

Segment-Level Register Analysis of the Multilingual Web

Erik Henriksson¹, Alireza Razzaghi¹, Tomasz Galica¹, Jiaqi Feng Guo¹,
Anastasia Zhiling Dong², Urho Kunnas¹, Erofilis Psaltaki¹, Tuomas Lundberg¹,
Antti Kanner¹, Veronika Laippala¹

¹University of Turku, ²University of Trier

Correspondence: erik.henriksson@utu.fi

Abstract

Web registers, the situationally defined varieties of online text, are almost always annotated at document level. A document, however, may contain multiple such varieties. We label 80,000 documents from the HPLT web corpus at segment level in seven languages using the open-weight Gemma 4, reaching Cohen’s $\kappa=0.73$ against 1,878 human-annotated segments, comparable to previously reported document-level agreement. We find that most documents have two or more registers, and that they recur in similar positions and combinations across the studied languages. Yet on average, a single register accounts for 73% of the labeled text, and a quarter to a third of segments receive no label. Document-level labels are therefore mostly correct but incomplete. Multi-dimensional analysis shows that the segment-level labels correspond to linguistic differences. In all seven languages, registers are ordered the same way, from informational to involved text.

1 Introduction

How people communicate depends on the situation: news, recipes, encyclopedias and song lyrics each call for a different kind of language. These differences span all linguistic levels from vocabulary and morphosyntax to the organization of discourse, and have been documented across modalities, historical varieties, and languages (Pescuma et al., 2023). Variation of this kind is studied under the term *register variation*, where a register is a variety of text marked by the linguistic features typical of its situation of use (Halliday, 1978; Biber and Conrad, 2019; Matthiessen, 2019; Biber and Egbert, 2023). Register studies aim to describe empirically how the situation of use relates to linguistic form.

The web has been used as a language corpus since the early 2000s (Kilgariff and Grefenstette, 2003; Baroni et al., 2009), offering textual variation far beyond traditional corpora (Biber and Egbert, 2018). In most previous work, web registers

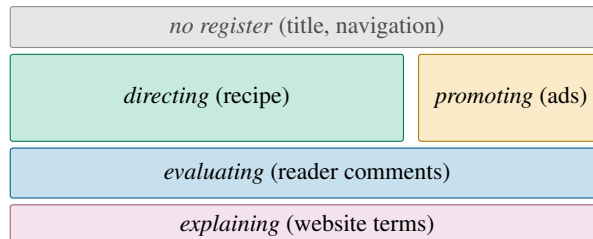


Figure 1: Schematic of register mixing on a recipe page.

are assigned to whole documents (Kuzman and Ljubešić, 2025). Yet a document may consist of situationally distinct units (Egbert et al., 2024; Egbert and Gracheva, 2022): a recipe page, for example, may combine instructions, advertising, reader comments and website terms, as well as titles and navigation that have no register (Figure 1). A single label for such a document does not distinguish between these units, which may in part explain problems reported in web register research (Biber and Egbert, 2023), including fuzzy categories (Biber et al., 2020; Henriksson et al., 2024), low inter-annotator agreement (Egbert et al., 2015), and classification performance that does not improve with larger models (Henriksson et al., 2025b).

Previous work has addressed these issues in two ways. The first is to let annotators assign many register labels to a single document (Laippala et al., 2023; Skantsi and Laippala, 2023). Such labels, however, only show whether a document is mixed, not whether one passage mixes two registers or two passages each have a different one. The second is to move below the document. Henriksson et al. (2025a) split English web documents recursively at the points where a register classifier predicted different registers on each side, reporting a cleaner register signal for the resulting segments. Henriksson et al. (2026) labeled paragraphs from the High Performance Language Technologies (HPLT) English web dataset (Burchell et al., 2025) with a proprietary LLM, finding that agreement is higher for segments than for documents and that over 80%

of documents contain more than one register.

In this paper, we apply segment-level register analysis to seven languages, using Gemma 4 (Gemma Team et al., 2026), an open-weight local LLM, and the HPLT 4.0 corpus (HPLT Consortium, 2026). The labels agree with 1,878 segments annotated by native speakers at Cohen’s $\kappa=0.73$ on average. We find that most web documents mix registers in positions and combinations that recur across languages, but that a single register usually covers most of a document’s text. The segment-level labels also correspond to linguistic differences. We release our code and annotated data at <https://github.com/TurkuNLP/multilingual-web-register-segments>.

2 Background

2.1 Web register and the unit of observation

The dominant corpus-linguistic approach to web registers labels documents using the CORE (Corpus of Online Registers of English) taxonomy, which has eight main registers and optional subregisters (Laippala et al., 2023). The approach was originally developed for analyzing English web register variation (Egbert et al., 2015) but has since been extended to other languages and to train multilingual classifiers (Laippala et al., 2019; Repo et al., 2021; Skantsi and Laippala, 2023; Henriksson et al., 2025b; Razzaghi et al., 2026). CORE labels have also been used in register-aware analyses of LLM-pretraining corpora (Myntti et al., 2025; Burchell et al., 2025; Galica et al., 2026).

The limits of the document as an observation unit are recognized within this tradition. Biber et al. (2020, p. 33) note that web texts can “integrate multiple communicative purposes”. Biber and Egbert (2023) observe that texts within a register vary situationally and linguistically, and that some texts belong to no register at all. Egbert and Gracheva (2022) show that situational parameters behave differently depending on the granularity of the unit. The problem is thus recognized from two directions (the unit and the categories) but this work leaves open how a smaller unit could be obtained at scale.

2.2 Register segmentation

Dividing text into coherent units has a long history in NLP (Hearst, 1994, 1997; Hearst and Plaunt, 1993; Galley et al., 2003; Liu et al., 2021). In register studies, Egbert et al. (2021) developed a method for dividing transcribed conversations into

discourse units, where boundaries require a shift in communicative goal. Biber et al. (2021) applied it to 2.3 million words of the British National Corpus 2014, coding each unit for communicative purpose and identifying 16 discourse types from 9 purposes.

Henriksson et al. (2025a) segment English CORE documents with a register classifier. They fine-tune ModernBERT on full documents, then split a document recursively at the sentence boundary where the classifier’s predictions for the two sides are most distinct. The segments carry a clearer register signal than the documents: a classifier trained on them reaches a micro-F1 of 0.89 against 0.86, and linguistic features vary less within a register. The method needs no segment-level training data, but it requires a document-level classifier fine-tuned for the target language, and the boundaries depend on its accuracy.

In another approach, Henriksson et al. (2026) labeled the paragraphs extracted from the XML representations in the HPLT 3.0 (Oepen et al., 2025) English dataset with a proprietary LLM, and merged consecutive segments with the same label into a single span. This approach assumes that XML-derived paragraphs of text are small enough to be roughly register-homogeneous, so that boundaries fall wherever consecutive labels differ. We adopt a similar approach here, with an open-weight LLM and further differences described in Section 3.

