ParsCORE: The Persian corpus of online registers

Alireza Razzaghi and Erik Henriksson and Veronika Laippala
TurkuNLP, University of Turku
{alireza.razzaghi, erik.henriksson, mavela}@utu.fi

Abstract

Despite recent advances in automatic web register (genre) labeling and its applications to web-scale datasets and LLM development, the effectiveness of these tools for digitally low-resource languages remains unclear. This study introduces ParsCORE, the first large-scale collection of Persian web registers (genres), and evaluates deep learning models for register classification and keyword analysis across major registers. Using 2,000 human-annotated documents, the models achieved a micro F1-score of 0.76. The findings provide a foundation for future research on the linguistic and cultural specificities of Persian registers.

1 Introduction

Register, sometimes referred to as genre, denotes text classes such as news articles, advertisements, and how-to pages (Biber, 1988; Biber and Conrad, 2019). Registers are defined based on the relationship between the typical linguistic features of a text and its situational context. Recently, register has also been reconceptualized as a cultural construct, highlighting the need to understand the specificities of the communicative context in the modeling of registers (Biber and Egbert, 2023).

Recent advances in web register and web genre identification have resulted in reliable register identification tools (Lepekhin and Sharoff, 2021; Kuzman and Ljubešić, 2023; Henriksson et al., 2024) that can be used to add register metadata to Web datasets such as HPLT (Burchell et al., 2025). This substantially improves the usability of such datasets. In line with FineWeb-Edu (Lozhkov et al., 2024), register information provides an effective mechanism for web data sampling, including the selection of documents for LLM training (Myntti et al., 2025). Additionally, register-labeled web datasets offer unprecedented possibilities for linguistics, with particular relevance for lower-resourced and minority languages.

Existing register and genre identification tools have demonstrated significant multilingual capacities (Henriksson et al., 2024; Repo et al., 2021; Lepekhin and Sharoff, 2021; Kuzman et al., 2023). However, these tools are largely based on European languages, such as English, Finnish, and Slovenian, with other languages represented only in small evaluation examples. The characteristics of online registers and their variation in non-European languages still remain largely unknown. In this paper, we target this lack by presenting resources for examining Web registers in Persian.

We present ParsCORE v0.1, the first large-scale manually annotated corpus of web registers in Persian. ParsCORE targets the entire unrestricted web, following the annotation standards established in the Multilingual CORE Corpora (Henriksson et al., 2024; Erten-Johansson et al., 2024; Skantsi and Laippala, 2023). These standards define a hierarchical register scheme of 9 main and 14 subregister classes. This provides a solid basis for both automatic web register identification and the linguistic analysis of language use online. In this paper, we present the corpus compilation process, its register distribution, and the linguistic characteristics of the registers using text dispersion keyness (Egbert and Biber, 2018). Finally, we report initial experiments applying the corpus for automatic register identification. The ParsCORE, codes, and results are available on GitHub.

2 Related Work

Web registers (or genres) have been studied in many corpora (see Kuzman and Ljubešić 2023 for a survey). Notably, the Corpus of Online Registers of English (CORE) (Biber and Egbert, 2018) was the first to target the full unrestricted web rather than predetermined categories. CORE's data-driven annotation scheme allows hybrid documents to be assigned multiple registers simultane-

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 60–73

ously (Biber et al., 2020, 2015). This design addresses challenges posed by the unrestricted web, where many documents combine features of several registers.

Despite the general underrepresentation of digitally low-resource languages, several web register corpora have been developed for such languages. Examples include GINCO, a web genre corpus for Slovenian (Kuzman et al., 2022), the Finnish Corpus of Online Registers (Skantsi and Laippala, 2023), the Turkish web register corpus (Erten, 2023), and Swedish and French web registers (Hellström et al., 2025). Multilingual CORE (Henriksson et al., 2024) extended this line of work by applying the CORE taxonomy across 16 languages, integrating several language-specific resources. Similarly, Sharoff (2018) proposed the Functional Text Dimensions (FTD) approach. This method models texts by their similarity to functional prototypes to enable analysis of hybrid documents and varying degrees of register specificity in five languages.

