Analyzing register variation in web texts through automatic segmentation

Erik Henriksson, Saara Hellström, Veronika Laippala
TurkuNLP, University of Turku
{erik.henriksson, sherik, mavela}@utu.fi

Abstract

This study introduces a novel method for analyzing register variation in web texts through classification-based register segmentation. While traditional text-linguistic register analysis treats web documents as single units, we present a recursive binary segmentation approach that automatically identifies register shifts within web documents without labeled segment data, using a ModernBERT classifier fine-tuned on full web documents. Manual evaluation shows our approach to be reliable, and our experimental results reveal that register segmentation leads to more accurate register classification, helps models learn more distinct register categories, and produces text units with more consistent linguistic characteristics. The approach offers new insights into document-internal register variation in online discourse.

1 Introduction

Text-linguistic analysis of registers—text varieties with shared situational characteristics and functionally related linguistic features—has greatly advanced our understanding of language variation in different situations and domains (Biber, 1988; Biber and Conrad, 2009; Biber and Egbert, 2023). In the domain of online discourse, recent advances in NLP techniques such as Transformer models (Vaswani et al., 2017; Devlin et al., 2019) have enabled automatic classification of web texts into registers across various languages with near-human level performance (Henriksson et al., 2024b). These automatic web register classifiers now serve valuable roles in many research areas, from large-scale linguistic analyses of online discourse (Myntti et al., 2024) to the curation of web-crawled datasets for Large Language Model (LLM) training (Burchell et al., 2025).

Despite recent progress in web register classification schemes (Egbert et al., 2015; Madjarov et al., 2019; Laippala et al., 2022; Kuzman and Ljubešić, 2023), web registers remain relatively fuzzy categories with substantial internal variation (Biber et al., 2020; Henriksson et al., 2024a). As Egbert and Gracheva (2023) have recently suggested, at least part of this unexplained variance may stem from the definition of *text*, the fundamental unit of observation. Critically, in all previous studies on web registers, this unit has always been defined as the full document. However, web documents are often too diverse in content to fit neatly into a single register category. For example, news texts (belonging to the *Narrative* register) are frequently followed by comments (*Interactive Discussion* register) (Biber and Egbert, 2018, p.39); similarly, narrative blogs often contain family recipes (*Instructional* register) (Biber and Egbert, 2018, p.158). Registers can also appear blended, as in sports reports that incorporate detailed sports data, combining elements of the *Narrative* and *Informational* registers (Biber et al., 2020, p.32).

In this article, we investigate whether an automatic register classifier, trained on full web documents, can be used to detect register shifts within documents, and assess whether segmenting documents based on these shifts produces more distinct web register categories. Specifically, we fine-tune a ModernBERT (Warner et al., 2024) register classifier and develop a segmentation algorithm that leverages the predicted probabilities from the classifier to detect document-internal register units. Using recursive binary splitting, our algorithm analyzes potential boundary points within web documents and selects segmentations with maximally distinct register predictions. We evaluate this method on the English Corpus of Online Registers (CORE) (Egbert et al., 2015; Laippala et al., 2022), which includes eight main register classes. As a preliminary step, we use Cleanlab (Northcutt et al., 2021) to remove noisy and ambiguous labels from the data, aiming for an enhanced model suitable for segmentation.

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 7–19
May 3-4, 2025

To evaluate our register segmentation approach, we assess it manually and compare segment-based and document-based analyses through classification performance, clustering, and linguistic feature analysis. Our results show that segment-based analysis produces more consistent register units. Additionally, we examine register distributions within documents, revealing patterns of register shifts in online discourse. The code and data used in this study are available at https://github.com/TurkuNLP/CORE-segmentation.

2 Background

Text segmentation is the task of dividing texts into coherent, non-overlapping units such as paragraphs or topics (Hearst, 1994). It has applications in discourse analysis, summarization, and information retrieval, among others (e.g. Hearst and Plaunt, 1993; Galley et al., 2003; Liu et al., 2021).

Existing approaches to text segmentation fall into two main categories: unsupervised and supervised. Unsupervised methods measure coherence between segments using features such as term co-occurrences (Hearst, 1997), topic vector shifts (Riedl and Biemann, 2012), or semantic embedding similarities (Solbiati et al., 2021; Yu et al., 2023). Supervised approaches learn segmentation from labeled data (e.g. Koshorek et al., 2018; Badjatiya et al., 2018; Xing et al., 2020; Glavaš and Somasundaran, 2020; Lukasik et al., 2020; Lo et al., 2021; Nair et al., 2023). Fine-tuned Transformer models (Vaswani et al., 2023) generally achieve higher accuracy than unsupervised methods (Inan et al., 2022), although unsupervised approaches can still perform well in contexts where labeled data is scarce or not available (Solbiati et al., 2021).

Register-labeled web datasets (e.g. Laippala et al., 2022; Henriksson et al., 2024a) are annotated at the document level, with no finer-grained register datasets available. While these often include *hybrid* texts—documents annotated with multiple register labels—they do not specify whether these labels correspond to separate sections or mixed content (see Section 1). This means we cannot directly use hybrid documents to inform segmentation models. Moreover, in contrast to structured platforms like Wikipedia, where documents have clear structural markers indicating content shifts (Koshorek et al., 2018; Arnold et al., 2019), web texts in general lack explicit register indicators in their HTML structure, complicating automatic boundary detection.

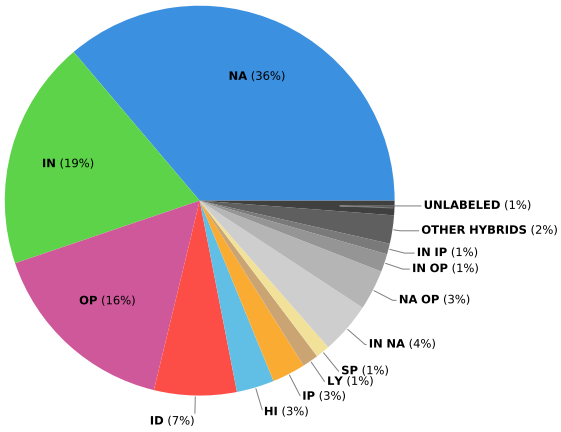

Figure 1: Register distribution in the CORE dataset after filtering out texts exceeding 8,192 tokens (N = 47,319).

