Alignment Quality Degradation Across the Parallel–Comparable Spectrum: A Comparative Analysis

Audrey Mash, Jonathan Ayebakuro Orama, Marc Juvillà Garcia, Maite Melero
Barcelona Supercomputing Center
audrey.mash, jonathan.ayebakuro, marc.juvilla, maite.melero @bsc.es

Abstract

Sentence-level alignment systems have been developed and evaluated primarily on parallel data, leaving their behaviour across the broader parallel–comparable spectrum of real web content poorly understood. We present a stratified empirical study of alignment quality for Catalan–English using 300 document pairs across three parallelism bands defined by mean-max LaBSE cosine similarity. We compare four systems: a hierarchical alignment pipeline (DocAlign), an ablation with paragraph pre-filtering disabled (DocAlign-NoFilter), the flat aligner Vecalign (Thompson and Koehn, 2019), and a flat LaBSE greedy baseline. Evaluation uses human-annotated sentence pairs and coverage-weighted quality. Quality degrades at different rates by system type: hierarchical systems maintain usable-pair rates ranging from 25% to 51% on comparable data while flat systems collapse to 2–7%. Paragraph pre-filtering reduces output volume on comparable data while *raising* pair quality relative to the unfiltered ablation. Vecalign is statistically indistinguishable from the greedy baseline at all parallelism levels, suggesting that LaBSE embedding discrimination is the binding constraint on flat alignment quality. Failure mode analysis of 550 low-rated pairs identifies topical mismatch as the dominant failure mode, with structural noise concentrated in flat systems.

1 Introduction

The availability of large parallel corpora has been central to the progress of statistical and neural machine translation. As MT systems increasingly target document-level translation, the demand has shifted toward document-aligned data that preserves long-range discourse structure. Web-crawled collections such as DocHPLT (O'Brien et al., 2025) offer scale, but web content does not sit neatly at one point on the parallelism spectrum: the same crawl that yields direct translations in one language pair yields independently authored topically related documents in another, and often yields both within the same collection.

Existing sentence alignment systems were designed for the parallel end of this spectrum. Vecalign (Thompson and Koehn, 2019), among the most widely used, was evaluated on known-parallel test sets. When practitioners apply these tools to web-crawled data without verifying parallelism level, the degradation in alignment quality is real but largely uncharacterised. There is no systematic evidence for how much quality drops as document pairs become less parallel, whether different alignment architectures degrade at different rates, or which failure modes are responsible.

This paper addresses that gap. We make the following contributions:

- A stratified empirical analysis of alignment quality across three parallelism bands for 300 Catalan–English document pairs, with human evaluation at sentence-pair level.
- An ablation design that isolates the contribution of paragraph-level pre-filtering by comparing a full hierarchical pipeline against an otherwise identical system with pre-filtering

disabled.

- A characterisation of the failure modes that emerge on comparable data, grounded in annotation evidence, with mechanistic explanations for why each mode occurs in particular system types.

- A coverage-weighted quality metric that accounts for the precision–recall trade-off across systems with different selectivity.

The instrument we use is DocAlign, a hierarchical alignment pipeline combining paragraph-level Dynamic Time Warping (DTW) with sentence-level fuzzy span matching. Catalan–English is chosen for institutional relevance and because the first author can annotate reliably in both languages.

2 Related Work

Sentence alignment has a history spanning more than three decades, progressing through three broad generations. Length-based methods, from the foundational probabilistic model of (Gale and Church, 1991) to the hybrid length-and-lexical-similarity approach of (Varga et al., 2007), treat alignment as a global sequence optimisation and work well on close-to-parallel texts. A later strand introduced machine translation as a bridge: Bleualign (Sennrich and Volk, 2011) bootstraps alignment from MT output, avoiding dependence on pre-existing lexical resources. The most recent systems ground alignment in massively multilingual sentence embeddings, enabling embedding-based tools such as Vecalign (Thompson and Koehn, 2019) and SentAlign (Steingrímsson et al., 2023) that score sentence pairs in a shared multilingual space. Across all three generations, alignment systems have been designed and evaluated primarily on parallel data. Early systems explicitly assumed few reorderings and limited insertions and deletions between documents (Munteanu and Marcu, 2005), and while embedding-based systems relax these structural assumptions, the core evaluation practice of testing on known-parallel data has persisted, leaving behaviour on comparable texts largely uncharacterised.

The construction of large parallel corpora from web crawls has been an active research area since at least (Resnik and Smith, 2003), who proposed the first systematic approach to mining parallel text from multilingual web pages. Subsequent work scaled this programme substantially: OPUS (Tiedemann, 2012) aggregated data across dozens of domains and language pairs, while ParaCrawl (Bañón et al., 2020), CCAligned (El-Kishky et al., 2020), and CCMatrix (Schwenk et al., 2021) demonstrated that web crawls could yield parallel data at previously infeasible scale. Across this progression, however, parallelism was treated as a property to be filtered for rather than measured: the WMT 2023 parallel data curation shared task (Sloto et al., 2023) frames quality recovery from noisy web-crawled data as an open challenge. DocHPLT (O'Brien et al., 2025) marks a departure, preserving complete document structure including unaligned content and annotating each document pair with alignment density metrics — making variation in parallelism explicit and quantifiable rather than filtered away — and providing the data infrastructure on which the present study is built.

