20min-XD: A Comparable Corpus of Swiss News Articles

Michelle Wastl[1] **Jannis Vamvas**[1] **Selena Calleri**[2] **Rico Sennrich**[1]
[1]Department of Computational Linguistics, University of Zurich [2]20 Minuten (TX Group)
{wastl,vamvas,sennrich}@cl.uzh.ch, {selena.calleri}@20minuten.ch

Abstract

We present *20min-XD* (**20 Min**uten **cross**-lingual **d**ocument-level), a French-German, document-level comparable corpus of news articles, sourced from the Swiss online news outlet *20 Minuten*/*20 minutes*. Our dataset comprises around 15,000 article pairs spanning 2015 to 2024, automatically aligned based on semantic similarity. We detail the data collection process and alignment methodology. Furthermore, we provide a qualitative and quantitative analysis of the corpus. The resulting dataset exhibits a broad spectrum of cross-lingual similarity, ranging from near-translations to loosely related articles, making it valuable for various NLP applications and broad linguistically motivated studies. We publicly release the dataset in document- and sentence-aligned versions and code for the described experiments[1,2].

1 Introduction

Cross-lingual datasets play a crucial role in Natural Language Processing (NLP), supporting a range of tasks such as bitext mining, machine translation, and cross-lingual information retrieval. Among these, comparable corpora—datasets containing text pairs with related but non-identical content across languages—are particularly valuable. Unlike parallel corpora, which consist of direct translations, comparable corpora naturally contain a mix of exact translations, paraphrases, and loosely related content, reflecting the linguistic and cultural variations between languages. This makes them a rich resource for training and evaluating multilingual NLP models (Lewis et al., 2020; Liu et al., 2020; Philippy et al., 2025).

However, existing document-level, cross-lingual corpora remain limited in scope. Many available resources are English-centric, primarily covering English alongside another high-resource language and/or are restricted to sentence-level alignments rather than full documents (Zweigenbaum et al., 2017; Artetxe and Schwenk, 2019). At the same time, large language models (LLMs) and modernized encoder architectures are advancing in their ability to process longer texts and numerous languages, further increasing the demand for multi-/cross-lingual, document-level corpora (Hengle et al., 2024; Wang et al., 2024; Zhang et al., 2024).

Beyond their NLP applications, cross-lingual document-level datasets also facilitate more linguistically motivated studies such as cross-cultural discourse analyses (Carbaugh and Cerulli, 2017) or comparative journalism research (Hanitzsch, 2019). More specifically, a German-French news article corpus could be used to examine how news narratives and framing strategies vary between the Germanophone and Francophone regions.

Given these potential interdisciplinary use cases, we collect comparable news articles in German and French from the online Swiss news outlet *20 Minuten*/*20 minutes*. As both editions are produced by the same publisher, with an internal article transfer workflow from one language to the other, they share a high degree of topical overlap, making them well-suited for comparable corpus creation. Our dataset comprises 15,000 article pairs, spanning nearly a decade (2015–2024). Each article pair consists of a German and a French news article published on the same day, covering the same or a highly related event. In addition to the document-level alignments, we release a sentence-aligned version of the dataset, which contains 117,126 sentences per language.

We release the dataset to the research community for non-commercial, scientific purposes[3].

[1]Dataset: https://huggingface.co/datasets/ZurichNLP/20min-XD

[2]Code: https://github.com/ZurichNLP/20min-XD

[3]See Appendix A for the detailed Copyright notice.

	Validation Set		Full Dataset		Top 15k	
Statistics	**German**	**French**	**German**	**French**	**German**	**French**
Total # of aligned articles	14	14	73,085	73,085	15,000	15,000
Total # of sentences	401	358	1,888,323	1,608,497	357,071	327,628
Total # of tokens	9,087	9,690	43,559,153	43,256,366	8,378,874	8,956,116
Total # of characters	38,523	38,519	189,598,932	174,789,207	36,924,383	36,387,070
Avg. title length in characters	59	54	51	53	51	54
Avg. title length in tokens	18	18	15	17	15	17
Avg. lead length in characters	146	155	152	146	152	150
Avg. lead length in tokens	39	43	39	40	38	41
Avg. content length in characters	2,547	2,542	2,391	2,192	2,258	2,222
Avg. content length in tokens	706	753	650	649	612	655
Avg. content length in sentences	29	26	26	22	24	22

Table 1: Detailed statistics of the validation, full, and top-15k subsets. The sentence segmentation was performed with spaCy '[de/fr]_core_news_sm' (Honnibal and Montani, 2017) models for sentence segmentation and tokenization with the `paraphrase-multilingual-mpnet` tokenizer.

