# Computing Semantic Similarity for Aligning Bilingual Semi-parallel Texts: A Case Study

**Steffen Frenzel, Maximilian Krupop, Manfred Stede**
University of Potsdam, Applied Computational Linguistics
{steffen.frenzel, maximilian.krupop, stede}@uni-potsdam.de

**Abstract**

*Semi-parallel text* refers to versions of the same text that have to some extent been edited by authors, translators, or others. They are of relevance especially in the social sciences and in literary genres. In this paper, we consider the bilingual (English/German) variant of the problem. The philosopher Hannah Arendt, for example, wrote political essays that often exist in multiple versions and in both languages. She repeatedly modified her texts, added or deleted parts, and framed topics differently for target audiences. For researchers to explore the history of such material in detail, and at the same time at scale, automatic alignment (i.e., finding the best match of semantically similar sentences) is a very valuable preprocessing step. In this paper, we compare the performances of a range of methods for this task, based on computing semantic similarity. We present the results and conduct a qualitative error analysis to identify recurring sources of error.
**Keywords:** Semi-parallel Text, Alignment, Semantic Similarity

## 1. Introduction

Automatic text alignment has been studied for a long time in the context of machine translation, where the aim is to induce translation models from texts known to be parallel, i.e., close translations of one another. The two widely-used levels in the alignment problem are sentences and words [Jurafsky and Martin, 2026].

In our work we are concerned with pairs of text that are similar to each other but decidedly *not parallel*. These arise in a variety of scenarios. For instance, the translation of literary texts into other languages should intuitively be "close" to the original; however, there are several reasons why such translations may contain different types of changes. Obviously, the language pair itself can cause syntactic or semantic mismatches [Dorr, 1994]; e.g., idiomatic expressions or names often cannot be translated literally and may therefore take a different shape in the target language.

Further, individual translators or publishers (or also government authorities) may invoke changes. An example of this can be found in J.D. Salinger's novel *The Catcher in the Rye*, which we will analyze in more detail below. While the original in American English contains numerous swear words and vulgar language, the German translation from the 1960s shows almost no offensive terms.

Another source of changes can be the authors themselves when they deliberately produce variants of their texts. This is rare in literary works but not uncommon in the social sciences, where adaptations can be done, e.g., in tailoring to a new publication medium or to a different target audience.

For example, the German-American philosopher Hannah Arendt wrote several political essays that were published around and after the Second World War. Depending on the language version, topics such as the persecution of Jews and fascism are framed differently in these texts. In addition, there are revisions that were intended, for example, as radio lectures: Arendt restructured the texts for this purpose, shortening them in places and adding new aspects in others.

Such text adaptations are the subject of debate in literary and translation studies, as well as in the social sciences. Because scholars are interested in studying the nature and potential reasons for adaptions, an interesting problem arises for computational linguistics: determining the "best" alignment of text variants, on the level of sentences. A pre-aligned pair of texts can, after all, be compared and analyzed much more effectively than two raw versions.

Technically, text pairs such as those by Salinger and Arendt provide a good basis for testing the suitability of automatic measurements of *semantic similarity*. For example, the widely-used approach of embedding sentences and comparing them with the cosine metric is assumed to distinguish (mono- or multi-lingual) paraphrases from "unrelated" sentences, but can it capture degrees of shift in meaning that arise in semi-parallel sentences?

In this paper, we test the capabilities of various models for computing semantic similarity as vehicles for automatically aligning semi-parallel text on the sentence level. We create manually-aligned gold data for text material from the two sources mentioned above (Arendt, Salinger) and then compare the similarity models within a window-based alignment algorithm that we designed for handling semi-parallel texts. We report all results using stan-



*Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026*, pages 9–19
11 May 2026. 

dardized evaluation metrics and provide error analyses for both sets of text.

Section 2 discusses relevant related work. Section 3 introduces the data we are using, and Section 4 explains the methods, in particular the semantic similarity measurements. Section 5 reports the results of the experiments, and Section 6 concludes.

## 2. Related Work

### 2.1. Semi-parallel Text

When machine translation turned to statistical methods in the 1990s, the term 'parallel corpora' was used to refer to pairs of texts that were understood as direct translations into another language [Wołk and Marasek, 2017]. These corpora formed the foundation for the induction of statistical translation models.

In between 'different' texts and 'parallel' texts, however, there is effectively a continuum of 'similar' texts, often regarded as a range of 'comparable' corpora [Cheung and Fung, 2004]. These materials have sparked quite a bit of research, which often focused on the task of extracting parallel sentences (e.g., Tillmann [2009], Rauf and Schwenk [2011], Smith et al. [2010], Chu et al. [2013]). In this research the term 'quasi-comparable' is sometimes used for texts that address the same or a similar topic [Cheung and Fung, 2004].