Keyword analysis—used here to examine Persian registers (Section 5.1)—is a corpus-linguistic method for identifying distinctive features of a text collection (Scott, 1997; Bondi, 2010; Culpeper and Demmen, 2015). We apply the dispersion-based method (Egbert and Biber, 2018), which identifies vocabulary distributed across documents rather than concentrated in a few outliers (Gries, 2021; Sönning, 2023). Previous cross-linguistic studies demonstrate that register variation can involve language-specific grammatical features, such as honorific marking in Korean (Biber, 1995). Register variation can also reflect culturally specific features, such as the strategic use of a perfective suffix in Turkish (Erten, 2023). Hellström et al. (2025) showed that cross-linguistic divergences often arise from grammatical realizations and register-specific traits. These findings highlight the importance of cross-linguistic register analysis. The present study contributes a Persian perspective.

Recent studies show strong performance on automatic register identification using manually annotated corpora. Henriksson et al. (2024) reported a micro F1 of 79% on the Multilingual CORE. This performance is obtained using multi-label classification with 25 registers and outperforms previous approaches such as Kuzman et al. (2022). Their results show that multilingual training benefits languages and register classes with limited data. Kuzman et al. (2022) reported a micro F1 of 78% on twelve genre classes aligned with the CORE schema using XLM-R (Conneau et al., 2020). In a cross-lingual experiment, they trained models on machine-translated English versions of GINCO and evaluated them on CORE texts, achieving a micro F1 of 63%. Kuzman et al. (2023) further examined different annotation schemes, including CORE, FTD (Sharoff, 2018), and GINCO. The best performance was achieved on the X GENRE dataset, which integrates all three corpora into nine classes. Using XLM-R Base with single-label classification, the model reached a micro F1 of 80%.

Building on this work, the present study focuses on Persian, a digitally low-resource language lacking extensive web register data and multi-label models. The Multilingual CORE (Henriksson et al., 2024) contains only 69 Persian instances. In contrast, our study presents a larger Persian web register corpus and implements multi-label classification, representing an initial step toward large-scale multilingual register analysis that incorporates Persian.

Narrative NA	359 | 569
News report, Sports report, Narrative blog, Other narrative	
Opinion OP	149 | 353
Review, Opinion blog, Advice, Denominational religious blog / sermon, Other opinion	
Informational Description IN	251 | 439
Description of a thing or a person, Research article, Encyclopedia article, FAQ, Legal terms and conditions, Other Informational description	
Interactive Discussion ID	43 | 190
How-to HI	35 | 146
Recipe, Other how-to	
Informational Persuasion IP	287 | 438
Description with intent to sell, News & opinion blog or editorial, Other informational persuasion	
Lyrical LY	28 | 199
Spoken SP	12 | 153
Interview, Other spoken	
Machine Translated MT	20 | 32

Table 1: Main registers and sub-registers with the number of documents before and after upsampling.

3 Data

We employed a randomized 2,000-document sample from the Persian portion of HPLT 3.0, a large-scale web-crawled monolingual dataset covering 198 languages and drawing on both general web content and internet archives. It contains approximately 3.3 petabytes of so-called "wide crawls" from the Internet Archive, covering the period from 2012 to 2020 and 57 full snapshots from Common Crawl spanning 2014 to 2025. Together, these sources contribute to a total volume of about 7.2 petabytes of raw web archive data. The Persian section of HPLT 3.0 contains 124.02M documents, and a MinHash-based global near-deduplication is implemented. Moreover, HPLT 3.0 provides automatic register annotations for Persian, assigning each document a confidence score between 0.0 and 1.0 for each register. We chose HPLT over raw Common Crawl data because it is cleaned and deduplicated, yielding less noisy and more reliable text for register analysis. The register annotations reported here apply only to the specific sample used in this study.

