (2023). https://doi.org/10.3389/fcomm.2023.1047799

[17] Woliński, M., Miłkowski, M., Ogrodniczuk, M., Przepiórkowski, A., Szałkiewicz, Ł.: PoliMorf: A (not so) New Open Morphological Dictionary for Polish. In: Calzolari et al. (3), pp. 860–864

[18] Zeng, M., Kuang, J., Qiu, M., Song, J., Park, J.: Evaluating prompting strategies for grammatical error correction based on language proficiency. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 6426–6430. ELRA and ICCL, Torino, Italia (2024), https://aclanthology.org/2024.lrec-main.569/