Two threads of prior work are most directly relevant to DocAlign's architecture, and together they define the gap this paper addresses. Works that apply hierarchical structure to cross-lingual data, such as the Hierarchical Document Encoder of (Guo et al., 2019) and the sentence-order-aware document alignment of (Thompson and Koehn, 2020), deploy document-level organisation to improve the retrieval of parallel document pairs, but neither investigates sentence-level alignment quality or how it degrades as documents become less parallel. The structural precedent for using paragraph-level organisation to constrain sentence matching in non-parallel data comes from (Barzilay and Elhadad, 2003), who address monolingual comparable corpora by first inducing topical structure at the paragraph level and then performing local sentence alignment within matched paragraph pairs - a structural insight that, to our knowledge, has not been extended to cross-lingual alignment. The foundational cross-lingual finding that standard alignment methods fail on comparable data comes from (Munteanu and Marcu, 2005), who showed that length- and position-based aligners cannot reliably identify translation pairs in comparable newspaper corpora. Their concern was the reliable extraction of sentence pairs from comparable data, a different research question from characterising alignment quality as a graded function of parallelism degree. DocAlign's hierarchical pipeline, which uses LaBSE-based paragraph

grouping and DTW to constrain sentence-level search, is, to our knowledge, the first such system to be evaluated systematically across stratified parallelism bands using both automatic metrics and human annotation. This evaluation reveals how structural hierarchy mediates the coverage–quality tradeoff as parallelism diminishes.

3 Methodology

3.1 Band Definition and Assignment

To enable a controlled comparison across the parallelism spectrum, document pairs are stratified into three bands based on a reproducible, model-agnostic similarity measure. For each document pair, we compute the *mean-max cosine similarity* between paragraph embeddings: for every source paragraph, we take its maximum cosine similarity to any target paragraph using LaBSE embeddings (Feng et al., 2022), then average these maxima across all source paragraphs. This mean-max formulation is preferable to a simple mean of all pairwise similarities, as it correctly penalises documents containing paragraphs with no counterpart on the other side — a common feature of web-crawled comparable content — without being sensitive to the absolute number of paragraphs.

Three bands are defined as follows:

- **Band 1 (Parallel):** mean-max similarity > 0.75. Documents in this range are typically direct translations with high correspondence throughout.
- **Band 2 (Moderate):** mean-max similarity 0.50–0.75. Web-parallel content with variable correspondence; structural drift is common.
- **Band 3 (Comparable):** mean-max similarity 0.30–0.50. Documents share a topic but are not direct translations.

These thresholds were established prior to any alignment experiments and held fixed throughout (Band 1: 1,136 pairs; Band 2: 1,727; Band 3: 135; 2 below threshold).

Inspection of sampled document pairs revealed relationships ranging from independently authored coverage of the same news event, to pages describing the same entity from different perspectives, to topically adjacent content with no direct correspondence. We treat this heterogeneity as ecologically valid: it is precisely the kind of variation a deployed alignment system must handle.

Band	Sim. range	n	Mean sim.	Mean paras
1	> 0.75	100	0.840	93.96
2	0.50–0.75	100	0.630	100.18
3	0.30–0.50	100	0.458	46.15
Total		300	—	—

Table 1: Corpus statistics by parallelism band for the 300-pair experimental dataset. Similarity is mean-max LaBSE cosine similarity computed from raw paragraph embeddings (Section 3.1). Mean paras is the average source paragraph count. The 100 pairs per band were drawn by stratified random sampling from a scored pool of 3,000 pairs.

3.2 Data and Corpus Construction

Document pairs are drawn from DocH-PLT (O'Brien et al., 2025), a large-scale multilingual document collection constructed from web crawls and organised by document-level language pairs.

A minimum length threshold of 20 paragraphs per side is applied before band assignment. This threshold is motivated by the properties of the DTW alignment used in DocAlign: on very short documents, the DTW path is near-diagonal regardless of actual correspondence, producing alignment behaviour essentially indistinguishable from flat sentence alignment. Documents below this threshold do not provide a meaningful test of hierarchical alignment.

After applying this floor, approximately 1.48 million document pairs (41% of the full ca-en subset of 3.58 million pairs) were available for band assignment. From these, 3,000 were randomly sampled for similarity scoring. The final dataset consists of 100 document pairs per band, 300 pairs in total, selected by stratified random sampling within each band. All four alignment systems are run on exactly the same 300 pairs; no document pair is modified or replaced after the dataset is fixed.

3.3 Systems

We compare four systems, all run on the same 300 document pairs. All systems produce sentence-pair output, which is the evaluation unit (Section 3.4). Table 2 summarises the systems and the component each pairwise comparison isolates.

DocAlign[1] is a hierarchical two-stage align-

[1]Anonymised code and data available at: https://github.com/docalign2026-hue/alignment-paper.

ment pipeline. In the first stage, all paragraphs are embedded using LaBSE (Feng et al., 2022) and aligned using DTW, which maps the source paragraph sequence to the target paragraph sequence while allowing many-to-one and one-to-many groupings. Paragraph groups with mean cosine similarity below a threshold are discarded before proceeding. In the second stage, sentences within each surviving paragraph group are split using a multilingual spaCy model (Honnibal et al., 2020)[2], embedded with LaBSE, and aligned using a greedy fuzzy span matching algorithm that supports 1:1, 1:2, and 2:1 sentence alignments.

DocAlign-NoFilter is an ablation of DocAlign in which the paragraph pre-filter is disabled by setting `para_sim_threshold: 0.0`. All other components are identical, including the sentence splitter, embedding model, and fuzzy span matching algorithm. Because the pre-filter is bypassed, the sentence aligner receives the full document as input, placing DocAlign-NoFilter on equal footing with the flat systems with respect to input scope — though it still uses DTW-derived paragraph grouping to structure its sentence-level search, unlike the flat systems which impose only positional constraints on the search space rather than semantically-derived structural groupings. The comparison between DocAlign and DocAlign-NoFilter directly isolates the contribution of paragraph-level pre-filtering.

