From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations

Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo, Veronika Laippala
TurkuNLP, University of Turku
{erik.henriksson, amanda.a.myntti, sherik, selcen.s.erten, aeeske, liina.t.repo, mavela}@utu.fi

Abstract

In corpus linguistics, registers—language varieties suited to different contexts—have traditionally been defined by their situations of use, yet recent studies reveal significant situational variation within registers. Previous quantitative studies, however, have been limited to English, leaving this variation in other languages largely unexplored. To address this gap, we apply a quantitative situational analysis to a large multilingual web register corpus, using large language models (LLMs) to annotate texts in English, Finnish, French, Swedish, and Turkish for 23 situational parameters. Using clustering techniques, we identify six situational text types, such as "Advice", "Opinion" and "Marketing", each characterized by distinct situational features. We explore the relationship between these text types and traditional register categories, finding partial alignment, though no register maps perfectly onto a single cluster. These results support the quantitative approach to situational analysis and are consistent with earlier findings for English. Cross-linguistic comparisons show that language accounts for only a small part of situational variation within registers, suggesting registers are situationally similar across languages. This study demonstrates the utility of LLMs in multilingual register analysis and deepens our understanding of situational variation within registers.

1 Introduction

Language varies with context as people adapt their linguistic choices to different situations. *Register variation* refers to the distinct forms of language functionally related to specific situations and communicative purposes (Biber, 1988, 2012; Biber and Conrad, 2019). In the text-linguistic approach to register analysis, the frequent use of linguistic features is assumed to be directly functional for the requirements of the situation (Biber and Egbert, 2023). As a result, text-linguistic register analyses typically start with situational descriptions of registers (e.g. Biber and Egbert, 2018, Section 2).

Nevertheless, register studies have traditionally focused on the linguistic features characterizing different registers, and much less attention has been given to analyzing the communicative situations in which texts are produced (Biber and Egbert, 2023). Furthermore, existing situational analyses often describe entire registers using the same categorical characteristics—such as medium, setting, communicative purpose, interactivity, and topic (Biber and Conrad, 2019)—which are then used to define register categories. These resulting classes are typically assumed to be situationally discrete.

Some recent studies, however, have provided strong evidence for register-internal situational variation (e.g. Gray, 2015; Biber et al., 2020; Egbert and Gracheva, 2023; Wood, 2024), casting doubts on the possibility of defining registers by any essential situational attributes. This has led to a reconceptualization of registers as *continuous* rather than discrete categories—categories that can be recognized but not strictly defined by linguistic features or situational context (Biber and Egbert, 2023). However, previous research on this variation has been limited to English texts, and its extent in other languages is largely unknown.

In this study, we address this gap by adopting the continuous approach to situational analysis introduced by Biber et al. (2020) and applying it to a large multilingual register-annotated corpus. In this framework, texts are coded for 23 parameters that capture situational variables such as purpose, background assumptions, and source of information, using an ordinal scale from 1 to 6. These annotations allow texts to be viewed within a continuous situational space and grouped into new *situational text types* based on their proximity within this space. The coding scheme was designed to capture the full range of situational factors identified in previous studies, including Biber (1994), Biber and Egbert (2018), and Biber and Conrad (2019).

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 308–318
November 16, 2024.

To annotate a large multilingual corpus for its situational characteristics, we apply a new approach: Instead of manually annotating the texts, we use multilingual large language models (LLMs) for the task. Specifically, we utilize GPT-4o-mini (OpenAI, 2024) and LLaMA 3.1 8B (AI@Meta, 2024) to annotate 8,406 texts from the register-labeled Multilingual CORE corpus (Henriksson et al., 2024) in English, Finnish, French, Swedish, and Turkish. We evaluate the LLM-generated annotations against each other and against a human-annotated sample corpus, demonstrating that the LLMs achieve good accuracy. By integrating these situational annotations with the texts' existing register and language labels, we conduct multilingual analyses on the relationships between situational context and register, as well as cross-linguistic comparisons of situational variation between registers.

Our analyses show that while registers are partially distinguishable by their situational characteristics, considerable register-internal variation exists across all included languages. These findings align with those reported by Biber et al. (2020) for English. We identify six situational text types—"Advice", "Information", "Marketing", "Personal", "Opinion", and "Speech"—each characterized by specific contextual features. These clusters partially align with established register categories but more often reveal situational overlap between registers. Moreover, our cross-linguistic comparisons show that language accounts for only a small portion of the total variance in each register, suggesting that the situational characteristics of registers are generally similar across languages. The code and data used in this study are available at `https://github.com/TurkuNLP/situational-analysis-llm`.

We start by describing the corpus and the LLM-based annotation process, including an evaluation of the LLM annotations against a human-labeled subcorpus. We then explore the situational variation within web registers and identify situational text types that emerge from the data. Next, we examine how these text types align with traditional register categories. Finally, we analyze cross-linguistic situational variation within registers.

2 The register-annotated CORE data

We utilize data from the Multilingual CORE corpus (Henriksson et al., 2024), a large manually register-annotated collection of unrestricted web content spanning 16 languages. The texts in the language subcorpora have been collected using different methodologies at different times.