Our approach to register segmentation combines elements of both supervised and unsupervised methods: we first fine-tune an encoder model on full documents, then use the fine-tuned model in an unsupervised manner to segment texts. Additionally, our algorithm employs recursive segmentation, which repeatedly divides text into smaller parts. This recursive approach creates a tree-like structure of segments and sub-segments, making it more similar to hierarchical segmentation approaches (e.g. Bayomi and Lawless, 2018; Hazem et al., 2020) than to linear segmentation methods (e.g. Hearst, 1997; Yu et al., 2023) which simply divide text into a flat sequence of adjacent segments.

3 Data

We use data from the English CORE corpus (Laippala et al., 2022), a manually register-annotated collection of unrestricted English web content comprising 48,435 documents. The corpus was collected via Google searches based on frequent English 3-grams (Egbert et al., 2015) and annotated through Amazon Mechanical Turk. Each document was labeled by four coders, with a register assigned if at least two chose the same label. In cases of an even split between two registers, both labels were assigned. When all four annotators selected different labels, no label was assigned.

The CORE scheme (Biber and Egbert, 2018) defines eight main register categories and 47 subcategories. In this study, we focus on the main classes: *How-to/Instructional* (HI), *Informational Description* (IN), *Informational Persuasion* (IP), *Interactive Discussion* (ID), *Lyrical* (LY), *Narrative* (NA), *Opinion* (OP), and *Spoken* (SP).

Due to our model's token limit of 8,192 (see Section 4.1) and our goal to segment entire documents, we exclude documents exceeding this limit, removing 1,116 documents (2.30%) from the dataset. Figure 1 shows the register distribution within the remaining documents: *Narrative* (36%), *Informational* (19%), and *Opinion* (16%) are the most common categories, with hybrid cases being mostly different combinations of these three registers.

4 Web register segmentation model

Our approach to web register segmentation consists of two stages: (1) fine-tuning a supervised register classifier on labeled CORE data and (2) recursively splitting documents into binary segments, using the classifier's output to find optimal bounds.

4.1 A ModernBERT register classifier

We begin by fine-tuning a ModernBERT (Warner et al., 2024) model for register classification using labeled CORE data (see Section 3). We choose ModernBERT for its extended 8,192-token limit, which enables segmentation of long documents—unlike previous encoders with a 512-token limit (e.g. Devlin et al., 2019; Liu et al., 2019)—and for its performance improvements.

We split the CORE dataset into training (70%), development (10%), and test (20%) sets and fine-tune the model using a multi-label classification approach with the HuggingFace Transformers library (Wolf et al., 2020). To address label imbalance, we use focal loss (Lin et al., 2017) with α=0.5 and γ=1.0. The model is trained for up to five epochs with early stopping based on the micro-F1 score on the development set, using a learning rate of 3e-5.

The model achieves a micro-F1 score of 0.76 and a macro-F1 score of 0.73, closely matching previous results on this dataset (Henriksson et al., 2024b). While these scores are reasonable given the well-known complexities of web register classification (Biber and Egbert, 2018; Laippala et al., 2022), our manual inspection suggests that some errors stem from noisy labels, including annotation mistakes, ambiguous cases, and hard-to-classify texts. Since our sequential segmentation approach could propagate classification errors, we attempt to improve the model by cleaning the dataset.

We use Cleanlab (Northcutt et al., 2021) to remove noisy labels from CORE. This algorithm has been shown effective for dataset cleaning across tasks (Goh et al., 2022; Thyagarajan et al., 2023; Chen and Mueller, 2024) and provides theoretical guarantees for label noise estimation. It uses predicted probabilities from a trained classifier on the test set; to obtain these for the full dataset, we perform 10-fold cross-validation (Kohavi, 1995) with iterative stratification (Sechidis et al., 2011; Szymański and Kajdanowicz, 2017), fine-tuning each model using the same settings as in Section 4.1.

Register	CORE	Cleaned	Diff (%)
Single Registers			
Narrative (NA)	17,125	15,308	-10.6
Informational Description (IN)	8,997	7,392	-17.8
Opinion (OP)	7,579	6,301	-16.9
Interactive Discussion (ID)	3,237	2,923	-9.7
How-to/Instructional (HI)	1,477	1,130	-23.5
Informational Persuasion (IP)	1,308	851	-34.9
Lyrical (LY)	635	598	-5.8
Spoken (SP)	555	482	-13.2
Hybrid Registers			
IN NA	2,027	1,184	-41.6
NA OP	1,577	868	-44.9
IN OP	703	329	-53.2
IN IP	420	318	-24.3
Other hybrids	1,109	764	-31.1
Unlabeled	570	0	-100.0

Table 1: Comparison of register distributions in the full CORE dataset and the cleaned version.

The Cleanlab process identifies 8,301 texts with potential label issues (see Appendix A for examples). Table 1 compares the full CORE dataset to the cleaned version, showing distributions for single-register texts and the most frequent hybrids. The cleaned dataset shows a significant drop in hybrid categories (by 24–53%) and eliminates all unlabeled texts, while preserving roughly the same distribution of the main single-register categories. This suggests that the cleaning process targets both noisy labels and inherently ambiguous texts—specifically, unlabeled documents (where no annotators agreed) and hybrids (where only half agreed; see Section 3). Removing these difficult-to-classify texts aligns with our goal of improving segmentation, as our model can be expected to better identify register shifts when trained on examples with clear register signals.

We fine-tune ModernBERT on the cleaned dataset, with results compared to the original model in Table 2. The cleaned model shows performance gains across all registers, with the most substantial improvements in previously underperforming categories: *Opinion* improves by 14 percentage points (0.68 to 0.82), *Informational Persuasion* also by 14

Register	All	Clean
How-to/Instructional (HI)	0.67	0.78
Interactive Discussion (ID)	0.85	0.91
Informational Description (IN)	0.71	0.84
Informational Persuasion (IP)	0.50	0.64
Lyrical (LY)	0.89	0.93
Narrative (NA)	0.84	0.91
Opinion (OP)	0.68	0.82
Spoken (SP)	0.71	0.80
Micro Average	0.76	**0.86**
Macro Average	0.73	**0.83**

Table 2: Comparison of F1 scores between the original and cleaned models.

points (0.50 to 0.64), and *Spoken* by 9 points (0.71 to 0.80). The increases in both micro-F1 (0.76 to 0.86) and macro-F1 (0.73 to 0.83) indicate that the cleaned model improves performance across the board; given these improvements, we integrate this model into our segmentation algorithm.