Vecalign (Thompson and Koehn, 2019) is a widely used flat sentence aligner that uses overlapping sentence-block embeddings (averaging adjacent sentences to represent multi-sentence spans) and a recursive dynamic programming approximation for alignment. It supports 1:0, 0:1, 1:1, 1:2, and 2:1 alignments. We run Vecalign with default configuration throughout and do not tune it for comparable data; the intent is to observe how a standard tool behaves when applied beyond its design conditions. To enable score comparability across systems, Vecalign's internal alignment scores are replaced with LaBSE cosine similarity scores computed over the aligned sentence pairs. In Section 4.5 we additionally apply post-hoc score filtering to Vecalign's output to test whether the quality gap can be recovered without retraining or re-configuration.

LaBSE baseline is a flat greedy 1:1 matcher that serves as a lower bound. For each source sentence, the highest-scoring unmatched target sentence within a position-proportional window of ±15 sentences is selected, subject to a minimum similarity threshold of 0.30 and strict monotonic ordering. The same sentence splitter (spaCy `xx_ent_wiki_sm`) is used as in DocAlign, ensuring that differences in output are attributable to the alignment algorithm and not to sentence segmentation.

DocAlign uses `para_sim_threshold : 0.30`; DocAlign-NoFilter sets this to 0.0. Both use approximate DTW with paragraph embeddings cached as `.npy` files.

[2]spaCy version 3.8.11, model `xx_ent_wiki_sm`.

Comparison	Component isolated
DocAlign vs. DocAlign-NoFilter	Paragraph pre-filtering
DocAlign-NoFilter vs. Vecalign	Paragraph-level search structure (DTW grouping vs. flat)
Vecalign vs. LaBSE baseline	Many-to-one span matching
DocAlign vs. Vecalign	Full hierarchical vs. flat (headline comparison)

Table 2: Pairwise comparisons and the component each isolates.

3.4 Evaluation Design

Evaluation: sentence-pair level. The unit of annotation is the sentence pair: one or two source sentences aligned to one or two target sentences (1:1, 1:2, or 2:1). This is the natural output unit for all four systems and allows direct cross-system comparison without system-specific post-processing affecting the items being judged.

For each of the 12 band–system cells (3×4), up to 100 sentence pairs are sampled from the system output, with one item per source document to avoid within-document clustering. Not all cells contain 100 usable documents: DocAlign produces output for 75 documents on Band 3 and DocAlign-NoFilter for 62, while all other cells contain 97–100. Items are stratified by automatic alignment score to ensure the sample covers the full quality range rather than being biased toward high-confidence output. System identity is blinded: annotators see a source text, a target text, and an alignment type label (1:1, 1:2, or 2:1), but not which system produced the pair.

Annotation uses a 5-point scale ranging from 1 (unrelated content; the alignment is incorrect) to 5 (essentially the same content in both languages; a translator would not change it). Scores of 4–5

are considered usable for MT training; scores of 1–2 are considered wrong or poor. Full annotation guidelines are provided in Appendix A.

Annotation is carried out by three annotators: a professional translator and native Catalan speaker (Annotator A), an independent native Catalan colleague (Annotator B), and the first author (Annotator C), a native English speaker with Catalan reading proficiency. Annotators A and C split the main annotation items; Annotator B rates only the inter-annotator agreement (IAA) subset. The IAA subset consists of 114 items distributed across all 12 band–system cells, rated independently by all three annotators. Cohen's κ is reported for all three pairwise combinations; the A–B pair is methodologically strongest as neither annotator has a stake in the results. Annotation uses a browser-based tool hosted on GitHub Pages with a Google Sheets backend for session persistence. System identity is blinded via randomised item IDs decoded post-hoc. Each annotator receives the full IAA subset followed by main items in auto-allocated batches of 50.

3.5 Metrics

We report three types of metric, applied per system per band.

Manual quality (from sentence-pair annotation): mean quality score with 95% confidence intervals; percentage of pairs rated 4–5 (usable for MT training); percentage rated 1–2 (incorrect or poor). These are computed over the annotated items per band–system cell.

Sentence coverage (automatic, over all 300 document pairs): the percentage of source sentences appearing in at least one output alignment pair. Coverage is computed per document then averaged within each band. Coverage is reported alongside quality scores because the two measures are not independent: a system that outputs only its highest-confidence pairs will achieve a higher mean quality score than one that aligns more of the document. Reporting quality without coverage would systematically favour selective systems.

Coverage-weighted quality (CWQ) combines the automatic alignment score and sentence coverage as their product:

$$\mathrm{CWQ} = \bar{s}_{\mathrm{auto}} \times c$$

where $\bar{s}_{\mathrm{auto}}$ is the mean LaBSE cosine similarity of aligned sentence pairs and c is mean sentence coverage, both computed automatically over all documents in the band. CWQ can be interpreted as the expected quality of a randomly drawn source sentence: the product of the probability that the sentence is aligned at all (coverage) and its expected quality given alignment (mean score). CWQ rewards systems that are both precise and recall-oriented, and penalises systems that improve one measure at the cost of the other. It is our primary ranking metric for cross-system comparison when coverage and quality must be considered jointly. We use automatic LaBSE scores rather than human scores in this formula because human annotation covers only a stratified sample of ∼75–85 pairs per band–system cell, while automatic scores are available for all sentence pairs across all 300 documents; using human scores would require extrapolating from a small sample to produce document-level coverage estimates. The limitations of automatic scores as a proxy for translational equivalence on comparable data are noted in Section 4.2.