	En	Fi	Fr	Sv	Tr	**Total**
News report	200	200	200	200	200	**1,000**
Description of a thing or person	200	200	200	200	124	**924**
Description with intent to sell	105	200	200	200	200	**905**
Other informational description	200	200	166	94	200	**860**
Narrative blog	200	200	200	200	52	**852**
Interactive discussion	200	200	200	118	50	**768**
Opinion blog	200	163	92	155	58	**668**
How-to or instructions	159	178	113	95	62	**607**
Encyclopedia article	64	104	132	200	18	**518**
Review	169	118	112	49	66	**514**
Sports report	200	166	66	39	30	**501**
Spoken	58	30	25	6	32	**151**
Lyrical	70	13	23	16	16	**138**
Total	**2,025**	**1,972**	**1,729**	**1,572**	**1,108**	**8,406**

Table 1: Composition of web register dataset.

For the English CORE, data was collected through Google searches targeting highly frequent English 3-grams (Egbert et al., 2015), and annotations were performed via Amazon Mechanical Turk, where each document was labeled by four coders, with a label assigned if at least two coders agreed. The Finnish corpus was sourced from a random sample of the Finnish Internet Parsebank (Luotolahti et al., 2015). The remaining subcorpora were derived from Common Crawl data, following the methodology described in Laippala et al. (2022), including steps such as sampling from various time periods, removing boilerplate content, and deduplication. All register annotations were made by trained experts, using a hierarchical taxonomy with 9 main categories and 16 subcategories.

In this study, we focus on the five largest language datasets in Multilingual CORE: English, Finnish, French, Swedish, and Turkish. We include 13 registers, listed in Table 1, based on the following criteria. Registers must have at least one example from each language and a minimum of 500 examples overall, except for the smaller *Spoken* and *Lyrical* registers, which are included for their situational distinctiveness. Secondly, we treat the *Spoken* and *How-to or instructions* categories as non-hierarchical, as their subcategories are small. For simplicity, we exclude texts with multiple labels or no label at all. Given the class imbalance in the original dataset across categories and languages (see Henriksson et al., 2024, Section 4.4), we randomly sample up to 200 examples from each language-register to balance the data while avoiding excessive downsampling. The resulting dataset is shown in Table 1.

3 Situational annotation using LLMs

Large language models (LLMs) (Brown et al., 2020) have emerged as powerful tools for textual analysis and annotation, with some studies suggesting that their accuracy can even surpass that of human annotators (e.g. Gilardi et al., 2023; Törnberg, 2023; Rathje et al., 2023). In this study, we experiment with two recent models—GPT-4o-mini and Llama 3.1 8B—for the situational coding task. We access GPT-4o-mini via the OpenAI API and deploy Llama 3.1 8B on the Mahti Supercomputer (CSC — IT Center for Science Ltd), using it with PyTorch through the HuggingFace Transformers library. For both models, we set the temperature to 0.01 for consistent responses.

We use the two LLMs to code the 8,406 documents for 23 situational parameters, as listed in Figure 1. For each text, we provide the first 5,000 characters as input to the models, along with the system prompt provided in Appendix A, which instructs the models to rate each parameter from 1 (completely disagree) to 6 (completely agree) based strictly on the given text. We also ask the models to briefly explain each scoring decision, which, in preliminary tests, significantly improved both models' performance. Both models generated the data in the requested output format without any issues.

We compare the inter-annotator agreement (IAA) of the LLM-generated annotations with each other and with a human-annotated sample consisting of 150 documents across all five languages, annotated by multiple human coders. The human annotators, all experts in the CORE label scheme, were given the parameters and documents without any additional guidance on how to annotate them and without being shown the texts' register labels. The results of these IAA evaluations are presented in Table 2.

	Kappa	Pearson's R	Support
GPT4-o-mini vs. Llama 3.1 8B (*full data*)	**0.73**	**0.76**	8,406
GPT4-o-mini vs. Llama 3.1 8B (*subset*)	**0.72**	**0.75**	150
Humans vs. GPT4-o-mini	0.50	0.56	150
Humans vs. Llama 3.1 8B	0.43	0.48	150
Biber et al. (2020)	0.46	0.52	1,002

Table 2: Inter-annotator agreement (IAA) scores.

The agreement between the two LLMs is strong across both the full dataset (8,406 documents) and the human-annotated subset (150 documents), with Cohen's kappa values of 0.72–0.73 and Pearson correlations of 0.75–0.76. In comparison, the agreement between the LLMs and human annotators is moderate (Kappa 0.43–0.50, Pearson's R 0.48–0.56), but these scores are similar to those reported by Biber et al. (2020) for human-made annotations of the same 23 parameters. GPT-4o-mini proves to be slightly more reliable than Llama 3.1 8B when compared to human annotations.