4.2 Recursive binary splitting segmentation

Our segmentation algorithm recursively partitions documents into register segments based on sentence boundaries and classifier predictions. It evaluates potential split points by comparing the register predictions of candidate segments. The process is illustrated in Figure 2.

The input document is first segmented into sentences using spaCy's sentence segmenter (Honnibal et al., 2020), with sentence boundaries serving as potential split points. For each split point, we assess register distinctness between the left and right segments using three window sizes: (1) full segments, comparing the entire left and right parts; (2) short, two-sentence windows on each side of the boundary; and (3) longer, five-sentence windows.

The optimal segmentation is determined using two metrics. First, we assess whether segmentation is necessary by checking if the predicted registers of the left and right segments differ and are not both identical to the parent text's registers. This decision is based on the classifier's threshold for positive predictions (0.70), optimized using micro-F1 scores on full documents during fine-tuning.

For qualifying split points, we then evaluate their quality by measuring differences between the classifier's predicted probabilities across the three scopes (full segments and the two- and five-sentence windows around the boundary). These differences are computed using cosine distance. To discourage oversegmentation, each cosine distance

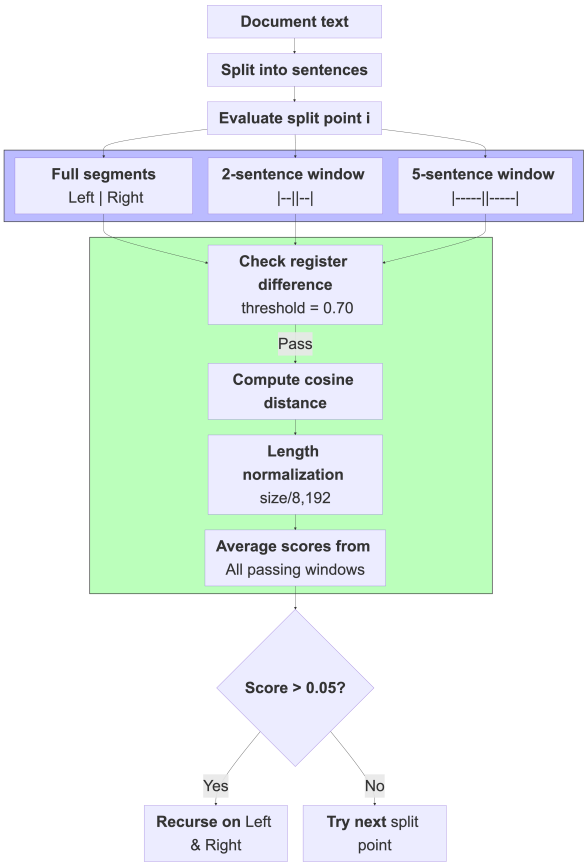

Figure 2: The recursive segmentation process.

is normalized by the ratio of the smaller segment's (left or right) token length to the model's maximum token limit (8,192). The final segmentation score for each split point is the average of these three normalized cosine distances. We select the split point with the highest score that exceeds our threshold (0.05). The process continues recursively on the resulting segments until no valid splits remain or we reach our recursion depth limit (4).

The selection of these parameters was guided by qualitative analysis during development. The two window sizes (2 and 5 sentences) complement the full-segment comparison by providing more precise boundary detection—using only full segments often missed local register transitions. The segmentation threshold (0.05) was calibrated to balance between oversegmentation and missed transitions. The recursion depth limit of 4 was set after observing that deeper recursion rarely produced meaningful additional segments while increasing computational cost.

4.3 Assigning segment labels

The segmentation algorithm maintains register predictions across all recursive levels, from the full

	A1	A2	κ
Labels	4.21 ± 0.82	4.13 ± 0.91	0.56
Segments	4.13 ± 0.81	4.29 ± 0.84	0.67

Table 3: Evaluation results for 75 randomly sampled segmentations. Scores range from 1 (incorrect) to 5 (correct/nearly correct).

document down to the smallest segments. This allows us to integrate register information from different granularities when labeling segments.

Each segment is labeled using the final recursion level for maximum specificity. However, we observe that certain registers function as broader *container* categories that frame the overall communicative context. In particular, *Interactive Discussion* (ID) and *Spoken* (SP) serve this role since they are defined primarily by their mode of communication rather than content—a forum post may contain narratives or opinions while remaining fundamentally interactive, and spoken text can similarly incorporate various sub-registers. To reflect this hierarchical relationship, whenever ID or SP appear as positive classes in the recursive hierarchy, we propagate them to the final label.

5 Evaluation and results

In this section, we evaluate our segmentation approach and present the results. We begin with a manual evaluation of a sample of segmented CORE documents, followed by descriptive statistics of the segmented corpus. Next, we assess the produced register segments by comparing them to full-document registers in terms of classification distinctiveness, embedding-space separation, and linguistic cohesion. Finally, we explore document-internal register structures using the segmentations.

5.1 Manual evaluation

To assess segmentation quality, we manually evaluate a random sample of 75 documents, including 55 documents with at least two segments and 20 documents that remained unsegmented. We assess segmentation and labels separately using a 5-point scale, from 1 (incorrect) to 5 (perfect/nearly perfect). Two annotators, both experts in web register research and the CORE scheme, conduct the evaluation. Inter-annotator agreement (IAA) is measured using Cohen's κ with quadratic weights.

Table 3 presents the evaluation results, including mean scores for segment boundaries and labels, along with IAA. The evaluation shows moderate to substantial agreement between annotators, with κ = 0.56 for labels and 0.67 for boundaries. The higher agreement on boundaries suggests that identifying web register segments is more objective than assigning register labels.

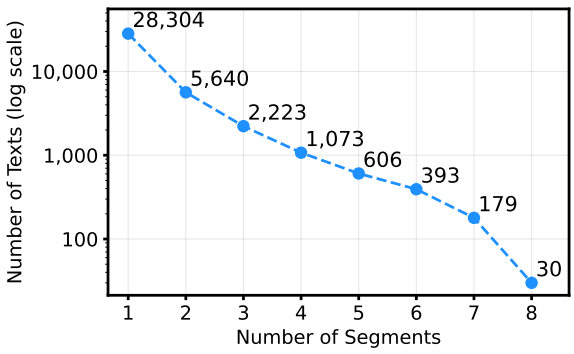

Figure 3: Distribution of segment counts across texts.

