How Good is AI on Swiss Voting Booklets? A Multilingual OCR and Alignment Benchmark

Elina Stüssi and **Jannis Vamvas**
Department of Computational Linguistics
University of Zurich
elina.stuessi@uzh.ch, vamvas@cl.uzh.ch

Abstract

Swiss federal voting booklets are an interesting resource for natural language processing due to their high editing standards and coverage of the four national languages of Switzerland (German, French, Italian, and Romansh Grischun). In this paper, we present VotingBooklets, an automatically extracted and aligned dataset, as well as VotingBooklets-Diamond, a subset that was manually corrected and verified by multiple annotators. We use the latter to benchmark a range of open and closed AI systems on two interdependent tasks: optical character recognition (OCR) and cross-lingual text alignment. Gemini 2.5 Flash Lite achieves the best OCR performance across all conditions, while a hybrid alignment approach using Sentence-SwissBERT for initial embedding-based alignment and Gemini for targeted post-hoc correction of low-confidence pairs yields the most accurate results. Applying these systems to the full collection of Swiss federal voting booklets, we release a large-scale four-language parallel corpus as a resource for low-resource NLP, multilingual representation learning, and the computational study of Swiss political discourse.

1 Introduction

Multilingual archives hold significant potential for computational research, but realizing this potential requires converting physical documents into machine-readable digital text. For historical administrative documents, this involves two core tasks: optical character recognition (OCR) to extract text from scanned pages, and cross-lingual alignment to link parallel content across languages.

In this paper, we investigate both tasks jointly using Swiss federal voting booklets (*Abstimmungsbüchlein*) as our document collection. These booklets are distributed to all eligible voters prior to each federal referendum and present identical political content in the four Swiss national languages. Their parallel structure makes them a uniquely suitable resource for evaluating multilingual document processing pipelines. We introduce VotingBooklets-Diamond[1], a carefully curated, manually corrected test set covering three voting dates (1977, 1985, and 2007) and use it to benchmark a range of open and proprietary AI systems on both OCR and cross-lingual alignment.

German	Bundesrat und Parlament empfehlen den Stimmberechtigten, am 11. März 2007 wie folgt zu stimmen:
French	Le Conseil fédéral et le Parlement vous recommandent de voter, le 11 mars 2007:
Italian	Consiglio federale e Parlamento vi raccomandano di votare come segue l'11 marzo 2007:
Romansh	Il cussegl federal ed il parlament recumondan a las votantas ed als votants da votar ils 11 da mars 2007 sco suonda:

Figure 1: Example of a parallel sentence from the VotingBooklets-Diamond dataset, illustrating all four Swiss national languages (German, French, Italian, Romansh) from the federal vote of 11 March 2007.

Our VotingBooklets-Diamond test set presents two central challenges. First, it includes languages with markedly different levels of resource availability. German, French, and Italian are well supported in modern NLP systems, whereas Romansh Grischun, spoken by roughly 60,000 people, remains a low-resource language with limited training data and few dedicated tools. Second, the two tasks we consider are inherently interdependent. Errors introduced during OCR propagate to downstream alignment and degrade its quality. Figure 1

[1]The dataset is available at https://huggingface.co/datasets/eljuanina/VotingBooklets-Diamond-v1

Vote	Languages	Pages	PDF Format	Notes
1977	DE, FR, IT	8	Scanned	No Romansh; two content pages per scan page
1985	DE, FR, IT, RM	4	Scanned	Romansh added; two content pages per scan page
2007	DE, FR, IT, RM	16	Born-digital	High-quality machine-readable PDFs

Table 1: Overview of the Swiss voting booklets included in VotingBooklets-Diamond. The *Pages* column refers to PDF scan pages; the 1977 and 1985 booklets contain 15 and 8 content pages respectively, with two content pages printed per scan page. Language abbreviations: DE=German, FR=French, IT=Italian, RM=Romansh.

Year	German (de) Tokens	French (fr) Tokens	Italian (it) Tokens	Romansh (rm) Tokens
1977	3,994	4,846	4,366	–
1985	1,618	2,193	1,821	2,040
2007	2,000	2,595	2,413	2,766
Total	7,612	9,634	8,600	4,806

Table 2: Token count per language and vote in the **VotingBooklets-Diamond** dataset. Romansh (`rm`) is only available for 1985 and 2007.

shows an example sentence from a Swiss federal voting booklet in all four national languages, illustrating the highly parallel structure that underlies our dataset.