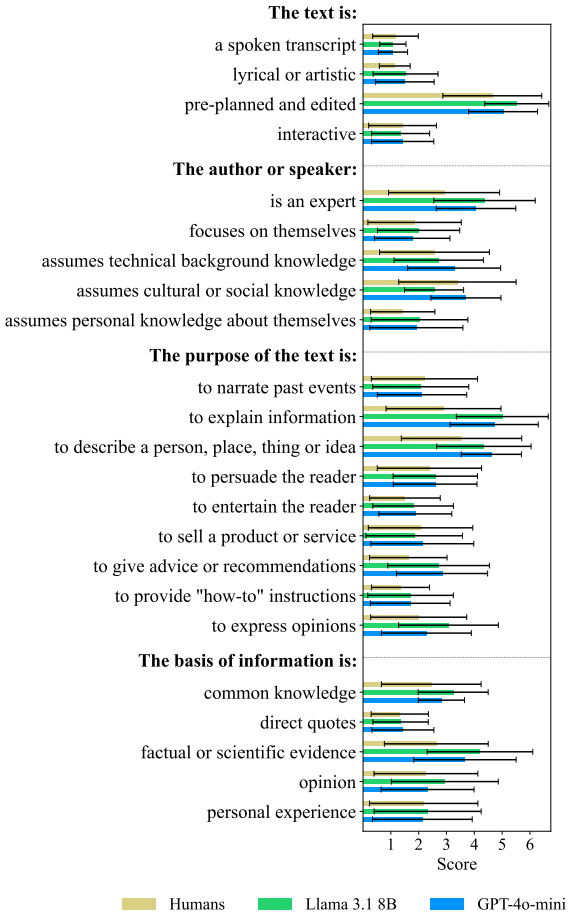

Figure 1: The 23 situational parameters with means and standard deviations, as annotated by GPT-4o-mini, Llama 3.1 8B (full corpus), and human annotators (150 text sample).

Figure 1 shows the means and standard deviations of the parameter scores from the two LLMs and human annotators. The means of most parameters are relatively close, and all parameters show similar dispersion. This suggests that the moderate-to-strong IAA scores are not simply due, for instance, to the LLMs uniformly selecting the same scores across the dataset. Overall, our results demonstrate that LLMs are well-suited for this annotation task.

In the following sections, we use the dataset of 8,406 texts annotated for the 23 situational parameters, register, and language, to conduct a series of

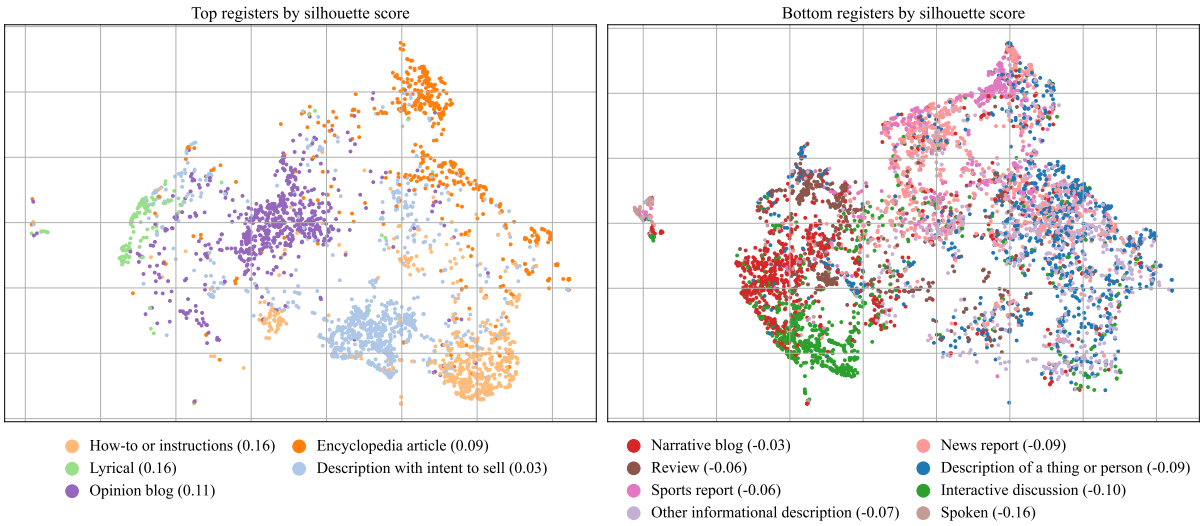

Figure 2: UMAP plots of the registers with the highest (left) and lowest (right) silhouette scores.

analyses. We start by evaluating how well registers are defined by their situational characteristics. Then, we identify distinct *situational text types* directly from the situational data and compare these data-driven categories to traditional register classifications to assess their alignment. Finally, we examine register-internal variation across languages.

4 Quantitative analyses of situation and register

4.1 Registers in a continuous situational space

To begin, we use the LLM-generated annotations to examine how well the register categories in our dataset are situationally defined. To evaluate how well the situational features distinguish each register, we calculate silhouette scores (Shahapure and Nicholas, 2020). This metric measures how similar a text is to the average of other texts within its own register (*cohesion*) compared to the closest instances in other registers (*separation*). Silhouette scores range from -1 to 1, where higher values indicate that instances are better aligned with their own register and more distinct from others.

To visualize the results, we present two UMAP plots (McInnes et al., 2018) in Figure 2. The plot on the left shows the five registers with the highest silhouette scores, while the plot on the right shows the remaining eight with the lowest scores. In both plots, each point represents a text, color-coded by register, with the 2D representation produced by applying UMAP dimensionality reduction to the 23-dimensional situational data.

