

Towards more reliable Web metadata: Do register classifiers learn register or genre features?

Antti Kanner, Tuomas Lundberg, Erik Henriksson, Veronika Laippala

TurkuNLP, University of Turku

antti.kanner@helsinki.fi

{tuomas.lundberg, erik.henriksson, veronika.laippala}@utu.fi

Abstract

Transformer-based web register classifiers are widely used to provide metadata for web crawls and to focus web-data sampling on specific register classes, such as news articles or advertisements. However, it remains unclear what these classifiers actually learn: do they detect pervasive linguistic features characteristic of registers, or do they rely on structural cues more indicative of genre? We examine this question through two hypotheses: the *text-macrostructure hypothesis* proposes that classifiers rely on structural features more typical of genres, whereas the *feature-density hypothesis* predicts that they identify typically register-associated features. Using English, Finnish, and French data from HPLT 3.0, we segment documents into thirds and compare classifier predictions for full documents and segments through five complementary analyses: main-label agreement, cosine similarity, confusion profiles, SACX keyword-weight correlations, and SACX keyword stability. As a more direct control, we compare predictions for a sample of untruncated documents with predictions for the same documents after their sentence order has been randomised. Top-label assignments and prediction vectors remain highly similar after shuffling, although the positional gradients are attenuated, indicating that sentence order contributes to classifier behaviour without determining its broader prediction profile. Overall, the results are more consistent with the feature-density hypothesis, although both hypotheses retain explanatory value.

1 Introduction

Web register and genre classifiers are already widely used to provide metadata for large web corpora such as HPLT and CLASSLA-web (Burchell et al., 2025; Pungertsek et al., 2026). This metadata supports the efficient use of heterogeneous web corpora by enabling informed sampling of their contents. Such sampling is beneficial in several ap-

plications, including the selection of training data for large language models (Myntti et al., 2025).

Automatic web-register identification has also advanced substantially in recent years. For example, the MultiCORE classifiers, developed using the Multilingual CORE corpora, achieve near-human performance (Henriksson et al., 2025). As these classifiers become more accurate and their outputs are used more widely, it becomes increasingly important to understand what they learn. From a practical perspective, this is necessary for evaluating the validity of the metadata they produce. From a theoretical perspective, automatic classification provides a new setting in which to examine the complex relationship between register and genre (Kuzman and Ljubešić, 2025; Laippala et al., 2022; Henriksson et al., 2025).

Register and genre are theoretically distinct notions, but they are not always explicitly defined in contrast to each other. Their boundaries also vary across theoretical frameworks. Consequently, their practical application may involve considerable analytical overlap, and register and genre categories may be strongly correlated in annotated data.

In this article, we approach this distinction by contrasting two hypotheses about the kinds of features on which a classifier bases its predictions. The first concerns pervasive features associated with registers, whereas the second concerns macrostructural properties associated with genres. We assume that a transformer-based classifier will exploit the most readily available structure in the data that corresponds to the categories on which it has been trained. Determining which kind of structure is more readily available to the classifier is therefore relevant not only to the present model but also to automatic text classification more broadly.

We therefore ask to what extent a multilingual web-register classifier relies on features that are pervasive across document positions, as opposed to features associated with the positional struc-

ture of texts. We test two competing hypotheses about what the MultiCORE classifier learns. The *feature-density hypothesis* proposes that the classifier identifies varying densities of register-typical features distributed throughout texts. The *text-macrostructure hypothesis*, by contrast, proposes that the classifier relies on cues associated with text macrostructure and therefore focuses more strongly on genre-typical than register-typical features.

This experimental design cannot categorically determine whether the classifier learns registers or genres. Pervasive and macrostructural features do not map exclusively onto either concept, and the distinction between register and genre varies across theoretical frameworks. Instead, the analysis characterises the distributional organisation of the evidence used by the classifier. In doing so, it captures one commonly recognised distinction between registers and genres without treating that distinction as exhaustive.

Our analysis tests the two hypotheses by dividing each document into three segments—the beginning, middle, and end—and examining prediction consistency across these positions. We compare the predictions for each segment with those obtained for the corresponding full document. In addition, we analyse the stability of linguistic features across the three segments using SACX keywords (Rönqvist et al., 2022). These keywords identify linguistic features that make important contributions to the classifier’s predictions and therefore provide complementary evidence for evaluating the two hypotheses.

2 Background

In linguistics, register and genre are among the most common analytical frameworks for characterising and classifying texts and, more broadly, written communication (Lee, 2001; Biber and Conrad, 2019). The division of labour between the two concepts is not always clear-cut. Different theoretical frameworks define them in different terms, dispense with the distinction altogether, or favour one concept over the other.

One common distinction associates registers with functionally interpretable linguistic characteristics that are pervasive, meaning that they occur throughout a text. The sets of linguistic features associated with particular registers are often referred to as their repertoires (Biber and Conrad, 2019). Genres, by contrast, are commonly described in

terms of their macrostructural organisation. This organisation may be characterised as a sequence of functionally distinct stages, sometimes—but not always—analysed as moves (Swales, 1990). The contrast is compatible with the text-linguistic approach presented by Biber and Conrad (2019), in which register refers to situationally and culturally defined varieties of language, such as news articles, advertisements, and student essays.

This theoretical distinction between register and genre raises a corresponding question about what automated register classifiers learn. Previous research has shown that register classifiers trained only on the beginnings of web documents can perform as well as classifiers trained on whole texts (Laippala et al., 2023). This result admits at least two interpretations. The classifiers may identify registers through features that occur pervasively across documents. Alternatively, they may learn cues associated with text structure, such as opening sequences characteristic of particular genres (Biber and Conrad, 2019; Swales, 1990; Dijk, 1980; Martin and Rose, 2005).