Beyond benchmarking, we apply our findings to the full collection of Swiss federal voting booklets. Gemini 2.5 Flash Lite achieves the best OCR performance on VotingBooklets-Diamond and is therefore used to process the full collection of Swiss federal voting booklets. For alignment, we employ a two-stage approach that combines embedding-based alignment with Sentence-SwissBERT (Grosjean and Vamvas, 2024) and targeted post-hoc correction using Gemini 2.5 Flash Lite to refine low-confidence matches. The resulting dataset, VotingBooklets[2], forms a large-scale four-language parallel corpus that serves as a new resource for research in low-resource NLP, multilingual representation learning, and the computational study of political discourse in multilingual societies.

2 Corpus Design

2.1 Document Collection

Swiss federal voting booklets are official documents issued by the Swiss Federal Chancellery to all eligible voters before each federal referendum.[3] Under Switzerland's system of direct democracy, referendums are held several times per year, and each booklet presents the same content in all four national languages: German, French, Italian, and Romansh Grischun. This makes them a naturally occurring resource of highly parallel multilingual text produced under real-world institutional conditions.

Our VotingBooklets-Diamond dataset comprises eleven voting booklets drawn from three federal votes: 12 June 1977, 1 December 1985, and 11 March 2007 (Table 1). These dates were selected to span three decades while maximizing linguistic and technical variation. The earliest booklet available online, from 1977, predates the introduction of Romansh editions and is therefore limited to German, French, and Italian. The 1985 vote marks the first inclusion of Romansh Grischun, while the 2007 materials are distributed as born-digital PDFs rather than scanned documents. Taken together, these three time points capture substantial variation in document quality, scan resolution, typography, and layout.

Each booklet follows a structured format comprising vote overviews, official proposals, arguments for and against, parliamentary recommendations, legal texts, and voting instructions. At the same time, the documents exhibit considerable layout complexity, with multi-column text, tables, and footnotes that pose challenges for OCR and layout analysis systems. Despite minor translation-level phrasing differences, the content is substantively equivalent across all language versions, enabling precise cross-lingual alignment.

[2]The dataset is available at https://huggingface.co/datasets/eljuanina/VotingBooklets-v1

[3]Available from the Swiss Federal Chancellery: https://www.bk.admin.ch/bk/de/home/dokumentation/abstimmungsbuechlein.html.

Language	German (de)		French (fr)		Italian (it)		Romansh (rm)	
	Tokens	Pages	Tokens	Pages	Tokens	Pages	Tokens	Pages
Collection	1,119,494	5,037	1,420,348	5,089	1,065,848	5,108	859,802	4,058

Table 3: Token and page counts per language in the **VotingBooklets** dataset (complete collection of Swiss federal voting booklets).

2.2 Gold Standard Annotation

We created a gold-standard transcription for the eleven booklets through manual correction. The scanned PDFs from 1977 and 1985 carry an existing OCR layer, but its quality is poor and served only as a rough starting point. For the 2007 born-digital PDFs, we extracted text directly from the PDF. In both cases, every passage was fully reviewed and corrected by hand to ensure fidelity to the original document. Transcription follows the physical page sequence of the PDF files rather than inferred reading order, so that OCR and layout analysis tools can be evaluated fairly regardless of whether they correctly reconstruct logical flow from non-linear layouts. For dual-page scans, left-hand pages were transcribed in full before right-hand pages, preserving physical layout structure. Non-textual elements, such as photographs, logos, and purely illustrative graphics, were not transcribed or described in the final dataset. Each transcription was independently checked by two additional annotators, who compared the transcriptions directly with the original PDF documents to ensure accuracy. Any discrepancies found were due to minor typographical errors or omitted words in the initial transcription. All such issues were carefully reviewed and corrected through discussion, achieving full agreement among the annotators.

Cross-lingual alignment was performed at the paragraph level, using the physical paragraph boundaries present in the PDF as the primary segmentation unit. Where structural divergence across languages required it, adjacent paragraphs were merged to ensure semantic equivalence across aligned pairs. Alignment decisions were verified by two annotators, both native German speakers with basic knowledge of French and Italian. They could infer the meaning of Romansh segments based on their knowledge of the other languages, and in cases of uncertainty, they consulted the original booklets. There were no disagreements between the annotators.

2.3 Corpus Statistics

Table 2 shows token counts for VotingBooklets-Diamond across all three voting dates and four languages. Each booklet is represented as a separate file, with German (`de`), French (`fr`), Italian (`it`), and Romansh (`rm`) where available.

2.4 Availability and License

VotingBooklets-Diamond and VotingBooklets are available on Hugging Face. All code and scripts used to preprocess, perform OCR, and align the documents are provided via our GitHub repository, enabling full reproducibility. The dataset is released under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, redistribution, and modification, provided appropriate credit is given.