We observe that the silhouette scores, displayed next to the register names in Figure 2, are generally very low, ranging from 0.16 to -0.16. The registers *How-to or instructions* and *Lyrical* (0.16) are relatively the most situationally well-defined, showing some separation based on their situational characteristics. *Opinion blog* (0.11) also shows some degree of separation. In contrast, registers like *Interactive discussion*, *Description of a thing or person* (-0.10), and *Spoken* (-0.16) have very low scores, indicating strong overlap with other registers and poor situational definition. In addition to overlap, the low silhouette scores are likely influenced by noise and the presence of numerous outliers in the data. For example, although the *Lyrical* category appears mostly clustered on the left edge of the UMAP plots, there are multiple texts from this register dispersed throughout the plot.

The low score for *Spoken* is particularly notable, given that the situational parameters explicitly include one for spoken transcripts (see Table 2). A manual inspection of situational outliers from this register reveals that many of these outliers are written in formal language (e.g. political speeches, presentations), which the LLMs have interpreted as lacking clear markers of direct speech. This issue likely stems from the LLM prompt not providing clear instructions on how to interpret the parameters. We suspect similar inconsistencies may exist for other parameters as well, and plan to address these in future work.

Visual inspection of the UMAP plots suggests that the registers cluster somewhat better than the low silhouette scores indicate (*Opinion blog*, for

instance, is relatively distinguishable at the center), though this may be partly due to UMAP's compression and focus on preserving local structure. Finally, we note that the positioning of the texts in the plot generally aligns with intuitive expectations; for instance, *Encyclopedia articles* are mostly grouped on the opposite side from *Lyrical* texts, reflecting their situational and communicative differences.

To summarize, there is some situational delimitation between registers, but the extent of this separation varies, and generally, the situational boundaries between registers are blurry.

4.2 Identifying clusters based on situational parameters

We apply K-means clustering on the LLM-annotated situational data to identify distinct situational categories in our multilingual dataset. This approach offers a new perspective on the contextual distinctions within the web-sourced texts, complementing the similar but English-only analysis presented by (Biber et al., 2020). The resulting *situational text types* represent groups that are maximally similar in their situational characteristics.

The K-means algorithm (MacQueen et al., 1967) partitions the data into clusters by minimizing the sum of squared distances between data points and their respective cluster average points (centroids). Since K-means requires the number of clusters to be specified in advance, evaluating a range of cluster numbers is a necessary preliminary step to determine the optimal number.

We evaluate situational clusters ranging from 3 to 15 using standard metrics. The silhouette score (as already explained in Section 4.1) measures cluster cohesion and separation, with higher scores indicating better-defined clusters. To compare different cluster sizes with the true register labels, we also calculate the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), which measures the agreement between the clusters and the labels. Additionally, we use the Davies-Bouldin Index (Davies and Bouldin, 1979) to assess the average similarity of each cluster to its most similar counterpart, where lower values indicate better separation. Finally, we calculate the within-cluster sum of squares (WCSS) for each cluster size, which helps identify the optimal number of clusters by potentially revealing an "elbow" point where the rate of decrease in WCSS drops, indicating a good balance between cluster number and compactness.

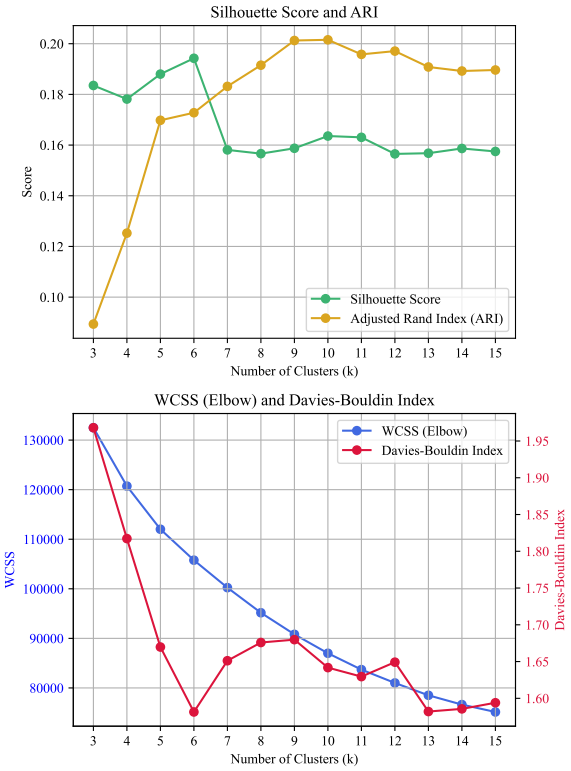

Figure 3: Evaluation of 3 to 15 clusters using silhouette score, ARI, WCSS, and Davies-Bouldin index.

As shown in Figure 3, the average silhouette score peaks at 0.19 with 6 clusters, suggesting optimal cluster definition at this number. At the same point, the Davies-Bouldin Index is also at its most optimal value (1.58), indicating optimal separation. On the other hand, the ARI score is highest with 9 and 10 clusters (0.20), and the WCSS method does not reveal a distinct "elbow" point, as the curve descends smoothly.

Based on these metrics, we select 6 clusters for subsequent analyses, prioritizing cluster cohesion and separation over similarity to the true labels (i.e. high ARI scores), as our goal here is to identify the natural groupings of the texts, independent of their predefined register labels.