Both annotators gave high scores for segmentation quality. For register labels, annotator scores averaged 4.21 and 4.13, with most texts (83% and 76% respectively) receiving scores of 4 or 5. Segment boundaries received similarly high ratings, with means of 4.13 and 4.29, and a large majority of texts (83% and 77%) scored 4 or 5. The small standard deviations (0.81-0.91) and consistent distribution of scores indicate reliable performance across different types of web documents.

For the 20 documents that remained unsegmented by the model, evaluation scores were higher (labels: 4.29/4.38; segments: 4.57/4.71) with strong inter-annotator agreement (κ = 0.87 for labels, 0.83 for segments). This indicates the model rarely misses necessary segmentation points, accurately identifying documents that genuinely represent a single register.

5.2 Descriptive statistics and an example

Figure 3 shows the distribution of segment counts across the dataset. Most texts (28,304 or 73.6%) remain unsegmented, and the number of texts decreases exponentially with segment count. On average, each text contains 1.49 segments.

Figure 4 compares register distributions in document-level vs. segment-level data, with lighter bars representing segments. The top panel shows distributions for single-register texts, and the bottom shows hybrids with at least a 0.1 percentage point difference between the two datasets.

The register distribution shows *Narrative* (NA) as dominant but decreasing from 39.8% to 33.0%

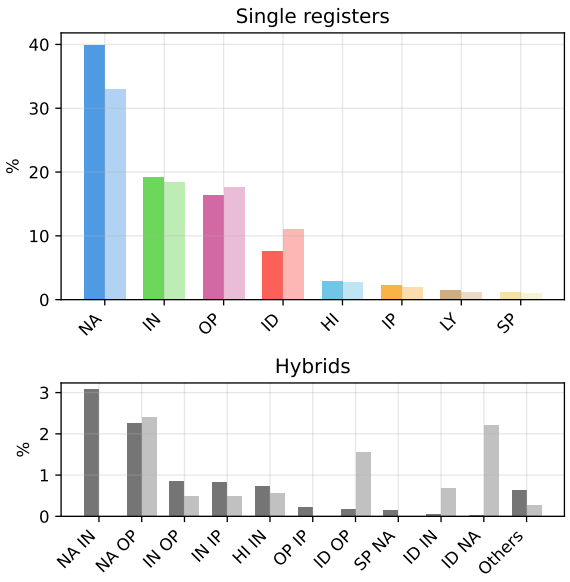

Figure 4: Register distributions in document-level vs. segment-level data. Lighter bars show segmented data.

Segment 1: *Narrative* (NA)

The return of the Taco Dog The time had come to revisit and old classic, in fact my first ever drunch dish...the taco dog. Now regular readers of the drunch blog may be aware of it but for the new little drunchlings out there allow me to tell you of its history. [...*narrative continues*...] However before I went all Dr Drunchenstien on the Taco Dog it occurred to me that one of the drunchards hadn't tried the original and there was no point in exposing him to potentially lethal levels of tasteyness without letting him limber up first.

Segment 2: *How-to/Instructional* (HI)

Now the taco dog is very simple to make but this time I made my own seasoning. All it requires is: Hot dogs (bratwurst kind, none of your piddly wee ones, they insult the gods of taco dogs & will curse you to 7 years and 3 months of odd socks). Mince. Taco seasoning. Cheese sauce made thickly with red peppers mixed in (or in a pinch Cheese nacho dip). Baguettes (hot dog buns are useless don't even waste your time). and nachos to use as cutlery. The preparation is mince as per instructed on package. add seasoning to mince. Cook hot dogs. [*recipe continues*...]

Figure 5: Register shift in a blog post, as segmented by our algorithm (manually annotated label: HI).

in the segmented corpus, while *Informational* (IN) and *Opinion* (OP) texts remain relatively stable around 19% and 17% respectively. The most notable change is in *Interactive Discussion* (ID), increasing from 7.6% to 11.0%. This increase occurs because our segmentation process identifies and separates discussion sections (such as comments) that were previously embedded within longer documents and labeled with the register of the main text (e.g. as part of a narrative blog). The remaining registers (HI, IP, LY, SP) each constitute less than 3% of either corpus with minimal variation.

In multi-label text units, the emergence of ID-NA (2.2%), ID-OP (1.6%), and ID-IN (0.7%) combinations in the segmented corpus results from our ID propagation approach (Section 4.3), where ID is retained in the final label if detected at any level of recursive segmentation. Overall, single-label units remain prevalent in both corpora, comprising 91.5% of document-based texts and 86.9% of segment-based texts.

Figure 5 illustrates a typical segmented document. This food blog post starts with a *Narrative* (NA) segment about discovering a "taco dog" at a takeaway place, then shifts to a *How-to/Instructional* (HI) segment providing a recipe. Our algorithm successfully detects this shift and partitions the document accordingly.

Register	Doc.	Seg.
How-to/Instructional (HI)	0.78	0.84
Interactive Discussion (ID)	0.91	0.87
Informational Description (IN)	0.84	0.89
Informational Persuasion (IP)	0.64	0.75
Lyrical (LY)	0.93	0.94
Narrative (NA)	0.91	0.93
Opinion (OP)	0.82	0.88
Spoken (SP)	0.80	0.76
Micro Average	0.86	**0.89**
Macro Average	0.83	**0.86**

Table 4: Comparison of F1 scores between a full-document based model vs. a segment-based model.

5.3 Segment-based register classification

We evaluate segment quality by comparing how well CORE registers can be learned from segments versus full documents. Intuitively, if fine-tuning a register classifier on segments improves performance over full documents, it suggests that segments provide a clearer register signal that the model can better differentiate.

We fine-tune a ModernBERT model on segmented data using the same configuration as the full-document classifier (Section 4.1). The segments are shuffled and stratified into 70% training, 20% test, and 10% development sets. We then compare the F1 scores of both models, using results from the cleaned full-document model (see Section 4.1) as a baseline.