3 Baseline Systems

To demonstrate the utility of the dataset and identify the best-performing systems for corpus creation, we benchmarked a range of open and closed AI models on both OCR and cross-lingual alignment.

3.1 OCR Evaluation

We evaluated three OCR approaches against the gold-standard VotingBooklets-Diamond transcriptions using Word Error Rate (WER) and Character Error Rate (CER): direct OCR with `gemini-2.5-flash-lite` (Gemini2.5), Pytesseract alone (v0.3.13), a Python wrapper for Google's Tesseract-OCR engine (Smith, 2007), and a post-correction pipeline in which Pytesseract output was corrected by `gemini-2.5-flash-lite` (Py+Gem2.5). To ensure that evaluation scores reflect only transcription quality and not segmentation mismatches, OCR output was manually aligned with the gold standard before computing WER and CER. Full results are shown in Table 4.

Gemini2.5 dominates OCR Direct OCR with Gemini2.5 achieves the lowest WER and CER in almost every condition, across all three decades and all four languages. The margin over Pytesseract is substantial: for the 2007 booklets, Pytesser-

act reaches WER values above 0.20 for all languages, while Gemini2.5 stays below 0.08. Even for the older scanned documents from 1977 and 1985, Gemini2.5 performs consistently well, suggesting strong robustness to varying scan quality.

Post-correction does not help Despite evidence in prior work that LLM-based post-correction can improve OCR output (Greif et al., 2025), especially in low-resource settings (Hebbalalu, 2026; Kanerva et al., 2025), the gains are limited in our case. While post-correction does improve over raw Pytesseract output, Gemini2.5 still outperforms the post-correction pipeline in nearly every condition. This suggests that for this document type, end-to-end vision-language models are a more effective approach than pipeline-based post-correction.

Romansh is not a weak point Somewhat surprisingly, Romansh does not consistently lag behind the other languages. For 1985 and 2007, Gemini2.5 achieves competitive CER for Romansh (0.0024 and 0.0095 respectively), comparable to German, French, and Italian. This may reflect the lexical similarity of Romansh to other Romance languages rather than any specific model training coverage.

3.2 Alignment Evaluation

We evaluate four alignment approaches using F1, precision, recall, and character error rate against VotingBooklets-Diamond gold-standard alignments. The evaluated systems include embedding-based alignment with `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019) (MiniLM), `Sentence-SwissBERT` (Grosjean and Vamvas, 2024) (SSB), a LLM-based alignment approach using `gemini-2.5-flash-lite` (Gemini), and hybrid methods that combine Sentence-SwissBERT with targeted post-hoc correction using either `gemini-3-pro-preview` (SSB+Gemini3) or `gemini-2.5-flash-lite` (SSB+Gemini2.5).

The Gemini-based system performs alignment by directly matching German anchor segments to target-language fragments using constrained prompting, allowing segments to be merged while enforcing one-to-one coverage of the anchor sequence. In contrast to embedding-based approaches, this method operates purely at the text level without explicit similarity scoring. Implementation details are provided in Appendix A.3.

To isolate alignment quality from OCR noise, all systems are evaluated on gold-standard transcriptions rather than raw OCR output. German serves as the anchor language, with paragraph-level segmentation from the gold standard used as the reference structure. French, Italian, and Romansh are provided as segmented inputs, which must be aligned to the German paragraphs. Full results are reported in Table 5.

Post-hoc correction improves alignment quality The combined Sentence-SwissBERT and Gemini post-hoc correction approaches achieve the best F1 scores in the majority of conditions and yield the lowest CER almost universally. The improvements are particularly pronounced for the low-resource language Romansh: in 2007, SSB+Gemini3 reaches an F1 of 0.986 and a CER of 0.003, compared to 0.903 and 0.072 for MiniLM, highlighting the benefit of combining embedding-based alignment with targeted LLM-based correction.

Comparing the two post-hoc variants, we find that `gemini-2.5-flash-lite` performs almost on par with `gemini-3-pro-preview`, with only small differences in F1 and CER across all languages and years. Given this near-identical performance, we adopt the SSB+Gemini2.5 configuration in our final pipeline, combining Sentence-SwissBERT alignment with post-hoc correction using `gemini-2.5-flash-lite`. This choice provides a substantially more cost-efficient solution without compromising alignment quality.

4 The Full Parallel Corpus

We construct the full VotingBooklets corpus by applying the best-performing systems identified in Sections 3.1 and 3.2, namely Gemini2.5 for OCR and SSB+Gemini2.5 for alignment, to the complete collection of Swiss federal voting booklets provided by the Federal Chancellery. The dataset comprises 144 booklets in German, 144 in French, 144 in Italian, and 97 in Romansh (RM) (91 obtained from the official webpage, with an additional 6 RM booklets acquired from the State Chancellery of Grisons). Other missing Romansh files could not be found in archives so far. The collection covers votes from June 1977 to March 2026 for German, French, and Italian, and from December 1985 to March 2026 for Romansh.