2 Related Work

Switzerland's multilingual landscape, with four official languages, provides fertile ground for cross-lingual corpus creation. Several prior works have leveraged this linguistic diversity to construct multilingual datasets. For instance, SwissAdmin (Scherrer et al., 2014) is a sentence-aligned corpus of official Swiss government press releases available in German, French, Italian, and English. Similarly, the Bulletin Corpus (Volk et al., 2016) aligns issues of the *Credit Suisse Bulletin* across the same four languages.

20 Minuten has also served as a resource for previous NLP-related studies. Rios et al. (2021) constructed a dataset for automatic text simplification by pairing original German *20 Minuten* articles with their simplified counterparts. More recently, Kew et al. (2023) created a dataset aimed at automatic news summarization in German, further expanding the utility of Swiss news data in NLP research.

With this work, we aim to bridge these two subjects by introducing *20min-XD*, a French-German document-level comparable corpus, sourced from *20 Minuten* (German) and *20 minutes* (French).

3 Data Acquisition

To construct our dataset, we first scrape a total of 593,897 online news articles from both `www.20min.ch/` and `www.20min.ch/fr/`, covering the period from 01.01.2015 to 01.12.2024. In the following subsections, we describe the process applied to identify and align the semantically related articles.

3.1 Validation Set

To establish a gold standard for alignment evaluation, we selected all articles from a single publication day, resulting in 87 German and 70 French articles. Each French article was manually compared against the German articles to identify comparable pairs. While we did not strictly prohibit n:n pairings, the resulting validation set only contains 1:1 pairings. Through this process, we aligned 28 articles into 14 pairs, forming our validation set. Detailed statistics can be found in Table 1.

3.2 Automatic Article Alignment

Since manually aligning comparable articles across languages is time-intensive and requires proficiency in both German and French, we automate the process leveraging multilingual embedding models. Specifically, we encode portions of each article as numerical vectors and compute cosine similarity scores, which range from -1 to 1 ($*100$), to quantify their semantic similarity.

In order to find the most appropriate alignment methods for the *20 Minuten* articles, we conduct experiments on our validation set with different embedding models, alignment approaches, and similarity thresholds.

We choose not to embed the full article texts to ensure a fair comparison across the tested models, some of which have a sequence length constraint (3 out of the tested 5). The results on our validation set suggest that concatenating the article's title and lead provides a sufficiently strong signal for document alignment. This enables resource-efficient experimentation with encoder-based embedding models while avoiding length limitations.

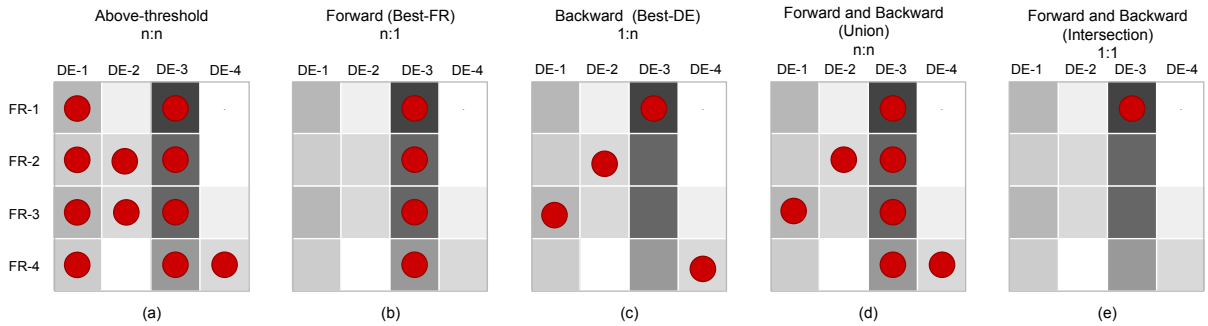

Figure 1: Matrix visualization of different alignment strategies.