If register and genre categories correlate in the training data, a classifier may perform equally well regardless of which of these types of evidence it has learned to use. Predictive performance alone therefore does not reveal whether the classifier relies primarily on pervasive register-associated features or positionally organised macrostructural cues. From the perspective of classifier development, however, it is important to understand what kinds of features these models rely on and with which theoretical constructs their outputs align most closely.

3 Data and register classifier

For this study, three samples of 100,000 documents were drawn from the English, Finnish, and French components of HPLT 3.0 (Oepen et al., 2025). To focus the analysis on relatively coherent texts, sampling was restricted to documents with a Word Document Score (WDS) of 9 or 10. WDS is a document-quality measure based on features associated with suitability for language-model training, including linguistic consistency and the structural cohesion and integrity of the document (Oepen et al., 2025). Each document was retained as a full-document representation and additionally divided into approximately equal beginning, middle, and final thirds according to whitespace-delimited word count. The classifier was applied separately

to the full-document representation and each of the three segments.

We also constructed an additional control sample from the 100,000-document English sample to assess the effect of document truncation and test how classifier predictions changed when sentence order was randomised. Candidate documents were restricted to 1,000–1,500 whitespace-delimited words. This provided documents with sufficient internal extent to form three substantial segments while reducing the likelihood that they would exceed the classifier’s 2,048-token input limit. Approximately 20,000 documents in the English sample met this word-count criterion, from which we randomly sampled 10,000. Token lengths were calculated using the FacebookAI/xlm-roberta-large tokenizer employed during prediction. Of the sampled documents, 658 exceeded the 2,048-token limit and were excluded. A further four documents were excluded because fewer than three sentences were detected, leaving a final paired control sample of 9,338 documents.

Sentence boundaries in the control sample were detected using the rule-based sentencizer in a blank English spaCy 3.8.16 pipeline (`spacy.blank("en")` with the sentencizer component). Each retained document was processed in two conditions. In the original-order condition, sentences were retained in their original sequence; their concatenation formed the full-document representation, and the sequence was divided into approximately equal beginning, middle, and final thirds according to sentence count. In the shuffled condition, the complete sentence sequence was randomly permuted once. The same permutation was used to construct both the shuffled full-document representation and its three successive thirds. The original and shuffled full documents therefore contained the same sentences and differed only in sentence order. Document identifiers and record order were preserved across conditions, enabling paired document-level comparisons.

All texts and segments were classified using the MultiCORE register classifier based on BGE-m3-retromae (Chen et al., 2024; Henriksson et al., 2025). BGE-m3-retromae is based on XLM-R (Conneau et al., 2020), has a maximum input length of 2,048 tokens, and has achieved high performance in register-identification tasks (Henriksson et al., 2025). The MultiCORE classifier performs multilabel classification using a simplified version

of the CORE scheme (Egbert et al., 2015; Laippala et al., 2022), comprising nine main registers and 16 sub-registers. The main registers are broad functional categories such as Narrative, Informational Persuasion, and How-to/Instructions, whereas the sub-registers are more specific categories such as Narrative blog and Recipe. The classification guidelines are available at <http://turkunlp.github.io/register-annotation-docs>.

In the main sample, full-document representations exceeding the model’s 2,048-token input limit were truncated during classification. This affected 15.5% of the English documents, 14.9% of the Finnish documents, and 12.8% of the French documents. Consequently, full-document predictions for these cases disproportionately represent document beginnings, which may increase agreement between the full-document and beginning-segment predictions. We account for this asymmetry when interpreting the main results. The English control sample addresses it directly because all retained full documents fall within the model’s input limit.

A register label was assigned whenever its predicted probability exceeded the optimised threshold of 0.35. Because the task is multilabel, a document can receive more than one register label. Documents commonly receive both a main-register label, such as Narrative, and a more specific sub-register label, such as News article.

As an informal validation check, we compared the classifier’s main-register predictions against manual annotations for 717 texts from the HPLT sample. The classifier achieved an accuracy of 63.6% and a macro-averaged F1 score of 52.0%. These figures are not directly comparable with the near-human performance reported by Henriksson et al. (2025), as the present evaluation was not constructed as a formal benchmark. It was used only to establish that the classifier captured a meaningful register signal in the present sample before its behaviour was examined comparatively. Accordingly, the analyses below characterise the behaviour of this classifier on the present HPLT samples rather than the behaviour of register classifiers generally.

4 Methods for hypothesis testing

Our primary question is whether the classifier relies predominantly on register-associated features that recur throughout texts or on features associated with their sequential macrostructural organisation. We refer to these alternatives as the

text-macrostructure hypothesis (1) and the feature-density hypothesis (2). The analyses do not provide a categorical test of whether the classifier learns genres or registers. Rather, they characterise the positional distribution of the evidence used by the classifier and assess which hypothesis provides the better account of the observed patterns.

The comparison assumes that beginnings, middle sections, and endings tend to perform different functions within the organisation of extended texts. A classifier relying strongly on sequential macrostructure should therefore exhibit positional differentiation in its predictions or attributed features. By contrast, if the classifier relies predominantly on features that recur throughout a text, its behaviour should remain comparatively stable across positions. This does not require complete positional invariance: the density and salience of pervasive features may vary over the course of a document even when the classifier relies on a broadly shared repertoire.