In total, VotingBooklets contains 19,292 pages and 4,465,492 tokens across all languages. Table 3 provides a detailed overview of the number of tokens and pages for each language in the collection.

The corpus is distributed as JSON Lines (JSONL) files, with one file per voting booklet containing aligned segments across all available languages. Each line in a file represents a paragraph-level alignment, with German as the anchor, alongside the corresponding segments in French, Italian, and Romansh Grischun where available. Languages that are missing for a segment are represented as empty strings.

Quality was assessed through manual evaluation on a held-out sample of aligned segments drawn from booklets not included in VotingBooklets-Diamond. Overall, quality is somewhat lower than on VotingBooklets-Diamond, which is expected: unlike the diamond set, the full pipeline operates on uncorrected OCR output and must handle the greater document diversity of the full corpus. Errors in alignment, OCR, or paragraph splitting occur more frequently, though they remain largely isolated to individual rows and do not propagate across entire documents.

5 Use Cases and Future Work

5.1 Use Cases

Benchmarking OCR and alignment systems VotingBooklets-Diamond enables fair comparison of OCR approaches on historical Swiss administrative documents across varying scan quality, typography, and document structure. The verified paragraph-level alignments allow evaluation of cross-lingual alignment methods independently of OCR quality, including both open-source embedding models and commercial APIs.

Low-resource NLP for Romansh VotingBooklets is one of the few large parallel resources available for Romansh Grischun, and can support machine translation, cross-lingual transfer, and language modeling for this language.

Multilingual representation learning The four-language parallel structure of VotingBooklets makes the corpus well-suited for training and evaluating multilingual embeddings, particularly for Swiss national languages in institutional and political domains.

Computational analysis of political discourse Covering several decades of federal referendum material, VotingBooklets enables longitudinal studies of political language, argumentation, and framing across languages and time periods.

5.2 Future Work

Our results reveal two interesting dependencies: between OCR and alignment, and between alignments across language pairs. These dependencies suggest promising directions for future work. An agentic pipeline that iterates between OCR post-correction and alignment, passing information in both directions, could leverage these relationships to improve performance on both tasks simultaneously. This methodology could also be extended to other Swiss multilingual document collections, such as parliamentary proceedings or cantonal administrative documents, to create richer resources for Swiss NLP.

Beyond the pipeline itself, the scope of the corpus remains an open direction for expansion. The current version omits non-textual elements such as images. Future versions could incorporate these alongside structural layout information, enabling layout-aware analysis and opening the corpus to a broader range of research applications in document understanding and multimodal NLP.

6 Related Work

OCR Technology Traditional OCR systems such as Tesseract (Smith, 2007) rely on pattern recognition and handcrafted linguistic heuristics. These systems work well on clean and modern documents but are less robust when faced with complex layouts, degraded scans, or low-resource languages (Ignat et al., 2022; Greif et al., 2025).

Recent deep learning approaches improve recognition accuracy considerably. Transformer-based models such as TrOCR (Li et al., 2023) jointly model visual and textual features and achieve strong performance on both printed and handwritten text. Large multimodal language models push this further. Gemini 2.0 Flash reaches a character error rate of 1.27% on historical German documents, which can be reduced to 0.84% with multimodal post-correction (Greif et al., 2025).

Beyond direct recognition, recent work explores agentic and LLM-based pipelines that refine OCR outputs iteratively. OCR-Agents use structured reasoning loops to improve recognition quality in difficult settings (Wen et al., 2026), and LLM-based post-correction has proven effective for historical and low-resource languages, where contextual information helps resolve uncertain character sequences (Greif et al., 2025; Hebbalalu, 2026; Kanerva et al., 2025).

The Swiss Multilingual Landscape There are relatively few resources of parallel multilingual text that cover all the Swiss national languages. Notable resources are *Swiss Law Translations* (Niklaus et al., 2025), which aligns federal laws on the level of documents, articles and paragraphs, and the trilingual *Allegra* corpus (Scherrer and Cartoni, 2012), which is composed of press releases by the canton of Grisons. Other corpora that include all Swiss national languages, but which are not necessarily parallel, pertain to domains such as web text (Krasselt et al., 2020; Penedo et al., 2025), news (Graën et al., 2023), social text (Dürscheid and Stark, 2011; Ueberwasser and Stark, 2017), and alpine yearbooks (Göhring and Volk, 2011). In addition, recent work has contributed parallel test sets for machine translation evaluation that cover all four Swiss national languages (Deutsch et al., 2025; Vamvas et al., 2025; Andrews et al., 2025).