2.3 LLMs for multilingual register annotation

Recent work has explored LLM-based register annotation (Henriksson et al., 2024, 2026), but has relied on proprietary API models and focused on English. The multilingual capabilities of LLMs for register analysis remain largely untested (though see Idahl et al., 2026, for related work). Segment-level annotation multiplies the number of labeling calls relative to full documents, making proprietary APIs costly at scale and, since endpoints and model versions change, unreliable for replication. In this work, we instead use a local open-weight model.

3 Methods

3.1 Data and sampling

We use the cleaned monolingual datasets from HPLT 4.0, where documents are distributed in shards ordered by a document quality score (Pavón et al., 2023). We take documents in descending order of quality score until we have a pool of 10 million per language. From this pool we draw a ran-

dom subsample of 10,000 documents per language. Finally, we draw 250 segments per language for human annotation, sampling over all segments of these documents. Every segment in these documents has the same probability of being sampled.

We target seven languages, chosen to vary writing system, family and corpus size: English, Finnish, Polish, Greek, Persian, Swedish and Mandarin Chinese. HPLT distributes Chinese as two separate collections by script. We sample from both, giving eight samples for seven languages.

3.2 Segments

We take the segments provided by HPLT as our labeling unit. A document’s segments are the newline-separated lines of its text field. HPLT describes them as “paragraph-like text passages” whose boundaries come from the markup of the original document (HPLT Consortium, 2026): text is extracted with Trafilatura (Barbaresi, 2021), which outputs one line per block-level element, so that each paragraph, heading, list item, block quote, and code block begins a new segment.

In the register analyses reported below, we merge consecutive segments with identical annotations across all fields (for the fields, see Section 3.3), and these merged spans are our unit of analysis (“register segments”). Following Henriksson et al. (2026), we treat segments with no register as transparent for the merge, so that a heading or label does not split a continuous stretch of the same register. Registerless segments are excluded from the merged span but kept as separate units in the analyses.

3.3 Annotation scheme

We annotate each segment along three dimensions, summarized in Table 1: *medium*, how the text was produced; *turn*, whether others can reply; and *purpose*, the communicative act the segment performs. The dimensions are independent, except that segments labeled *cannot_rate* for medium (headings, labels, boilerplate, and other similar text) receive no turn or purpose. Segments are annotated in context: a segment that is fragmentary on its own but continues a passage is rated as part of that passage. The same codebook is given to both the human annotators and to the LLM.

To relate our annotations to previous work, we map each segment to a single CORE main register (Figure 2): *transcribed* segments become *Spoken*, the remaining *dialogic* ones *Interactive discussion*, and all others are mapped by purpose, with *explain-*

Medium — <i>how the text was produced</i>	
written	composed to be read
transcribed	a record of words actually spoken
cannot_rate	not self-standing communication: headings, labels, boilerplate, metadata, fragments, machine-generated or garbled text
Turn — <i>whether others can reply</i>	
dialogic	a real exchange where others can reply
monologic	the audience cannot reply
Purpose — <i>the communicative act performed</i>	
explaining	conveying how something is or works
supporting	defending a specific claim with evidence
recounting	relating particular real events
directing	telling the reader (how) to do something
evaluating	arguing a position, backed by reasons
expressing	a feeling or reaction for its own sake
promoting	marketing own product or organisation
creating	literary or artistic text as the main content

Table 1: The annotation scheme. *Turn* and *purpose* are left empty when the medium is *cannot_rate*. Glosses are abridged; Appendix A gives the codebook in full.

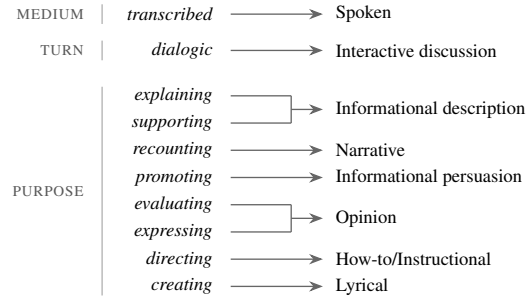


Figure 2: Mapping of segment annotations to CORE main registers.

ing and *supporting* collapsing into one register and *evaluating* and *expressing* into another. Although CORE registers were originally defined for full documents (e.g. *Narrative* covers blogs, news articles, and stories), the situational parameters underlying them (Egbert et al., 2015) are mostly comparable to those used here. The exception is *creating* (literary or artistic text), which we map to the more restrictive *Lyrical* CORE register (songs and poems).

3.4 Human annotation

For each language, 148–250 sampled segments were annotated by one native speaker of that language, using a browser-based annotation tool. Annotators followed the same codebook as the LLM (Appendix A) and saw the same context: the target segment and its ± 5 neighboring segments, the first five segments of the document, and the page URL.

The annotations were also checked for codebook consistency. We used Claude Opus 5 (Anthropic,

2026) to flag labels that were in clear conflict with the guidelines, and one author reviewed the flagged cases, using machine translation as necessary, and fixed obvious errors. This post-correction step reduces the impact of annotator errors on the figures.

3.5 LLM annotation

We annotate with Gemma 4 31B Instruct (Gemma Team et al., 2026), a dense open-weight instruction-tuned model, served with vLLM (Kwon et al., 2023) on two AMD MI250 GCDs of the LUMI supercomputer. Decoding is greedy (temperature 0) and constrained to a JSON schema with XGrammar (Dong et al., 2024), so the model can only output the six fields of the scheme.

The prompt gives the model the codebook, the page URL, and a context block in which the target segment is marked and its neighbors are shown (Appendix B). Before outputting the labels, the model writes a one-sentence rationale for its decision. The schema places this reasoning field first, and because the grammar generates the fields in order, the model must state its reason before labeling.

To check whether the open-weight model is competitive on this task, we repeat the annotation with Claude Opus 5 and compare the two sets of labels. The setup is identical, except that the constrained JSON is obtained through forced tool use, the equivalent mechanism in the Anthropic API.

3.6 Evaluation

We score the model against the human annotation per dimension. *Medium* and the CORE-mapped register are scored over all segments, since both are defined for every segment. *Turn* and *purpose* are scored over the segments that both sides judged ratable, because they are undefined where medium is *cannot_rate*. The joint *all* column requires the model to match the human on all three dimensions. We report F1 and Cohen’s κ (Cohen, 1960).

