Register Mixing Is the Norm on the Web

Erik Henriksson, Alireza Razzaghi, Tuomas Lundberg, Antti Kanner, Veronika Laippala
TurkuNLP, University of Turku
{eihenr,alireza.razzaghi,tuomas.w.lundberg@utu.fi,antti.kanner,mavela}@utu.fi

Abstract

Nearly all studies on web registers—online text varieties associated with characteristic social contexts and linguistic features—use full documents as the unit of analysis. However, web documents often contain sections in different registers. A cooking blog, for instance, may combine personal storytelling, recipe instructions, user comments, and promotional text within a single URL. This internal variation raises doubts about the validity of document-level register labeling. In this paper, we propose an LLM-based approach that identifies register-homogeneous segments within documents and apply it to a 10,000-document English sample from HPLT 3.0. We show that segmentation addresses persistent problems in register analysis, including low inter-annotator agreement and category fuzziness. Strikingly, it also reveals that most web documents contain more than one register, making register mixing the norm rather than the exception on the web.

1 Introduction

Register, the linguistic variety associated with a particular communicative situation and social context, has long been recognized as a central organizing principle of language use (Halliday, 1978; Biber, 1988; Matthiessen, 2019; Biber and Conrad, 2019). On the web, register variation is particularly diverse, ranging from news articles and recipes to forum discussions and product reviews, each with distinct situational configurations and linguistic features. Corpus linguists have focused on identifying and classifying these varieties, with the Corpus of Online Registers of English (CORE; Egbert et al., 2015; Laippala et al., 2023) being the foundational large-scale example. CORE established a taxonomy of web registers and a document-level annotation methodology widely adopted in corpus linguistics and NLP (e.g., Laippala et al. 2019; Biber et al. 2020; Repo et al. 2021; Henriksson et al. 2024, 2025a; Myntti et al. 2025; Burchell et al. 2025).

However, document-level analysis oversimplifies web text by treating individual documents as instances of a single register. Within a single URL, a cooking blog post may consist of multiple sections with distinct communicative purposes. These sections may include promotional text, information about the author, cooking instructions, or a comment section. Some studies have addressed this document-internal register mixing by allowing individual annotators to assign multiple register labels to a single document (e.g., Skantsi and Laippala, 2023), which better reflects the mixed nature of web text. However, multi-labeling at the document level lacks the granularity needed to distinguish true register-mixing within a passage from register shifts between distinct text segments (Egbert and Gracheva, 2022). The mismatch between theory and annotation unit has concrete empirical consequences. Inter-annotator agreement on CORE reaches only $\kappa = 0.47$ (Egbert et al., 2015), and no machine learning model has achieved above 80% F1 on document-level web register classification (Henriksson et al., 2025b).

In this paper, we introduce a pipeline for analyzing the internal register composition of web documents (`github.com/TurkuNLP/register-segments-hplt-en`). We utilize the XML structure of the HPLT 3.0 web corpus (Oepen et al., 2025) to segment documents into paragraphs and use a large language model (LLM) to label each paragraph, merging consecutive same-register segments. Human evaluation shows that the LLM achieves near-human annotation quality, and a classifier fine-tuned on the resulting labels outperforms document-level baselines. Applied to a 10,000-document English sample, the pipeline reveals that the majority of web documents contain many registers, with characteristic positional patterns, co-occurrences, and transition sequences. Linguistic analysis confirms that the segment-level labels represent genuine dimensions of register variation.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 138–149

2 Background

2.1 Document-level register annotation

The Corpus of Online English (CORE) (Egbert et al., 2015; Laippala et al., 2023) established a widely used approach for corpus-linguistic web register research, in which annotators assign a register label to documents from a taxonomy of eight main registers with optional subregisters. In this context, a *document* refers to the content retrieved from a single URL by a web crawler and stored as a single record in a corpus. In document-level analysis, register labels often correspond to what are sometimes interchangeably called web *genres* (Kuzman and Ljubešić, 2025). The sub-document approach we take here arguably makes the *genre* vs. *register* distinction clearer and comes closer to a traditional understanding of register as a functional text variety (Goulart et al., 2020). In our approach, a web genre such as *recipe* or *blog* may contain multiple register varieties, each with their own communicative purpose within the document.

The CORE approach has been applied to various languages besides English and used to develop automatic register classifiers (e.g., Repo et al., 2021; Skantsi and Laippala, 2023; Henriksson et al., 2025b). However, no register classifier has exceeded 80% micro F1 (Henriksson et al., 2025a). This ceiling may partly reflect the inherent fuzziness of register categories (Biber et al., 2020; Henriksson et al., 2024), but another likely factor is the unit of analysis itself. Many web documents are composites made of multiple discourse units with different communicative purposes rather than unified wholes (Biber et al., 2020; Biber and Egbert, 2023). This within-document variation is invisible to document labels and contributes to annotator disagreement when different raters attend to different sections of the same page (Egbert et al., 2015).

2.2 Sub-document register segmentation

Text segmentation is the task of dividing a text into coherent, non-overlapping units (Hearst, 1994, 1997; Hearst and Plaunt, 1993; Galley et al., 2003; Liu et al., 2021). In the context of web register analysis, the goal is to identify spans of text that are internally consistent in terms of register.