Alignment Methods Early alignment methods rely on sentence length models (Gale and Church, 1993). Bleualign (Sennrich and Volk, 2010) uses machine translation and BLEU scores for alignment and achieves strong results on noisy OCR corpora. Later work removes the dependency on pre-trained translation models using iterative bootstrapping (Sennrich and Volk, 2011). Vecalign (Thompson and Koehn, 2019) replaces translation signals with sentence embeddings from LASER (Artetxe and Schwenk, 2019) and uses approximate dynamic programming for alignment. CroCoAlign (Molfese et al., 2024) extends this with context-aware sentence embeddings computed at document level using a transformer encoder and improves alignment quality across language pairs.

Sentence representations are a crucial factor in alignment quality. For Swiss languages, SentenceSwissBERT is introduced by Grosjean and Vamvas (2024). It is based on SwissBERT (Vamvas et al., 2023) and fine-tuned using contrastive learning on Swiss news data in German, French, Italian, and Romansh. It improves performance over multilingual Sentence-BERT (Reimers and Gurevych, 2019), especially for Romansh. This supports the use of domain-specific sentence embeddings in cross-lingual alignment.

7 Conclusion

We presented VotingBooklets-Diamond, a carefully curated gold-standard test set of Swiss federal voting booklets spanning three decades and all four Swiss national languages, and used it to benchmark open and closed AI systems on OCR and cross-lingual alignment. Gemini 2.5 Flash Lite emerged as the strongest OCR system across all conditions, Sentence-SwissBERT (SSB) alignment with Gemini-based post-hoc correction achieved the best alignment performance, particularly for Romansh Grischun.

Applying these systems to the full collection of Swiss federal voting booklets, we release VotingBooklets, a large-scale four-language parallel corpus as a new resource for the Swiss NLP community. We hope both the benchmark and the corpus will support future work on multilingual document processing, low-resource NLP, and the computational study of Swiss political discourse.

Acknowledgments

We thank Sophia Conrad and Giuanna Caviezel for their helpful advice. We also thank the State Chancellery of Grisons, the Federal Chancellery (BK), the Federal Department of Home Affairs, and the Swiss National Library (NB) for their assistance in searching for and locating missing voting booklets.

References

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre 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. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Christa Dürscheid and Elisabeth Stark. 2011. sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland. In Crispin Thurlow and Kristine Mroczek, editors, *Digital Discourse. Language in the New Media*, Oxford Studies in Sociolinguistics, page 299–320. Oxford University Press.

William A. Gale and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102.

Anne Göhring and Martin Volk. 2011. Le corpus Text+Berg Une ressource parallèle alpin français-allemand (The Text+Berg Corpus An Alpine French-German Parallel Resource). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 97–102, Montpellier, France. ATALA.

Johannes Graën, Igor Mustac, Nikolina Rajovic, Jonathan Schaber, Gerold Schneider, and Noah Bubenhofer. 2023. Swissdox@LiRI. A large database of media articles made accessible to researchers. In Krister Linden, Jyrki Niemi, and ThalassiaEditors Kontino, editors, *CLARIN Annual Conference Proceedings*, CLARIN Annual Conference Proceedings, page 111–115. CLARIN ERIC.

Gavin Greif, Niclas Griesshaber, and Robin Greif. 2025. Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents. *Preprint*, arXiv:2504.00414.

Juri Grosjean and Jannis Vamvas. 2024. Fine-tuning the SwissBERT Encoder Model for Embedding Sentences and Documents. In *Proceedings of the 9th edition of the Swiss Text Analytics Conference*, pages 41–49, Chur, Switzerland. Association for Computational Linguistics.

Vishwambhara Hebbalalu. 2026. Dual-Stage OCR Correction for Classical Languages using LLMs: A Comparative Evaluation. Preprint, ResearchGate.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022. OCR Improves Machine Translation for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

Jenna Kanerva, Cassandra Ledins, Siiri Käpyaho, and Filip Ginter. 2025. OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 38–47, Tallinn, Estonia. University of Tartu Library, Estonia.

Julia Krasselt, Philipp Dressen, Matthias Fluor, Cerstin Mahlow, Klaus Rothenhäusler, and Maren Runte. 2020. Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4145–4151, Marseille, France. European Language Resources Association.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: transformer-based optical character recognition with pre-trained models. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Francesco Maria Molfese, Andrei Stefan Bejgu, Simone Tedeschi, Simone Conia, and Roberto Navigli. 2024. CroCoAlign: A Cross-Lingual, Context-Aware and Fully-Neural Sentence Alignment System for Long Texts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2209–2220, St. Julian's, Malta. Association for Computational Linguistics.

Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill A. H. Chevalley, Claude Humbel, Christophe Gösken, Lorenzo Tanzi, Thomas Lüthi, Stefan Palombo, Spencer Poff, Boling Yang, Nan Wu, Matthew Guillod, Robin Mamié, Daniel Brunner, Julio Pereyra, and Niko Grupen. 2025. SwiLTra-Bench: The Swiss Legal Translation Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14894–14916, Vienna, Austria. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All — Adapting Pre-Training Data Processing to Every Language. In *Second Conference on Language Modeling*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Yves Scherrer and Bruno Cartoni. 2012. The Trilingual ALLEGRA Corpus: Presentation and Possible Use for Lexicon Induction. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2890–2896, Istanbul, Turkey. European Language Resources Association (ELRA).

Rico Sennrich and Martin Volk. 2010. MT-based Sentence Alignment for OCR-generated Parallel Texts.

In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based Sentence Alignment of Parallel Texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Raymond W. Smith. 2007. An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2:629–633.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved Sentence Alignment in Linear Time and Space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Simone Ueberwasser and Elisabeth Stark. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 84(5):online.

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. SwissBERT: The Multilingual Language Model for Switzerland. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 54–69, Neuchatel, Switzerland. Association for Computational Linguistics.

Jannis Vamvas, Ignacio Pérez Prat, Not Soliva, Sandra Baltermia-Guetg, Andrina Beeli, Simona Beeli, Madlaina Capeder, Laura Decurtins, Gian Peder Gregori, Flavia Hobi, Gabriela Holderegger, Arina Lazzarini, Viviana Lazzarini, Walter Rosselli, Bettina Vital, Anna Rutkiewicz, and Rico Sennrich. 2025. Expanding the WMT24++ Benchmark with Rumantsch Grischun, Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1028–1047, Suzhou, China. Association for Computational Linguistics.

Shimin Wen, Zeyu Zhang, Xingdou Bian, Hongjie Zhu, Lulu He, Layi Shama, Daji Ergu, and Ying Cai. 2026. OCR-Agent: Agentic OCR with Capability and Memory Reflection. *Preprint*, arXiv:2602.21053.

A Corpus Creation and Technical Details

A.1 Data Collection

The raw Swiss federal voting booklets were downloaded from the official Swiss Federal Chancellery website.[4] A Python scraper was used to traverse language-specific navigation menus and extract all available PDF files. Filenames were sanitized to remove illegal characters and spaces were replaced with underscores.

A.2 OCR Processing

All scanned PDFs were converted to images using `pdf2image`, and each page was processed individually. OCR extraction was performed using Gemini 2.5 Flash Lite via a LangChain ChatOpenAI interface.

For native-born PDFs, text was also extracted using Gemini; the quality of these outputs was sufficiently high that little post-processing was required. In both cases, extraction followed the same segmentation and formatting rules to ensure consistency across booklets for alignment.

Critical extraction rules included:

- Rejoining hyphenated words split across lines.
- Joining non-hyphenated line breaks within sentences with a single space.
- Producing one line per sentence or list item.
- Omitting repeated headers, footers, page numbers, and other boilerplate.

The output of each OCR pass was stored as a text file, named according to the PDF source.

Scope and Constraints The current version of the corpus focuses exclusively on textual content. Visual elements, such as campaign photography and statistical charts, were systematically ignored during the extraction process. We did not perform any image transcription or alternative text generation; therefore, the dataset does not contain metadata or transcriptions for non-textual components.

A.3 Cross-Lingual Alignment

We implemented a two-stage alignment pipeline to align parallel text segments across languages.

Stage 1: Embedding-based alignment with Sentence-SwissBERT Each segment in the German anchor text and the target language is encoded into a vector representation using Sentence-SwissBERT (Grosjean and Vamvas, 2024) with the appropriate language adapter (`de_CH`, `fr_CH`, `it_CH`, `rm_CH`).