Model	Above-threshold	Intersection	Union	Best-DE	Best-FR	Avg.
`paraphrase-multilingual-mpnet`	54.1	**64.7**	54.1	57.8	55.8	57.3
`gte-multilingual-base`	55.6	62.1	55.6	60.0	58.5	58.3
`LaBSE`	53.3	48.5	56.5	60.0	46.2	52.9
`sentence-swissBERT`	62.9	62.5	62.9	61.1	62.5	62.4
`gte-modernbert-base`	45.5	53.3	50.0	54.1	50.0	50.6

Table 2: F1 performance comparison of different models and different alignment approaches on the validation set. The corresponding thresholds are presented in Appendix B.

3.2.1 Models

We experiment with the set of models presented in Table 2: `paraphrase-multilingual-mpnet` is a state-of-the-art multilingual sentence-level paraphrase recognition model (Song et al., 2020); `gte-multilingual-base`, a long-context multilingual text representation model (Zhang et al., 2024); `sentence-swissBERT`, a sentenceBERT-based (Reimers and Gurevych, 2019) model trained on in-domain (*20 Minuten*) data (Grosjean and Vamvas, 2024); `gte-modernbert-base`, modernized, more efficient, long-context version of BERT that has been trained on predominantly English data (Warner et al., 2024).

Preliminary experiments with an LLM-based model (Wang et al., 2024) have shown that they outperform encoder-based models while also being able to process longer input sequences. They do, however, also increase the computational complexity of the embedding process, making it rather resource-intensive and barely feasible in terms of memory and time if scaled to a larger number of documents.

3.2.2 Alignment Strategies

Previous work in cross-lingual alignment has considered multiple possible alignment strategies that either expand or restrict the resulting number of alignments according to different categories as described in e.g. Jalili Sabet et al. (2020) for cross-lingual word alignment. Similarly to Hämmerl et al. (2024), we experiment with strategies that result in a range from weak to strong alignment, where strategies for weaker alignments typically allow a higher range of semantic similarity and multiple possible alignments, while strategies for stronger alignments are more restrictive towards a high semantic similarity and may only include one good alignment (Figure 1).

Above-Threshold considers all document pairs with a similarity score above a certain threshold as alignable, allowing for many-to-many (n:n) alignments. This means that any number of French articles can be linked to any number of German articles without additional constraints beyond the similarity threshold. While this approach captures a broad range of potential alignments, it does not enforce uniqueness or best-match constraints, leading to a higher number of alignments (Figure 1a).

Best-FR applies a many-to-one (n:1, German:French) constraint, where each FR article is aligned to the single DE article with which it has the highest cosine similarity, provided that the similarity exceeds the threshold. This ensures that each FR document has a single best-matching DE counterpart, but multiple French articles can still be mapped to the same German article. This approach prioritizes French articles selecting their closest German equivalent while allowing asymmetry in alignments (Figure 1b).

Best-DE follows the same principle as Best-FR

but from the German perspective, enforcing a one-to-many (1:n, German:French) constraint. This results in a setting where a single German article may be linked to multiple French articles, capturing scenarios where a single German document is the best translation candidate for multiple French counterparts (Figure 1c).

Union takes the union of Best-DE and Best-FR alignments, allowing many-to-many (n:n) alignments, but in a more restrictive manner than the Above-Threshold approach. Instead of considering all pairs above the threshold, it only retains document pairs where at least one side selects the other as its most similar document above the threshold (Figure 1d).

Intersection is the most restrictive strategy, enforcing a one-to-one (1:1) constraint. A valid alignment occurs only when the French article is the best match for the German article and vice versa provided their similarity score exceeds the threshold. This method forms the intersection of Best-DE and Best-FR, ensuring that alignments are bidirectional and mutually optimal (Figure 1e).