The beginning–middle–end division is used as a coarse, cross-register proxy for positional macrostructure. It does not identify genre-specific moves or functional stages. Manually annotating such units would require different structural schemes for research articles, narratives, recipes, advertisements, discussions, and the other categories represented in the data; units produced under these schemes would not necessarily be comparable across registers. Such an analysis would answer the narrower question of whether the classifier recognises predefined structural units within particular genres, rather than the present question of whether its predictions generally depend on sequential document organisation. The positional division therefore provides a theory-neutral basis for comparison across the heterogeneous sample. The design rests on the assumption that, if reliance on sequential macrostructure is sufficiently systematic to explain the classifier’s register predictions across this sample, it should produce aggregate differences among broad document positions. Finer genre-specific effects may remain undetected when their locations vary or when they occur within individual thirds, and the conclusions are limited accordingly.

We evaluate the hypotheses using five complementary tests:

1. agreement between full-document and segment main-register labels;

2. cosine similarity between full-document and segment-level prediction vectors;
3. confusion patterns between full-document and segment-level main-register predictions;
4. Spearman correlation of SACX keyword weights across document segments; and
5. overlap of SACX keyword sets across document segments and full documents.

Label agreement and cosine similarity measure the consistency of segment-level predictions with the corresponding full-document prediction. Because main-sample full-document inputs exceeding 2,048 tokens are truncated, these comparisons use a reference that disproportionately represents document beginnings in 12.8–15.5% of cases. The English untruncated control sample is therefore used to repeat the main-label agreement and cosine-similarity analyses under original and shuffled sentence order. The confusion matrices provide a more detailed view of whether positional disagreement involves stable or position-specific relationships between register categories. Finally, the two SACX analyses examine the positional stability of the features attributed importance by the classifier.

Under the text-macrostructure hypothesis, we expect comparatively strong positional differentiation: lower agreement between full-document and segment predictions, position-specific confusion patterns, weaker correlations between attributed-feature rankings, and less overlap between keyword sets. Under the feature-density hypothesis, we expect prediction patterns to remain broadly consistent and attributed features to exhibit a substantial shared core across positions. These expectations, summarised in Table 1, are comparative interpretive criteria rather than categorical outcomes required by either hypothesis.

5 Results

5.1 Main-label agreement

This test measures how many of the highest predicted register labels assigned to the full documents are also assigned to the document segments. In order to focus the analysis on the broader, functional patterns and limit the number of comparisons, this test focuses on the nine main register labels.

Table 2 shows that the overlap of register labels assigned to full documents and to segment

Test	Text macrostructure	Feature density
Label agreement	Lower overlap and stronger positional differences	High overlap and modest positional differences
Cosine similarity	Lower similarity and stronger positional differences	High similarity and modest positional differences
Confusion patterns	Position-specific reorganisation	Broadly stable patterns
SACX correlation	Weaker correlations between feature rankings	Stronger correlations within the shared feature core
SACX stability	More position-specific keywords	Substantial keyword overlap across positions

Table 1: Expected patterns under the text-macrostructure and feature-density hypotheses. These expectations are comparative rather than categorical.

	Beginning	Middle	End
Finnish	78.49%	71.62%	69.54%
French	73.98%	70.00%	68.93%
English	77.98%	72.20%	69.61%

Table 2: Label overlap between full documents and segments. Values are means. 95% CI is attained with $\pm 0.3\%$ for each item

labels ranges from 69.54–78.49%. The progression is monotonic: the overlap is the highest in the beginnings and decreases consistently towards the endings. The overlap with full documents and beginnings is exaggerated by the classifier truncation procedure as they might overlap fully in the case of longer documents. With all languages, the change is gradual and predictable, rather than abrupt. The results correspond better with hypothesis (2) than (1), as the differences are moderate and rather show gradual weakening of the register signal than markedly different profiles for each segment, even after taking into account the possible effect of truncation.

The main-label agreement analysis was repeated on the untruncated English control sample under original and shuffled sentence order (Table 3). In the original-order condition, agreement between full-document and segment predictions declined from 83.77% for beginnings to 75.90% for middle sections and 72.46% for endings. After sentence shuffling, the positional gradient was substantially

Condition	Beginning	Middle	End
Original order	83.77%	75.90%	72.46%
Shuffled order	85.51%	80.62%	79.99%

Table 3: Agreement between the highest-scoring register labels assigned to full documents and their corresponding segments in the untruncated English control sample ($n = 9,338$). All 95% bootstrap confidence intervals have margins below 0.94 percentage points.

attenuated: agreement was 85.51% for beginnings, 80.62% for middle sections, and 79.99% for endings. The difference between beginning and final segments therefore decreased from 11.31 to 5.52 percentage points. Direct comparison of the original and shuffled full documents showed that 85.80% retained the same highest-scoring main-register label (95% CI [85.05%, 86.51%]). Sentence order therefore affects classification in a minority of cases and contributes to the stronger positional gradient found in naturally ordered texts, but most full-document classifications remain unchanged when macrostructural sequencing is disrupted.

5.2 Cosine similarity

The distribution of cosine similarities between the prediction vectors of full documents versus each of the document segments tells how similar the prediction patterns are overall regardless of which class is assigned as the main label. Especially in documents where the classifier assigns near equal probability to two register candidates, relying only on the main-label agreement might mask considerable underlying similarity of the prediction vectors.

All nine cosine similarity distributions (three languages times three segments) showed high agreement with regards to the shape of their distributions. All nine follow clear power law curve with the histogram peak in high similarity end bin. Thus, reporting averages for brevity is grounded. These are given in Table 4. The table shows that mean cosine similarities are very high overall. Consistent, gradual but modest decrease along text progression is visible on all languages. Despite the greater textual overlap between document beginnings and the truncated full-document inputs, cosine similarity with full-document predictions was only slightly higher for beginnings than for middle and final segments.