In the aforementioned continuum, we see *semi-parallel* text as one step removed from the 'direct translation': Two texts have the same author, the overall intention is the same in both versions, but occasional modifications have been made. This covers both the monolingual and the multilingual case; though in this paper we focus on a bilingual English/German scenario.

For the semi-parallel setting, research on paraphrase detection and paraphrase generation can provide relevant background. The concept of paraphrases describes the possibilities to change sentences on a lexical, morphological or syntactic level without affecting the meaning [Wahle et al., 2023]. The problems of paraphrase detection (e.g., Gold et al. [2019], Liu and Soh [2022]) and paraphrase generation (e.g., Bandel et al. [2022], Yang et al. [2022]) have lea to manifold research efforts, including the conception of shared tasks based on publicly-available data.

Furthermore, paraphrases are also being analyzed as a phenomenon of 'intertextuality' in the context of digital humanities (e.g., Sier and Wöckener-Gade [2019]). Intertextuality refers to connections between texts that were not necessarily intended by the author. It is described as a phenomenon of reception; that is, the relation of two paraphrases is established by the reader and is analyzed from this perspective.

Recently, Frenzel and Stede [2025] tackled the task of sentence alignment in semi-parallel monolingual (German) text, which was taken from news texts and their simplifications, encyclopedia articles on writers, and also an essay by Hannah Arendt. Apart from sentences, they also tried to utilize the concept of Elementary Discourse Units (cf. Frenzel et al. [2026]) as a level for text alignment.

Our notion of semi-parallel text is based on these various strands of research. In our current experiments, we work with bilingual versions of texts written by Hannah Arendt and J.D. Salinger; the former is associated with social science, the latter with fictional writing. In both cases the text versions in our corpus are not literal translations, but include adaptations made by re-writing and by translating, respectively. Detailed information on the corpus is provided in Section 3.

### 2.2. Text Alignment and Semantic Similarity

In the early days of sentence alignment, the assumption of strong parallelism led to rather successful scoring functions that merely compared the number of words or characters [Brown et al., 1991, Gale and Church, 1993]. Later on, lexical features and heuristics have been utilized to improve the alignment quality while still being temporally efficient (e.g. Moore [2002]). More recently, LERA [Pöckelmann et al., 2022] is an example of a system that models the alignment problem in a graph-theoretic fashion and makes its alignment decisions with a distance function based on the Jaccard index [Jaccard, 1901].

Following the introduction of BERT by Devlin et al. [2019], the use of sentence embeddings has become increasingly popular in this field of research. Notably, Reimers and Gurevych [2019] improved the computation of sentence embeddings with their Sentence-BERT (SBERT) model, which mitigates the computational effort of the classical BERT model.

For comparing sentence embedding vectors to each other, classical similarity calculations such as cosine similarity or Euclidean distance have been employed. One of the first systems that implemented this approach to sentence alignment was VecAlign [Thompson and Koehn, 2019]. Both VecAlign and SentAlign [Steingrimsson et al., 2023] are based on bilingual sentence representations such as LASER [Artetxe and Schwenk, 2019] and LaBSE [Feng et al., 2022]. Building on this work, Molfese et al. [2024] introduced CroCoAlign, an algorithm that incorporates more contextual information for disambiguating possible sentence map-

pings.

In the field of neural information retrieval (IR), recent work commonly follows a retrieve-and-rerank approach: an efficient retriever proposes a small candidate set using independent query/document representations (e.g., dense bi-encoders with cosine similarity), which is then refined by a stronger reranker that jointly models query and candidate text [Nogueira and Cho, 2020]. Beyond single-vector representations, late-interaction models such as ColBERT [Khattab and Zaharia, 2020] compute relevance via token-level matching after independent encoding and provide a middle ground between dense retrieval and full cross-encoders. More recently, listwise rerankers allow richer interactions by processing a query together with multiple candidates in one context window and producing per-candidate relevance scores [Wang et al., 2025]. We adopt this IR perspective to model bilingual alignment as retrieving the best-matching target sentence for each source sentence in semi-parallel text.

## 3. Data

We are using two datasets for our alignment experiments: a book chapter from *The Human Condition*, written by Hannah Arendt, and a part of J. D. Salingers *The Catcher in the Rye*. These texts differ in terms of style, syntactic complexity and also in the level of 'semi-parallelism' and are therefore well suited to test our alignment approaches.

Hannah Arendt's works are currently studied as part of the online publication of a Critical Edition[1], which will allow users to follow the emergence of her works step by step. This is achieved by publishing all versions, including manuscripts, revisions and translations into other languages. Depending on the date of origin, target audience and publication format, these text versions contain numerous variations and differences. Apart from that, Hannah Arendt is known to use long, syntactically complex sentences to express her thoughts. For these reasons, the alignment of the text versions is expected to be challenging.