3.3 Setting a threshold

Since not every article has a comparable counterpart in the other language, we define a similarity score threshold above which two articles are considered alignable. This threshold must be exceeded in each of the alignment strategies described above. To determine the optimal threshold θ, we iterate through the range of 0 and 100 in steps of 0.5, selecting the one that maximizes the F1 score on our validation set:

$$\hat{\theta} = \arg\max_{\theta \in \{0, 0.5, \ldots, 100\}} F_1(\theta)$$

And we define F1 as follows, where P denotes predicted pairs and G gold pairs:

$$Prec = \frac{|P \cap G|}{|P|}$$

$$Recall = \frac{|P \cap G|}{|G|}$$

$$F_1 = 2 \cdot \frac{Prec \cdot Recall}{Prec + Recall}$$

This process is repeated for each of the embedding models described above. Our results show that `paraphrase-multilingual-mpnet` with the alignment strategy *intersection* at a similarity score threshold of 46, outperforms all other models on the validation set (see Table 2), making it our approach for article alignment.

It is worth noting that the number of samples in our validation set is small (87 German and 70 French articles). This could lead to statistical noise, exaggerating the apparent differences in the results, making them seem larger/smaller than they truly are.

3.4 Choosing A Time Window

To ensure precise alignment and reduce computational complexity, we restrict comparisons to articles published on the same date. This approach minimizes spurious matches between articles that discuss similar topics but are unrelated in terms of specific events or developments.

3.5 Post-Processing

After aligning the French and German articles, we clean the resulting corpus. Manual inspection indicates that faulty articles usually have a suspiciously high similarity score and contain an error message or the same text in the same language. We remove such pairs.

3.6 Sentence Alignment

To provide more fine-grained insights into the dataset, we conduct sentence-level analyses. To achieve this, we first segment articles into sentences using the spaCy '[de/fr]_core_news_sm' (Honnibal and Montani, 2017) models for German and French.

Once segmented, we perform cross-lingual sentence alignment, once again, applying the best performing approach described above: `paraphrase-multilingual-mpnet` with the *intersection* alignment strategy. While we consider only sentence pairs with a similarity score above 46 for our analyses, we release the sentence-aligned version of our corpus on all aligned sentences, including those whose similarity score does not exceed the threshold. This allows for more holistic future analyses, capturing not only the most strongly aligned sentences but also those with the weakest still detectable semantic similarity.

We post-process the sentence-level version of the dataset by removing sentence pairs that contain less than 30 characters, which entails names, trailing characters and source abbreviations.