Henriksson et al. (2025a) presented the first attempt to segment web documents by register, using a recursive binary splitting algorithm applied to ModernBERT-based register predictions (Warner et al., 2024). Working with the plain-text CORE documents, which have no structural markup, the algorithm identified register boundaries by iterating over sentence boundaries and splitting wherever the predicted register distributions of the two resulting halves diverged sufficiently. They found that sub-document segments carry a cleaner register signal than full documents: classifier F1 scores improved, embedding-based silhouette scores showed more distinct register clusters, and within-register linguistic variance decreased. The approach was limited, however, by the absence of structural markup in the source documents and by its dependence on an imperfect classifier for segmentation.

The present work extends this line of research by using an LLM to perform both segmentation and labelling in a single pass, and by applying the approach at scale to a modern web dataset.

3 Methods

3.1 Data

We sample 10,000 English documents from HPLT 3.0 (Oepen et al., 2025), a large multilingual web corpus that provides documents in an XML format encoding structural information such as paragraphs, headings, lists, and comment sections. The HPLT 3.0. corpus is sharded by Web Document Score (WDS; Pavón et al., 2023), a measure of text quality, and only documents with an overall WDS of 9–10 are selected.

We convert each document to a simplified plain-text format that retains structural markup as section headers (`main`, `comments`) and numbered paragraph indices, while stripping character-level formatting tags (e.g. `<hi>`). This yields a representation that is both human-readable and compact enough for LLM processing.

3.2 Classification scheme

We annotate three orthogonal facets per segment:

- **Source**: whether the text is *written*, *transcribed* speech, or *cannot_rate* (boilerplate, navigational, or incoherent content).
- **Interactive**: either *true* (participatory context: forums, comment sections, interviews) or *false* (monologic). Null for *cannot_rate* segments.
- **Function**: the primary communicative function, chosen from: *narrating*, *describing*, *instructing*, *opinionated*, *promoting*, or *lyrical*. Null for *cannot_rate* segments.

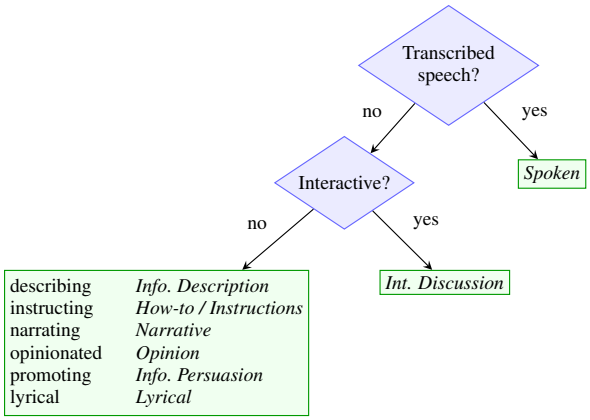

Figure 1: Mapping from the three annotation facets to CORE register labels, applied in hierarchical order.

The faceted scheme maps directly onto the CORE taxonomy following the decision tree in Biber and Egbert (2018, p. 17), as shown in Figure 1. Specifically, any segment with the source *transcribed* maps to CORE's *Spoken*; if the segment is *interactive*, it maps to *Interactive Discussion*; and the remaining written, non-interactive segments map directly via their function. In our analyses, we report results for both the original three-facet scheme and the CORE mappings.

3.3 LLM annotation pipeline

We annotate segments using two LLMs: Gemini 3 Flash (Google DeepMind, 2025) and Qwen3 235B A22B (Qwen Team, 2025). These models were selected for their strong performance, low inference cost, and availability through the OpenRouter API[1], which we used for the task. Each model receives a zero-shot prompt of approximately 1,100 words that defines the classification scheme (see Appendix B). Documents are converted into numbered segments following a simplified markdown format that distinguishes headings, paragraphs, blockquotes, lists, tables, and code snippets. The model classifies paragraphs, blockquotes, and list items by register; headings, tables, and code snippets are used as context but assigned *cannot_rate*. Segments are labeled 20 at a time with a sliding window of ± 5 segments of surrounding context.

After labeling, consecutive segments with identical register labels are merged into single spans, treating *cannot_rate* segments as transparent during merging so that boilerplate or navigational content does not artificially fragment register spans. The resulting register segments thus range from single paragraphs to multi-paragraph stretches of consistent register. The paragraph represents the resolution limit of the approach (i.e., sub-paragraph register shifts are ignored); however, the granularity is substantially finer than document-level analysis.

3.4 Evaluation and analysis

We evaluate our approach in three ways. First, to validate annotation quality, two human annotators independently labeled 100 randomly sampled segments; we compute Cohen's κ and F1 scores between the two humans (H1 vs. H2) and between each LLM and the human average. Second, to assess the learnability of the resulting labels, we finetune a ModernBERT (Warner et al., 2024) model on the LLM-generated segment labels and compare its performance against document-level baselines. Third, to verify that the register labels correspond to genuine linguistic differences, we tag the segments with 96 Biber-style features using BiberPlus (Alkiek et al., 2025) and apply PCA to the standardized register means.

Beyond validation, we use the segmented annotations to analyze the internal register structure of web documents, including register distributions, within-document combinations, positional patterns, co-occurrences, and sequential transitions.