The cosine-similarity analysis was also supplemented with the sentence-shuffling control and it produced the same overall pattern as the main la-

	Beginning	Middle	End
Finnish	0.932	0.905	0.897
French	0.919	0.901	0.897
English	0.941	0.921	0.912

Table 4: Mean cosine similarity by segment position.

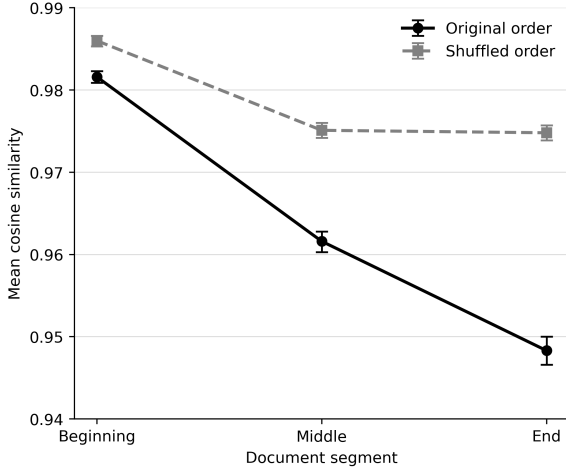


Figure 1: Mean cosine similarities between full-document prediction vectors and the beginning, middle and final thirds under the original and sentence-shuffled conditions. Error bars show 95% bootstrap confidence intervals with 2,000 resamplings.

bel agreement control analysis (Figure 1). The results also corroborate the patterns observed above. Reordering the sentences produced only a small change in the full-document prediction profiles: cosine similarity between the original and shuffled full-document vectors averaged .987 (95% CI [.986, .987]). Similarity between full-document and segment predictions also remained high in both conditions. In the original documents, however, mean similarity declined from .982 for beginnings to .962 for middle sections and .948 for endings. After sentences were shuffled and repartitioned, this positional gradient was substantially attenuated, with corresponding similarities of .986, .975 and .975. These results indicate that the classifier relies primarily on pervasive features distributed across the document, although naturally ordered texts appear to concentrate some diagnostically relevant evidence toward their beginnings.

5.3 Confusion profiles

Figure 2 reports the confusion profiles for each register in each segment in each language. The heatmap rows correspond to segment registers and

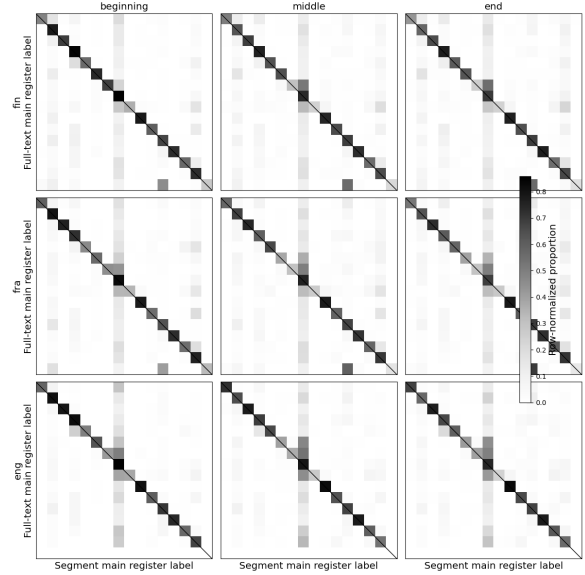


Figure 2: Confusion heatmaps for register predictions for full documents and segments.

columns to the full text registers.

The figure shows that in each heatmap, the diagonal line is strong, especially in the first registers. Some level of confusion is expected in the lower end for the sub-registers—these can be very detailed and easily confused, such as *encyclopedia article* and *description of a person or thing*.

All the heatmaps show strong vertical lines in the middle, in the same place. This is main register *opinion*, which captures registers like reviews, opinion blogs and advice pieces. The reason why this seems to be a strong fallback candidate for the classifier irrespective of language is difficult to assess. One explanation could be found from markers of evaluative language, which are crucial features of opinion registers, but can also be found in other registers (Koivunen et al., 2021).

Further, the figure clearly shows that there is no visible regime change in the later segments—the later segments are not systematically misinterpreted as texts belonging to other registers by virtue of their different structural properties. Quite the contrary, all segments show similar confusion profiles, with the gradual and modest loss of confidence visible as increase in confused mass, rather than its relocation. Further, there seems to be no major differences between the languages. All these findings suggest that feature density has the heavier impact on classifier behaviour.

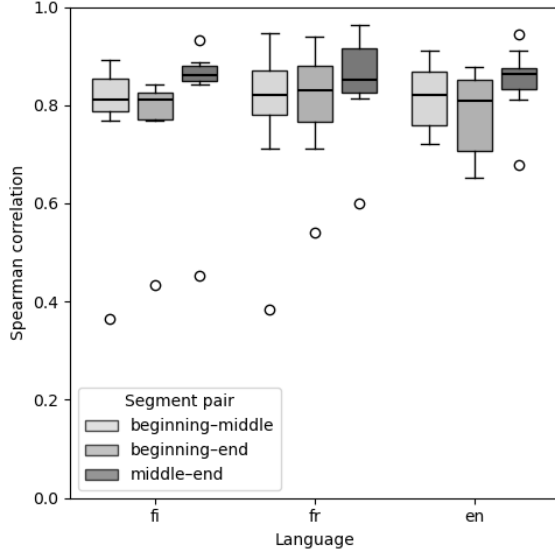


Figure 3: Spearman correlations of SACX feature weights between different text segments

5.4 SACX weight correlation across document segments

SACX (Rönnqvist et al., 2022) provides class-level keywords for text categories using the model explainability method Integrated Gradients (IG) (Sundararajan et al., 2017). For each document, and separately for each register predicted for it, IG accumulates the gradient of that register’s score along a straight path in the embedding space, running from a baseline input in which every non-special token has been replaced by a filler token to the actual input, and assigns the resulting attribution to each token. The word-level attributions are then averaged over documents, yielding an importance weight for each word for each modeled class, where larger positive values indicate a stronger contribution to the prediction. In practice, this creates a list of keywords and weights for each modeled text category.