J.D. Salingers *Catcher in the Rye* is an American coming-of-age novel that was partially published in 1945-46 before being novelized in 1951. This text is interesting for our work for different reasons as well: the novel contains heavily stylized language, containing slang, vulgar language, and many old-fashioned words and expressions. The differences between the American original and the German translation are striking – vulgar language has been considerably "softened" in the German version and the numerous idioms could not be translated literally.

[1]The Critical Edition is published online and can be found here: https://hannah-arendt-edition.net/home

| Corpus Statistics | | | | |
|---|---|---|---|---|
| **Datasets:** | **Arendt** | | **Salinger** | |
| **Language:** | DE | EN | DE | EN |
| **Sentences** | 71 | 59 | 135 | 148 |
| **Words** | 3055 | 2486 | 2089 | 2310 |
| **Avg. Words / Sentence** | 43.0 | 42.1 | 15.5 | 15.6 |

Table 1: Statistics for both corpora.

| Results of the Manual Annotation | | |
|---|---|---|
| **Datasets:** | **Arendt** | **Salinger** |
| **IAA (Cohen's Kappa)** | 0.75 | 0.97 |
| **Non-aligned Share** | 20.0% | 17.0% |

Table 2: Results of the manual annotation.

The basic corpus statistics in Table 1 show that the text versions have different lengths: for Arendt's text, the German version is longer than the English one. Average sentence length, on the other hand, is very similar across languages - with more than 40 words per sentence on average, the sentences are very long. For Salinger's text, the English version is longer. Again the sentence length is similar across languages, but with around 15 words per sentence the complexity is much lower compared to *The Human Condition*.

### 3.1. Manual Annotation and IAA

We conducted a manual annotation study to create gold data for the automatic alignment experiments. Two linguistically-informed annotators worked on both texts, and we can therefore compute Inter-Annotator-Agreement (IAA) for both datasets. Our annotation guidelines were derived from a study that was conducted for monolingual text versions by Frenzel and Stede [2025]. Their guidelines specify that the basis for alignment must always be semantic similarity rather than surface form. It is specified that multiple alignments of the same element should only be made in justified exceptions and that, in contrast, there is no obligation to align all elements. According to these guidelines, the following alignment patterns are allowed: `[1:0, 0:1, 1:n, n:1]`. However, `[n:m]` alignments are not possible.

We use Cohen's Kappa to measure chance-corrected agreement. The results in Table 2 show that the IAA is considerably higher for Salinger's text. This finding is not surprising, since the text is more parallel; this is underscored also by the amount of text that remains non-aligned: In Arendt's essay, 20% out of the 70 sentences from the source text are not aligned (14 sentences), but for Salinger only 17% remain non-aligned (23 sentences).

After both annotators had processed the texts independently and the IAA had been measured, all disagreements were discussed, and a gold standard was derived from both annotations.

# 4. Methods

## 4.1. VecAlign

`VecAlign` [Thompson and Koehn, 2019] is a widely-used alignment model based on contextualized embeddings. It uses a Dynamic Programming approach on the cosine distance of `LASER` sentence embeddings [Artetxe and Schwenk, 2019] to align bilingual text versions. However, this approach was originally designed to align parallel texts and cannot produce mappings that violate parallel sentence ordering. For our purposes, we use `VecAlign` as a baseline to test whether a more flexible approach leads to improvements in alignment quality.

## 4.2. Alignment Window

As our framework for testing different approaches to automatic alignment, we cast bilingual alignment as a retrieval problem: Given a source sentence $s_i$ as a query, the goal is to retrieve the most relevant target sentence(s) $t_j$ from the target text. In contrast to parallel sentence alignment, our semi-parallel setting contains insertions, deletions, and local rewrites. Therefore, we operationalize relevance in terms of semantic similarity instead of surface overlap between segments.

To reflect the largely (though not entirely) monotonic progression of both versions between source and target texts, we restrict the retrieval search space to a sliding window of size $w$ around the current source index. For each query sentence $s_i$ we construct a candidate set $\mathcal{C}_i = \{t_{i-w}, \ldots, t_{i+w}\}$ and compute a relevance score $f(s_i, t_j)$ for each candidate $t_j \in \mathcal{C}_i$. We then select the best-matching target sentence (`1:1`). Candidates can be reused across different queries, i.e. the same target element may be selected by multiple queries ($n$:1).

In our implementation, a strict `1:1` matching can be enforced but it may put a ceiling on model performance compared to the human annotation we use as gold data. Spans of consecutive target elements may be aligned together, resulting in `1:m` matching; however, span alignment does increase the number of candidates per source element linearly.

Our main interest is to compare several different similarity metrics inside the sliding window to find the best match in the respective candidate set. We test cosine similarity of sentence embeddings in conjunction with three different embedding models; BERTScore and NLI Entailment probability; a neural reranker model; and finally we also prompt GPT-5-Mini for this task. In the following we describe these alternative scorers in more detail.