4 Results and Discussion

4.1 Human evaluation of segments

Table 1 reports inter-annotator agreement and F1 scores for human–human and human–LLM comparisons for 100 randomly sampled segments.

Gemini 3 Flash reaches agreement comparable to the human annotators. Its κ of 0.77 on CORE-mapped labels and F1 of 0.81 are in the same range as the human–human κ of 0.76 and F1 of 0.80, though the small evaluation sample ($n = 100$) limits the precision of these estimates. Qwen3 235B A22B performs consistently lower, with a CORE-mapped κ of 0.63 and F1 of 0.69. The gap between the two models, despite Qwen3's relatively large size, suggests that register annotation is a non-trivial task where LLM choice matters. Based on these results, all subsequent analyses use Gemini-generated labels.

To assess whether the results in Table 1 generalize beyond the 100-segment sample, a single annotator independently labeled an additional 500 randomly sampled segments. Compared against Gemini 3 Flash, agreement was consistent with the smaller evaluation across all fields (κ = 0.88 /

[1] https://openrouter.ai

	Cohen's κ		F1	
Field	H1/H2	Gemini	H1/H2	Gemini
Source	0.84	0.74	0.96	0.94
Interactive	0.92	0.88	0.97	0.95
Function	0.70	0.74	0.76	0.80
CORE label	0.76	0.77	0.80	0.81

Table 1: Inter-annotator agreement (Cohen's κ) and F1 between human annotators (H1/H2) and between humans (averaged) and Gemini 3 Flash. Qwen3 235B A22B performs consistently lower (κ = 0.67, 0.76, 0.56, 0.63; F1 = 0.91, 0.90, 0.83, 0.69 for the same fields).

0.94 / 0.81 / 0.83 for source, interactive, function, and CORE label respectively; weighted F1 = 0.96 / 0.97 / 0.86 / 0.86). A human–human baseline at this scale was not conducted, however.

Importantly, both human–human and Gemini agreement substantially exceed the κ of 0.47 reported for document-level CORE annotation (Egbert et al., 2015), suggesting that paragraph-level segments constitute a more stable unit for register annotation than full documents.

Most disagreements between humans and the LLM involve conceptually proximate functions, most notably *describing* versus *narrating*. Consider the following example:

> *[Context -1]* Lila also received some money for her savings account from MeMe and Henry.
>
> **[TARGET]** These were the gifts that Andrew and I gave her: the Jesus Storybook Bible, a PraiseBaby CD and DVD set, and a pretty, wooden cross frame.
>
> *[Context +1]* We love you baby girl! Our hope is that you grow to love Jesus and know him as your friend and Saviour!

One human annotator labeled the target segment as *describing*, reading it as an informational list. The LLM and the other human labeled it as *narrating*, interpreting it as part of an account of an event. The segment does refer to a concrete past event, but it lacks temporal sequencing, making both readings defensible. The disagreement thus reflects genuine ambiguity in the text rather than misclassification. Most of the LLM's "errors" are of a similar kind.

4.2 Learnability of segments

To assess whether the segmented labels yield a learnable register signal, we fine-tuned a ModernBERT-large (Warner et al., 2024) model on the LLM-generated segment labels (segments $\geq$ 100 words; $n \approx$ 36k). The supervised model achieves 0.83 F1 on CORE-mapped labels. For reference, document-level English CORE classifiers have not exceeded 0.76 F1 (Henriksson et al., 2024), though the two settings differ in training data, label source, and text granularity, making direct comparison difficult. The higher figure is nonetheless consistent with the finding in Henriksson et al. (2025a) that segmentation produces a cleaner register signal, and suggests that register segments are an easier classification target than full documents. A caveat is that the fine-tuning data consist of LLM-generated labels, so systematic annotation biases may propagate into the model.

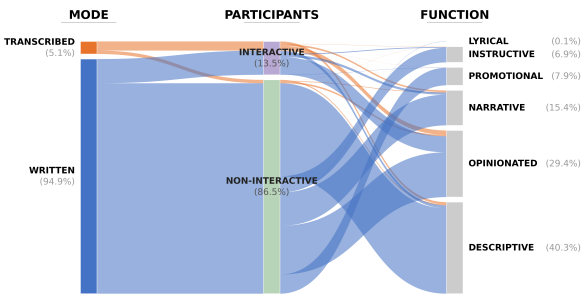

Figure 2: Distribution of register facets across all labelled segments (excluding *cannot_rate*). Flow widths are proportional to segment counts.

4.3 Distribution of register facets

Figure 2 shows how labeled segments (excluding *cannot_rate*) distribute across the three annotation facets source, interactivity, and function. The vast majority of segments are *written* (94.9%), with *transcribed* content accounting for only 5.1%.

The middle column reveals an asymmetry between the two source types. Written segments are predominantly non-interactive (86.5% overall), whereas transcribed segments flow almost entirely into the interactive category. This follows naturally from the nature of transcribed web content, which consists largely of interviews, podcast transcripts, and similar participatory varieties.