Similarity Scores	German	French
Cosine: 98.48 (**max**) SentLengthCorr: 0.75 AlignRatio DE: 0.68 AlignRatio FR: 0.56 Monotonicity: 1.0	**Title:** Mobilität.: «Ab 2030 bieten wir nur noch vollelektrische Fahrzeuge an» **Lead:** Die Elektro-Revolution rollt. Traditionelle Autohersteller haben derzeit einen schweren Stand. Wir haben bei Helen Hu, Geschäftsführerin des Schweizer Ablegers von Volvo, seit 2010 in chinesischer Hand, nachgefragt, wie sie die Zukunft der Mobilität sieht.	**Title:** Mobilité: «A partir de 2030, nous ne proposerons plus que des véhicules entièrement électriques» **Lead:** La révolution électrique est en marche. Les constructeurs automobiles traditionnels ont actuellement la vie dure. Nous avons demandé à Helen Hu, directrice de la filiale suisse de Volvo, en mains chinoises depuis 2010, comment elle voit l'avenir de la mobilité.
Cosine: 84.05 (**mean among top-15k**) SentLengthCorr: -0.78 AlignRatio DE: 0.23 AlignRatio FR: 0.21 Monotonicity: -1.0	**Title:** LKW kreuzte Lieferwagen und stürzte dann ab **Lead:** Ein Lastwagen stürzte am Dienstag 300 Meter in die Tiefe. Der 66-jährige Fahrer wurde schwer verletzt. Jetzt gibt es erste Erkenntnisse, wie es zum Unfall kam.	**Title:** Un camion chute de 300 mètres, le chauffeur survit **Lead:** Un chauffeur de poids lourd a été grièvement blessé, mardi, après que son véhicule est sorti de la route, dans le canton d'Uri.
Cosine: 78.65 (**min among top-15k**) SentLengthCorr: -0.47 AlignRatio DE: 0.07 AlignRatio FR: 0.2 Monotonicity: -0.3	**Title:** GP Brasilien - Bottas gewinnt das Sprintrennen – Hamilton nach irrer Aufholjagd auf Rang 5 **Lead:** Am Samstag stand beim GP von Brasilien die Sprint-Entscheidung an. Die 3 WM-Punkte und die Pole-Position für das Rennen am Sonntag sicherte sich Valtteri Bottas.	**Title:** Automobile – Bottas prive Verstappen de la victoire au sprint et de la pole **Lead:** Valtteri Bottas s'est offert la course sprint et partira de la première case dimanche au Grand Prix du Brésil. Max Verstappen sera placé derrière lui et Lewis Hamilton 10e.
Cosine: 46.00 (**min among full dataset**)	**Title:** Sein Zwillingsbruder brachte ihn vor Gericht **Lead:** Hochriskante Börsengeschäfte ihres Verwaltungsratspräsidenten haben eine renommierte Churer Treuhandfirma in den Ruin getrieben. Der Beschuldigte musste vor Gericht erscheinen.	**Title:** Plombé par Kairos, Julius Bär doit se rattraper **Lead:** La filiale italienne de Julius Bär apparaît presque comme la source de tous les maux du gestionnaire de fortune zurichois.

Table 3: Comparison of the title and lead text of the aligned articles receiving the lowest, mean and highest cosine similarity scores from the top 15,000 aligned articles as well as the aligned articles with the lowest overall score from the full set of aligned articles, which is filtered from the final dataset.

3.7 Additional Measures of Similarity

In the corpus description in Section 4 we make use of additional cross-lingual similarity measures apart from the cosine distance that are based on the sentence alignments:

Alignable sentences per document To estimate how much text within an article is highly similar, we compute the relative percentage of alignable sentences. This measure is particularly interesting, as the full document is not considered during automatic article alignment, as described in Subsection 3.2. For each article, we define the alignable sentence ratio as:

$$\text{AlignRatio} = \frac{\text{NumAlignedSentences}}{\text{TotalSentences}}$$

Sentence length correlation If the sentence length, measured as the number of characters in the sentence, differs between the two languages in a systematic way, a high correlation between sentence lengths in aligned articles could be an additional indicator of semantic similarity. Hence, we compute the sentence length correlation of an article as a Pearson correlation.

Monotonicity We measure the cross-lingual monotonicity (degree by which aligned sentences appear in the same order) between an aligned article pair by calculating the Kendall rank correlation of the aligned sentences' position.

4 Dataset

Our alignment process results in 74,507 article pairs. During post-processing the corpus is filtered down to 73,085 article pairs. By agreement with *20 Minuten*, our dataset release is limited to 30,000 articles. Consequently, we select the top 15,000 article pairs sorted by their similarity score for publication, which we refer to as top 15k dataset in the following. Nonetheless, in the remainder of this paper, we will consider both the full dataset and the top 15k article pairs as subject of analysis. The detailed dataset statistics for both are presented in Table 1.

Out of the total 300,000+ sentences in each language from the top 15k dataset, we align 133,693 sentences per language, from which 117,126 are left after filtering. For the correlation studies in Section 4.2, we consider all the sentence pairs with similarity score above 46, totaling to 109,871 sentence pairs.

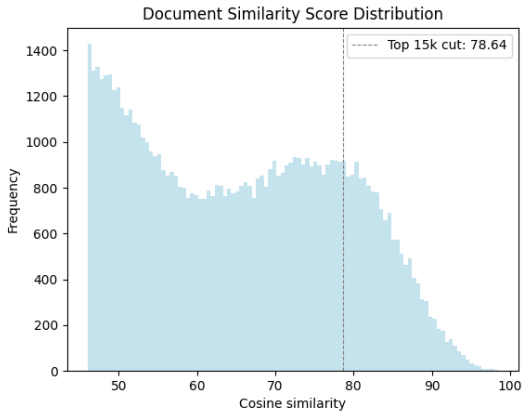

Figure 2: Document (cosine) similarity score distribution over all 74,085 article pairs divided into 100 bins ranging from the threshold of 46 to 100. The dashed line indicates the cut above which the top 15,000 article pairs form the final comparable dataset.