We note that all these metrics yield relatively low scores, indicating an overall weak clustering structure for the parameters. This outcome is expected, as the UMAP visualization discussed in Section 4.1 already suggested a lack of clear cluster separation in the data. Furthermore, the low ARI score is unsurprising, given that situational context is only one aspect of what characterizes registers, alongside their linguistic features. Nonetheless, as we show in the next section, the clustering still provides some meaningful differentiation.

Cluster 1: **"Marketing"** (sil: 0.21, N = 919	M	ΔM
The purpose of the text is to sell a product or service	2.93	3.39
The purpose of the text is to persuade the reader	0.76	1.06
The purpose of the text is to describe a person, place, thing or idea	0.86	0.78
Cluster 2: **"Information"** (sil: 0.24, N = 2975)		
The basis of information is common knowledge	0.51	0.57
The basis of information is factual or scientific evidence	0.85	0.55
The text is pre-planned and edited	0.66	0.44
Cluster 3: **"Personal"** (sil: 0.17, N = 1510)		
The author or speaker focuses on himself/herself	1.90	2.57
The basis of information is personal experience	1.79	2.49
The author or speaker assumes personal knowledge about himself/herself	1.83	2.45
Cluster 4: **"Advice"** (sil: 0.21, N = 1117)		
The purpose of the text is to provide "how-to" instructions	2.57	3.06
The purpose of the text is to give advice or recommendations	1.70	2.14
The author or speaker assumes technical background knowledge	0.61	0.65
Cluster 5: **"Speech"** (sil: 0.14, N = 157)		
The text is a spoken transcript	5.27	5.45
The basis of information is personal experience	1.24	1.94
The basis of information is direct quotes	1.24	1.62
Cluster 6: **"Opinion"** (sil: 0.12, N = 1728)		
The purpose of the text is to express opinions	1.32	1.75
The basis of information is opinion	1.29	1.67
The purpose of the text is to persuade the reader	1.11	1.42

Table 3: Six situational text clusters with silhouette scores (sil.) and number of examples (N). Listed parameters are those with the largest deviations in cluster medians (M) from their global medians (ΔM).

4.3 Interpreting the clusters as situational text types

We now identify the parameters that best characterize each cluster by ranking them based on their typical values within the clusters. Then, we use these rankings to interpret the clusters.

Since the parameter distributions are non-normal (as confirmed by Shapiro-Wilk tests, with p-values < 0.001 in each case), we measure their central tendencies using medians, which are relatively robust against outliers and skewed distributions. To further understand how each cluster stands out relative to the entire dataset, we calculate the deviation of each parameter's cluster median from the global median for each parameter. This lets us identify which parameters best define each cluster by seeing how much they deviate from the overall trend. The results are shown in Table 3, with descriptive names assigned to each cluster based on their top parameters.

This analysis produces clearly distinguishable situational text types. In "Marketing", all top parameters relate to the purposes of selling and persuading. "Information" focuses on common, factual, and scientific information. "Personal" centers on self-reflection and personal knowledge. "Advice" has high scores for instructions, advice, and technical background knowledge, often essential in following instructions. "Speech" includes spoken transcripts, personal experiences, and direct quotes, while "Opinion" is characterized by opinions and persuasion.

The silhouette scores, shown next to the cluster names in Table 3, are low across all clusters (0.12–0.24). This indicates that although the clusters are interpretable based on their top parameters, they are not highly distinct in the situational space. The blurred boundaries between clusters may be partly due to parameters that can be interpreted differently depending on the context. For example, the parameter "The purpose of the text is to persuade the reader" has a high median in both the "Opinion" and "Marketing" clusters, but it serves different functions within these contexts (e.g. arguments in a discussion vs. persuasion with the intent to sell).

4.4 Comparing situational text types and registers

Next, we compare the situational text types, identified in the previous section, with the register categories. The aim is to investigate the mapping between the six data-driven clusters and the 13 human-labeled registers from two perspectives: (1) the composition of each situational text type in terms of registers (*cluster purity*), and (2) the extent to which texts from each register are concentrated within a single situational text type (*register completeness*).

To visualize these alignments, we create a 2D UMAP plot with texts colored by register and overlay a Voronoi diagram (Aurenhammer, 1991), shown in Figure 4. This diagram divides the plot into regions representing each situational text type, with each region containing all points closest to the centroid of the corresponding situational text type.

As Figure 4 shows, there is some alignment between the situational text clusters and the register categories (e.g. *Description with intent to sell* is primarily found within cluster "Marketing", and *How-to or instructions* is largely in "Advice"). However, no situational text type aligns perfectly with any single register. This imperfect mapping is expected, as (1) the registers are not well-defined situationally, as discussed above in Section 4.1; (2) the clusters were created independently of the register categories by maximizing situational definition;

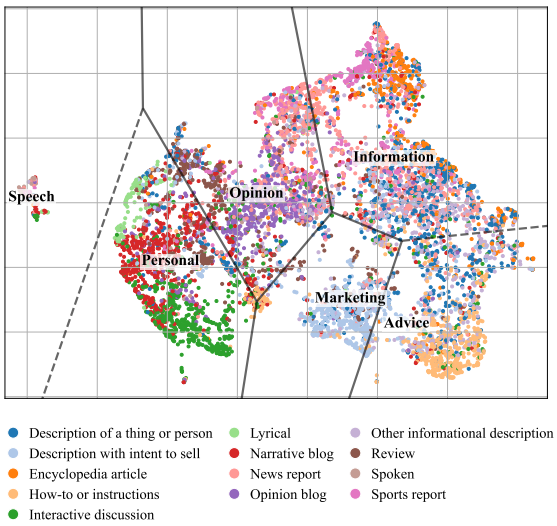

Figure 4: UMAP plot of the six bottom-up situational clusters and 13 registers.

and (3) the number of situational text types differs from the number of registers.