Turning to communicative function, the dominant category is *describing* (40.3%), followed by *opinionated* (29.4%) and *narrative* (15.4%). The prevalence of *opinionated* segments among interactive content is unsurprising: comment sections and forum replies are prototypically evaluative. More noteworthy is the relative scarcity of *narrative* at the segment level. In document-level CORE, *Narrative* is the single most frequent category, accounting for approximately 40% of documents (Egbert et al., 2015). One possible explanation is that documents labeled as narratives at document level are

Figure 3: Distribution of the number of distinct registers per document, under the full faceted scheme (blue) and CORE-mapped labels (orange).

internally heterogeneous, combining a narrative frame with stretches of other text varieties.

4.4 Within-document register combinations

Figure 3 shows the distribution of distinct register counts per document under the faceted and CORE labeling schemes. Since *cannot_rate* segments are treated as transparent during merging (Section 3.3), these counts reflect genuine register shifts rather than fragmentation by boilerplate or navigational content.

Fewer than 20% of documents contain a single register, and most often, documents contain two or three registers, with one and four registers being the next most common. The view offered by the segment analysis is therefore radically different from the document-level analysis in CORE, where 69% of documents received a single consensus register label from multiple annotators (Egbert et al., 2015, p. 10). More broadly, this perspective challenges the common assumption in corpus linguistics that register categories can be meaningfully assigned to entire documents.

Figure 4 zooms in on the specific combinations of registers, showing the 15 most frequent CORE-label co-occurrences across documents. Of these, only four are single-register: *Informational Description* (9.1%), *Informational Persuasion* (3.6%), *Opinion* (3.1%) and *Narrative* (2.3%). The most frequent combination is *Informational Description* + *Opinion* (9.7%), which alone outnumbers any single-register category. Three-way combinations are also common. *Informational Description* + *Opinion* + *Narrative* ranks third overall (8.6%); and the seventh most frequent pattern combines as many as four registers.

Figure 5 shows the number of distinct registers per document as a function of document length. Although longer documents tend to contain more registers, the relationship is weak: even short documents ($\sim 10^3$ characters) frequently contain two or three distinct registers, and the modal count remains at two across most of the length range. Register mixing is thus not simply a byproduct of document length, but a persistent feature of web text regardless of size.

4.5 Positional distribution of registers

To examine where registers tend to appear within documents, we computed the probability mass distribution of each CORE-mapped label as a function of relative character position, accumulated into 128 equal-width bins and smoothed with a Gaussian filter ($\sigma = 2$). Each label is normalized independently, so curves reflect positional *shape* rather than overall frequency.

The results reveal clear positional patterns (Figure 6). *Interactive Discussion* rises near-monotonically from near zero to a peak at position 100%, directly reflecting the convention that comment sections appear after main content on most webpages. Conversely, *Narrative* is front-loaded, peaking in the first 5–10% and declining steadily. *How-to/Instructions* peaks later (around 75–80%), likely following introductory or contextualizing material. *Spoken* content peaks early (15–20%) and declines gradually, plausibly reflecting embedded interview excerpts near document openings.

These patterns suggest that register mixing is not random but *structurally organized*: registers occupy characteristic positions that follow conventions of web genre composition.

4.6 Register co-occurrence patterns

To examine which registers tend to co-occur within the same document, we compute two measures. The conditional probability matrix shows $P(\text{col} \mid \text{row})$: given that a document contains the row register, the probability that it also contains the column register. The pointwise mutual information (PMI) matrix shows $\log \frac{P(A \cap B)}{P(A) \cdot P(B)}$: how much more often two registers co-occur than expected by chance. As PMI is symmetric, only the lower triangle is shown.

The conditional heatmap (Figure 7) shows that *Informational Description* co-occurs with almost every register at high rates, largely as a base-rate effect of its overall frequency. The PMI matrix (Figure 8) controls for this and reveals that *Informational Description* has near-zero PMI with most registers, suggesting it in fact functions as a register-neutral backdrop in web documents.

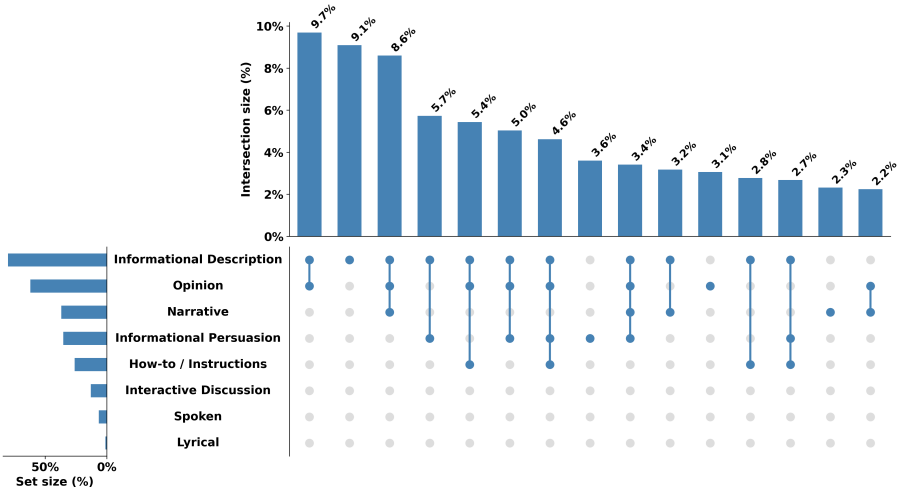

Figure 4: UpSet plot of the 15 most frequent CORE-mapped register combinations across documents. Horizontal bars show per-register document percentages; vertical bars show combination percentages.