3.7 Multi-dimensional analysis

To test whether the labels correspond to linguistic differences, we run a multi-dimensional analysis (Biber, 1988) over the register segments. Segments of at least 50 words are parsed with Stanza (Qi et al., 2020) and described by rates of universal part-of-speech (UPOS) tags, universal dependency relations (de Marneffe et al., 2021) and punctuation types per 1,000 content tokens, plus eight structural measures (sentence and word length, dependency distance, tree depth, long and head-final

Field	Cohen’s κ		F1	
	Gemma 4	Opus 5	Gemma 4	Opus 5
Medium	0.79	0.80	0.90	0.91
Turn	0.86	0.78	0.97	0.96
Purpose	0.70	0.68	0.74	0.72
CORE label	0.73	0.72	0.78	0.78

Table 2: Agreement between each LLM annotator and the human annotation, averaged over the eight language samples ($n=1,878$ segments). The *turn* κ excludes both Chinese samples, where too few segments are dialogic.

dependencies, lexical density, type–token ratio). Features are standardized within each sample, so that dimensions capture within-language covariation rather than cross-language contrast. After filtering (Appendix C), 47 features over 425k segments are factored by iterated principal axis factoring with promax rotation; a segment’s score on a dimension sums its standardized features loading at $|\lambda| \geq 0.35$.

4 Results and discussion

We first evaluate the LLM-generated labels against the human annotation, then use the labels to describe the corpus, looking at how registers combine and how they are distributed within documents. Finally, we examine how the segment-level registers differ linguistically.

4.1 Annotation quality

Gemma 4 reaches $\kappa=0.73$ on the CORE-mapped label on average, falling between 0.72 and 0.74 in every sample (F1 0.76–0.79). This resembles the human–human agreement in the CORE tradition: 83% for the FinCORE main registers (Skantsi and Laippala, 2023) and 78–84% F1 for FreCORE and SweCORE (Repo et al., 2021), against Fleiss’ $\kappa=0.47$ for English CORE (Egbert et al., 2015). At the segment level, Henriksson et al. (2026) report $\kappa=0.76$ for English. The labels thus appear to be close to the ceiling the task allows.

Claude Opus 5, a proprietary frontier model, does not do better on this task ($\kappa=0.72$), and the two models agree with each other on 81.2% of segments, more often than either agrees with the annotators (78.2% and 77.8%; $p=0.003$). This suggests the disagreements are a property of the task and the codebook rather than of the chosen model. Below, we report the results for Gemma 4, the model we use for the large-scale annotation (Section 4.2).

The model and the annotator disagree on at least one dimension in 26% of the segments. Disagree-

	model										
human	expl	supp	reco	dire	eval	expr	prom	crea	n.r.	<i>n</i>	
expl	308	5	11	7	9	–	8	1	11	360	
supp	16	10	6	–	20	–	–	–	1	53	
reco	15	–	181	1	6	2	1	–	1	207	
dire	13	–	3	116	5	–	2	–	6	145	
eval	16	1	5	8	120	1	2	–	–	153	
expr	12	–	24	4	36	39	1	–	3	119	
prom	29	–	5	5	9	–	83	–	3	134	
crea	1	–	8	1	2	1	–	48	2	63	
n.r.	66	1	10	19	9	10	8	5	516	644	
<i>n</i>	476	17	253	161	216	53	105	54	543	1878	

Table 3: Confusion between the human annotation and Gemma 4 31B on *purpose*, with the segments carrying no register as a ninth category, pooled over the eight samples ($n=1,878$ segments). "n.r." = cannot_rate.

ments fall mainly on *purpose* (63%) and ratability (32%); turn and the *written/transcribed* distinction account for the remaining 5%. Table 3 shows the first two together. Ratability errors are asymmetric: the model assigns cannot_rate less often than the annotator (precision 0.95, recall 0.80). Recall is 0.89 for segments under 40 characters but 0.62 for those over 150, so the missed cases are most often long texts that the model judges ratable.

Within *purpose*, two of the eight categories perform poorly. The model recovers 19% of the annotator’s *supporting* segments and 33% of the *expressing* ones (F1 0.29 and 0.45), against 0.65–0.82 for the other six. The scheme separates both from a close neighbour by degree: *supporting* is separated from *explaining* by whether the text is defending a claim or just stating it, and *expressing* from *evaluating* by whether the stance is argued for. Both are absorbed mainly by *evaluating*: 20 of the 53 *supporting* segments and 36 of the 119 *expressing* ones are labelled *evaluating*, which the model uses for 216 segments against the annotator’s 153. Many of these cases are genuinely ambiguous. In the comment below, the annotator read the target as a bare reaction (*expressing*) and the model as an argued position (*evaluating*):

[Context –1] ... My blog is the one place — besides with my oldest friends — where I am the closest to who I am at my core.

[Target] Finally, there are many, many ways to monetize blogs beyond ads, brands, and reviews. Good for the people that make it work, but we can all expand.

[Model rationale] “The author argues that bloggers should expand their monetization methods beyond ads and brands.”

Both readings follow the codebook (Appendix A):

Language	Segments	Mchar	no reg.	tra.	dia.
English	150,111	158.3	26.6	4.3	6.3
Finnish	142,626	76.2	31.0	1.6	10.0
Polish	124,585	75.1	35.4	3.5	6.1
Greek	139,457	84.7	30.2	5.2	6.6
Persian	167,024	91.4	32.6	7.1	6.0
Swedish	147,955	85.3	31.1	2.3	5.3
Chinese (S)	113,456	47.2	27.7	4.6	5.0
Chinese (T)	93,027	23.1	29.8	1.6	1.7
Total	1,078,241	641.4			

Table 4: The labeled corpus: 10,000 documents per sample. A segment is a merged span (Section 3.2), with cannot_rate segments transparent to the merge. *no reg.* = cannot_rate; *tra.* = transcribed and *dia.* = dialogic.

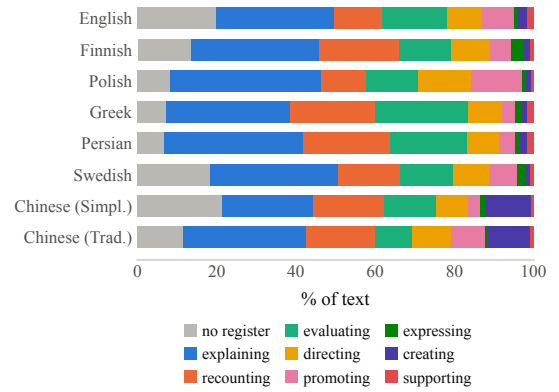


Figure 3: Composition of each sample. Each bar is all of that sample’s text: in grey the share of segments labelled cannot_rate, the rest divided by *purpose*.

the segment recommends a course of action but gives no reasons for it. The same confusions occur in all languages, but in different amounts: per-sample agreement on *purpose* ranges from $\kappa=0.61$ to 0.75, with the three lowest samples containing the most *supporting* and *expressing* segments. Under the CORE mapping, which merges *explaining* with *supporting* and *evaluating* with *expressing*, agreement is more stable (0.72–0.74).