**Cosine Similarity:** Calculating the cosine similarity of sentence embeddings is a very popular approach for text alignment, as it offers several practical advantages: The calculation is relatively inexpensive, as embeddings can be precomputed once, i.e., for the similarity computation no additional forward passes through the encoder are required. The approach can also be combined with different embedding models, making it flexible in terms of language coverage and computational complexity. We test cosine similarity within the window approach described above with three different multilingual embedding models: we use `LASER` [Artetxe and Schwenk, 2019] to be able to compare our window algorithm to `VecAlign` directly. In addition to that, we use the two multilingual, BERT-based models `LaBSE` [Feng et al., 2022] and `paraphrase-multilingual-MiniLM-L12-v2` [Reimers and Gurevych, 2019].

**BERTScore:** The BERTScore [Zhang et al., 2020] is very similar in principle to the cosine approach mentioned above. The algorithm was developed to evaluate the quality of translations or summaries by first calculating the pairwise cosine similarity for all tokens of a candidate text and a reference text. The final BERTScore is obtained by selecting the maximum similarity for all tokens using greedy matching and calculating an average per sentence from this. Optionally, IDF importance weighting can be activated to give rare words greater weight in the BERTScore.

**NLI Entailment Probability:** Natural Language Inference (NLI) refers to the modeling of inference relationships, which are expressed in the form of binary relations between two textual units (e.g., sentences). An entailment relation holds whenever the truth of one text fragment follows from another text. Therefore, the alignment is in this case modeled as a text classification task: the relation between source sentence and the target sentences is to be labeled as 'entailment', 'neutral'

or 'contradiction'. Our scorer selects the target sentence that is assigned the 'entailment' label with the highest probability. To perform the classification task, we used `joeddav/xlm-roberta-large-xnli`, a RoBERTa model that was explicitly finetuned on NLI-labeled datasets in 15 languages, including German and English.

**Neural Reranker:** We use `jina-reranker-v3` [Wang et al., 2025] as a multilingual neural reranker. The model scores each query segment against the corresponding set of candidate segments from our context window. Unlike the bi-encoder retrieval from the previous approaches, the reranker models query-candidate interactions during encoding. The model returns a relevance score for each candidate, where we select the highest-scoring candidate as the alignment for our query segment, thus giving us a `1:1` alignment.

**LLM Prompting:** We prompt OpenAI's GPT-5-Mini[2] via the Responses API. We set the reasoning effort to `low` and use a zero-shot prompt with minimal instructions, where we ask the LLM to align the sentences from German to English based on their indices. We include the query segment $s_i$ and the list of candidate sentences from our context window. To better compare the model to the previous approaches, we test two different alignment settings: in one approach we prompt the model to follow a strict `1:1` matching, while in the other we also specify rules that allow these different forms of alignment: `[1:1, 1:0, 1:n]`. In both cases we then prompt the model to output only valid `JSON`, which we can easily parse for further processing. The full system prompts for both alignment settings can be found in Appendix A.

## 5. Experimental Results

In the following section, we present the quantitative results of our experiments with the range of models described above. In Section 5.2, we provide a qualitative error analysis of the predictions of selected models.

### 5.1. Quantitative Evaluation

**Model Ranking:** In order to compare all models as fairly as possible, for the first set of experiments we reset our window algorithm to its default settings: For each sentence in the source text, exactly one sentence in the target text is required – `1:0` or `1:n` alignments are not permitted in this setup. There are several practical reasons for this: The amount of gold data is not large enough to form a dedicated validation and test set. However, testing `1:0` alignments, which are based on the definition of thresholds using the confidence scores of the models (see below), would require such a validation set. The same applies to span alignment, which in turn must be limited to varying degrees depending on the model. Within the scope of this paper, we therefore provide the performance on the default settings and explore potential improvements by fine-tuning the aforementioned hyperparameters directly on the test set.

Other parameters, such as window size, are not affected by this, as we specified them per dataset and they are therefore identical for all models. While a small window (10 elements in each direction) should suffice for the Arendt text, a larger window was chosen for the Salinger text (10 elements backward and 20 elements forward). This is due to the fact that this text contains more sentences overall and that the target text in this case is longer than the source text – the model should therefore be given a larger context forward than backward.

Recall that `VecAlign` is a standalone system that is not integrated into the window algorithm. It receives the entire German and English text as input and then aligns the sentences using its inherent logic.