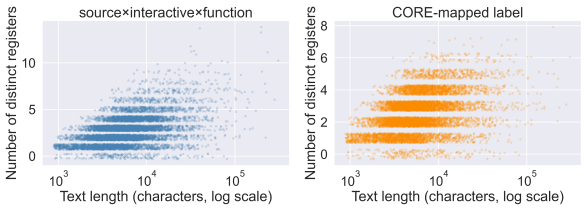

Figure 5: Number of distinct registers per document as a function of document length (log scale), under the faceted scheme (left) and CORE-mapped labels (right).

The clearest pattern is the isolation of *Informational Persuasion*, which has negative PMI with nearly every register except *How-to/Instructions* (0.25), plausibly because product and service pages often include usage instructions. *Interactive Discussion* is the opposite: positive PMI with *Narrative* (0.31), *Lyrical* (0.22), and *Opinion* (0.16), but negative with *Informational Persuasion* (–0.41) and *Informational Description* (–0.14). Comment sections attract subjective and narrative content but repel informational and promotional material.

The strongest positive associations are *Spoken+Lyrical* (1.07) and *Narrative+Lyrical* (0.71), though both should be interpreted cautiously given that *Lyrical* and *Spoken* are among the rarest registers (Figure 4), making PMI sensitive to small counts. The LLM may also conflate poetic prose with informal spoken-style writing, which would inflate their apparent association.

4.7 Register transitions

To examine sequential register structure, we compute a PMI matrix over register transitions, where each value reflects how much more or less likely a given register-to-register sequence is relative to base rates. Consecutive segments with identical labels have been collapsed before counting (Section 3.3), so the matrix captures genuine register shifts.

The results (Figure 9) reveal clear sequential preferences. The strongest avoidances involve *How-to/Instructions*, which almost never transitions into *Spoken* (-2.52) or *Lyrical* (-1.21), and *Informational Persuasion*, which strongly avoids *Lyrical* (-1.64) and *Spoken* (-1.15). In other words, promotional and instructional content is rarely followed by lyrical or conversational text.

On the positive side, *Informational Description* functions as a transitional register, showing above-chance transitions to and from most other registers, even when base rates are taken into account. This contrasts with the co-occurrence analysis, where *Informational Description* showed near-zero PMI with most registers. It appears that when documents shift between registers, they tend to pass through descriptive stretches.

Narrative and *Spoken* show near-symmetric mutual transitions (0.89 and 0.86), suggesting genuine alternation. This *Spoken*↔*Narrative* pattern likely reflects genre-level structure rather than register affinity per se: interview or podcast transcripts are a plausible host genre in which spoken framing and narrative content naturally interleave: for example, a guest might recount an experience (*Narrative*) within a conversational exchange (*Spoken*). This points to a theoretical limitation of flat register segmentation: genres such as interviews impose a hierarchical structure within which registers alternate in conventionalized ways. Modeling this hierarchy

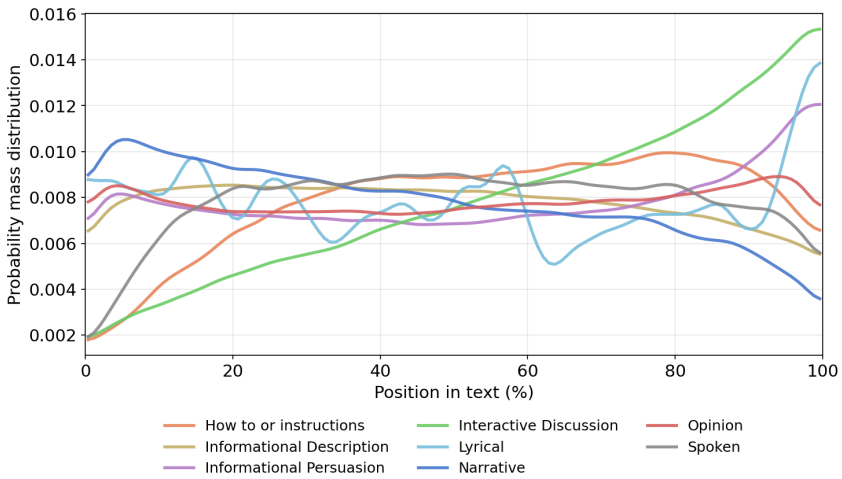

Figure 6: Distribution of CORE-mapped register labels by relative document position. Each label is normalised independently; curves reflect positional shape rather than overall frequency.

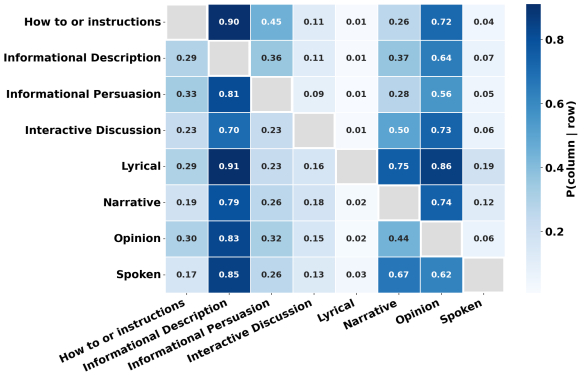

Figure 7: Conditional probability heatmap: $P(\text{col} \mid \text{row})$, the proportion of documents containing the row register that also contain the column register.