Alignment is performed using dynamic programming (DP) over the two segment sequences. The

[4] https://www.bk.admin.ch/bk/de/home/dokumentation/abstimmungsbuechlein.html

Year	Lang	Method	WER	CER	Ins	Del	Sub
1977	DE	Gemini2.5	**0.0462**	0.0262	40	63	76
		Pytesseract	0.1074	0.0217	185	15	216
		Py+Gem2.5	0.0860	**0.0136**	160	3	170
1977	FR	Gemini2.5	**0.0454**	**0.0142**	45	67	110
		Pytesseract	0.1332	0.0233	167	49	436
		Py+Gem2.5	0.1287	0.0199	158	48	424
1977	IT	Gemini2.5	**0.0321**	**0.0143**	27	43	71
		Pytesseract	0.1058	0.0224	97	24	344
		Py+Gem2.5	0.0930	0.0184	59	17	333
1985	DE	Gemini2.5	**0.0122**	**0.0015**	8	1	10
		Pytesseract	0.1312	0.0584	69	75	61
		Py+Gem2.5	0.0621	0.0462	12	73	12
1985	FR	Gemini2.5	**0.0181**	**0.0017**	1	18	20
		Pytesseract	0.1107	0.0534	65	101	72
		Py+Gem2.5	0.0689	0.0456	15	107	26
1985	IT	Gemini2.5	**0.0163**	**0.0052**	3	10	16
		Pytesseract	0.1536	0.1021	103	115	55
		Py+Gem2.5	0.1294	0.1075	40	168	22
1985	RM	Gemini2.5	**0.0171**	**0.0024**	3	9	22
		Pytesseract	0.1134	0.0661	58	96	71
		Py+Gem2.5	0.0977	0.0668	37	102	55
2007	DE	Gemini2.5	**0.0348**	**0.0192**	14	25	27
		Pytesseract	0.2294	0.1518	135	173	127
		Py+Gem2.5	0.1187	0.1004	17	167	41
2007	FR	Gemini2.5	**0.0720**	**0.0305**	31	40	109
		Pytesseract	0.2277	0.1476	149	234	186
		Py+Gem2.5	0.1509	0.1033	27	223	127
2007	IT	Gemini2.5	**0.0391**	**0.0081**	16	7	68
		Pytesseract	0.2094	0.1434	118	228	141
		Py+Gem2.5	0.1410	0.1058	29	203	96
2007	RM	Gemini2.5	**0.0568**	**0.0095**	15	12	125
		Pytesseract	0.2202	0.1489	132	261	196
		Py+Gem2.5	0.1787	0.1139	37	252	189

Table 4: Evaluation of three OCR methods on VotingBooklets-Diamond, reporting Word Error Rate (WER), Character Error Rate (CER), and raw error counts (insertions, deletions, and substitutions) across all voting booklets and languages. Gemini2.5 = direct OCR using Gemini-2.5-Flash-Lite. Pytesseract = baseline OCR without post-correction. Py+Gem2.5 = Pytesseract OCR with Gemini-2.5-Flash-Lite post-correction. Languages: DE=German, FR=French, IT=Italian, RM=Romansh. Bold values indicate the best result per year, language, and metric. Lower WER/CER values indicate better performance.

Year	Lang	Method	F1	Prec	Rec	CER
1977	FR	MiniLM	0.9187	0.9187	0.9187	0.0891
		SSB	0.9210	0.9234	0.9187	0.0763
		Gemini	**0.9688**	**0.9712**	**0.9665**	**0.0147**
		SSB+Gemini2.5	0.9210	0.9234	0.9187	0.0760
		SSB+Gemini3	0.9569	0.9569	0.9569	0.0192
	IT	MiniLM	0.8995	0.8995	0.8995	0.1681
		SSB	0.9187	0.9187	0.9187	0.0677
		Gemini	0.9474	0.9474	0.9474	0.0337
		SSB+Gemini2.5	0.9713	0.9713	0.9713	0.0171
		SSB+Gemini3	**0.9856**	**0.9856**	**0.9856**	**0.0151**
1985	FR	MiniLM	**1.0000**	**1.0000**	**1.0000**	**0.0008**
		SSB	0.9533	0.9533	0.9533	0.0211
		Gemini	0.9813	0.9813	0.9813	0.0105
		SSB+Gemini2.5	0.9720	0.9720	0.9720	0.0158
		SSB+Gemini3	0.9626	0.9626	0.9626	0.0191
	IT	MiniLM	**1.0000**	**1.0000**	**1.0000**	**0.0000**
		SSB	**1.0000**	**1.0000**	**1.0000**	**0.0000**
		Gemini	0.9671	0.9717	0.9626	0.0133
		SSB+Gemini2.5	**1.0000**	**1.0000**	**1.0000**	**0.0000**
		SSB+Gemini3	0.9907	0.9907	0.9907	0.0064
	RM	MiniLM	0.9626	0.9626	0.9626	0.0911
		SSB	0.9813	0.9813	0.9813	0.0795
		Gemini	0.9813	0.9813	0.9813	0.0130
		SSB+Gemini2.5	**1.0000**	**1.0000**	**1.0000**	**0.0000**
		SSB+Gemini3	**1.0000**	**1.0000**	**1.0000**	0.0001
2007	FR	MiniLM	0.9694	0.9667	0.9721	0.0257
		SSB	0.9805	0.9778	0.9832	0.0102
		Gemini	0.9776	0.9775	0.9777	0.0086
		SSB+Gemini2.5	**0.9916**	**0.9889**	**0.9944**	0.0004
		SSB+Gemini3	**0.9916**	**0.9889**	**0.9944**	**0.0003**
	IT	MiniLM	0.9749	0.9722	0.9777	0.0103
		SSB	0.9861	0.9833	0.9888	0.0037
		Gemini	0.9636	0.9663	0.9609	0.0122
		SSB+Gemini2.5	0.9805	0.9778	0.9832	0.0075
		SSB+Gemini3	**0.9916**	**0.9889**	**0.9944**	**0.0004**
	RM	MiniLM	0.9025	0.9000	0.9050	0.0716
		SSB	0.9526	0.9500	0.9553	0.0270
		Gemini	0.9494	0.9548	0.9441	0.0235
		SSB+Gemini2.5	0.9582	0.9556	0.9609	0.0250
		SSB+Gemini3	**0.9861**	**0.9833**	**0.9888**	**0.0027**