The results of all approaches on both datasets are listed in Table 3. While GPT-5-Mini shows the best performance on the Arendt data, the Cosine/LaBSE approach achieves the best f1-score on the Salinger data. Although the best values for both datasets are close to each other, several important differences can be observed: In the Salinger data, precision is higher than recall for all models, while this trend is reversed in the Arendt data. However, since alignment is enforced for each source element in this baseline run, these figures primarily allow conclusions to be drawn about the gold data: There are 8 source elements (11%) in the Arendt text that have not been aligned. The models cannot predict these cases correctly because they are forced to align – and so the recall increases (i.e., the models 'over-align'). One possible way to mitigate this evaluation problem would be to filter the outputs before calculating the scores: if all items that are not aligned `1:1` in the gold data are deleted, possible problems with false negatives and false positives in the evaluation could be avoided. However, the `1:0` and `1:n` alignments in the gold data are indicating a high alignment difficulty - and therefore the evaluation scores would be too generous if these items were excluded from the calculation.

This problem does not exist with Salinger, as all source elements in the gold data have been aligned. The f1 scores therefore confirm the pecu-

[2] `gpt-5-mini-2025-08-07`

| Scorer | P | R | f1 |
|---|---|---|---|
| **Salinger** | | | |
| GPT-5-Mini | 0.89 | **0.88** | 0.88 |
| jina-reranker-v3 | 0.84 | 0.84 | 0.84 |
| BERTScore | 0.87 | 0.79 | 0.83 |
| NLI Entailment Prob | 0.90 | 0.81 | 0.86 |
| Cosine Sim / MiniLM | 0.90 | 0.81 | 0.86 |
| Cosine Sim / LaBSE | **0.95** | 0.86 | **0.90** |
| Cosine Sim / LASER | 0.83 | 0.79 | 0.81 |
| VecAlign | 0.78 | 0.81 | 0.79 |
| **Arendt** | | | |
| GPT-5-Mini | **0.86** | **0.95** | **0.90** |
| jina-reranker-v3 | 0.85 | 0.92 | 0.88 |
| BERTScore | 0.84 | 0.92 | 0.88 |
| NLI Entailment Prob | 0.74 | 0.81 | 0.78 |
| Cosine Sim / MiniLM | 0.80 | 0.87 | 0.83 |
| Cosine Sim / LaBSE | 0.85 | 0.93 | 0.89 |
| Cosine Sim / LASER | 0.77 | 0.80 | 0.78 |
| VecAlign | 0.74 | 0.79 | 0.76 |

Table 3: Results for different datasets and alignment algorithms using the basic settings - an exact 1:1 alignment is enforced for all models.

liarities of the two datasets, but they tend to be too poor for the predictions on the Arendt text, as the models are forced into some incorrect alignments here.

Apart from that, it is striking that VecAlign achieves the worst results in both cases. Since approaches using cosine similarity generally produce good results, this is probably due to the logic of the aligner, which was developed for parallel data and does not allow alignments that contradict the sentence ordering in the two texts. Cases of such "crossing alignments" are not frequent, but do occur in both datasets.

The NLI Entailment Probability delivers good results on the Salinger data, but falls behind on the Arendt data. One possible explanation for this is the high sentence complexity in this text, which can blur entailment relations and thus lead to incorrect alignments.

**Aligning Spans:** Our window algorithm allows the alignment of multiple target elements per source element, provided that the target elements follow each other directly. The maximum length of such spans can be controlled by the user; to test the potential of this span annotation, we ran several trials with different span lengths. However, the results were ambivalent: All models based on cosine similarity almost always chose the longest allowed span in this scenario, even though there were only a few cases in the gold data where spans were aligned at all.

This points to a general weakness of cosine similarity: as long as the sentences are semantically related, they seem to have almost exclusively positive effects on cosine similarity when combined. Individual semantically distant words therefore have a much smaller negative effect than related words have a positive effect. This observation is also supported by the fact that even the cosine similarity between widely differing sentences within our datasets never reaches a negative value, even though the cosine scale ranges from -1 to 1. Other approaches, such as NLI Entailment Probability, were affected by this problem to a much smaller extent.

In order to use span alignment in a meaningful way, it must therefore be penalized to an appropriate extent. We set different levels of penalties for all metrics in our experiments in order to achieve positive results in the end. However, since very few spans were aligned in the gold data and the models occasionally predicted incorrect span alignments, this option did not result in any significant improvements.

**Thresholds for 1:0 Alignments:** Another option for improving the quality of automatic alignments is to allow 1:0 alignments, i.e., cases where no matching element is found in the target text for a source element. The only way to implement this option is to define thresholds: based on the confidence scores, a threshold value can be set for each metric that must be exceeded in order for the alignment to be allowed.

However, since the confidence scores of the various metrics are not comparable with each other, they must be set individually for each model. In the course of our experiments, we also found that, ideally, the scores need to be adjusted in relation to the datasets. Table 4 shows the thresholds with which we achieved the best results for each model and dataset and it also shows the change in the f1 score compared to the values in Table 3. Additionally, we also tested GPT-5-Mini with the 1:0 alignment option. In this case, we adjusted the prompt instead of using a threshold.