Statistical testing: pairwise system comparisons within each band use the Mann-Whitney U test, appropriate for ordinal annotation data with no distributional assumptions. p-values are Bonferroni-corrected for the six pairwise tests per band (18 total across all bands), and corrected p-values below 0.05 are considered significant. Effect sizes are reported as rank-biserial correlation r.

4 Results

Table 3 summarises automatic metrics across all 300 document pairs; human evaluation results follow in Section 4.1.

4.1 Human Evaluation: Quality Degradation Across Bands

Figure 1 shows mean human quality scores across parallelism bands for all four systems. The two hierarchical systems (DocAlign and DocAlign-NF) diverge sharply from the two flat systems (Vecalign and the LaBSE baseline) as document parallelism decreases. On Band 1, all systems score above 3.3; by Band 3, the flat systems fall below 1.5 while the hierarchical systems remain above 2.7. The gap widens monotonically, confirming that the choice of alignment architecture has greater consequences on comparable data than on parallel data.

		DocAlign	DocAlign-NF	Vecalign	Baseline
Band 1	Docs w/ output	100	100	100	100
	Sentence pairs	5,802	5,806	9,324	9,330
	Coverage (%)	58.1	58.2	86.8	84.9
	Mean auto-score	0.906	0.906	0.700	0.655
Band 2	Docs w/ output	97	97	100	100
	Sentence pairs	1,566	1,568	10,074	8,914
	Coverage (%)	17.1	17.1	82.2	70.7
	Mean auto-score	0.831	0.831	0.397	0.464
Band 3	Docs w/ output	75	62	100	100
	Sentence pairs	483	397	6,541	5,120
	Coverage (%)	4.7	3.6	71.4	51.5
	Mean auto-score	0.839	0.890	0.321	0.409

Table 3: Automatic metrics by system and band, computed over all 300 document pairs. Coverage is the percentage of source sentences appearing in at least one output pair, averaged within each band. Auto-score is mean LaBSE cosine similarity of aligned pairs. DocAlign and DocAlign-NF produce identical output on Bands 1–2 because the paragraph pre-filter is non-binding at those similarity levels (Section 4.2).

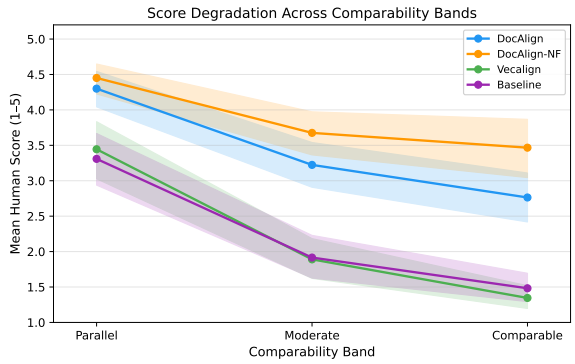

Figure 1: Mean human quality score (1–5) per system across parallelism bands. Shaded regions show 95% bootstrap confidence intervals. The hierarchical–flat divergence widens monotonically as parallelism decreases.

Table 4 reports the full primary results. On Band 1, DocAlign-NF achieves the highest mean score (4.45, 82% usable), closely followed by DocAlign (4.30, 77% usable). The flat systems score 3.44 (Vecalign, 57% usable) and 3.31 (baseline, 52% usable). A quality gap between hierarchical and flat systems is already present on parallel data, though effect sizes at Band 1 are moderate compared to what follows.

On Band 2, the divergence becomes pronounced. DocAlign-NF scores 3.68 (55% usable, 27% poor); DocAlign drops to 3.22 (40% usable, 41% poor). The flat systems collapse to approximately 1.9 (Vecalign 15% usable, 79% poor; baseline 17% usable, 81% poor). Over three-quarters of flat-system pairs on moderate documents are not suitable for MT training.

Band 3 shows the most extreme differentiation. DocAlign-NF achieves 3.47 (51% usable, 36% poor), while DocAlign falls to 2.76 (25% usable, 47% poor). Both flat systems score below 1.5 (Vecalign 1.35, 2% usable, 91% poor; baseline 1.48, 7% usable, 93% poor). These documents are within the nominal operating range of both flat systems, yet nearly all output is unusable.

4.2 Ablation: Role of the Paragraph Pre-Filter

DocAlign-NF removes the paragraph similarity pre-filter that DocAlign applies before sentence-level alignment. On Bands 1–2, both systems produce identical output: paragraph similarities on parallel and moderately parallel data are high enough that no paragraph groups fall below the default pre-filter threshold of 0.30, so the filter is non-binding. The ablation comparison is therefore valid only for Band 3, where paragraph similarities are routinely below threshold and the pre-filter actively discards candidate groups.

The automatic metrics produce a counterintuitive result: removing the pre-filter *reduces* both coverage (4.7% $\rightarrow$ 3.6%) and the number of documents that produce any output at all (75 $\rightarrow$ 62). (Table 5) The pre-filter is therefore not functioning primarily as a quality gate—it is functioning as a *failure prevention mechanism*. Without it, the pipeline processes paragraph groups that are too dissimilar for sentence-level alignment to succeed, triggering downstream threshold failures and producing zero output rather than low-quality output.

Human annotation tells a different story at the pair level. DocAlign-NF achieves a higher mean quality score on Band 3 (3.47 vs. 2.76) and a higher proportion of usable pairs (51% vs. 25%).