Figure 8: Pointwise mutual information (PMI) between register pairs. Positive values indicate above-chance co-occurrence; negative values indicate avoidance.

explicitly is left for future work.

4.8 Linguistic analysis

To verify that the segment-level register labels reflect genuine register variation, we analyze the

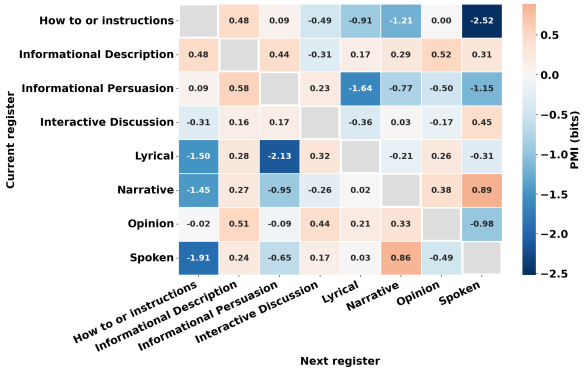

Figure 9: PMI matrix over register transitions (rows = current, columns = next). Positive values indicate above-chance sequences; negative values indicate avoidance.

Biber-style feature profiles of each register using BiberPlus (Alkiek et al., 2025), a neural tagger that estimates 96 lexico-grammatical features per text. Segments shorter than 50 words are excluded, and features with near-zero variance or pairwise correlations above 0.95 are removed to reduce redundancy. We compute mean feature values per register label and apply PCA to the z-scored means. We focus on PC1 and PC3, which together account for 52.3% of variance and yield the most interpretable dimensions; PC2 was less clearly interpretable and is omitted for brevity. Table 2 reports the five highest-loading features at each pole of each component.

PC1 (37.4%) corresponds to the involved versus informational opposition that has been identified as the primary dimension of register variation across numerous Multi-Dimensional studies (Biber, 1988; Biber and Egbert, 2016). The positive pole is defined by private verbs, contractions, that-deletion,

Dimension	Feature	Loading
PC1 (37.4%) – Involved (+) vs. Informational (−)		
+	Private verbs	+0.179
+	Contractions	+0.171
+	That-deletion	+0.170
+	*because*	+0.169
+	Direct WH-questions	+0.161
−	Mean word length	−0.168
−	Present participles	−0.165
−	Other nouns	−0.163
−	Present participial WHIZ deletion	−0.140
−	Attributive adjectives	−0.132
PC3 (18.4%) – Instructional (+) vs. Narrative (−)		
+	Second-person pronouns	+0.246
+	Infinitive verbs	+0.227
+	Predictive modals	+0.216
+	*if/unless*	+0.190
+	Possibility modals	+0.179
−	Sentence relatives	−0.218
−	By-passives	−0.197
−	Third-person pronouns	−0.196
−	Conjuncts	−0.195
−	Perfect aspect	−0.192

Table 2: Top 5 feature loadings at each pole of PC1 and PC3, from PCA on z-scored BiberPlus register means.

because, and direct WH-questions; the negative pole by mean word length, nouns, present participles, and attributive adjectives. In Figure 10, the registers spread along this axis as expected: *Interactive Discussion*, *Lyrical*, and *Spoken* on the involved side, *Informational Description* on the informational side.

PC3 (18.4%) separates directive from narrative discourse. The positive pole is loaded with second-person pronouns, infinitive verbs, predictive and possibility modals, and conditional subordinators; the negative pole with sentence relatives, by-passives, third-person pronouns, conjuncts, and perfect aspect. This component appears to merge aspects of Dimensions 5 (Irrealis vs. Informational Narration) and 3 (Oral Narrative vs. Written Information) from Biber and Egbert (2016) into a single axis. In the figure, *How-to/Instructions* and *Informational Persuasion* score highest, while *Narrative* occupies the extreme negative end. *Spoken* is worth noting: it is highly involved on PC1 but leans toward the narrative pole on PC3, which is consistent with interview and podcast transcripts that recount events conversationally.

Together, the two components account for over half the variance in 96 features and align with dimensions established in previous work, strongly suggesting that the segment-level labels reflect real linguistic differences rather than artifacts of the LLM annotation.

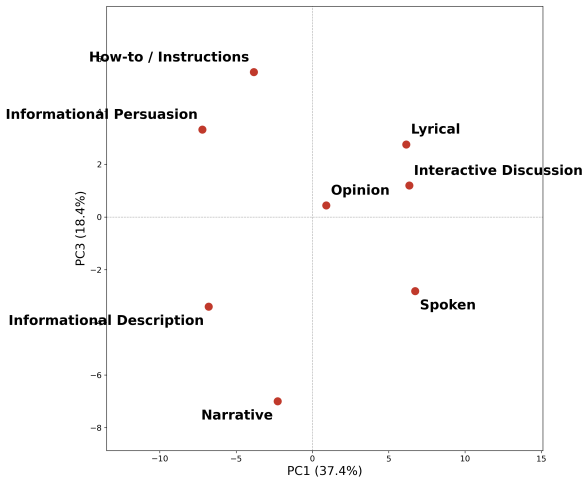

Figure 10: Mean register positions on PC1 (Involved vs. Informational) and PC3 (Directive vs. Narrative), from PCA on z-scored BiberPlus features.