It is striking that, for the Cosine/miniLM ap-

| Model | Threshold | f1 |
|---|---|---|
| **Salinger** | | |
| **Cosine / miniLM** | 0.5 | 0.85 (-0.01) |
| **Cosine / LaBSE** | 0.565 | 0.94 (+0.04) |
| **NLI** | 0.87 | 0.88 (+0.02) |
| **GPT-5-Mini** | - | 0.96 (+0.06) |
| **Arendt** | | |
| **Cosine / miniLM** | 0.45 | 0.80 (-0.03) |
| **Cosine / LaBSE** | 0.562 | 0.92 (+0.03) |
| **NLI** | 0.85 | 0.81 (+ 0.03) |
| **GPT-5-Mini** | - | 0.93 (+0.03) |

Table 4: Results with thresholds for seleceted models.

proach, the application of thresholds did not show any improvement for either dataset. Looking at the alignments in detail, it can be seen that, although some correct 1:0 alignments are predicted, false negatives also occur. This suggests that the boundary between correct and incorrect alignments is very blurred in this model, i.e., very similar confidence scores can underlie both cases.

It is also noticeable that the thresholds for the Arendt text have to be set slightly lower for all models than for Salinger. This could be due to the fact that the Salinger text (presumably due to its lower sentence complexity) is easier to align overall and the confidence scores are correspondingly higher.

**Efficiency:** Finally, we will briefly address the issue of efficiency on the theoretical level. Unfortunately, we cannot report comparable data on computing time for all models, since we cannot control hardware use for the LLM approaches, and other factors like internet speed would further blur the results. However, even from a theoretical standpoint, significant differences between the individual metrics become obvious. While approaches that use cosine similarity can generally be implemented very efficiently, the NLI Entailment Probability and the BERTScore in particular are very expensive. The reason for this lies in the underlying logic of these models.

Cosine-based approaches (e.g., Sentence-BERT embeddings, LaBSE, LASER) follow a two-stage computational paradigm: First, each sentence is mapped independently into a fixed-dimensional vector space using a pre-trained encoder. This embedding step is performed once per segment and the resulting vectors can be cached. For multilingual sentence encoders such as LaBSE or LASER, the embedding model processes each segment independently, i.e., no interaction between candidate pairs occurs at this stage. Alignment scoring is then reduced to computing cosine similarity between vector representations, which is computationally inexpensive.

In contrast, the BERTScore does not operate on precomputed sentence embeddings. Instead, it takes raw text as input and performs contextual token-level comparisons using a full transformer model. Thus, for every candidate span, a complete forward pass through a large transformer model is required. When multiple candidate spans are evaluated per source segment, this cost multiplies accordingly. Even when batched, the system must process all candidate pairs jointly through the model, which remains computationally expensive relative to the cosine similarity computations.

NLI models introduce an even stronger computational coupling between candidate pairs, because for each alignment candidate the candidate pair is concatenated as a premise–hypothesis input. The combined sequence is processed jointly by a transformer, and a classification head then predicts probabilities for entailment, contradiction, or neutrality. Unlike independent sentence embeddings, NLI models rely on cross-attention between the two texts, and even when batched, the cost remains proportional to the number of candidate spans, since cross-text attention must be computed for each pair.

## 5.2. Error Analysis

As already mentioned in Section 3, the Salinger text is syntactically less complex than the Arendt essay. Furthermore, there are no `1:0` alignments in the gold data here, which means that very few errors are enforced by the basic settings described in Section 5.1. Almost all errors that do occur are caused by span alignments in the gold data. Since the English text contains 13 more sentences than the German version, several English target sentences sometimes have to be aligned with one German source sentence. In the basic settings, all models are limited to `1:1` alignments, but even when span alignment is allowed, almost all errors occur there. A good example of such a case is shown in Example 1. In this case, all models align only the longer, first target sentence to the source sentence and miss the second English sentence. However, this second sentence is very short and does not contain semantically strong words. The fact that it is not a literal translation makes the alignment even more difficult in this case.

In contrast, the Arendt essay contains more errors that are also more diverse. As mentioned

- Dafür hat Pencey einen guten Ruf als Schule, das muss man sagen.
  (Pencey does have a good reputation as a school, that has to be said.)
- It has a very good academic rating, Pencey.
- It really does.

Examples 1: Alignment error in Salingers *Catcher in the Rye*. Literal translations of the German sentences are provided in round brackets.

- Es handelt nur von den allerelementarsten Gliederungen, in die das Tätigsein überhaupt zerfällt, also von denjenigen, die der Überlieferung wie unserer eigenen Meinung zufolge offenbar innerhalb des Erfahrungshorizonts jedes Menschen liegen sollten.
  (It deals only with the most basic categories into which human activity can be divided, that is, those which, according to tradition and our own judgment, should clearly lie within the scope of every person's experience.)
- It deals only with the most elementary articulations of the human condition, with those activities that traditionally, as well as according to current opinion, are within the range of every human being.
- This, obviously, is a matter of thought, and thoughtlessness – the heedless recklessness or hopeless confusion or complacent repetition of "truths" which have become trivial and empty–seems to me among the outstanding characteristics of our time.