To explore the alignment between clusters and registers in more detail, we present two heatmaps. Figure 5 illustrates *cluster purity*, showing the register composition of each situational text type. Each row represents a situational text type, with columns showing the percentages of registers within each text type.

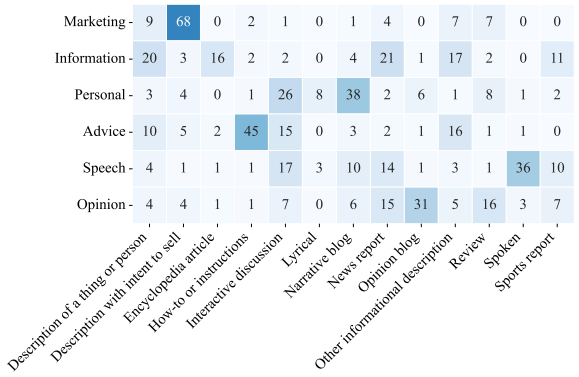

Figure 5: Cluster purity: percentages of registers (columns) in the situational text type clusters (rows).

We find that the register compositions of the clusters generally match the cluster descriptions, though there is significant variation. The "Marketing" cluster, the least variable, includes 68% of texts labeled as *Description with intent to sell*. The "Information" cluster aligns well with informational registers such as *News report* (21%), *Description of a thing or person* (20%), and *Other informational description* (17%). The "Personal" text type

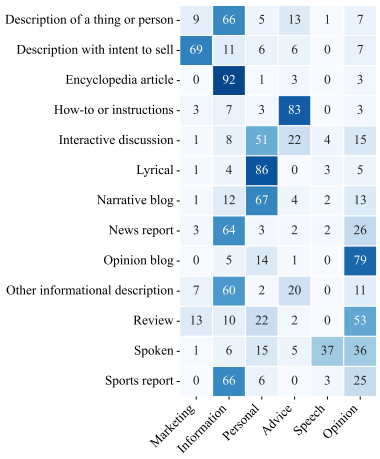

Figure 6: Register completeness: percentages of registers (rows) in the situational clusters (columns).

is primarily composed of *Narrative blogs* (38%) and *Interactive discussions* (26%), where personal matters are often the focus. In the "Advice" cluster, 45% of texts belong to the *How to or instructions* register, followed by *Other informational description* (16%) and *Interactive discussion* (15%), likely providing various forms of advice. The "Speech" cluster includes *Spoken* registers (36%) along with other registers that may contain speech-like elements, such as *Interactive discussion* (17%). Finally, the "Opinion" cluster contains opinionated registers like *Opinion blogs* (31%) and *Reviews* (16%), but also includes *News reports* (15%) and other registers not usually associated with opinion.

The second heatmap (Figure 6) illustrates *register completeness*, showing how registers are distributed across different situational text types. As expected, registers that are more situationally well-defined (see the UMAP plots in Figure 2, Section 4.1) generally map more completely to a single text type. For example, 92% of *Encyclopedia articles* map to "Information", 86% of *Lyrical* to "Personal", and 83% of *How to or instructions* to "Advice". Less well-defined registers, such as *Interactive discussion*, *Review*, and *Spoken*, are more spread across many situational clusters. Notably, the *Spoken* register performs the worst, with texts dispersed across all clusters (1–37%), likely because spoken texts are defined as much by their *purpose* (e.g. expressing opinions) as by the fact that they are spoken.

Interestingly, our multilingual results on the mapping of registers onto situational text types is largely in line with the findings of Biber et al. (2020) for English, which were based on human-

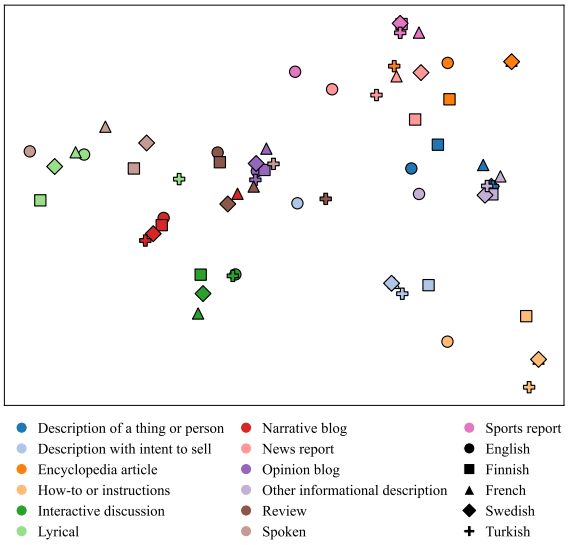

Figure 7: UMAP plot showing situational centroids of registers (colors) across languages (markers).