5 Conclusions

We have presented an LLM-based pipeline for register segmentation that achieves annotation quality comparable to human annotators. Applied to 10,000 web documents from HPLT 3.0, the pipeline reveals that register mixing is the norm online: fewer than 20% of documents contain a single register, contrasting sharply with the document-level view in CORE where 69% received a single consensus label. Segmentation also improves inter-annotator agreement (κ = 0.77 vs. 0.47) and classifier learnability (F1 = 0.83 vs. 0.76). Moreover, the segmented annotations reveal that registers occupy characteristic document positions, show systematic co-occurrence and transition patterns, and align with known dimensions of linguistic variation. Together, the findings suggest that the web document is not a natural unit for register analysis.

Limitations

Our human evaluation covered only 100 segments with two annotators; a single-annotator evaluation on 500 segments yielded similar results, but larger-scale evaluation is left for future work. As the ModernBERT classifier is trained on LLM-generated labels, systematic biases may propagate into it. The segmentation also relies on a simplified version of HPLT's (itself noisy) XML structure, and our sample was restricted to high-quality documents (WDS 9–10), leaving open whether register mixing is equally common in lower-quality text. Future work should explore the relationship between genre-level structure and register segmentation.

Acknowledgements

Alireza Razzaghi received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101177564 (HAIF). Cofunded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Furthermore, 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), "Green NLP: Controlling the Carbon Footprint in Sustainable Language Technology" (project 353167), and the "Centre of Excellence Human Diversity through Contacts" (project 374223). We also wish to acknowledge CSC – IT Center for Science Ltd. for providing computational resources.

References

Kenan Alkiek, Anna Wegmann, Jian Zhu, and David Jurgens. 2025. Neurobiber: Fast and interpretable stylistic feature extraction. *Preprint*, arXiv:2502.18590.

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

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Douglas Biber and Jesse Egbert. 2016. Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, 44(2):95–137.

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

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(1):1–22.

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, and 16 others. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, volume 1: Long papers, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

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. 2022. Linguistic variation within registers: Granularity in textual units and situational parameters. *Corpus Linguistics and Linguistic Theory*, 19.

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.

Google DeepMind. 2025. Gemini 3 flash model card. Technical report, Google DeepMind.

Larissa Goulart, Bethany Gray, Shelley Staples, Amanda Black, Aisha Shelton, Douglas Biber, Jesse Egbert, and Stacey Wizner. 2020. Linguistic perspectives on register. *Annual Review of Linguistics*, 6(Volume 6, 2020):435–455.

Michael Alexander Kirkwood Halliday. 1978. *Language as Social Semiotic*. Arnold, London.

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, Saara Hellström, and Veronika Laippala. 2025a. Analyzing register variation in web texts through automatic segmentation. In *Proceedings of the 5th International Conference on Natural*

Language Processing for Digital Humanities, pages 7–19, Albuquerque, USA. Association for Computational Linguistics.

Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo, and Veronika Laippala. 2024. 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. 2025b. Automatic register identification for the open web using multilingual deep learning. *Preprint*, arXiv:2406.19892.

Taja Kuzman and Nikola Ljubešić. 2025. Automatic genre identification: A survey. *Language Resources and Evaluation*, 59(1):537–570.

Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079.

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

Christian MIM Matthiessen. 2019. Register in systemic functional linguistics. *Register Studies*, 1(1):10–41.

Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. 2025. Register always matters: Analysis of LLM pretraining data through the lens of language variation. In *Conference on Language Modeling (COLM)*, Montreal, Canada.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. Hplt 3.0: Very large-scale multilingual resources for llm and mt. mono- and bi-lingual data, multilingual evaluation, and pre-trained models. *Preprint*, arXiv:2511.01066.

Pablo Pavón and 1 others. 2023. Web docs scorer. `https://github.com/pablop16n/web-docs-scorer`. GitHub repository.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. *arXiv preprint arXiv:2102.07396*.

Valtteri Skantsi and Veronika Laippala. 2023. Analyzing the unrestricted web: The finnish corpus of online registers. *Nordic Journal of Linguistics*, 48(1):1–31.

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.

A Appendix: Classification scheme details

The three-facet classification scheme assigns the following CORE-mapped labels based on combinations of Source, Interactive, and Function:

- Written + Non-interactive + Describing → **Informational Description**
- Written + Non-interactive + Instructing → **How-to/Instructional**
- Written + Non-interactive + Narrating → **Narrative**
- Written + Non-interactive + Opinion → **Opinion**
- Written + Non-interactive + Promoting → **Informational Persuasion**
- Written + Non-interactive + Lyrical → **Lyrical**
- Transcribed (any) → **Spoken**
- Written + Interactive (any function) → **Interactive Discussion**

B Appendix: Annotation prompt

The following prompt was used for zero-shot LLM annotation, prepended to each batch of 20 segments presented to the model.

Line format conventions Each line is numbered like [1], [2], etc. Lines use simplified markdown: `# text` = heading; plain text = paragraph; `> text` = blockquote; `- item1; item2; item3` = list items joined by semicolons; `TABLE: ...` = table data; `CODE: ...` = code snippet. Use headings, tables, code, and surrounding lines to understand the document's overall purpose, but classify every line in the [CLASSIFY THESE LINES] section.