We conclude that the disagreement is concentrated in two category boundaries, and not in any particular language. Once *supporting* and *expressing* are merged with their neighbors under the CORE mapping, agreement is nearly identical across the eight samples.

4.2 Register distribution across languages

The eight samples have similar register compositions (Table 4, Figure 3). A segment here is a merged span of consecutive identical annotations (see Section 3.2); the 2.1 million annotated HPLT segments are merged into the 1.1 million regis-

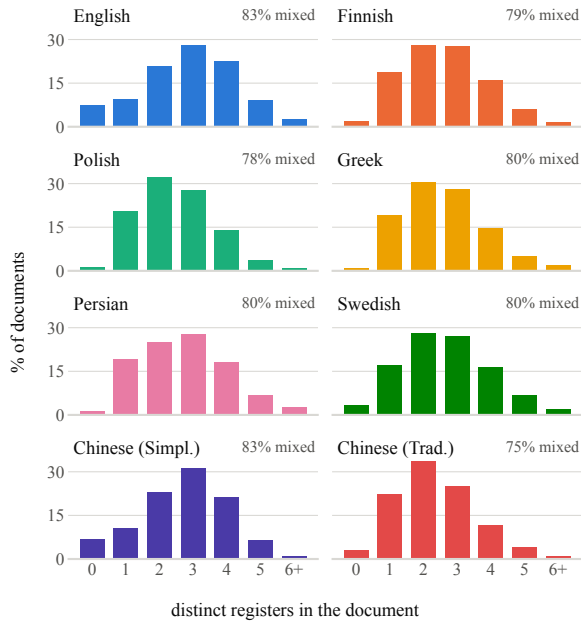


Figure 4: Distinct CORE registers per document. The scheme has eight registers, but counts of ≥ 6 are pooled, as together they hold 0.7–2.6% of documents.

ter segments reported in the table (9–17 per document). Between 26.6% and 35.4% lack register (including headings, labels, navigation, metadata and fragments). Measured in text (Figure 3) they are smaller, holding 6.8% (Persian) to 21.4% (Chinese Simpl.). Transcribed text accounts for 1.6–7.1% of all text and dialogic text for 1.7–10.0%.

Figure 3 shows the distribution of purposes in the labeled segments. *Explaining* is the largest in every sample, from 29.3% (Chinese Simpl.) to 41.6% (Polish) of ratable text, and *explaining*, *recounting*, *evaluating* and *directing* together account for 76.5–91.6%. *Creating* is the most different between samples, taking 14.7–11.9% of ratable text in the two Chinese samples against at most 2.7% in any other.

4.3 Register mixing within documents

Most documents contain more than one register, in all eight samples. Figure 4 counts the distinct CORE registers (Figure 2) in a document. We count CORE registers here rather than the three original dimensions to make the figure comparable with previous work (e.g. Biber and Egbert, 2018; Henriksson et al., 2025b). *cannot_rate* is the absence of a register, so a document with only *cannot_rates* is counted as having 0 registers. Between 75.0% (Chinese Trad.) and 83.1% (English) of documents have two or more, most commonly two or three, with a mean of 2.4 to 2.9. Only 9.5% (English) to

22.1% (Chinese Trad.) are register-homogeneous, and 0.9% (Greek) to 7.4% (English) lack register. For English, the figure agrees with the mixing rate Henriksson et al. (2026) report at segment level.

4.4 Positional distribution of registers

To see where registers fall within a document, we count characters in each category by relative position in 128 equal-width bins (Figure 5). Each category is normalized over its own bins and expressed as a ratio to an even spread, so the curves show positional shape. We plot *dialogic* and *transcribed* but not their opposites, since *monologic* (93%) and *written* (95%) reproduce the corpus average. *Supporting* is omitted (1.3% of the text).

The patterns are clear and nearly identical across the eight samples. Registerless text concentrates at page edges, reaching 2.3 times an even spread in the first bin and 1.8 in the last, likely reflecting titles and navigation at the top and boilerplate at the end. *Recounting* tends to appear early, peaking in the first quarter and falling to 0.7 by the end, while *explaining* is spread almost evenly, varying only between 1.1 and 0.8. *Dialogic* and *expressing* instead concentrate at the end: *dialogic* rises from 0.5 to 1.3 across the document, and *expressing* dips to 0.7 at the quarter mark before rising to 2.6. Both point to document-final comments, where opinionated text is common.

4.5 Register dominance

Section 4.3 above reports register *counts* per document (Figure 4). Weighting by *content* shows a different picture: it is dominated by a single register. Figure 6 shows, averaged over documents, the share of register-bearing text held by the largest CORE register, the 2nd largest, and so on. Across the eight samples, the largest holds 69.0–74.6%, the second 18.3–21.4% and the third 5.2–7.4%.

The profiles are similar across languages and slightly more concentrated than chance. We compare each page against a random division of its text among the registers it contains: for k registers, the expected largest share is $(1/k) \sum_{i=1}^k 1/i$, or 75% for two registers and 61% for three. The observed share exceeds this baseline in all eight samples, by 1.6–5.5 points. The comparison holds the number of registers fixed, so it tests how evenly a page’s text is divided among its registers. Although most pages are multi-register, they are usually organized around one dominant register.

The text outside the dominant register follows

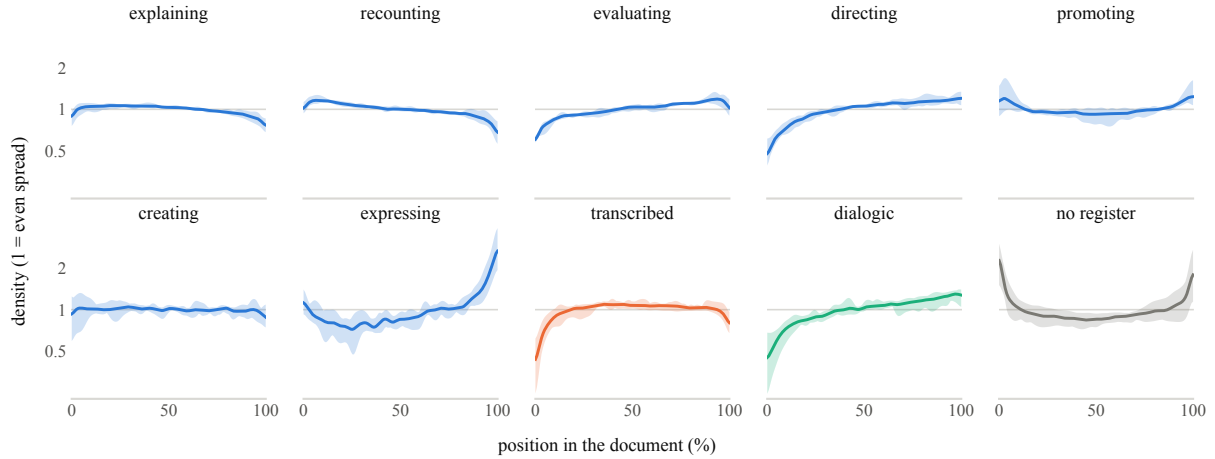


Figure 5: Positional distribution of registers within documents, as a ratio to an even spread (1, logarithmic axis) over 128 bins, Gaussian-smoothed ($\sigma = 2$). Lines are means over the eight samples, bands show the full range.