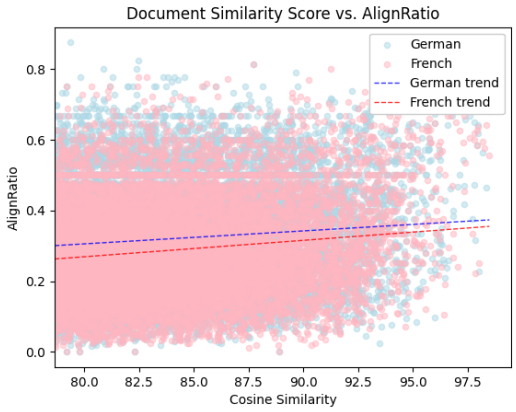

Figure 3: The document cosine similarities in comparison to the AlignRatio of each aligned article in German and French. Both languages show a positive trend line with weak positive correlation (FR: Pearson correlation coefficient $r = 0.145$; DE: $r = 0.103$).

4.1 Qualitative Analysis

Table 3 provides a qualitative comparison of article pairs with the lowest, mean, and highest cosine similarity scores in the top 15k dataset as well as the article pair with the lowest similarity score of all 75,085 initially aligned articles. The highest-scoring pairs exhibit strong lexical and syntactic similarities. The mean-scoring pairs effectively convey the same meaning but demonstrate noticeable differences of the order in which the information is presented. Only the last sentence in the German lead as well as the last phrase in the French lead introduce different information. The lowest-scoring pair in the top 15k dataset covers the same event but differs strongly in word choice and the order in which the information is conveyed. The lowest-scoring pair of the full set of aligned articles, while still loosely related (financial crises), differs in the actual event that is described (e.g., court case leading to a company's collapse vs. corporate struggle with subsidiary).

These results suggest that our dataset mostly consists of articles covering the same topic but with varying degrees of semantic overlap, text structure and length. In order to gain further insight into these features and their relationship to semantic similarity, we conduct a correlation study between the cosine scores of the aligned articles and the different measures described in Section 3.7.

4.2 Quantitative Analysis

4.2.1 Cosine Similarity Distribution

Figure 2 presents the distribution of cosine similarity scores among the aligned articles. The distribution exhibits a right-skewed pattern, suggesting that among the collection of scraped articles, French and German articles with moderate semantic relatedness are more prevalent than those with extremely high similarity scores. The number of articles first drops and then rises again with a rising cosine similarity before reaching a small peak at around 80, located almost exactly at our top 15,000 cutoff point. Following the cutoff, the frequency of article pairs declines sharply to a relatively low level towards higher similarity scores. This pattern loosely suggests the presence of two clusters of article pairs: one representing moderately related articles and another, less prominent, group of more closely related articles.

4.2.2 Correlation with AlignRatio

As a further measure of semantic similarity, we employ the alignment ratio (AlignRatio), which measures the proportionality of aligned sentences between the articles in the two languages, and examine how document similarity scores correlate. As shown in Figure 3, both German and French exhibit weak positive correlations between cosine similarity scores and AlignRatio ($r = 0.145$ for French, $r = 0.103$ for German). These findings suggest that articles with more alignments in the full text tend to have slightly higher semantic similarity. This supports our assumption that relying

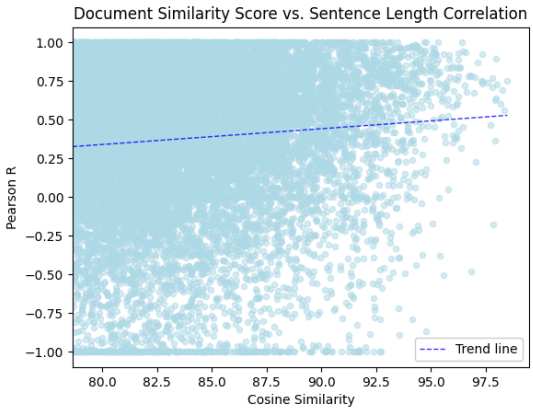

Figure 4: The document cosine similarities in comparison to the sentence length correlation of each aligned article. There is a very weak positive trend of correlation detectable between the two variables (Pearson correlation coefficient $r = 0.084$).