		DocAlign	DocAlign-NF	Vecalign	Baseline
Band 1	n	70	71	72	81
	Mean quality $\pm$ CI	4.30 ± 0.25	$\mathbf{4.45 \pm 0.21}$	3.44 ± 0.40	3.31 ± 0.36
	% usable (4–5)	77	**82**	57	52
	% poor (1–2)	**7**	6	38	40
	CWQ†	0.527	0.527	**0.608**	0.556
Band 2	n	76	74	82	83
	Mean quality $\pm$ CI	3.22 ± 0.32	$\mathbf{3.68 \pm 0.31}$	1.89 ± 0.28	1.92 ± 0.30
	% usable (4–5)	40	**55**	15	17
	% poor (1–2)	41	**27**	79	81
	CWQ†	0.142	0.142	**0.326**	0.328
Band 3	n	55	45	81	85
	Mean quality $\pm$ CI	2.76 ± 0.35	$\mathbf{3.47 \pm 0.41}$	1.35 ± 0.16	1.48 ± 0.20
	% usable (4–5)	25	**51**	2	7
	% poor (1–2)	47	**36**	91	93
	CWQ†	0.039	0.032	**0.229**	0.211

Table 4: Primary evaluation results per system per band. Mean quality from sentence-pair human annotation (1–5 scale). †CWQ = coverage-weighted quality (mean automatic LaBSE score $\times$ mean sentence coverage), computed over all 300 document pairs per band. All other rows are from human annotation. All hierarchical-vs-flat contrasts are significant at Bands 2–3 (Bonferroni-corrected Mann-Whitney U, $p_{\text{corr}} < 0.05$); at Band 1, DocAlign-NF vs. flat systems reaches significance but DocAlign vs. Vecalign does not ($p_{\text{corr}} = 0.129$). Within-class contrasts are non-significant at all bands (Section 4.4).

	DocAlign	DocAlign-NF
Docs w/ output	75	62
Coverage (%)	4.7	3.6
CWQ	0.039	0.032
Mean quality	2.76	**3.47**
% usable (4–5)	25	**51**

Table 5: Ablation: DocAlign vs. DocAlign-NF on Band 3. Coverage and CWQ are automatic; quality and usable rate are from human annotation. On Bands 1–2 the pre-filter is non-binding and both systems produce identical output (Section 4.2).

The pairs that DocAlign-NF does produce are rated higher than those produced by full DocAlign, even though it produces fewer of them.

These two findings are consistent: DocAlign-NF fails more often (fewer documents with any output), but when it succeeds, the individual pairs are cleaner. The likely mechanism is that marginal paragraph groups passed by DocAlign's pre-filter produce low-scoring pairs that nonetheless clear the chunk-level quality threshold; without the pre-filter, those same groups generate chunks that fall below it entirely, leaving only output from groups with genuine sentence-level correspondence.

Notably, CWQ (computed from automatic LaBSE scores) slightly favours DocAlign (0.039 vs. 0.032) while human annotation reverses this ranking. This divergence arises because automatic similarity conflates topical overlap with translational equivalence—the same bottleneck underlying the Vecalign–baseline non-result—while human scores reflect actual translation quality. The pre-filter retains paragraph groups in a middle similarity range: high enough to pass the 0.30 threshold but too low for genuine sentence-level correspondence. These marginal groups produce predominantly Type T failures (Section 5.1). Without the pre-filter, most such groups fail at sentence-level thresholds and produce zero output; those that succeed are the groups with genuine correspondence. The pre-filter thus converts what would be a coverage loss into a quality loss.

This finding is suggestive ($p_{\text{corr}} = 0.327$; Table **??**) and should not be over-interpreted given the sample sizes (DocAlign $n = 55$, DocAlign-NF $n = 45$). The practical direction is nonetheless clear: on web-crawled comparable data, disabling the pre-filter and accepting more document-level failures may be preferable to retaining it and accepting lower pair-level quality.

DocAlign's non-1:1 alignment rate increases from 5.9% at Band 1 to 28.6% at Band 3 (Appendix C), reflecting the DTW stage grouping structurally divergent paragraph pairs and forcing many-to-one resolution at sentence level; Vecalign's non-1:1 rate remains flat at 5–10% across all bands, suggesting it does not exploit span alignment where structural divergence is highest.

4.3 Coverage vs. Quality: The Precision–Recall Tradeoff

Figure 2 plots each system's mean sentence coverage against its mean human quality score per band, making the precision–recall tradeoff directly

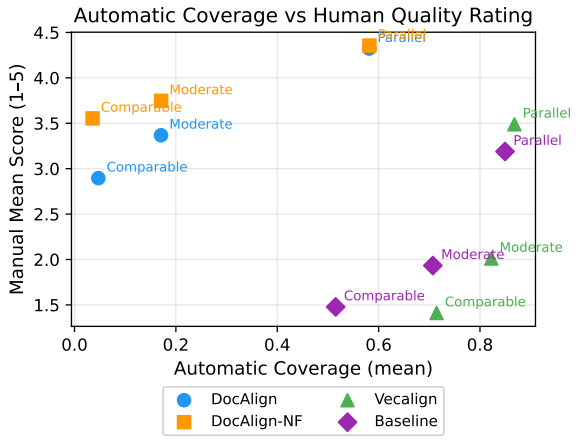

Figure 2: Mean automatic sentence coverage (x-axis) vs. mean human quality score (y-axis) per system per band. Each point is one system–band combination. No system achieves both high coverage and high quality on non-parallel bands.

visible. DocAlign and DocAlign-NF cluster in the high-score, low-coverage region at Bands 2–3, while Vecalign and the baseline occupy the high-coverage, lower-score region across all bands. No system is simultaneously high on both axes at non-parallel bands, which is the central quantitative finding of this paper.