Examples 2: Alignment error in Arendts *The Human Condition*. Literal translations of the German sentences are provided in brackets.

above, there are 8 cases (11%) in which no match for a source element was found in the gold annotations. In two other cases (3.4%), spans of two elements per source element are annotated, even though in this case the source text is longer than the target text. These statistics again indicate a lower degree of parallelism compared to the Salinger text.

However, errors in the model predictions arise not only from these special cases, but also from some incorrect 1:1 alignments. Such an incorrect alignment is shown in Example 2.

In this case the German source sentence was aligned with the third sentence by several models, although it should actually be aligned with the second sentence. The two English sentences are only three index positions apart and are semantically quite closely related: words such as *thoughts*, *opinions*, *articulations*, *range*, etc. come from similar semantic fields, making alignment not unreasonable. The general length of the sentences and their structure are also quite similar. Nevertheless, other words without a counterpart in the German sentence (e.g., *heedless recklessness*, *hopeless confusion*) and formal peculiarities (e.g., the dash or the quotation marks around the word *truth*) should actually interfere with the alignment. It is possible that errors like this one are happening because of sentence length and complexity (the German source sentence contains 32 words), but since most sentences in the Arendt text are equally long and get aligned correctly, this alone may not be the problem.

## 6. Conclusion

This paper reports a case study to systematically test various existing metrics for the task of aligning semi-parallel bilingual texts. The results are promising: both GPT-5-Mini as a generative language model, as well as cosine similarity approaches and the NLI entailment probability achieve f1 scores well above 0.8. Our window logic helps to make alignment more efficient without imposing severe restrictions on the alignment options.

Nevertheless, these results can only be the starting point for further work on aligning semi-parallel texts. First of all, going beyond the present case study by running the experiments on larger datasets is a central next step.

From an empirical perspective, an interesting experiment would be to swap the source and target texts to see if this results in changes to the alignment decisions, even though our annotation guidelines call for consideration of semantic similarities in both directions.

Each similarity metric currently has its own problems, which have become particularly apparent in the area of 1:0 and 1:n alignment. Another important future task will be the systematic testing of ensemble scores to overcome the weaknesses of individual approaches. Further work is also needed in the area of alternative alignment levels: for example, preliminary alignment of paragraphs could help to further narrow down the search areas. Segmenting long sentences into clauses could also simplify difficult cases where parts of sentences overlap while other parts do not match.

## Acknowledgments

We thank our student assistant Dietmar Benndorf for annotating training data, and we are grateful to the anonymous reviewers for their helpful feedback. Our work is supported by the Deutsche Forschungsgemeinschaft (DFG), project (524057241) "Semi-automatische Kollationierung verschiedensprachiger Fassungen eines Textes“.

## A. LLM System Prompts

```
You align German sentences to English
    sentences.
Align ONLY the current German sentence
    labeled "DE".
If DE is a fragment or clause, still
    choose the best matching English
    candidate.
Choose ONLY from the candidate IDs below
    .

German sentence:
{de_instance}

English candidates:
{candidates}

Return JSON only, no extra text. en_span
     must reference English candidate
    IDs:
{{
  "type": "1-1",
  "en_span": [j, k],
  "confidence": 0.0-1.0
}}

Rules:
- Choose exactly ONE candidate ID from
    the list.
- Output must be type "1-1".
- en_span must be [j, j] with j from
    candidate IDs.
- Do not output "none" or "1-many".

Do not include explanations. Output must
     be a single JSON object.
```

Listing 1: System prompt used for strict alignment setting

```
You align German sentences to English
    sentences.
Align ONLY the current German sentence
    labeled "DE".
If DE is a fragment or clause, still
    choose the best matching English
    candidate(s).
Return "none" if no clear match exists
    in the candidates.
Choose ONLY from the candidate IDs below
    .

German sentence:
{de_instance}

English candidates:
{candidates}

Return JSON only, no extra text. en_span
     must reference English candidate
    IDs:
{{
  "type": "1-1" | "1-many" | "none",
  "en_span": [j, k],
  "confidence": 0.0-1.0
}}

Rules:
- If 1-1: en_span must be [j, j].
- If 1-many: en_span must be contiguous
    [j, k] with (k-j+1) <= {max_span}.
- If none: en_span must be [-1, -1].

Do not include explanations. Output must
     be a single JSON object.
```

Listing 2: System prompt used for non-strict alignment setting

## B. References

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

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. Quality Controlled Paraphrase Generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 596–609, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.45.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, page 169–176, USA, 1991. Association for Computational Linguistics. doi: 10.3115/981344.981366.