Register	Doc.	Seg.	Δ
How-to/Instructional (HI)	0.712	0.773	+0.061
Interactive Discussion (ID)	0.666	0.774	+0.108
Informational Description (IN)	0.634	0.626	-0.008
Informational Persuasion (IP)	0.186	0.572	+0.386
Lyrical (LY)	0.856	0.856	0.000
Narrative (NA)	0.475	0.631	+0.156
Opinion (OP)	0.500	0.601	+0.101
Spoken (SP)	0.811	0.754	-0.057
Overall	0.541	0.650	+0.109

Table 5: Embedding silhouette scores by register: full documents vs. segments

As shown in Table 4, the segment-based model outperforms the document-based model, achieving a micro-F1 of 0.89 (vs. 0.86) and a macro-F1 of 0.86 (vs. 0.83). Several registers see notable improvements: *How-to/Instructional* (+0.06), *Informational Description/Explanation* (+0.05), *Informational Persuasion* (+0.11), and *Opinion* (+0.06). However, performance slightly decreases for *Interactive Discussion* (-0.04) and *Spoken* (-0.04)—precisely the registers propagated from the hierarchy when assigning final segment labels (see Section 4.3). This suggests that our propagation approach may need refinement in future work, though we do not explore it further here.

Overall, these results indicate that our segmentation method identifies more homogeneous register units than document-based analysis.

5.4 Evaluating register segment embeddings

To further evaluate whether our segmentation approach produces more distinct register units, we compare the embedding spaces of segments and full documents. Specifically, we compute register-averaged silhouette scores (Shahapure and Nicholas, 2020) to measure intra-register cohesion and inter-register separation. This analysis focuses on single-register texts, using embeddings from: (1) the full-document model (Section 4.1) and (2) the segment-trained model (Section 5.3). In both cases, we use *true* labels—human-annotated gold labels for document embeddings and segmentation-derived labels for segment embeddings.

Table 5 shows that segmentation consistently improves silhouette scores, with the largest gains for *Informational Persuasion* (IP) (+0.386) and *Narrative* (NA) (+0.156); overall improvement is +0.109.

To visualize how registers cluster in the two approaches, we reduce the 1024-dimensional embeddings to 2D using UMAP (McInnes et al., 2018). Figure 6 compares the full-document (top) and segment-based (bottom) embeddings, showing clearer register separation in the latter. Notably, *Narrative* and *Opinion*, which overlap in the document-based plot, are more distinct in the segment-based representation.

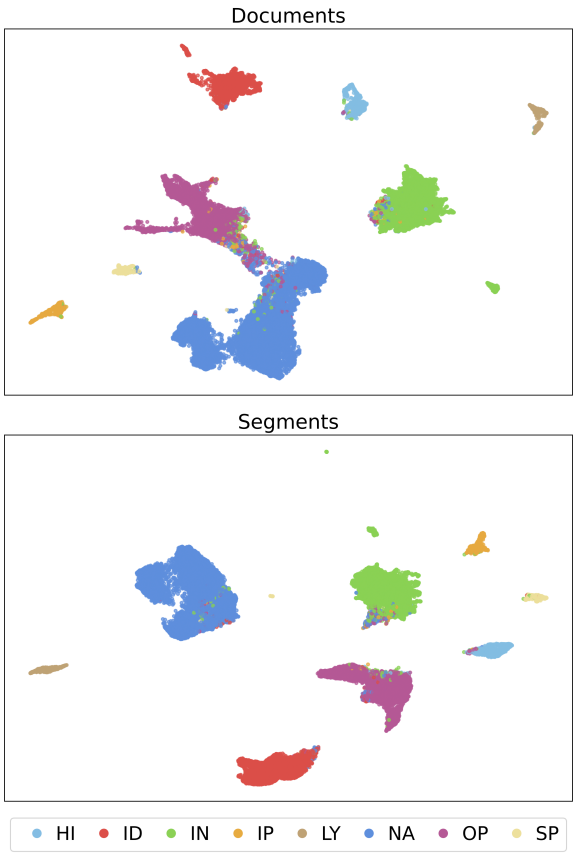

Figure 6: UMAP visualization of register embeddings: full documents (top) vs. segments (bottom).

5.5 Evaluating linguistic cohesion

We examine whether segmentation results in more clearly defined linguistic characteristics within registers compared to full texts. We process both segments and full documents using Trankit (Nguyen et al., 2021), chosen for its state-of-the-art performance on dependency parsing and morphological analysis.

We use Trankit's `posdep` function to extract three categories of linguistic features: (1) part-of-speech distributions (nouns, verbs, adjectives, etc.), (2) syntactic dependency relations (subject, object, modifiers), and (3) morphological features (number, tense, case). These surface-level features are established indicators of register variation (Biber, 1988; Biber and Egbert, 2018). For each text (full document or segment), we count the frequency of

	Variance		Pairwise dist.	
Register	Seg.	Doc.	Seg.	Doc.
How-to/Instructional (HI)	0.87	1.23	13.81	15.60
Interactive Discussion (ID)	1.07	1.46	14.76	16.22
Informational Description (IN)	0.76	0.87	12.94	13.40
Informational Persuasion (IP)	0.84	1.12	13.65	15.42
Lyrical (LY)	0.98	1.04	14.64	15.03
Narrative (NA)	0.93	1.58	13.97	15.23
Opinion (OP)	1.06	1.47	14.66	16.32
Spoken (SP)	1.11	1.42	15.50	17.38
Average	**0.95**	1.27	**14.24**	15.57

Table 6: Linguistic cohesion metrics by register in full documents vs. segments (lower is better).

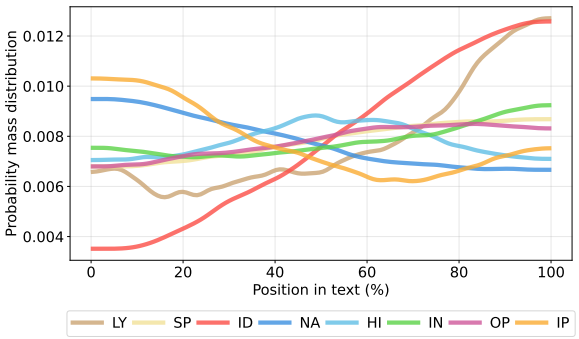

Figure 7: Register probability distributions across document positions.

each linguistic feature and then divide by the total token count in that text, yielding a normalized feature vector for each text.

To assess linguistic cohesion, we compute two metrics: (1) the average within-register variance of linguistic features and (2) the mean Euclidean distance between all text pairs within each register, serving as an intra-register similarity measure.