	p	R^2		p	R^2
Description with intent to sell			English vs. Finnish	0.003	0.07
English vs. Finnish	0.001	0.10	French vs. Swedish	0.031	0.06
English vs. Turkish	0.001	0.08	**Narrative blog**		
English vs. French	0.001	0.07	French vs. Swedish	0.002	0.05
English vs. Swedish	0.001	0.07	**News report**		
How-to or instructions			English vs. Finnish	0.001	0.05
English vs. Finnish	0.001	0.07	**Review**		
English vs. French	0.001	0.06	English vs. Turkish	0.002	0.07
English vs. Swedish	0.001	0.06	Swedish vs. Turkish	0.002	0.06
Finnish vs. Turkish	0.001	0.05	Finnish vs. Turkish	0.002	0.05
Lyrical			**Spoken**		
Finnish vs. Turkish	0.003	0.25	English vs. Turkish	0.005	0.14
Finnish vs. French	0.003	0.21	English vs. Finnish	0.005	0.09
Finnish vs. Swedish	0.016	0.11	French vs. Turkish	0.033	0.06
Swedish vs. Turkish	0.016	0.10	**Sports report**		
French vs. Turkish	0.010	0.08	English vs. Finnish	0.003	0.08

Table 4: Register comparisons with p-values and R^2 values showing language-explained variance (where p-values < 0.01 and R^2 >= 0.05).

made annotations. For example, in their analysis, 97% of *Encyclopedia articles* mapped onto a single cluster (vs. our 92%), as did 91% of *Lyrics* (vs. our 86%), and 76% of *How-to* texts (vs. our 83%). This lends further support to the use of LLMs for situational annotation and suggests possible language-independent patterns in register characteristics, though more detailed analysis is needed.

4.5 Cross-linguistic comparisons

Finally, we investigate the similarities of the registers across the included languages—English, Finnish, French, Swedish, and Turkish—in the continuous situational space. As an intuitive way to compare the language-specific registers, we plot their centroids, representing the average position of each register's data points. We plot the centroids in 2D using UMAP, as shown in Figure 7.

We observe notable variation in how tightly the registers from different languages cluster together in the plot; overall, the registers are not well-separated (consistent with their low silhouette scores; see Section 4.1). The most clearly grouped centroids are those of *Opinion blogs* and *Interactive discussions*, indicating that these registers share similar situational characteristics across languages. Likewise, the *Description of a thing or person* and *Other informational description* registers are also relatively close. *Narrative blogs* are clustered closely in all languages except French. Other registers show more variability, with four of the five languages typically positioned close together, while the remaining language (often English or Turkish) is more distant. In sum, based on a visual examination of the centroids, there is some situational consistency in registers across languages, but the degree of this consistency varies.

To test whether the situational differences between the language-registers are significant, we use PERMANOVA (Permutational Multivariate Analysis of Variance; Anderson 2017), an alternative to ANOVA that does not assume normality, as our data is not normally distributed. We conduct pairwise PERMANOVA tests across all language pairs within each register, applying Bonferroni correction for multiple tests. Additionally, we calculate R^2 scores to measure the proportion of situational variance explained by language.

Table 4 presents the results for comparisons with p-values < 0.01 and R^2 >= 0.05. While the tests reveal statistically significant differences across several language pairs, the R^2 values are generally very low, typically around 0.05 to 0.10, indicating that language explains only a small portion of the total variance in each register. The relatively higher R^2 values in the *Lyrical* and *Spoken* registers (e.g. 0.25 for Finnish vs. Turkish in *Lyrical)* should be interpreted cautiously due to very small sample sizes (only 6–70 examples per language). The majority of the comparisons (86 of 110), omitted from Table 4, yielded nonsignificant results.

Overall, while there are statistically significant language-specific differences in how registers appear in the situational space, they generally account for only a small part of the total variance. This suggests that most of the situational variance within registers is influenced by factors other than lan-

guage. These factors are worth exploring in future research, though it is beyond the scope of this article.

5 Conclusion

This study explored the situational variation of web registers across multiple languages by utilizing LLM-generated situational annotations alongside manual register labels. Analyzing 8,406 texts in English, Finnish, French, Swedish, and Turkish, we identified six situational text types—such as "Advice" and "Opinion"—that cut across the traditional register categories in the dataset. Our findings indicate that while some registers correspond to specific situational clusters, there is significant variation within registers, supporting the view that registers are better described as situationally continuous rather than discrete. Cross-linguistic comparisons further suggest that situational variance within registers is more influenced by internal variation than by language differences, implying that registers are similarly varied across languages rather than distinctly different. The successful use of LLMs for annotation in this study demonstrates their potential in corpus-linguistic register studies.

Limitations

We excluded texts with multiple or missing register labels for simplicity, which limits the scope of our findings. Future work could explore how such texts are positioned within the situational space using cluster analysis and UMAP plots, offering a new method to analyze hybrid or difficult-to-classify texts (Biber et al., 2020). Another limitation of this study is that we focused solely on situational analysis, without addressing linguistic variation. Given the well-established link between linguistic patterns and situational context, comparing these dimensions presents an interesting direction for future research (Egbert et al., 2024). One approach we plan to explore is analyzing how the situational characteristics of texts align with their positioning in Transformer-based (Vaswani et al., 2017) semantic embedding spaces. Finally, in this study, we could only briefly explore the role of language in accounting for situational variation within registers. In future work, we plan to include more languages and conduct detailed statistical analyses to better understand the situational differences and similarities of registers across languages.