Table 5: Alignment results for five methods on VotingBooklets-Diamond across all voting booklets and languages. F1, Precision (Prec), and Recall (Rec) are computed via fuzzy matching against the gold standard. CER measures character error rate of the aligned text against gold (lower is better). Bold values indicate the best result per year, language, and metric. Languages: FR=French, IT=Italian, RM=Romansh. MiniLM: paraphrase-multilingual-MiniLM-L12-v2. SSB: Sentence-SwissBERT. Gemini: Alignment using Gemini-2.5-Flash-Lite. SSB+Gemini2.5: Sentence-SwissBERT with post-hoc correction using Gemini-2.5-Flash-Lite. SSB+Gemini3: Sentence-SwissBERT with post-hoc correction using Gemini-3-pro-preview.

algorithm searches for the globally optimal alignment by assigning scores to different operations: a correct match between a German segment and a target segment earns a score proportional to their cosine similarity; skipping an unmatched segment incurs a penalty of -0.3; and merging up to five consecutive segments on either side into a single unit is allowed but penalised by -0.05 per additional segment merged. When multiple segments are merged, their embeddings are averaged before computing similarity, and merges are only considered if the resulting similarity exceeds 0.30. The algorithm thus supports one-to-one, one-to-many, and many-to-one alignments. The best alignment path is recovered by backtracking through the DP grid.

Stage 2: Post-hoc correction with Gemini Aligned pairs with a cosine similarity below 0.65 are flagged as uncertain and sent to Gemini 2.5 Flash Lite for review in batches of five. The model receives the German segment, the proposed target segment, and up to 150 candidate fragments from the target language. It is instructed to either confirm the alignment or replace the target segment using only verbatim text from the available fragments - no translation or paraphrasing is permitted. The prompt used is shown in Appendix A.4. A correction is applied only if Gemini returns a non-empty target segment and leaves the German text unchanged.

A.4 Post-hoc Correction Prompt

The following prompt was used for post-hoc alignment correction with Gemini. Placeholders in curly braces (`{lang_name}`, `{lang_key}`, etc.) are filled dynamically at runtime.

```
You are a multilingual text alignment expert for
Swiss official documents.

You will receive a list of aligned segment pairs
(German DE and {lang_name} {lang_key}).
Some alignments may be incorrect — the {lang_name}
text may be misaligned, incomplete, or merged
incorrectly.

You also receive the full list of original {lang_name}
fragments the aligner had available.

For each pair:
- If the alignment looks correct, keep it as-is.
- If the {lang_name} text is clearly wrong or
misaligned, find the correct fragment(s) from the
available fragments and replace it.
- If no good match exists, return an empty string
for that pair.
- Do NOT translate or paraphrase — only use text
from the original fragments verbatim.

You must return VALID JSON ONLY.

Rules:
- No explanations
- No markdown
- No comments
- No trailing commas
- Escape all quotes
- Output must parse with json.loads()

Return exactly:

[
  {{
    "de": "...",
    "fr": "...",
    "it": "...",
    "rm": "..."
  }}
]

--- ALIGNED PAIRS TO REVIEW ---
{pairs}

--- AVAILABLE {lang_key_upper} FRAGMENTS ---
{fragments}
```

A.5 Output Format

The final aligned corpus is provided as JSON Lines files (`.jsonl`), one per vote, with each line containing a paragraph-level alignment across all available languages. All scripts and code used to create this corpus, including OCR extraction and multilingual alignment, are publicly available on our GitHub repository.[5]

[5] https://github.com/Eljuanina/VotingBooklets