The results in Table 6 show that segments exhibit more defined linguistic characteristics than full texts. Register-internal variances are consistently lower for segments across all registers, averaging 0.95 compared to 1.27 for full texts. Similarly, pairwise distances indicate greater cohesion in segments, with an average distance of 14.24 versus 15.57 in full texts. The effect is most notable in *Spoken* (15.50 vs. 17.38) and *Opinion* (14.66 vs. 16.32) texts. Overall, these findings suggest that segmentation produces text units with more consistent linguistic patterns.

5.6 Analyzing document-internal register variation

We end with two brief analyses on document-internal register variation on the segmented CORE

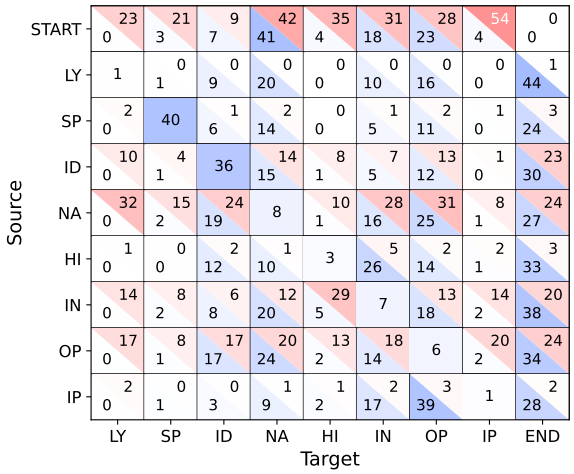

Figure 8: Register transitions between adjacent segments. Blue triangles represent row-to-column percentages and red ones to-column-from-row percentages.

data, to illuminate the benefits of segmentation.

First, we examine register distribution within documents. We divide each document into 128 equal-length bins and track character counts at each position, weighted by predicted register probabilities. As shown in Figure 7, this reveals clear document-internal patterns in register distribution. *Narrative* (NA) and *Informational Persuasion* (IP) peak early in documents. *How-to/Instructional* (HI) shows a noticeable increase in the middle, likely reflecting the typical placement of instructional content such as recipes and guides. Most strikingly, *Interactive Discussion* (ID) rises sharply toward the end, aligning with the common placement of comment sections in web documents. Similarly, *Lyrical* (LY) content increases noticeably toward the ends of documents.

Second, we analyze document-internal register transitions. Figure 8 presents a split-cell heatmap where cells show transitions from a source register (row) to a target register (column). Blue triangles show the percentage of transitions from the row register to the column register, while the red ones show the percentage of the column register following the row register. `START` and `END` indicate the beginnings and endings of documents, respectively.

Several clear patterns emerge from this analysis. *Narrative* (NA) typically opens documents (41% of beginnings, 42% of all NA segments), followed by *Opinion* (OP, 23%) and *Informational Description* (IN, 18%). Document endings favor different registers, with *Informational Description* (38%), *Opinion* (34%), and *Interactive Discussion* (ID,

30%) being most common.

For document-internal transitions, there is clear register mixing between certain categories: *Informational Persuasion* (IP) frequently transitions to *Opinion* (39%), with OP also often preceding IP (20%). Similar relationships exist between *How-to/Instructional* (HI) and *Informational Description*. *Interactive Discussion* (ID) and *Spoken* (SP) are commonly self-transitioning (36-40%), partly due to our labeling approach (see Section 4.3). *Narrative* segments commonly lead to *Opinion* (25%) or *Interactive Discussion* (19%), and these registers in turn most frequently follow *Narrative* (31% and 24% respectively), suggesting a strong pattern of narrative content followed by commentary.

6 Conclusion

This paper has introduced a new way to analyze register variation within web texts by segmenting documents rather than treating them as single units. We combined a ModernBERT classifier with a recursive binary segmentation algorithm that detects document-internal register shifts without requiring pre-labeled segment data.

Our results show that segmentation improves register analysis in several ways. Models trained on segments outperform those trained on full documents, with micro-F1 scores rising from 0.86 to 0.89 and macro-F1 from 0.83 to 0.86. Registers cluster more closely in embedding space when analyzed as segments, and they have more consistent linguistic characteristics.

By segmenting texts, we uncovered patterns that document-level analyses miss. Different registers tend to occur in specific positions within documents: *Narrative* and *Informational Persuasion* texts typically appear at the beginning, *How-to/Instructional* content is favored in the middle, and *Interactive Discussion* and *Lyrical* content usually appear at the end.

Our approach opens up new possibilities for studying online discourse. By examining texts at a more granular level than full documents, we get a more detailed view of how registers are used in web communication. This could benefit not only register studies but also applications like summarization systems and web corpus curation.

Limitations and future work

Although our segmentation approach demonstrably benefits register analysis, several limitations should be acknowledged. First, the segmentation parameters (recursion depth, cosine distance threshold, window sizes) were selected through qualitative analysis. Future research should systematically tune these parameters on manually segmented data.

Second, our method relies on sentence boundaries for potential segmentation points, which may not always align with actual register shifts. In web texts, non-textual elements like horizontal lines or headings often signal register transitions without corresponding sentence breaks. Future implementations should incorporate HTML structural elements and other visual markers, although these were not available in the CORE corpus used in this study.

Third, this study focused exclusively on English texts from the CORE corpus. Cross-linguistic validation, and testing on other web corpora such as HPLT 2.0 (Burchell et al., 2025), would be required to assess the generalizability of our method.

Finally, our label propagation approach for *Interactive Discussion* and *Spoken* registers led to worse performance for these categories in classification experiments. This suggests that the modeling of hierarchical register relationships through propagation should be reconsidered in future work.

Acknowledgments

This work was supported by the Research Council of Finland through several projects: FIN-CLARIAH research infrastructure (project 358720, which has also received funding from the European Union – NextGenerationEU instrument), "Mechanisms of register variation in massively multilingual web-scale corpora" (project 362459), "Massively multilingual modeling of registers in web-scale corpora" (project 331297), and "Green NLP - controlling the carbon footprint in sustainable language technology" (project 353167). We also wish to acknowledge CSC – IT Center for Science Ltd. for providing computational resources.

References

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Pinkesh Badjatiya, Litton J Kurisinkel, Manish Gupta, and Vasudeva Varma. 2018. Attention-based neural text segmentation. *Preprint*, arXiv:1808.09935.

Mostafa Bayomi and Seamus Lawless. 2018. C-HTS: A Concept-based Hierarchical Text Segmentation approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press.

Douglas Biber and Jesse Egbert. 2023. What is a register?: Accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies*, 5.