Acknowledgements

We wish to acknowledge FIN-CLARIAH (Common Language Resources and Technology Infrastructure), and CSC – IT Center for Science for computational resources. This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Academy of Finland under grant numbers 358720 and 331297.

References

AI@Meta. 2024. Llama 3.1.

Marti J. Anderson. 2017. *Permutational Multivariate Analysis of Variance (PERMANOVA)*, pages 1–15. John Wiley & Sons, Ltd.

Franz Aurenhammer. 1991. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23(3):345–405.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

Douglas Biber. 1994. An analytical framework for register studies. In Douglas Biber and Edward Finegan, editors, *Sociolinguistic Perspectives on Register*, Oxford studies in sociolinguistics, pages 31–56. Oxford University Press, Oxford.

Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus linguistics and linguistic theory*, 8(1):9–37.

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

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

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020.

Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

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, Douglas Biber, Daniel Keller, and Marianna Gracheva. 2024. Register and the dual nature of functional correspondence: Accounting for text-linguistic variation between registers, within registers, and without registers. *Corpus Linguistics and Linguistic Theory*.

Jesse Egbert and Marianna Gracheva. 2023. Linguistic variation within registers: Granularity in textual units and situational parameters. *Corpus Linguistics and Linguistic Theory*, 19(1):115–143. Publisher Copyright: © 2022 Walter de Gruyter GmbH, Berlin/-Boston.

Fabrizio Gilardi, Meysam Alizadeh, and Mael Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120(30):1–3.

Bethany Gray. 2015. *Linguistic Variation in Research Articles*. Studies in Corpus Linguistics. John Benjamins Publishing Company.

Erik Henriksson, Amanda Myntti, Anni Eskelinen, Selcen Erten-Johansson, Saara Hellström, and Veronika Laippala. 2024. Untangling the unrestricted web: Automatic identification of multilingual registers.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2:193–218.

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.

Juhani Luotolahti, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter. 2015. Towards universal web parsebanks. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 211–220, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

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

OpenAI. 2024. Gpt-4o-mini.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J. Van Bavel. 2023. GPT is an effective tool for multilingual psychological text analysis.

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

Petter Törnberg. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning.

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

Margaret Wood. 2024. Linguistic variation in functional types of statutory law. *Applied Corpus Linguistics*, 4(1):100081.

A Appendix: LLM system prompt

We used the following system prompt for the LLMs for situational coding of texts:

You are an expert in describing multilingual web pages for their situational characteristics. The web pages can be written in any language. There are 23 different situational parameters listed below. Your task is to read the document I give to you and code register characteristics based on the content of the web-scraped text.

For each item, select the number that best represents the text. The scale runs from 1 (Disagree completely) to 6 (Agree completely).

Guidelines:

*1. **Read Carefully**: Base your coding only on the text's content.*
*2. **Absence of Features**: Assign a score of 1 if you do not observe any relevant features for a parameter.*
*3. **Objective vs. Subjective Content**: Score as opinion only if the text clearly expresses personal views or judgments. Otherwise, give very low scores for "opinion" related parameters.*

Here are the 23 parameters you will be coding for:

[P1] the text is a spoken transcript [1-6] (explanation)
[P2] the text is lyrical or artistic [1-6] (explanation)
[P3] the text is pre-planned and edited [1-6] (explanation)
[P4] the text is interactive [1-6] (explanation)
[P5] the author or speaker is an expert [1-6] (explanation)
[P6] the author or speaker focuses on himself/herself [1-6] (explanation)
[P7] the author or speaker assumes technical background knowledge [1-6] (explanation)
[P8] the author or speaker assumes cultural or social knowledge [1-6] (explanation)
[P9] the author or speaker assumes personal knowledge about himself/herself [1-6] (explanation)
[P10] the purpose of the text is to narrate past events [1-6] (explanation)
[P11] the purpose of the text is to explain information [1-6] (explanation)
[P12] the purpose of the text is to describe a person, place, thing or idea [1-6] (explanation)
[P13] the purpose of the text is to persuade the reader [1-6] (explanation)
[P14] the purpose of the text is to entertain the reader [1-6] (explanation)
[P15] the purpose of the text is to sell a product or service [1-6] (explanation)
[P16] the purpose of the text is to give advice or recommendations [1-6] (explanation)
[P17] the purpose of the text is to provide 'how-to' instructions [1-6] (explanation)
[P18] the purpose of the text is to express opinions [1-6] (explanation)
[P19] the basis of information is common knowledge [1-6] (explanation)
[P20] The basis of information is direct quotes [1-6] (explanation)
[P21] The basis of information is factual or scientific evidence [1-6] (explanation)
[P22] The basis of information is opinion [1-6] (explanation)
[P23] The basis of information is personal experience [1-6] (explanation)

For each of the 23 points, give a score from 1 to 6 based on the text you read. For each point, explain your given score very briefly, in one short sentence.

In your output, strictly adhere to the following format:

[P1-23] Parameter Name [Your score] (Your explanation)

In the first brackets, write the parameter number [P1 to P23], followed by the parameter name. Then, write your given score in brackets [1-6]. Finally, write your explanation in parentheses ().

Strictly adhere to this output format in all parameter responses. Make sure to fill in all parameters exactly as instructed above.