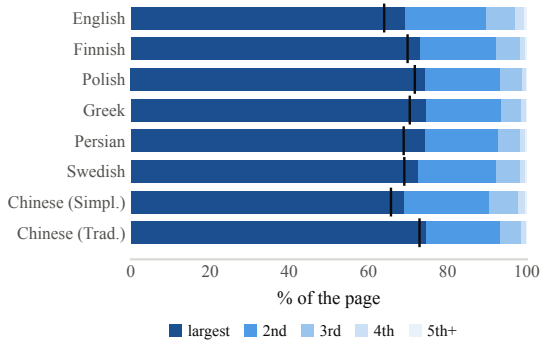


Figure 6: Register profiles of documents. Mean share of labeled text held by the largest register, the second largest, and so on. Black marks give the expected share of the largest register under random division of a page into the number of registers it actually contains, computed per sample. `cannot_rate` segments are excluded.

regular patterns too. Figure 7 groups documents by their dominant register and shows the composition of the rest of the page. The share of pages with text outside the dominant register ranges from 64% to 89%, depending on which register dominates. On a page dominated by Informational description, the rest is mostly How-to (32%), Opinion (27%) and Narrative (22%). On a How-to page, it is mostly Informational description (60%).

The figures above partly reflect how common each register is overall. Figure 8 therefore compares each register’s share of a remainder against its share of remainders elsewhere (1 = no difference). Informational description, How-to and Informational persuasion co-occur more often than expected (1.1–1.6), and the last two co-occur with Narrative, Spoken and Lyrical *less* often than ex-

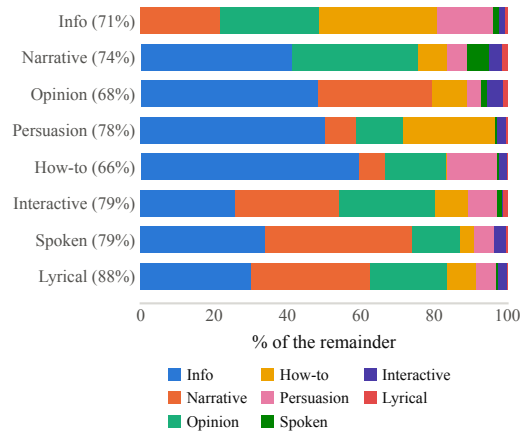


Figure 7: The mean composition of the text on the page accompanying the main register’s share (shown in brackets). Shares are of the remainder of the page, averaged over documents and pooled across languages.

pected (0.2–0.4). Narrative forms a second group with Opinion, Interactive discussion, Lyrical and Spoken (1.3–2.5, and 1.3–1.9 in the reverse direction). The association is strongest with Spoken, but these four do not co-occur consistently with each other. The two groups correspond to the Informational and Involved poles reported in multi-dimensional analyses (Biber and Conrad, 2019).

4.6 Linguistic dimensions

The segment-level labels correspond to linguistic differences. We show this with a multi-dimensional analysis of the register segments (Biber, 1988). The features are taken from Stanza parses (see Section 3.7) and the dimensions are derived from their co-variation. Four dimensions were extracted, account-

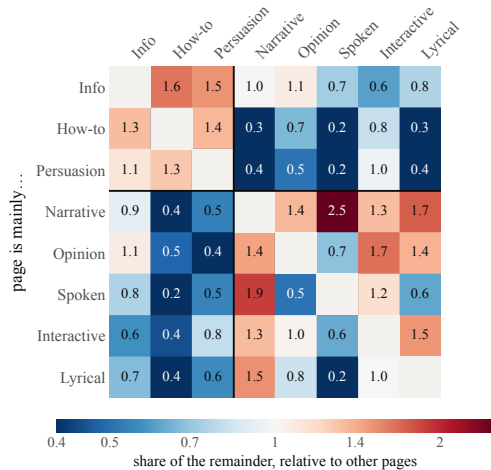


Figure 8: Remainder composition (rows: dominant register; columns: registers in the remainder). Each cell gives a register’s share of that remainder. Values are relative to the same register’s share of remainders on pages with a different dominant register, so 1 means no difference. The rule separates the informational registers from the narrative and interpersonal ones.

ing for 33.3% of the variance. We report the first, which accounts for 15.4%; the remaining three account for 5.5–6.6% each.

Dimension 1 corresponds to the opposition between *Informational* and *Involved* production that is reported throughout the multi-dimensional literature (Biber, 1988, 1995; Biber and Conrad, 2019). Its positive pole consists of verbs, pronouns, auxiliaries, adverbs, subordinators, subjects and subordinate clauses of three kinds, and its negative pole of lexical density, nominal modifiers, nouns and word length (Table 5). Although no register information enters the analysis, register accounts for 18.3% of the variance in the resulting scores.

Figure 9 shows each register’s mean score in each of the eight samples, together with the pooled mean. The involved pole is occupied by Lyrical and Interactive discussion (both +1.0), followed by Spoken (+0.74); the informational pole by Informational persuasion (−0.43) and Informational description (−0.22), with Opinion, How-to and Narrative between them. The ordering is largely consistent across the eight samples: Spearman’s ρ between each sample’s ordering and the pooled one ranges from 0.88 to 1.00, and every register at either pole stays there, Lyrical, Interactive discussion and Spoken positive in all eight samples and the two informational registers negative. How strongly register predicts the score is less

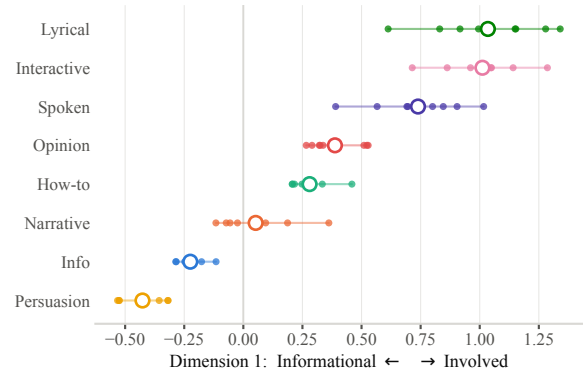


Figure 9: Register means on Dimension 1, in standard deviations from the corpus mean. Small marks give the eight sample means and the line their range; the open circle is the mean over samples.