Our setup departs from the original SACX formulation in two respects: we use a single classifier rather than several independently trained ones, and the keyword lists are calculated with a 0.5 prediction threshold rather than the optimised 0.35 used for the label assignments reported elsewhere in this paper. Appendix A gives the details, along with the attribution settings.

For each language and main register, we selected words that occurred in at least 50 documents and had a positive mean attribution in all three segments. We ranked these shared-positive words by

their mean SACX attribution separately in beginnings, middle sections, and endings, and calculated pairwise Spearman correlations between the three rankings.

The results of the analysis are presented in Figure 3. These show how similar or dissimilar the SACX weights are across the three document segments.

Overall, the correlations are very high. This means that the rankings of most important keywords across the different segments are essentially the same. The boxplots show the variation between the main registers. The low end outlier, especially prominent in Finnish and French, is Informational Description (IN).

5.5 SACX keyword stability across document segments

For this analysis, we used the keyword lists produced by SACX for each main register and language. The documents of each language and segment were processed in successive groups of 600, and a word was retained as a keyword for a register only if it was attributed to that register in at least 60% of those groups. This is the stability criterion for which the method is named: it keeps the words that are important across the whole set of documents and discards those that matter in only a few of them. We kept at most the 200 highest-scoring keywords per register. We call a keyword *pervasive* if it is selected in all three segments and *position-specific* if it is selected in exactly one. As an additional check, we compared the pervasive keywords against the keyword lists extracted from the full documents.

Table 5 reports the results. Averaged over the eight main registers with non-empty keyword lists, 42% (English), 40% (Finnish) and 46% (French) of the keywords are pervasive, and of these 71%, 65% and 60% respectively also occur among the full-document keywords. For the six registers whose segment lists are truncated at the 200-keyword limit these shares are lower bounds, since a keyword can fail the pervasiveness test simply by falling below the 200-word cut-off in one segment. The untruncated SP (Spoken) lists reach 73% (English) and 76% (French). The two analyses rank the registers differently — ID (Interactive Discussion) has the lowest pervasive share here, whereas the rank correlations place IN (Informational Description) lowest — which reflects the different sensitivities of set overlap and rank correlation rather than a

Register	English		Finnish		French	
	3/3	$\geq 2/3$	3/3	$\geq 2/3$	3/3	$\geq 2/3$
IN	40	59	29	56	40	62
NA	42	67	40	62	51	70
OP	47	69	44	70	44	67
HI	30	52	40	62	31	60
IP	41	66	50	67	46	69
MT	50	69	47	63	45	64
SP	73	89	55	82	76	86
ID	15	56	15	43	36	72
Mean	42	66	40	63	46	69

Table 5: Keyword pervasiveness % by language and main register. 3/3 and $\geq 2/3$ denote the number of segments where the keywords were identified. For the first six registers, all three segment lists are truncated at the 200-keyword limit, so the 3/3 shares are lower bounds. The SP (Spoken) lists are untruncated in every language. LY (Lyrical) is omitted, as its keyword lists are empty.

substantive disagreement.

The position-specific keywords are distributed in the way the test setup anticipates. Excluding ID (Interactive Discussion), whose segment lists differ substantially in length, beginning-only keywords account for 39% (English), 42% (Finnish) and 40% (French) of all position-specific keywords, middle-only keywords for 30%, 27% and 27%, and end-only keywords for 32%, 32% and 34%. The middle sections thus contribute the fewest position-specific keywords in all three languages, which is consistent with the assumption that beginnings and endings carry more structural signal in a text compared to middle parts.

Relaxing the criterion from three segments to at least two changes the picture substantially. Keywords stable in at least two of the three segments account for 66% (English), 63% (Finnish) and 69% (French) of all keywords, so around two thirds of what the classifier relies on is shared by most of a document rather than being confined to one part of it.

The keywords steadily identified in exactly two segments are not, however, merely borderline cases of pervasiveness: disproportionately often, the segment they are missing from is the beginning. Restricting this count to the six registers whose segment lists are all truncated at the same 200-keyword limit, so that differences in list length cannot bias the comparison, the missing segment is the beginning for 45% of these keywords, the end for 32%, and the middle for only 23%. Middle and end sections thus share vocabulary more closely with each

other than either does with beginnings, and beginnings carry additional distinctive material on top of a shared core.

All of these observations point at the same conclusion. The text macrostructure hypothesis would predict that each part of a document signals its register by its own distinct means. What we see instead in our segment comparison is that there is a substantial shared core running through the whole document, with beginnings carrying additional material, being structurally the most marked part.