Douglas Biber, Jesse Egbert, and Daniel Keller. 2020. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16(3):581–616.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, and Jaume Zaragoza-Bernabeu. 2025. An expanded massive multilingual dataset for high-performance language technologies. *Preprint*, arXiv:2503.10267.

Jiuhai Chen and Jonas Mueller. 2024. Automated data curation for robust language model fine-tuning. *arXiv preprint arXiv:2403.12776*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Jesse Egbert and Marianna Gracheva. 2023. Linguistic variation within registers: granularity in textual units and situational parameters. *Corpus Linguistics and Linguistic Theory*, 19(1):115–143.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. Association for Computational Linguistics.

Goran Glavaš and Swapna Somasundaran. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. *Preprint*, arXiv:2001.00891.

Hui Wen Goh, Ulyana Tkachenko, and Jonas Mueller. 2022. Crowdlab: Supervised learning to infer consensus labels and quality scores for data with multiple annotators. *arXiv:2210.06812*.

Amir Hazem, Beatrice Daille, Dominique Stutzmann, Christopher Kermorvant, and Louis Chevalier. 2020. Hierarchical text segmentation for medieval manuscripts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6240–6251, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marti A. Hearst. 1994. Multi-paragraph segmentation expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Marti A. Hearst and Christian Plaunt. 1993. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, page 59–68, New York, NY, USA. Association for Computing Machinery.

Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo, and Veronika Laippala. 2024a. From discrete to continuous classes: A situational analysis of multilingual web registers with LLM annotations. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 308–318, Miami, USA. Association for Computational Linguistics.

Erik Henriksson, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. 2024b. Automatic register identification for the open web using multilingual deep learning. *Preprint*, arXiv:2406.19892.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hakan Inan, Rashi Rungta, and Yashar Mehdad. 2022. Structured summarization: Unified text segmentation and segment labeling as a generation task. *Preprint*, arXiv:2209.13759.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, page 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Taja Kuzman and Nikola Ljubešić. 2023. Automatic genre identification: A survey. *Language Resources and Evaluation*.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2022. Register identification from the unrestricted open Web using the Corpus of Online Registers of English. *Language Resources and Evaluation*.

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, Los Alamitos, CA, USA. IEEE Computer Society.

Yang Liu, Chenguang Zhu, and Michael Zeng. 2021. End-to-end segmentation-based news summarization. *Preprint*, arXiv:2110.07850.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray Buntine. 2021. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3334–3340, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716, Online. Association for Computational Linguistics.

Gjorgji Madjarov, Vedrana Vidulin, Ivica Dimitrovski, and Dragi Kocev. 2019. Web genre classification with methods for structured output prediction. *Inf. Sci.*, 503(C):551–573.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Amanda Myntti, Liina Repo, Elian Freyermuth, Antti Kanner, Veronika Laippala, and Erik Henriksson. 2024. Intersecting register and genre: Understanding the contents of web-crawled corpora. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 386–397, Miami, USA. Association for Computational Linguistics.

Inderjeet Nair, Aparna Garimella, Balaji Vasan Srinivasan, Natwar Modani, Niyati Chhaya, Srikrishna Karanam, and Sumit Shekhar. 2023. A neural CRF-based hierarchical approach for linear text segmentation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 883–893, Dubrovnik, Croatia. Association for Computational Linguistics.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research (JAIR)*, 70:1373–1411.

Martin Riedl and Chris Biemann. 2012. TopicTiling: A text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. *Machine Learning and Knowledge Discovery in Databases*, pages 145–158.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 747–748.

Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. 2021. Unsupervised topic segmentation of meetings with bert embeddings. *Preprint*, arXiv:2106.12978.

Piotr Szymański and Tomasz Kajdanowicz. 2017. A network perspective on stratification of multi-label data. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74 of *Proceedings of Machine Learning Research*, pages 22–35. PMLR.

Aditya Thyagarajan, Elías Snorrason, Curtis Northcutt, and Jonas Mueller. 2023. Identifying incorrect annotations in multi-label classification data. In *ICLR Workshop on Trustworthy ML*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linzi Xing, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi. 2020. Improving context modeling in neural topic segmentation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 626–636, Suzhou, China. Association for Computational Linguistics.

Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, and Wen Wang. 2023. Improving long document topic segmentation models with enhanced coherence modeling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5592–5605, Singapore. Association for Computational Linguistics.

A Appendix: Sample texts with labels identified as noisy by Cleanlab

This appendix presents examples of texts from the CORE corpus that Cleanlab identified as noisy. These include mislabeled texts (where human annotators assigned an apparently incorrect register) and ambiguous hybrid-labeled cases (texts where annotators were split between two registers, as explained in Section 3). For each example, we show the original human-assigned label as well as the more appropriate register category based on content analysis.

Mislabeled as *Interactive Discussion* (ID)

The People of West Cork and Kerry It seems to me the people of West Cork and Kerry They seem to understand the ways of my soul They seem to recognise the healing ways of a young lad Born into pain for the song and to roam And now though still not old, I live alone in the garden The pen it is slow but my heart is at rest And when I see the world now, I see a world without turmoil And all things I see now, I look for the best Chorus I know all the towns and I know all the places I have kissed your lips and I have held your hand Been all around the world but have not found such graces For the people of West Cork and Kerry were grand But when I was a young lad, the world was heavy on me You gave me plain talk, and you made me feel blessed You gave me the magic of all that went before me When I needed to lay low, you gave me the nest Chorus And now though still not old, I live alone in the garden The pen it is slow but my heart is at rest You gave me the magic of all that went before me When I needed to lay low, you gave me the nest

Appropriate label: *Lyrical* (LY)

Mislabeled as *Narrative* (NA)

But It Was Only A 2 Stair So after a year of beating, my charmer snapped right in half. Now I need to decide on a new frame, I was thinking either BB17 "Serpent" or Hold Fast "Converter29? . Does anyone know any other 29er frame(s) out there that has a mid or negative bb? I'm quite glad this question came up. I've been looking at 29er frames for what feels like ages now. Thanks Nelson. Why don't you just get a new Charmer? Another two questions by the way: 1. Why does it look like Mike Chacon was the only one riding (his signature frame..) the Leader Hurricane? Anything wrong with that frame but the BB drop? I mean he does pretty much everything on that frame but still everyone else seems to prefer breakbrake17's or Hold Fast's frames.. I am sure a bunch of cali kids rock Mike Chacon's frame, but I think everyone doing pro-level FGFS stuff wants that higher bb. Mike definitely has The Hurricane dialed in for his style of riding though. Nelson Definitely good considerations there, I appreciate all the input! I'll let you know what I end up with

Appropriate label: *Interactive Discussion* (ID)

Mislabeled as *Spoken* (SP)

Do you have a strong trademark? A trademark is one of your most important business assets, and the selection of your mark needs to be done with care. At the outset of a trademark application, your trademark agent or trademark lawyer can and should explain to the you the strengths and weaknesses of your proposed mark. The selection of trademarks can be broken down into five broad categories: inherently strong marks, inherently weak marks, suggestive marks, compound word marks and marks that have acquired a second meaning, each of which are discussed in this video.