Positive	verbs (.78), clause markers (.71), pronouns (.71), subordinators (.70), auxiliaries (.65), adverbial modifiers (.64), subjects (.60), adverbs (.60), auxiliary dependents (.54), clausal complements (.53), adverbial clauses (.52), open clausal complements (.46), copulas (.39), objects (.38)
Negative	lexical density (−.66), nominal modifiers (−.65), nouns (−.60), word length (−.55), adjectival modifiers (−.37)

Table 5: Features loading at ≥ 0.35 on Dimension 1.

even, from 12.8% of the dimension’s variance in Chinese (Trad.) to 25.9% in Finnish. The same Informational–Involved opposition was inferred in Section 4.5 from the registers that co-occur within a document; it is obtained here from the linguistic features directly.

Segments labeled `cannot_rate`, omitted from the figure, fall below the informational pole (−0.83; −1.61 in Persian), indicating that the text carrying no register is a distinct, densely nominal variety. This holds in six of the eight samples; in both Chinese samples they are just above Informational persuasion. However, only 11.5% of registerless segments reach the 50-word floor (Section 3.7) against 59–72% for the eight registers, so this analysis characterizes only the longer minority.

5 Conclusion

We annotated the register of 80,000 web documents at segment level in seven languages and four scripts, using Gemma 4 and validating the labels against native-speaker annotation. Agreement reaches $\kappa=0.73$ for the CORE mapping and varies little across the eight samples (0.72–0.74). Claude Opus 5 does not perform better, and the two models agree with one another more often than ei-

ther agrees with the annotators. The disagreements are therefore not specific to the model we use.

Two distinctions in the scheme, between *explaining* and *supporting* and between *evaluating* and *expressing*, are not applied reliably at segment level, and merging them under the CORE mapping gives stable agreement across the samples.

In general, the language samples are very similar. Two or more registers occur in 75–83% of documents; *explaining*, *recounting*, *evaluating* and *directing* account for most of the text; the registers occupy similar positions within the document, with expressing and dialogic text concentrated at the end; and the multi-dimensional analysis yields the same *Informational–Involved* opposition, ordering the registers in the same way, in all eight samples.

Although most documents mix registers, a single one covers 73% of the register-bearing text in an average document, and the remainder follows regular patterns, with informational registers grouping apart from narrative and interpersonal ones.

Therefore, we conclude that document-level labeling, typical of previous work, describes only part of most pages. It identifies the dominant register, missing on average a quarter of the register-bearing text and all that lack one. For register-aware curation of LLM corpora (Myntti et al., 2025; Burchell et al., 2025; Galica et al., 2026), segment-level labels allow for finer-grained data selection.

6 Limitations

With one annotator per language, we report model–human rather than human–human agreement, and comparisons to published IAA scores are indicative, as they come from different studies and label sets. The human labels were post-corrected with model assistance, which may inflate the agreement.

We use HPLT’s segmentation, which is not linguistically motivated: boundaries may not fall where the register changes, and segments can have internal shifts. Merging same-label segments corrects the first problem, but a mixed segment still receives a single label; locating these shifts, with multi-label assignment, is left for future work. The languages were chosen by annotator availability and are well resourced in HPLT, and we use only high-quality text. Other languages and lower-scoring text might yield different results.

The multi-dimensional analysis omits morphology for cross-language comparability, and depends on parser accuracy that varies by language. Finally,

we use a single model and a prompt developed informally rather than tuned against a development set; future work could optimize prompts systematically, e.g., with DSPy (Khattab et al., 2023).

Ethics statement

Our data comes from HPLT 4.0, which is built from web crawls. Web text of this kind contains personal information. We use the corpus as distributed and have not removed such content, since that would alter the distribution of text we set out to describe. HPLT drops documents from domains whose robots.txt disallows the crawlers it uses, flags any personal information detected, and gives rights holders an address for take-down requests, acting on them in subsequent releases (HPLT Consortium, 2026). Our own annotations describe the communicative situation of passages of text; we add no personal information to what HPLT already publishes.

Acknowledgments

This work was supported by the Research Council of Finland through the FIN-CLARIAH research infrastructure (project 358720, which has also received funding from the European Union NextGenerationEU instrument), the project “Mechanisms of Register Variation in Massively Multilingual Web-Scale Corpora” (project 362459), and the Centre of Excellence “Human Diversity through Contacts” (project 374223). Further support came from the OpenEuroLLM and LLMs4EU projects under Digital Europe Programme grant agreements No 101195233 and 101198470, and from the European Union’s Horizon Europe research and innovation programme under Marie Skłodowska-Curie grant agreement No 101177564 (HAIF), co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. We also wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

- Anthropic. 2026. *Claude opus 5*.
- Adrien Barbaresi. 2021. *Trafilatura: A Web Scraping Library and Command-Line Tool for Text Dis-*

- covery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.
- Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.
- Douglas Biber. 1995. *Dimensions of register variation: a cross-linguistic comparison*. Cambridge University Press, Cambridge.
- Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.
- Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press, Cambridge.
- Douglas Biber and Jesse Egbert. 2023. What is a register?: Accounting for linguistic and situational variation within—and outside of—textual varieties. *Register Studies*, 5(1):1–22.
- Douglas Biber, Jesse Egbert, and Daniel Keller. 2020. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16(3):581–616.
- Douglas Biber, Jesse Egbert, Daniel Keller, and Stacey Wizner. 2021. *Towards a taxonomy of conversational discourse types: An empirical corpus-based analysis*. *Journal of Pragmatics*, 171:20–35.
- Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. *An expanded massive multilingual dataset for high-performance language technologies (HPLT)*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, volume 1: Long papers, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. *Universal Dependencies*. *Computational Linguistics*, 47(2):255–308.
- Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2024. XGrammar: Flexible and efficient structured generation engine for large language models. *arXiv preprint arXiv:2411.15100*.
- 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*, 20(3):505–538.
- Jesse Egbert and Marianna Gracheva. 2022. *Linguistic variation within registers: Granularity in textual units and situational parameters*. *Corpus Linguistics and Linguistic Theory*, 19.
- Jesse Egbert, Stacey Wizner, Daniel Keller, Douglas Biber, Tony McEnery, and Paul Baker. 2021. *Identifying and describing functional discourse units in the bnc spoken 2014*. *Text & Talk*, 41(5-6):715–737.
- Tomasz Galica, Amanda Myntti, Erik Henriksson, Fedor Vitiugin, Sampo Pyysalo, and Veronika Laippala. 2026. Register-based data selection for multilingual llms. Accepted at EMNLP 2026.
- Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. *Discourse segmentation of multi-party conversation*. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. Association for Computational Linguistics.
- Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. *Gemma 4 technical report*. Preprint, arXiv:2607.02770.
- Michael Alexander Kirkwood Halliday. 1978. *Language as Social Semiotic*. Arnold, London.
- Marti A. Hearst. 1994. *Multi-paragraph segmentation expository text*. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, USA. Association for Computational Linguistics.
- Marti A. Hearst. 1997. *Text tiling: Segmenting text into multi-paragraph subtopic passages*. *Computational Linguistics*, 23(1):33–64.
- Marti A. Hearst and Christian Plaunt. 1993. *Subtopic structuring for full-length document access*. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in*