Qualitatively, the pervasive keywords are register vocabulary. For NA (Narrative), the top keywords contain clear register markers such as *blog*, *post*, *said*, *yesterday*, *comment*, *readers*, *news*, and *announced*, together with weekday names; for HI (How-to / Instructions), words like *recipe*, *tips*, *instructions*, *guide*, *steps*, *method*, and *follow*; IP (Informational Persuasion) includes *clients*, *product*, *prices*, *customers*, *affordable*, *offer* and *guarantee*; SP (Spoken) has *tell*, *yes*, *talk*, *say*, *feel*, *think*, and *yeah*. The position-specific keywords are by contrast recognisably macrostructural: the end-only keywords of IN (Informational Description) include *disclaimer*, *conclusion* and *methodology*, those of IP (Informational Persuasion) *consultation*, *appointment*, *booking* and *pricing*, and those of HI (How-to / Instructions) *finally*, *advice* and *tip*. The keywords named here are illustrative; Appendix B gives the ten highest-scoring keywords of each class for every main register and all three languages, without applying manual selection.

The above results provide the linguistic counterpart to the quantitative results presented in Sections 5.1–5.5, though the picture here is more mixed than in those sections. Under the strict criterion requiring a keyword to be selected in all three segments, a minority of retained keywords (40–46%) qualify as pervasive, and more than half are therefore position-specific to some degree — a pattern that on its own would be at least as consistent with the text-macrostructure hypothesis as with feature density. Two further observations tip the balance towards feature density. First, relaxing the criterion to keywords stable in at least two of three segments raises the shared share to 63–69%, indicating that most position-specific keywords under the strict criterion are missing from only one segment rather than confined to a single position — closer to noise at the margins of a shared core than to distinct per-position vocabularies. Second, the qualitative content of the keywords is asym-

metric in a way the text-macrostructure hypothesis does not predict: pervasive keywords are overwhelmingly register vocabulary (blog, recipe, tips, clients, product), while genuinely structural markers (disclaimer, conclusion, booking, pricing) are concentrated in the smaller position-specific set, and disproportionately at document endings. Taken together, these results support the feature density hypothesis as the better fit to the SACX evidence, though the strict-criterion overlap figures alone do not settle the question.

6 Conclusions

The results are more consistent overall with the feature-density hypothesis (2) than with the text-macrostructure hypothesis (1), although both account for part of the observed behaviour. Across English, Finnish, and French, the classifier's predictions remained broadly similar across document positions, while agreement with full-document predictions declined gradually from beginnings to middle and final segments. This pattern suggests attenuation of a shared register signal rather than a categorical change in the kinds of register assigned to different parts of a document.

The prediction-based analyses supported this interpretation. Main-label agreement and cosine similarity declined progressively over the course of documents, but cosine similarities remained high throughout. Confusion profiles were also broadly stable across positions: disagreement increased in later segments, but the alternative labels followed similar distributions rather than showing systematic position-specific reorganisation. These findings indicate that document position affects prediction consistency without producing a general shift to qualitatively different register profiles.

The SACX analyses produced a more mixed but compatible pattern at the feature level. Keyword rankings were strongly correlated across positions, indicating a stable core of features attributed importance by the classifier. Under the strictest criterion, only 40–46% of retained keyword types occurred in all three segments, a result that alone is at least as consistent with positional macrostructure as with pervasive feature density. However, 63–69% occurred in at least two segments, and the qualitative distribution was asymmetric: pervasive keywords predominantly represented register-associated vocabulary, whereas recognisably structural markers were concentrated in the smaller position-specific

set, particularly at document endings. The SACX evidence therefore indicates a substantial shared feature core alongside genuine positional variation.

The sentence-order control provided a more direct test of the role of sequential organisation. In the untruncated English control sample, the same highest-scoring main-register label was assigned to the original and shuffled full documents in 85.80% of cases, while their prediction vectors had a mean cosine similarity of .987. Shuffling also reduced the beginning-to-end gradient in main-label agreement from 11.31 to 5.52 percentage points. Cosine similarity displayed the same attenuation: similarities between full documents and their beginning, middle, and final segments were .981, .961, and .948 in original order, compared with .986, .975, and .975 after shuffling. Sentence order therefore contributes to positional differences and changes the top prediction in a minority of documents, but disrupting document-level sequencing generally leaves the classifier's broader prediction profile intact.

Taken together, the results suggest that the MultiCORE classifier relies primarily on register-associated evidence distributed across documents while also exploiting positional and structural cues where these are available. The findings do not establish that register classifiers generally learn register rather than genre, nor do they exclude finer genre-specific structural effects that are not captured by the three-way positional division. They instead characterise the behaviour of this classifier on the present HPLT samples: its decisions are predominantly supported by a recurring feature repertoire, with sequential macrostructure providing a secondary contribution.

Limitations

The study operationalises sequential macrostructure through a three-way positional division rather than through annotation of genre-specific moves or functional stages. This makes comparison possible across heterogeneous register categories without requiring a separate structural framework for each category, but it cannot detect structural variation occurring within individual thirds or effects whose positions vary substantially across documents. The conclusions therefore concern broad positional organisation rather than all possible forms of genre structure. A study using manually annotated structural units would address the complementary but

narrower question of whether the classifier recognises predefined stages within particular genres.

In the main multilingual sample, full-document inputs exceeding 2,048 tokens were truncated during classification. This affected 15.5% of the English documents, 14.9% of the Finnish documents, and 12.8% of the French documents, causing their full-document predictions to represent document beginnings disproportionately. Comparisons with full-document predictions may therefore favour beginning segments. The untruncated English control sample addresses this confound directly and reproduces the overall pattern, but it does not remove the asymmetry from the main multilingual analysis.