Appropriate label: *Informational Description* (IN)

Mislabeled as *Narrative* (NA) + *Opinion* (OP)

A bit about Clark, Jane-Michele ... Jane-Michele Clark is president of The Q Group (www.theQgroup.com), a strategic positioning and marketing firm with a 30 year history. In addition to being a business/marketing strategist, Jane-Michele teaches MBA level marketing at the Schulich School of Business, is a corporate trainer, author

and speaker. She is also a 9-time nominee for the Canadian Woman Entrepreneur of the Year Award. Jane-Michele can be reached at jmc@theQgroup.com or 416-424-6644

Appropriate label: *Informational Description* (IN)

Mislabeled as *How-to/Instructional* (HI) + *Informational Description* (IN)

'Tiara Oranye' at Telco Company Hi Marta.. may I discuss more about this with you..? this is from the community manager's side, how about if the community is a brand community, what's the tips and trick for the brand owner who manage the community? 3 months ago Reply Are you sure you want to Yes No matter33 My name is Miss matter Garba,i saw your profile today on (slideshare.net) and became intrested in you,i will also like to know you the more,and i want you to send an email to my email address (mattergarba56@yahoo.com) so i can give you my picture for you to know whom i am. However i believe we can move on from here! I am waiting for your mail to my email address above.(Remeber the distance, colour or language does not matter but love matters alot in life miss matter. (mattergarba56@yahoo.com) 4 months ago Reply [...]

Appropriate label: *Interactive Discussion* (ID) (?)

Mislabeled as *Interactive Discussion* (ID) + *Narrative* (NA)

Still, having tried to watch the show myself, I can't say I'm surprised. Saying this epi was the best sure ain't sayin' a lot. And what was up with that not-so-amazing singer/songwriter they kept showcasing? It's not like Will & Grace having a guest star. I don't like any form of media which tries to shove another medium down my gullet. I saw parts of two of the shows, and it appeared to me that they were schlepping some artists. The "love" part was totally absent. And really, how freakin' exciting is being a music A&R rep? It was like Ed without the humor... or the plot. too bad for the actor. He seems like a good enough guy. Comments are now closed on this post. Like what you're reading? To view other posts at Signifying Nothing , please visit the BlogFront . Signifying Nothing formerly featured the stylings of Brock Sides , a left-leaning philosopher turned network administrator currently residing in Memphis, Tennessee who now blogs at Battlepanda , and Robert Prather , a libertarian-leaning conservative economist and occasional contributor at OTB .

Appropriate label: *Opinion* (OP)

Ambiguous, labeled as *Informational Persuasion* (IP) + *Opinion* (OP)

Discuss this article with... The 'slippery slope to murder' argument must not prevail. Canada has shown mercy to sufferers and we must too Death for Tony Nicklinson will have come as a blessed relief. Anyone who watched the footage of the moment when he learnt that his appeal to the High Court had failed – and I defy anyone to do so with dry eyes – will have seen a man of astonishing courage, broken by the immutability of the law. His final act of bravery was to start refusing food, rather than to put his loved ones at risk of prosecution. Pneumonia, fortunately, did the rest. But this was not the ending he deserved.

Appropriate label: *Opinion* (OP)/*Informational Persuasion* (IP)

Ambiguous, labeled as *Informational Description* (IN) + *Opinion* (OP)

Tuesday, November 6, 2012 Stressing about things? Stare at these for a few moments.... This blog post is offered as a moment of quiet serenity on the day before a pretty serious election. There is a lot of the stuff, from the national races to some local propositions, that will certainly have a direct effect on my life, if not yours. But it is stressful. We live next to the Tuolumne River, and there is a river walk with a lot of shrubs where a colony of cats has taken up residence. It probably doesn't do much for the local squirrel population, but the local residents probably don't mind the relative absence of mice and rats. The cats are pretty suspicious of strangers, but they always come out to see of we are bringing catfood... Sooo...imagine the purring, and feel your blood pressure go down a few points. Say "ahhh..." a couple of times, and the stress lines will leave your forehead... But as this one is clearly saying..."don't forget to vote tomorrow"... THANKS TO ALL WHO VOTED TO SUPPORT EDUCATION! About Me I am a teacher of geology at Modesto Junior College and former president of the National Association of Geoscience Teachers, Far Western Section. I have led field trips all over the western United States, and a few excursions overseas, but my homebase is the Sierra Nevada, the Great Valley, and the Coast Ranges of California.

Appropriate labels: Begins with *Narrative* (NA)/*Opinion* (OP) and transitions to *Informational Description* (IN)

Ambiguous, labeled as *Lyrical* (LY) + *Opinion* (OP)

you heard it here first "Intimate but grand, Crybaby is a triumph" **** THE GUARDIAN FILM & MUSIC "Unafraid to be both beautiful and sad, songs such as Shame and Misery Of Love are like Roy Orbison tackling Scott Walker" **** Q MAGAZINE "A Bristolian tunesmith with as much heart as Richard Hawley" NME Bristols newcomers Crybaby head out on their first headline tour in support of their latest single 'We're Supposed To Be In Love' (out Sept 24th), which is the third single to be taken from their critically acclaimed eponymous debut album. September gig dates 15th Edinburgh, Electric Circus; 16th Glasgow, King Tuts; 17th Leeds Nation of Shopkeppers; 18th Manchester, The Castle; 19th London, Lexington; 20th Birmingham, Hare & Hounds; 21st Leicester, The Cookie Jar; 22nd Brighton,The Hope; 27th Bristol, Louisiana

Appropriate label: *Opinion* (OP)/*Informational Description* (IN)/*Informational Persuasion* (IP) (?)