- Information Retrieval*, SIGIR '93, page 59–68, New York, NY, USA. Association for Computing Machinery.
- Erik Henriksson, Saara Hellström, and Veronika Laippala. 2025a. [Analyzing register variation in web texts through automatic segmentation](#). In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 7–19, Albuquerque, USA. Association for Computational Linguistics.
- Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo, and Veronika Laippala. 2024. [From discrete to continuous classes: A situational analysis of multilingual web registers with LLM annotations](#). In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 308–318, Miami, USA. Association for Computational Linguistics.
- Erik Henriksson, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. 2025b. [Automatic register identification for the open web using multilingual deep learning](#). *Preprint*, arXiv:2406.19892.
- Erik Henriksson, Alireza Razzaghi, Tuomas Lundberg, Antti Kanner, and Veronika Laippala. 2026. [Register mixing is the norm on the web](#). In *Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities*, pages 138–149, San Diego, USA. Association for Computational Linguistics.
- HPLT Consortium. 2026. HPLT 4.0 Monolingual Data Pool. <https://hplt-project.org/datasets/v4.0>. Version 4.0, accessed 2026-08-05.
- Maximilian Idahl, Benedikt Droste, Björn Plüster, and Jan Philipp Harries. 2026. [propella-1: Multi-property document annotation for llm data curation at scale](#). *Preprint*, arXiv:2602.12414.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. [Dspy: Compiling declarative language model calls into self-improving pipelines](#). *Preprint*, arXiv:2310.03714.
- Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.
- Taja Kuzman and Nikola Ljubešić. 2025. [Automatic genre identification: A survey](#). *Language Resources and Evaluation*, 59(1):537–570.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*.
- Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297.
- Veronika Laippala, Samuel Rönqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079.
- Yang Liu, Chenguang Zhu, and Michael Zeng. 2021. [End-to-end segmentation-based news summarization](#). *Preprint*, arXiv:2110.07850.
- Christian MIM Matthiessen. 2019. Register in systemic functional linguistics. *Register Studies*, 1(1):10–41.
- Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. 2025. [Register always matters: Analysis of LLM pretraining data through the lens of language variation](#). In *Conference on Language Modeling (COLM)*, Montreal, Canada.
- Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, et al. 2025. Hplt 3.0: Very large-scale multilingual resources for llms and mt. mono-and bi-lingual data, multilingual evaluation, and pre-trained models. *arXiv preprint arXiv:2511.01066*.
- Pablo Pavón et al. 2023. Web docs scorer. <https://github.com/pablop16n/web-docs-scorer>. GitHub repository.
- Valentina N. Pescuma, Dina Serova, Julia Lukassek, Antje Sauermaun, Roland Schäfer, Aria Adli, Felix Bildhauer, Markus Egg, Kristina Hülk, Aine Ito, Stefanie Jannedy, Valia Kordon, Milena Kuehnast, Silvia Kutscher, Robert Lange, Nico Lehmann, Mingya Liu, Beate Lütke, Katja Maquate, and 24 others. 2023. [Situating language register across the ages, languages, modalities, and cultural aspects: Evidence from complementary methods](#). *Frontiers in Psychology*, Volume 13 - 2022.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Alireza Razzaghi, Erik Henriksson, and Veronika Laippala. 2026. [ParsCORE: The Persian corpus of online registers](#). In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*, pages 60–73, Rabat, Morocco. Association for Computational Linguistics.
- Liina Repo, Valtteri Skantsi, Samuel Rönqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas

Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. *Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers*. *arXiv preprint arXiv:2102.07396*.

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

A Annotation codebook

The following is the codebook given to the human annotators and embedded verbatim in the model prompt. Only its typesetting has been adapted here.

Order of decisions. Start with `medium`. If it is `cannot_rate`, leave the other values empty. Web pages contain a great deal of text that is not self-standing communication; expect roughly a third of segments to be `cannot_rate`. Do not treat *written* as a fallback for anything you cannot otherwise place — decide `cannot_rate` on its own merits first.

mode_medium (always pick one). A segment is `cannot_rate` when it does not work as a self-standing piece of human communication: titles, headings, labels, menus, link lists, boilerplate, metadata, captions, dates, bare fragments, machine-generated text, and garbled text. A list item qualifies only if it is a complete statement on its own.

Headings are headings whatever their grammar. A segment that names what the following segments are about is a heading even when it is a full sentence or a question. The test is positional, not grammatical: the next segment starts the exposition rather than continuing the target’s sentence, and the target announces a topic instead of saying anything about it. Question-form headings (“Why is X worth having?”, “How do I do Y safely?”) are common on generated articles and are `cannot_rate`.

Judge the segment on its own text, not on the page around it. A segment is `cannot_rate` only when its own words do not form coherent communication. A segment that reads as a coherent statement stays `rateable` even when the surrounding page is machine-generated, even when it is logically unrelated to the segments around it, and even when it contains typos or a wrong word. Spun or machine-translated text is `cannot_rate` only where the segment itself is word salad: unrelated words, product names, or other-language fragments spliced together inside the one segment; implausible synonyms that leave the segment itself meaningless (“Bond release of Forex Megadroid”, “the

too much to handle clutter”); the search keyphrase stuffed into the segment; or leftover spinner markup (“{Reciprocal}”) and words run together (“and-our”). A blurb on a directory or listing page is `rateable` when the blurb itself is a coherent statement, even when its neighbours are unrelated entries.

Judge this in context, not in isolation. A segment may look like a fragment on its own but read as part of a continuous passage with the segments before and after it (for example, a single sentence split across segments). When it is coherent in context, rate it as part of that passage. This is the test for list items too: a list item is `rateable` when it reads as a complete statement, whether on its own or together with its surrounding context. Dates, headings, and labels stay `cannot_rate`.

For anything that is `rateable`, pick how it was made. *Transcribed*: a record of words actually spoken aloud — transcripts, interviews, speeches, captions or subtitles, or text carrying speaker labels or turn markers; use this only when the text is clearly a transcription of speech. *Written*: everything else, text composed to be read; casual or chatty tone does not make it transcribed (comments, posts, messages, and emails are all written).

mode_turn (skip if `cannot_rate`). *Dialogic* only when the text is part of a real exchange where others can reply: forums, comments, social posts, chats, interviews. *Monologic* otherwise: one author addressing an audience that cannot reply in place (articles, guides, FAQs). This is the default. Dialogue inside a fictional or scripted work is still monologic, since it has one author and is not a real interaction.

field_purpose (skip if `cannot_rate`).