The sentence-order control is also limited to English documents containing 1,000–1,500 whitespace-delimited words and meeting the study’s high document-quality criterion. Its results may not generalise fully to shorter, noisier, or non-English web documents. Furthermore, the classifier achieved an accuracy of 63.6% and a macro-averaged F1 score of 52.0% against the available manual annotations. The study consequently characterises the distribution and stability of the evidence used by this particular classifier; it does not establish that its predictions are correct or that other register classifiers rely on the same features.

Acknowledgements

This work was supported by the Research Council of Finland through several projects: FINCLARIAH 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), 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

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

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,

Bhavivyva 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.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. *Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation*. Preprint, arXiv:2402.03216.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. *Unsupervised cross-lingual representation learning at scale*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Teun Adrianus van. Dijk. 1980. *Macrostructures : an interdisciplinary study of global structures in discourse, interaction, and cognition / Teun A. van Dijk*. Lawrence Erlbaum Associates, Hillsdale (N.J.).

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.

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

Anu Koivunen, Antti Kanner, Maciej Janicki, Auli Harju, Julius Hokkanen, and Eetu Mäkelä. 2021. *Emotive, evaluative, epistemic: A linguistic analysis of affectivity in news journalism*. *Journalism*, 22(5):1190–1206.

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

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.

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo. 2022. *Towards better structured and less noisy web data: Oscar with register annotations*. In *Proceedings of the Eighth Workshop on*

Noisy User-generated Text (W-NUT 2022), pages 215–221, Gyeongju, Republic of Korea. Association for Computational Linguistics.

David Y. W. Lee. 2001. *Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle*. *Language Learning & Technology*, 5(3):37–72.

J. R. Martin and David. Rose. 2005. *Genre relations : mapping culture*. Equinox textbooks and surveys in linguistics. Equinox Pub, London ;.

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 *Proceedings of the Second Conference on Language Modeling (COLM 2025)*.

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.

Taja Kuzman Pungeršek, Peter Rupnik, Vít Suchomel, and Nikola Ljubešić. 2026. *The Growing Gains and Pains of Iterative Web Corpora Crawling: Insights from South Slavic CLASSLA-web 2.0 Corpora*.

Samuel Rönqvist, Aki-Juhani Kyröläinen, Amanda Myntti, Filip Ginter, and Veronika Laippala. 2022. *Explaining classes through stable word attributions*. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1063–1074, Dublin, Ireland. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. *Axiomatic attribution for deep networks*. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

John Swales. 1990. *Genre analysis : English in academic and research setting / John M. Swales*. Cambridge applied linguistics series. Cambridge University Press, Cambridge.

A SACX configuration

Attributions were computed at the embedding layer using ten integration steps, summed over the embedding dimensions, and normalised per document. Sub-word attributions were merged to the word level by retaining the largest absolute value within each word.

Two properties of the setup should be taken into account when comparing our results to other applications of SACX. First, in its original formulation, SACX assesses the stability of the attributions

across several independently trained classifiers. Here, we use a single classifier, and stability is instead assessed with respect to variation across the document groups described in Section 5.5. The criterion therefore controls for document-level rather than model-level variation. Second, attributions were computed for every register whose prediction probability exceeded 0.5, which is stricter than the optimised threshold of 0.35 used for the label assignments reported elsewhere in this paper. In other words, the keyword lists are created using more confident predictions.

B Keyword lists

Tables 6–8 give the ten highest-scoring keywords in each of the three classes discussed in Section 5.5 — pervasive, beginning-only, and end-only — for every main register and each of the three languages. Keywords are ordered by mean SACX importance and no manual selection was applied, so these lists show the classes as the method returns them.

Register	Pervasive (all three segments)	Beginning-only	End-only
IN	abstract, appointed, references, employed, article, characters, authors, writers, allowing, diagnosis	episode, here's, contents, nowadays, discusses, phenomenon, definitions, bound, entrepreneurs, excellence	therapies, disclaimer, yes, litigation, gov, conclusion, html, legacy, associates, mentioned
NA	blog, post, said, yesterday, comment, writer, added, comments, photos, readers	blogger, blogging, blogs, writers, jan, week's, posting, appointed, bbc, baseball	concluded, bridge, scene, crew, videos, media, johnson, highlight, mike, match
OP	review, reviews, blog, post, writers, writer, album, episode, book, reader	yesterday, october, april, november, folks, recent, morning, dear, guide, writing	rating, amen, corruption, wise, tips, authors, drama, reminds, magazine, artist
HI	recipe, tips, instructions, guide, note, steps, step, method, try, ways	we, found, getting, home, taking, making, better, i, give, few	advice, tip, ingredients, finally, techniques, mix, helpful, useful, fun, carefully
IP	book, today, contact, clients, product, prices, price, customers, affordable, client	author, sale, stock, incredible, we've, course, recommend, days, reasons, master	consultation, appointment, specialist, booking, guaranteed, pricing, we'll, packages, package, proven
MT	winning, customer, mobile, video, device, app, post, website, software, quality	via, requirements, improve, penis, sort, betting, training, single, latest, choose	user, solution, enterprise, policy, european, china, version, collection, difference, digital
SP	tell, yes, you, talk, what, say, your, feel, us, experience	career, today, her, she	thank, next, future, best, day
ID	question, think, i'm, sure, help, like, am, really, my, i	been, all, even, time, for, be, how, out, from, make	luck, yeah, pm, advice, wow, questions, hey, game, personally, comments

Table 6: English: the ten highest-scoring keywords in each class, per main register, ordered by mean SACX importance. No manual selection was applied. LY (Lyrical) is omitted, as its keyword lists are empty.