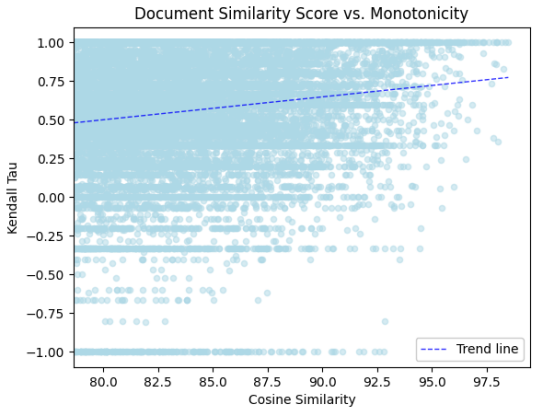

Figure 5: The document cosine similarities in comparison to the monotonicity score of each aligned article. A weak positive correlation trend is detectable between the two variables (Pearson correlation coefficient $r = 0.147$).

solely on the title and lead for the automatic alignment is sufficient but not perfect.

4.2.3 Correlation with Sentence Length

To analyze the relationship between document similarity scores and sentence length variations in aligned articles, we compute the correlation between cosine similarity scores and the sentence length correlation of each article pair. As illustrated in Figure 4, the results indicate a very weak positive correlation ($r = 0.084$).

4.2.4 Correlation with Monotonicity

We also investigate the relationship between document similarity scores and monotonicity, which quantifies the extent to which the order of information (= sentences) is preserved between aligned articles. Figure 5, presents the correlation between cosine similarity scores and monotonicity, showing a weak positive correlation ($r = 0.147$). This suggests, similarly to the previous results, that while higher document similarity scores are slightly associated with a more monotonic alignment of information, the effect is not strong. The clusters near -1.00 and 1.00 may indicate a high number of articles with only one or two aligned sentences — a pattern that could be worth to investigate further.

Given our qualitative analysis and correlation studies, we are confident our dataset maintains an adequate quality for a comparable corpus, covering the full range between direct translations and fairly unrelated text sequences. However, further work with these metrics could provide more insight. Specifically, the alignment ratio may serve as an indicator on which pieces of information are considered essential in both linguistic regions and which are missing from one or the other. Similarly, sentence length correlation could offer valuable perspectives in news-specific translation research. Lastly, monotonicity could be explored further by analyzing topic-specific trends, potentially revealing which topics tend to be translated in a more monotonic fashion than others.

5 Future Work

5.1 Comparing similarity of full text

While using only titles and leads was sufficient for aligning comparable articles, incorporating full article content into the similarity score calculation could provide a more granular and accurate insight into the degree of semantic similarity and relatedness of the articles. This approach could provide a more nuanced representation of narrative structure, argumentation, and topical emphasis. Although computationally intensive, modern embedding models such as `e5-instruct-7b` or `gte-multilingual-base` can theoretically process longer text spans, making full-text comparison increasingly feasible.

5.2 Multilingual long-context embedding models

Encoder-based embedding models are currently going through a renaissance with modernized imple-

mentations, such as ModernBERT (Warner et al., 2024), with significantly improved efficiency and ability to process longer text sequences. At this point in time, multilingual versions of this model specified for the text similarity task are scarce. Future work could explore extending ModernBERT to a multilingual setting and/or optimization for cross-lingual document alignment. Another potential direction is leveraging these modern architectures to develop a document-level counterpart to the (sentence-)swissBERT model.