- *explaining* — conveying how something is or works so the reader understands it: concepts, definitions, mechanisms, background. The information is presented as established fact, not as a claim the author is out to prove. E.g. an encyclopedia-style paragraph on how a process works.
- *supporting* — establishing a specific factual claim by offering evidence for it: data, study results, sources, or firsthand findings presented as proof. Typical of research and evidence-based writing. If the text just presents accepted information without defending a contested claim, it is explaining, not supporting.

- *recounting* — relating specific real events or experiences presented as having actually happened: a trip report, news account, incident writeup, or personal anecdote. May be a single event and need not be in time order — what matters is that these are particular real happenings. Not general facts (explaining), not a summary of fiction (a plot recap is explaining, not recounting), and not events still to come.
- *directing* — telling the reader to do something, or how to do it: instructions, steps, recipes, how-to guidance, advice on what to choose. Impersonal advice still directs (“do X”, “one should do X”), but a statement that something is important or beneficial, or a document’s own rules and legal terms, is explaining — directing tells the reader to act.
- *evaluating* — arguing a position: a judgment, recommendation, or claim about what is good or bad or what should be done, backed by reasons. Covers reviews, opinion pieces, and advocacy — anything that makes a case the reader could agree or disagree with.
- *expressing* — voicing a feeling, reaction, or stance for its own sake, without arguing for it. The point is how the author feels, not a case they make. E.g. a comment that simply says the author loves or hates something.
- *promoting* — the author marketing their own product, service, or organization.
- *creating* — literary or artistic text as the main content: poetry, lyrics, scripture, fiction, crafted narrative prose. Story form alone is not recounting; if it reads as fiction or literature, it is creating.

Prefer the simplest purpose that fits; only reach for a richer label when its defining feature is actually present. Use *evaluating* only when the author argues a position (not for merely opinionated or topical text), *supporting* only when evidence is offered for a claim (not for plain description), and *recounting* only for specific real events (not general facts, and not summaries of fiction).

When two are close. *Evaluating* vs. *expressing*: reasons given = evaluating; bare reaction = expressing. Ask whether the reader could argue with the reasoning or only with the taste. Opinion pieces and advocacy that give reasons are evaluating, even when passionate. *Explaining* vs. *supporting*: neutral, accepted information you are meant to under-

stand = explaining (an encyclopedia entry); a specific claim defended with evidence so you will believe it = supporting (a study’s results). Facts alone are not supporting — ask whether the passage is proving a point or just informing. *Explaining* vs. *recounting*: general, timeless facts = explaining; specific things that actually happened = recounting. *Explaining* vs. *directing*: how something works, told so you understand it = explaining; what you should do about it = directing. Advice phrased as a fact about what is important or effective is still directing. *Explaining* vs. *promoting*: promoting is the author marketing their own product, service, or business to attract customers — a seller’s own description of what they offer is promoting even when the tone is factual. Two limits keep it from over-firing: a third party describing an organisation it does not own is explaining, and text describing a product’s drawbacks or harms is not promoting. *Recounting* vs. *creating*: really happened = recounting; fiction/literary = creating. *Recounting* vs. *expressing*: events for their own sake = recounting; events as setup for a present feeling or plea = expressing.

B Prompt

The user prompt is the codebook of Appendix A, followed by the output instructions below, the page URL, and the rendered context block. The system message is: “You are an expert linguistic annotator. You label one TARGET text segment at a time along three dimensions, using the surrounding segments only as context. You respond with a single JSON object and nothing else.”

Output instructions. These state in words what the annotation tool enforces through its interface for the human annotators.

Output:

- Label the TARGET segment only; use the surrounding segments as context.
- Always give mode_medium. If mode_medium is cannot_rate, set the other five fields to null.

Respond with a single JSON object with exactly these keys and nothing else:
mode_medium, mode_turn, field_purpose

In the *reasoning* configuration used throughout the paper, the schema additionally carries a leading reasoning string and the prompt ends with: “First fill the reasoning field with ONE short sentence (25 words max) naming the deciding cue for the target’s purpose (and its medium if unclear). Then give the labels. Do not exceed one sentence.”

Context block. Segments are rendered with their offset from the target; where the local window does not reach the start of the document, the first five segments are shown separately as the document opening. The following is a real example, illustrating why segments must be judged in context: HPLT’s segmentation has split a single sentence of an abstract across several segments, and the document opening consists of navigation labels.

URL: [https://www.researchgate.net/profile/\[anonymized\]](https://www.researchgate.net/profile/[anonymized])

```
[Document start]
[seg 0] Network
[seg 1] Followers
[seg 2] Following
[seg 3] Research
[seg 4] Research Items (196)
[... segments omitted ...]
[Context around the target]
[-5] calibration and archival at the
    scale of HUC 12 watersheds
[-4] (Hydrologic Unit Code) anywhere in
    the continental US. Our prototype
[-3] for model-data integration focuses
    on creating tools for fast data
[-2] storage from selected national
    databases, as well as the
[-1] computational resources necessary
    for a dynamic, distributed
    watershed simulation. The paper
>>> [TARGET] will describe cyber-
    infrastructure tools and workflow
    that attempts to
[+1] resolve the problem of model-data
    accessibility and scalability such
[+2] that individuals, research teams,
    managers and educators can create a
[+3] WVO in a desired context. Examples
    are given for the NSF-funded Shale
[+4] Hills Critical Zone Observatory and
    the European Critical Zone
[+5] Observatories within the SoilTrEC
    project. In the future
```

Both the human annotator and the model labelled this segment *written / monologic / explaining*; the model’s rationale was “The text describes the contents and goals of a paper, which is a factual explanation of the work’s scope.”

C Multi-dimensional analysis: features and filtering

Parsing. All eight samples are parsed into Universal Dependencies (de Marneffe et al., 2021) with Stanza (Qi et al., 2020). We parse only segments whose segment-level language, identified with OpenLID, matches the collection’s, which keeps 87.9–97.4% of them.

Feature selection. A feature enters the analysis only if it is attested in all eight samples. This ex-

cludes the entire morphological layer: the morphological feature layer FEATS is language-specific, and the eight treebanks share none of it; Chinese annotates 13 feature–value pairs against the 84 in Polish, and Swedish marks no verb person. We then remove features correlating above $r = 0.99$, features that are exact linear functions of the rest, and features whose squared multiple correlation falls below 0.20, leaving 47.

Factoring. Features are standardized within their own sample before the samples are pooled, so that a dimension is a pattern of within-language covariation rather than a contrast between languages. The correlation matrix is factored by iterated principal axis factoring and rotated with promax ($\kappa = 4$). Bartlett’s test rejects sphericity ($p < 10^{-16}$) and the Kaiser–Meyer–Olkin measure is 0.665. A segment’s score on a dimension is the sum of its standardized features loading at $|\lambda| \geq 0.35$, with negative loadings subtracted (Biber, 1988).