Percy Cheung and Pascale Fung. Sentence Alignment in Parallel, Comparable and Quasi-comparable Corpora. In *LREC Workshop on the amazing utility of parallel corpora*,

2004. URL https://api.semanticscholar.org/CorpusID:29228186.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. Chinese–Japanese Parallel Sentence Extraction from Quasi–Comparable Corpora. In Serge Sharoff, Pierre Zweigenbaum, and Reinhard Rapp, editors, *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 34–42, Sofia, Bulgaria, 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2505/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Bonnie J. Dorr. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633, 1994. URL https://aclanthology.org/J94-4004/.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62.

Steffen Frenzel and Manfred Stede. Sentence-Alignment in Semi-parallel Datasets. In Anna Kazantseva, Stan Szpakowicz, Stefania Degaetano-Ortlieb, Yuri Bizzoni, and Janis Pagel, editors, *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 87–96. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.latechclfl-1.9.

Steffen Frenzel, Maximilian Krupop, and Manfred Stede. Discourse segmentation of german text with pretrained language models. *Journal for Language Technology and Computational Linguistics*, 2026.

William A. Gale and Kenneth W. Church. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102, 1993. URL https://aclanthology.org/J93-1004.

Darina Gold, Venelin Kovatchev, and Torsten Zesch. Annotating and analyzing the interactions between meaning relations. In Annemarie Friedrich, Deniz Zeyrek, and Jet Hoek, editors, *Proceedings of the 13th Linguistic Annotation Workshop*, pages 26–36, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4004.

Paul Jaccard. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, page 547–579, 1901.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. Pearson, 3rd edition, 2026. URL https://web.stanford.edu/~jurafsky/slp3/. Online manuscript released January 6, 2026.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert, 2020. URL https://arxiv.org/abs/2004.12832.

Timothy Liu and De Wen Soh. Towards Better Characterization of Paraphrases. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8592–8601, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.588.

Francesco Molfese, Andrei Bejgu, Simone Tedeschi, Simone Conia, and Roberto Navigli. CroCoAlign: A Cross-Lingual, Context-Aware and Fully-Neural Sentence Alignment System for Long Texts. In Yvette Graham and Matthew Purver, editors, *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, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.135.

Robert C. Moore. Fast and accurate sentence alignment of bilingual corpora. In Stephen D. Richardson, editor, *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 135–144, Tiburon, USA, 2002. Springer. URL https://link.springer.com/chapter/10.1007/3-540-45820-4_14.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert, 2020. URL https://arxiv.org/abs/1901.04085.

Marcus Pöckelmann, André Medek, Jörg Ritter, and Paul Molitor. LERA—an interactive platform for synoptical representations of multiple text witnesses. *Digital Scholarship in the Humanities*, 38(1):330–346, 2022.

Sadaf Abdul Rauf and Holger Schwenk. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, 25(4): 341–375, 2011. ISSN 09226567, 15730573. URL http://www.jstor.org/stable/41487466.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019. URL https://arxiv.org/abs/1908.10084.

Kurt Sier and Eva Wöckener-Gade. Paraphrase als Ähnlichkeitsbeziehung. Ein digitaler Zugang zu einem intertextuellen Phänomen. In *Platon Digital. Tradition und Rezeption*. Propylaeum, 2019.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In Ron Kaplan, Jill Burstein, Mary Harper, and Gerald Penn, editors, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411, Los Angeles, California, 2010. Association for Computational Linguistics. URL https://aclanthology.org/N10-1063/.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. SentAlign: Accurate and Scalable Sentence Alignment. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.22.

Brian Thompson and Philipp Koehn. Vecalign: Improved Sentence Alignment in Linear Time and Space. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1136.

Christoph Tillmann. A Beam-Search Extraction Algorithm for Comparable Data. In *Annual Meeting of the Association for Computational Linguistics*, 2009. URL https://api.semanticscholar.org/CorpusID:7798552.

Jan Wahle, Bela Gipp, and Terry Ruas. Paraphrase Types for Generation and Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 12148–12164. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.746. URL http://dx.doi.org/10.18653/v1/2023.emnlp-main.746.

Feng Wang, Yuqing Li, and Han Xiao. jina-reranker-v3: Last but not late interaction for listwise document reranking, 2025. URL https://arxiv.org/abs/2509.25085.

Krzysztof Wołk and Krzysztof Marasek. Unsupervised Construction of Quasi-comparable Corpora and Probing for Parallel Textual Data. In Aleksander Zgrzywa, Kazimierz Choroś, and Andrzej Siemiński, editors, *Multimedia and Network Information Systems*, pages 307–320, Cham, 2017. Springer International Publishing.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Haibo Zhang, Xue Zhao, Wenqing Yao, and Boxing Chen. GCPG: A General Framework for Controllable Paraphrase Generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4035–4047, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.318.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.