Categories For each line, assign three fields.

source (always required)

- *written* — originally composed as written text. This is the default for most web content. Informal or ungrammatical writing is still "written" — poor grammar does not mean speech.
- *transcribed* — text that was originally spoken aloud and then written down. Use ONLY when there is clear evidence of oral origin: explicit transcript labels, speaker turn markers (e.g. "Interviewer:", "Q:", "[Speaker]"), or the text is clearly a transcript of a speech, podcast, interview, or hearing. Do NOT use just because the writing is informal, conversational, or ungrammatical. A letter or written document that is quoted or embedded in a page is "written" even if introduced with "here is the full text."
- *cannot_rate* — use for: (1) machine-generated, boilerplate, navigational, or structural content with no substantive human-authored message (e.g. cookie notices, auto-generated footers, breadcrumb navigation, "Related posts" lists, copyright lines, "Share on Facebook", short bylines/datelines, section labels like "7 Parts:", calls to edit/contribute like "Click EDIT to write this answer", and metadata). Note: an author bio with full sentences describing qualifications is substantive — classify it by its function, not as cannot_rate. (2) Incoherent, garbled, or spam-like text that lacks coherent human intent, including bad machine translations and keyword stuffing. (3) Fragments too short to classify — isolated numbers, single words. A complete sentence with clear communicative intent is always classifiable as written or transcribed, not cannot_rate.

interactive (required if source is *written* or *transcribed*; null otherwise)

- *true* — the text is produced in a participatory context where responses from other participants are expected or possible. This includes: forum threads, comment sections, discussion boards, social media posts with replies, chat messages, and interviews with alternating speakers. A single comment or reply is interactive even without visible back-and-forth, because it is produced within a discussion context.
- *false* — monologic: one author/speaker addressing a general audience with no participatory context. Articles, blog posts (the post itself, not comments), encyclopedia entries, guides, stories.

Important: editorially structured Q&A content is NOT interactive. This includes FAQ pages, wiki-style Q&A, and how-to sites where questions and answers are organized as reference content rather than real conversation between participants.

function (required if written/transcribed; null otherwise) The key question: what communicative act is the author performing? Classify by the DOMINANT purpose of the line. If a line serves multiple functions (e.g. describes a feature while also promoting it), choose the function that best captures the author's primary intent. Comments and forum posts can serve any function — people in discussions also narrate events, describe facts, ask questions, and give instructions.

Valid values (no other values are permitted):

- *narrating* — recounting specific events or stories in temporal sequence. Requires specific events that actually happened (or are presented as having happened). This includes news reports of events, historical accounts, and biographical passages that recount what someone did. Factual reference material about a topic is NOT narration even if it mentions dates or timelines. Hypothetical scenarios and descriptions of typical situations are also NOT narration.
- *describing* — presenting factual information, concepts, states of affairs, or general knowledge. This covers neutral reference material, encyclopedic content, and explanatory or analytical writing. Minor framing or incidental

evaluative language does not disqualify a line from "describing" — the test is whether the primary purpose is to inform the reader about facts rather than to argue a position or express a judgment. However, if evaluative language is prominent or the line's purpose shifts toward advocacy or critique, prefer "opinionated."

- *instructing* — teaching the reader how to do something, or providing direct answers to how-to or problem-solving questions. How-to guides, tutorials, recipes, technical instructions, troubleshooting steps, and Q&A content where the answer provides information or guidance in response to a question. Recipe components such as ingredient lists, quantities, and preparation notes are "instructing" — they are part of the instructional act of telling the reader how to make something.

- *opinionated* — expressing subjective views, evaluations, arguments, complaints, praise, or commentary where the evaluative stance is the primary purpose of the line. Includes text that uses narrative framing primarily to support an evaluative point (when an anecdote serves to complain, critique, praise, or argue, the function is "opinionated"). Also includes short expressive utterances like "Thanks!", "Great post!", "Agreed" — these express a stance or sentiment, not information. Signals include: sustained value judgments, subjective adjectives ("essential", "crucial", "best"), rhetorical questions, prescriptive statements ("you should", "it's a must"), and recommendations. If text argues what should be done or advocates a position, it is "opinionated" even if the topic is factual.

- *promoting* — selling, advertising, or marketing a product, service, brand, or organization. This includes: explicit ads and calls to action; business pages describing their own services or qualifications to attract clients; SEO-style content stuffed with keywords for commercial purposes; fundraising appeals and donation requests; text where the author represents an organization and frames its offerings positively to potential customers. "Promoting" does not require explicit "buy now" language — if the text's purpose is to attract clients or donors, it is promoting.

- *lyrical* — poetry, song lyrics, verse, or artistic literary expression presented as primary content. A poem or song quoted within a review or critical essay should be classified by the outer framing purpose, not as lyrical.

Output format Return ONLY a JSON array with one object per line in [CLASSIFY THESE LINES]. Each classified line gets exactly one JSON object. No markdown fences, no commentary, no extra text.
Example:

```
[{"line":1,"source":"written","
    interactive":false,"function":"
    describing"},
{"line":2,"source":"cannot_rate","
    interactive":null,"function":null}]
```