CWQ captures this tradeoff in a single number. On Band 1, Vecalign leads (0.608) because it achieves high coverage without much score loss. DocAlign collapses by Band 3 (CWQ 0.039) due to near-zero coverage (4.7%), while Vecalign degrades more gracefully (CWQ 0.229, coverage 71.4%). The baseline also outperforms DocAlign on CWQ at Bands 2 and 3, despite lower pair-level quality, purely through coverage.

4.4 Statistical Significance

We applied pairwise Mann-Whitney U tests between all four systems within each band (six comparisons per band, 18 total), with Bonferroni correction; corrected p-values below 0.05 are considered significant. Full results are reported in Appendix D.

Across all three bands, hierarchical systems (DocAlign, DocAlign-NF) significantly outperform flat systems (Vecalign, Baseline), with effect sizes growing as parallelism decreases: moderate in Band 1 ($|r| = 0.29$–0.37 for significant contrasts only), large in Band 2 ($|r| = 0.55$–0.66), and very large in Band 3 ($|r| = 0.60$–0.80). Within-tier contrasts are consistently non-significant: DocAlign vs. DocAlign-NF does not reach significance at any band (Band 1 $p_{\text{corr}} = 1.0$; Band 2 $p_{\text{corr}} = 0.893$; Band 3 $p_{\text{corr}} = 0.327$), and Vecalign vs. Baseline is non-significant at all three bands ($p_{\text{corr}} = 1.0$).

The Vecalign–baseline non-result is substantively meaningful and warrants a mechanistic explanation. Vecalign's dynamic programming alignment and support for non-1:1 spans are designed to recover from sequencing errors and structural drift in imperfect parallel texts, and on parallel data these properties provide a modest advantage over greedy matching. On comparable data, however, the binding constraint is not alignment sequencing but embedding discrimination: LaBSE similarity is high between any two sentences drawn from documents on the same topic, regardless of whether they are translations, so the alignment algorithm is operating on noise and its sophistication provides no additional purchase. Improvements to flat alignment quality will therefore require either better discrimination at the embedding level or structural constraints that limit the candidate set before scoring.

4.5 Post-hoc Filtering Cannot Close the Quality Gap

Vecalign has no built-in output quality threshold: its tuneable parameters (deletion penalty, search width, alignment size limit) control the alignment algorithm, not output filtering. To test whether post-hoc score filtering could recover quality, we filtered Vecalign's existing output at LaBSE cosine similarity thresholds of 0.4, 0.5, 0.6, and 0.7, dropping all pairs below the cutoff. Filtering consistently improves mean score but destroys coverage: at T=0.5, Band 2 mean score rises from 0.45 to 0.72 but coverage falls from 85% to 25%; Band 3 coverage collapses to 6% (Table 6). Critically, filtering *reverses* Vecalign's CWQ advantage over the greedy baseline: at T=0.5, filtered Vecalign CWQ on Band 2 (0.183) already falls below unfiltered Baseline (0.328), and the gap widens at stricter thresholds. The conclusion is that the quality gap between flat and hierarchical systems on non-parallel data cannot be closed by post-hoc filtering: the underlying problem is embedding discrimination, not threshold setting.

4.6 Inter-Annotator Agreement

All 114 IAA items were rated independently by all three annotators. Table 7 reports Cohen's κ (unweighted and linear-weighted) and agreement rates for each pair.

Band	T	Pairs	Cov. (%)	CWQ
1	None	9,324	88.6	0.643
	0.5	6,568	67.1	0.569
	0.7	5,360	54.4	0.495
2	None	10,074	84.7	0.377
	0.5	2,317	25.4	0.183
	0.7	1,111	13.4	0.117
3	None	6,541	72.2	0.229
	0.5	683	6.2	0.041
	0.7	237	2.5	0.022

Table 6: Post-hoc quality filtering of Vecalign output at LaBSE cosine similarity threshold T. Filtering improves mean pair score but reduces coverage faster, collapsing CWQ below unfiltered Baseline at all non-trivial thresholds(T=0.5 and T=0.7 shown)

Pair	κ	κ_w	Exact	±1
A–B	0.688	0.821	78.1%	93.9%
A–C	0.818	0.907	86.8%	97.4%
B–C	0.729	0.875	80.7%	97.4%

Table 7: Inter-annotator agreement on 114 shared items. κ = Cohen's kappa (unweighted); κ_w = linear-weighted kappa. ±1 = proportion differing by at most one point.

Unweighted κ ranges from 0.688 to 0.818, corresponding to moderate to almost perfect agreement (Landis and Koch, 1977). Weighted κ is higher in all cases (0.821–0.907), indicating that disagreements tend to be by one point rather than extreme. Agreement is lowest for the A–B pair, which is also the pair with the least shared calibration time, and highest for A–C. The near-perfect within-1 agreement (97.4% for both pairs involving Annotator C) is particularly important for a 5-point scale: systematic one-point disagreements do not reverse the ordering of conditions. The IAA results support treating the annotation as reliable for cross-system comparison.

5 Discussion

5.1 Failure Mode Taxonomy

To characterise why quality degrades, we coded all low-rated annotation items (mean score ≤ 2.5, $n = 550$) into failure types by reading the source–target pairs directly.[3] Failure coding was performed by the primary annotator (Annotator C); see Limitations for discussion. Two primary types account for virtually all failures across all systems and bands.

[3]One item rated 1 appeared to be a correct alignment and was excluded from counts.

Type T: Topical mismatch. Both sides are well-formed sentences that share a topic or entity, but they are not translations of each other. The alignment system has matched sentences on the basis of embedding-level semantic similarity that does not correspond to translational equivalence. This is the dominant failure mode, accounting for 78% of coded Band 3 failures in flat systems and 89% in hierarchical systems.