Register	Pervasive (all three segments)	Beginning-only	End-only
IN	syntynyt, yhteystiedot, tietojen, helsingissä, henkilötietoja, nimellä, yksikkö, vaihtelee, syntyi, suhde	properly, tausta, ilmiö, ominaisuuksien, function, opt-out, as, lasketaan, necessary, peräisin	tehtävästä, etc, functionalities, henkilötiedot, evästeitä, advertisement, julkaisuja, osoite, lisätietoa, yrityksille
NA	toteaa, ottelussa, sanoo, kertoo, ottelua, kommentoi, ajattelin, ottelun, johtaja, blogissa	otteluun, haastattelussa, maanantai, ihmettelin, kuvittelin, kansanedustaja, törmäsin, lauantai, sunnuntai, käsiteltiin	kommentti, uutishuoneessa, blogspot, kirjoittaja, suunnitelmia, kommentit, postauksia, yötä, instagramissa, kerrottiin
OP	suosittelen, elokuva, raamatun, elokuvan, jeesus, teksti, esimerkki, jumala, uskon, kirjassa	viikolla, viikon, sanotaan, luin, näkökulmasta, perustuu, kävi, kertaa, olemme, toisin	kommentti, suositella, kommentit, kirjoittaja, blogin, kommentoi, ps, hauska, lähetä, lukija
HI	ohjeet, sekoita, vinkkejä, dl, g, kannattaa, ohje, minuuttia, muista, mikäli	artikkelissa, tällä, selvittää, läpi, oma, tietää, tapa, pois, tehdä, olet	leikkaa, tutustu, löytyvät, kokeile, lisätietoja, kerro, paista, huomioida, ruokaa, seuraavan
IP	kysy, tutustu, asiakkaan, löydät, tarjoamme, hinta, suomalainen, saadaan, tervetuloa, tuote	sijaitseva, vaihtoehdon, hotelli, kerromme, käytämme, oletko, yhdistyvät, materiaalit, upeat, vastaa	kanssamme, helsingissä, ammatilainen, asiakkaillemme, mieltä, lisätietoa, asiakkaidemme, puh, suomeen, yrityksesi
MT	ilmoitus, yleens, kumppani, sovellus, valikoima, kokemus, pelejä, suunnitelma, pelit, tuhan-sia	osoite, merkkejä, sarjakuva, suurta, elää, kaupunki, talo, xvideos, kuva, club	valikoiman, palveluiden, kyky, mahdollisuudet, ssa, tekniikka, ohjelman, teidän, tietokoneen, arvostelu
SP	miten, olen, mitä, nyt, niin, on, paljon, kanssa, kun, oli	—	en
ID	kiitos, hei, tuossa, mieltä, minulla, tosiaan, juttu, d, takia, varmaan	tähän, minun, vuotta, täällä, lisäksi, tuli, silloin, viime, heti, tullut	viestiä, haluaisin, moikka, kysyä, moi, muistaakseni, aiheesta, toivon, lähetä, lienee

Table 7: Finnish: the ten highest-scoring keywords in each class, per main register, ordered by mean SACX importance. No manual selection was applied. LY (Lyrical) is omitted, as its keyword lists are empty.

Register	Pervasive (all three segments)	Beginning-only	End-only
IN	étude, organisme, espèces, personnage, article, missions, l'étude, appelée, document, caractéristiques	née, l'organisme, agit, appareil, chimique, appelés, appelées, trouverez, présentation, table	prenant, oui, notes, essentiels, références, source, conclusion, résumé, consultez, améliorer
NA	blog, souligne, jeudi, communiqué, interview, explique, expliqué, annoncé, déclare, merci	organise, news, post, accueilli, exposition, spectateurs, ouvre, profite, l'annonce, artiste	espère, thierry, conclut, remercier, emplois, ailleurs, document, auteur, professeur, champion
OP	album, avis, blog, roman, commentaire, commentaires, recommande, série, message, critiques	croire, marque, propose, vais, joueurs, sérieux, parle, utile, textes, joie	conseille, l'auteur, étude, avantage, écouter, témoignage, avenir, apprécié, facebook, approche
HI	recette, article, conseils, procéder, ingrédients, voici, utilisez, méthode, étape, recommandé	modifier, allons, ici, éléments, permet, connaître, correctement, lors, fonction, quelques	recommandons, conseil, n'oubliez, formulaire, aidera, alimentation, obligatoire, commentaires, mesures, mettez
IP	ouvrage, proposons, exposition, découvrez, avantages, contactez-nous, roman, retrouvez, accompagnement, contacter	auteur, lecteurs, innovant, accueil, article, innovants, bienvenue, invitons, paradis, l'esthétique	newsletter, acteur, logo, commentaire, l'agence, témoignages, l'opportunité, spectacles, dédiée, promo
MT	utilisateur, fonctionnalité, avantage, tenant, appareil, website, appareils, utilisation, objectif, emploi	étoiles, confortable, sentir, décoration, note, télécharger, maison, population, basé, essentiel	alimentation, more, achats, assurance, efficacité, microgaming, écran, exercice, privée, augmentation
SP	oui, comment, question, votre, quel, projet, ans, avez, premier, france	interview, questions, directeur, président, politique, pendant, vie	projets, quels, niveau, mettre, personnes, sera
ID	question, bonjour, merci, mon, j'ai, j'ai, me, suis, par, depuis	—	forum, répondre, sujet, avis, semble, effet, plutôt, sinon, ta, france

Table 8: French: the ten highest-scoring keywords in each class, per main register, ordered by mean SACX importance. No manual selection was applied. LY (Lyrical) is omitted, as its keyword lists are empty.