5.3 Difference recognition

While semantic similarity has been a predominant focus in NLP, the ability to detect and quantify differences between texts—especially across languages—is an emerging research area (Vamvas and Sennrich, 2023). Inspired by diff-based operations in version control, this task could have implications for natural language versioning, collaborative document editing, and editorial workflows. Vamvas and Sennrich show that semantic similarity datasets can be repurposed for difference detection, but have to be synthetically altered to cover cross-linguality and longer text sequences.

Given the variation spectrum observed in our dataset (see Section 4), the diversity of near-translations and loosely related articles, an extension of our corpus with fine-grained annotations—at the paragraph, sentence, or even token level—could enable research into automatic cross-lingual difference recognition.

6 Conclusion

We introduce *20min-XD*, a new French-German document-level comparable dataset of news articles, sourced from the Swiss newspaper *20 Minuten/20 minutes*. The dataset consists of 15,000 aligned articles (or 117,126 aligned sentences) published over a ten-year period. To establish document-level and sentence-level alignment, we employ a multilingual paraphrase recognition model, which demonstrated strong performance during experiments on a manually curated validation set. Both qualitative and quantitative results show that our corpus captures a broad spectrum of cross-lingual similarity, from near-translations to more loosely related text pairs that still cover the same event, with varying degrees of alignable sentences, text lengths and monotonicity. We anticipate its use in future studies across a broad range of linguistically motivated studies.

Acknowledgments

This work was funded by the Swiss National Science Foundation (project InvestigaDiff; no. 10000503 for MW, JV, and RS, and project MUTAMUR; no. 213976 for RS). We sincerely thank everyone at 20 Minuten (TX Group) for their support and for making their data accessible to the research community, with special appreciation to Dean Cavelti for his patient communication. We are also grateful to Unitectra, particularly Peter Loch, for their valuable legal guidance. Finally, we extend our thanks to the Department of Computational Linguistics at the University of Zurich for their inspiring discussions and guidance, with special recognition to Sarah Ebling, Andrianos Michail, Patrick Haller and Anastassia Shaitarova.

References

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.

Donal Carbaugh and Tovar Cerulli. 2017. *Cultural Discourse Analysis*.

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.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Hanitzsch. 2019. Comparative Journalism Research.

Amey Hengle, Prasoon Bajpai, Soham Dan, and Tanmoy Chakraborty. 2024. Multilingual needle in a haystack: Investigating long-context behavior of multilingual large language models. *Preprint*, arXiv:2408.10151.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Tannon Kew, Marek Kostrzewa, and Sarah Ebling. 2023. 20 Minuten: A multi-task news summarisation dataset for German. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 1–13, Neuchatel, Switzerland. Association for Computational Linguistics.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems*, volume 33, pages 18470–18481. Curran Associates, Inc.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

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*. Association for Computational Linguistics.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

Yves Scherrer, Luka Nerima, Lorenza Russo, Maria Ivanova, and Eric Wehrli. 2014. SwissAdmin: A multilingual tagged parallel corpus of press releases. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1832–1836, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Jannis Vamvas and Rico Sennrich. 2023. Towards unsupervised recognition of token-level semantic differences in related documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13543–13552, Singapore. Association for Computational Linguistics.

Martin Volk, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel. 2016. Building a parallel corpus on the world's oldest banking magazine. In *Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19-21, 2016*, volume 16 of *Bochumer Linguistische Arbeitsberichte*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

A Copyright Notice

The resulting dataset is released with the following copyright notice:

German / Deutsch (original):

English / Englisch:

B Experiments on Validation Set

Model	Above-threshold	Intersection	Union	Best-DE	Best-FR
`paraphrase-multilingual-mpnet-base-v2`	61.5	46.0	61.5	47.0	46.0
`LaBSE`	66.0	50.5	50.5	50.5	50.5
`sentence-swissBERT`	74.5	69.5	74.5	73.0	74.5
`gte-multilingual-base`	65.0	65.0	65.0	60.0	56.0
`gte-modernbert-base`	66.0	66.0	66.0	66.0	63.0

Table 4: Optimal threshold values for different models and alignment approaches. The corresponding F1 scores are presented in Table 2.