The mechanism differs by architecture. In flat systems, topical mismatch arises because LaBSE similarity is high between any two sentences drawn from documents about the same subject, and without structural constraints the aligner has no way to distinguish topic overlap from translation equivalence. In hierarchical systems, topical mismatch arises in a narrower context: the DTW stage has identified a paragraph group with sufficient similarity to survive the pre-filter, but the sentence pairs within that group are only thematically related, not translationally equivalent. The pre-filter thus acts as a coarse topic gate rather than a translation gate, and Type T failures are what pass through it on comparable data.

Type S: Structural noise. At least one side of the aligned pair is a navigation element, date, URL, citation fragment, list item, or boilerplate string. These account for 22% of Band 3 failures in flat systems and 11% in hierarchical systems. The lower rate in hierarchical systems is consistent with the pre-filter discarding structurally weak paragraph groups before they reach the sentence aligner. Structural noise is most prevalent in Band 1 failures of flat systems (32% of Vecalign Band 1 failures), reflecting that even on parallel documents, web-crawled text contains substantial non-content material that flat aligners will match if it appears at compatible positions.

The T/S split is consistent with the architectural interpretation developed above. On Band 3, DocAlign produces almost exclusively Type T failures (33 of 35 coded items), with near-zero structural noise, because the pre-filter eliminates low-quality paragraph groups before sentence-level matching. Vecalign, processing the full document without structural constraints, produces a higher proportion of Type S failures (25 of 91), alongside a large Type T majority. The practical implication is that no architecture resolves the topical mismatch problem on comparable data: both hierarchical and flat systems are susceptible, but for

different structural reasons.

Illustrative examples of each failure type are provided in Appendix B.

5.2 Practical Implications

The results carry several concrete recommendations for corpus builders applying sentence alignment to web-crawled resources.

Use CWQ, not coverage or score in isolation. Table 4 shows that pair-level quality scores favour selective systems while raw coverage favours permissive ones; neither alone characterises a system's utility for corpus construction. On Band 1, Vecalign leads on CWQ (0.608) because it achieves high coverage with modest quality loss. By Band 3, DocAlign's CWQ collapses to 0.039—lower than both flat systems—despite producing the highest pair-level quality among documents where it does produce output. A practitioner relying on pair-level quality alone would systematically overestimate DocAlign's usefulness on non-parallel data.

Prefer DocAlign-NF over DocAlign on comparable data. The pre-filter reduces output volume without improving pair quality on comparable data (Section 4.2). Note that CWQ slightly favours DocAlign (0.039 vs. 0.032); human annotation reverses this ranking (Section 4.2). Where coverage is already low, the priority should be maximising the quality of whatever output is produced.

Flat systems are unreliable below Band 2. Both Vecalign and the greedy baseline produce output that is over 90% rated poor or wrong on Band 3. Critically, Vecalign is statistically indistinguishable from the greedy baseline at all three bands, despite its more sophisticated alignment algorithm. Applying Vecalign to comparable web content is no better, in terms of pair quality, than the simplest possible similarity-threshold matcher.

5.3 Limitations

This study has several limitations that bound the generalisability of its conclusions. We evaluate a single language pair (Catalan–English); whether the T/S failure distribution and the relative degradation rates generalise to other pairs, particularly those with less parallel web presence, remains an open question. Band 3 is the smallest stratum (100 document pairs) and the most internally heterogeneous; the failure mode counts should be interpreted as characterising the Band 3 population in our sample, not as precise prevalence estimates. Sub-band variation within Band 3—between, for example, independently authored news coverage and topically adjacent content with no direct correspondence—is not characterised here and may account for some of the variance in system performance at this level. The ablation is valid only for Band 3, limiting conclusions about the pre-filter's role on moderate data. Failure mode coding was performed by a single annotator; an independent reliability check on a sample of coded items would strengthen the taxonomy.

6 Conclusion

We have presented a stratified empirical study of alignment quality degradation across the parallel–comparable spectrum for Catalan–English, using 300 document pairs with human evaluation. Three findings stand out. First, hierarchical and flat systems degrade at markedly different rates: on Band 3 comparable data, DocAlign-NF maintains a 51% usable-pair rate and DocAlign 25%, while Vecalign and the greedy baseline collapse to 2% and 7% respectively. Failure mode analysis traces this to topical mismatch as the dominant failure across all systems, with structural noise concentrated in flat systems. Second, paragraph pre-filtering on comparable data prevents document-level failures but lowers pair quality relative to the unfiltered ablation. Third, Vecalign is statistically indistinguishable from a greedy LaBSE baseline at all bands, and post-hoc quality filtering of its output cannot close the gap: filtering improves pair-level scores but collapses coverage-weighted quality below even the unfiltered baseline, confirming that embedding discrimination rather than alignment algorithm sophistication or threshold setting is the binding constraint on flat alignment quality.

Acknowledgements

This work/research has been promoted and financed by the Government of Catalonia through the Aina project. This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Re- siliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

References

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July. Association for Computational Linguistics.

Barzilay, Regina and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

El-Kishky, Ahmed, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. Ccaligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In Muresan, Smaranda, 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, May. Association for Computational Linguistics.

Gale, William A. and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 177–184, Berkeley, California, USA, June. Association for Computational Linguistics.

Guo, Mandy, Yinfei Yang, Keith Stevens, Daniel Cer, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Hierarchical document encoder for parallel corpus mining. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 64–72.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Landis, J Richard and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

O'Brien, Dayyán, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. DocHPLT: A massively multilingual document-level translation dataset. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 286–300, Suzhou, China, November. Association for Computational Linguistics.

Resnik, Philip and Noah A Smith. 2003. The web as a parallel corpus. *American Journal of Computational Linguistics*, 29(3):349–380.

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500.

Sennrich, Rico 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.

Sloto, Steve, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the wmt 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 95–102.

Steingrímsson, Steinþór, Hrafn Loftsson, and Andy Way. 2023. Sentalign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263.

Thompson, Brian and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In Inui, Kentaro, 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, November. Association for Computational Linguistics.

Thompson, Brian and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online, November. Association for Computational Linguistics.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Varga, Dániel, Péter Halácsy, Andras Kornai, Viktor Nagy, László Németh, and Viktor Trón, 2007. *Parallel corpora for medium density languages*, pages 247–258. 01.

A Annotation Guidelines

Each item presents a source text in Catalan (one or two sentences) and a target text in English (one or two sentences), together with the alignment type (1:1, 1:2, or 2:1). Alignments were produced automatically; system identity was blinded. Annotators judged whether the two texts convey the same content. The task assesses alignment quality, not translation quality.

For multi-sentence spans (1:2 and 2:1), annotators evaluated the full span as a unit rather than rating individual sentences.

Rating scale.

5 Perfect. The Catalan and English say essentially the same thing.

4 Good. Same core content, with minor additions, omissions, or paraphrase that do not affect meaning. Dates or some proper nouns may differ.

3 Acceptable. Clearly about the same topic or event, but with notable differences in specific content. May have some aligned text but large parts unaligned.

2 Poor. Partially related — same general domain or topic, but different specific information.

1 Wrong. Unrelated content.

Scores of 4–5 were considered *usable* for MT training; scores of 1–2 were considered *poor* or incorrect.

Borderline guidance. For the 3 vs. 2 boundary, annotators were instructed to consider whether any of the content is genuinely aligned; if so, 3 is appropriate. Two sentences covering the same news event from different angles, or mentioning the same entity in different contexts, are not the same content and should be rated 1 or 2. When genuinely uncertain between two adjacent scores, annotators were told to use the lower score and flag the item for review.

What to ignore. Differences in register, formality, or style; differences in sentence order within a multi-sentence span; minor factual elaborations consistent with the core content; and whether the translation is fluent or grammatically correct.

Calibration. A 15-item calibration session was conducted with all three annotators before independent annotation began. Annotators were instructed to log brief notes on items rated 1 or 2 to support subsequent failure mode analysis.

B Failure Mode Examples

Type	System	Source	Target
T	Vecalign	La resposta a l'usuari es representa en el monitor. Es una forma de treball interactiva…	Microsoft Windows users will compare the Unix shell to DOS.
T	DocAlign	It crosses the 3 regions of Belgium: its main part (51.7 km) is situated in Flanders…	Brussel·les-Halle-Vilvoorde… és una circumscripció electoral de Bèlgica…
S	Vecalign	See also[edit] References[edit]	Referències[modifica]
S	Baseline	`http://www.themidniteblues.blogspot.com/`	`http://diaridunapoliticaestudiant.blogspot.com/`

C Alignment Type Distribution

System	Band	1:1	1:2	2:1
DocAlign	1	94.1	3.7	2.2
	2	73.7	18.3	8.0
	3	71.4	16.1	12.4
DocAlign-NF	1	94.1	3.8	2.2
	2	73.7	18.2	8.0
	3	83.6	9.6	6.8
Vecalign	1	90.2	3.6	1.8
	2	85.0	5.0	2.6
	3	87.3	1.8	3.4
Baseline	1–3	100	0	0

Table 8: Alignment type distribution (%) per system and band.

D Pairwise Statistical Tests

Band	Comparison	n_1	n_2	r	p_{corr}
Band 1	DocAlign vs. DocAlign-NF	70	71	$+0.066$	1.000
	DocAlign vs. Vecalign	70	72	-0.236	0.129
	DocAlign vs. Baseline	70	81	-0.316	$\mathbf{0.005}$
	DocAlign-NF vs. Vecalign	71	72	-0.290	$\mathbf{0.013}$
	DocAlign-NF vs. Baseline	71	81	-0.372	$\mathbf{< 0.001}$
	Vecalign vs. Baseline	72	81	-0.050	1.000
Band 2	DocAlign vs. DocAlign-NF	76	74	$+0.178$	0.893
	DocAlign vs. Vecalign	76	82	-0.554	$\mathbf{< 0.001}$
	DocAlign vs. Baseline	76	83	-0.549	$\mathbf{< 0.001}$
	DocAlign-NF vs. Vecalign	74	82	-0.659	$\mathbf{< 0.001}$
	DocAlign-NF vs. Baseline	74	83	-0.640	$\mathbf{< 0.001}$
	Vecalign vs. Baseline	82	83	-0.018	1.000
Band 3	DocAlign vs. DocAlign-NF	55	45	$+0.268$	0.327
	DocAlign vs. Vecalign	55	81	-0.652	$\mathbf{< 0.001}$
	DocAlign vs. Baseline	55	85	-0.599	$\mathbf{< 0.001}$
	DocAlign-NF vs. Vecalign	45	81	-0.805	$\mathbf{< 0.001}$
	DocAlign-NF vs. Baseline	45	85	-0.758	$\mathbf{< 0.001}$
	Vecalign vs. Baseline	81	85	$+0.077$	1.000

Table 9: Pairwise Mann-Whitney U tests (Bonferroni-corrected, 18 comparisons). r is rank-biserial effect size; negative values indicate System 1 > System 2. Significant results ($p_{\text{corr}} < 0.05$) are marked **bold**.