The sample is human-annotated based on the hierarchical CORE register scheme. The definition is available at the annotation guidelines, and the abbreviations are presented in Table 7 (Appendix). The annotation followed a two-step procedure, in which documents were first accepted or rejected and then assigned register labels. A total of 700 documents were rejected during annotation. The purpose of rejection was to restrict register annotation to full, coherent texts. Documents were rejected if they (1) consisted mainly of headline-style fragments, lists, download pages, or captions; (2) contained fewer than two complete sentences or lacked coherent text; (3) were dominated by boilerplate or 'junk' content; (4) were poorly extracted, not in the target language; or (5) consisted mostly of special characters or numbers.

As a document might have multiple registers, a multi-label tagging approach was taken with a holistic perspective. The proportion of a register across the whole document has been considered for the multi-label tagging. A hybrid document could be in multiple forms. One type is a single coherent document about a topic that has multiple registers. Another is a document consisting of short paragraphs about different topics or different aspects of a topic that are united.

In addition to register labels, a separate "Tags" column is defined for metadata. For accepted documents, this column can be empty, "OTHLI" (code-switching to other languages such as English or Arabic), or "FINGLISH" (Persian transliterated in the Latin alphabet). Rejected documents are tagged as "JNK" (junk), "MLPT" (more than three main registers), or "OTHL" (text in another language).

"MLPT" texts are longer sections where coherence may be lacking, and the register can shift, sometimes including more than three distinct registers. "JNK"texts are short, headline-like sentences that, while complete, cannot establish a register. Although texts with excessive main registers may still be suitable for next-token prediction tasks, they do not constitute coherent register instances and were therefore excluded from register annotation. Examples of the most frequent Tags are presented in Table 2. Figure 1 shows the proportion of tags within each main register, highlighting that code-switching varies across registers. Code-switching is less frequent in the Narrative register, but more common in How-to and Informational Persuasion.

In the first batch of 1,100 randomly extracted instances from HPLT3, we observed a highly skewed class distribution, with some main registers occurring fewer than 100 times.. To address this imbalance, we applied an upsampling strategy. URLs of rejected texts were excluded from further sampling, and a minimum of 145 instances was set for underrepresented register classes. To reach this threshold, we selected instances from HPLT3 with automatic labeling confidence between 0.2 and 0.4. These scores fall below the 0.4 confidence threshold established in Burchell et al. (2025) and were intended to target difficult cases for manual annotation. Table 1 presents the hierarchical register schema along with instance counts from random sampling and after upsampling; hybrid documents were counted multiple times, once for each register they contain.

4 Methods

4.1 Keyness

In this section, we focus on how keyness was calculated for the registers in our corpus. We use the term "token" rather than "word" because text was segmented by whitespace rather than processed using a linguistically informed tokenization method. Considering the nature of the Persian writing sys-

Translation	Text	TAG
*[…]*Crowdedness, reason for not burying Morteza Pashaei Ali Daei's resemblance to the Hollywood actor + Photo The Asian champion peddling on the street*[…]*	ازدحام جمعیت عامل عدم دفن مرتضی پاشایی شباهت علی دایی با بازیگر هالیود + عکس قهرمان آسیا در کنار خیابان دست فروشی می کند	JNK
Puffer fish contain a toxin called tetrodotoxin, which is 1,200 times more deadly to humans than cyanide*[…]* It's not fair that in this big city, you are the proverbial needle in a haystack. Three days ago, after making the necessary plans, Persepolis Club sent this program to the representatives of the Milan Stars team, and *[…]*	ماهی پف کننده حاوی سمی به نام تتراتوکسین است که برای انسانها 1200 بار کشنده تر از سم سیانور می باشد... اِنصاف نباشد که در این شهرِ دَرَندَشت ضرب المثل سوزن در کاه تو باشی. باشگاه پرسپولیس سه روز پیش پس از برنامه ریزي هاي لازم اين برنامه را براي نمايندگان تيم ستارگان ميلان ارسال كرد و...	MLPT
*[…]*The rotation speed of this hard drive is 7200 revolutions per minute (RPM) and uses a high-speed SATA III interface with a speed of 6 Gbps for data transfer. -Memory: 2 TB -Size: 3.5 inch -Connection: SATA 3.0	...سرعت چرخش این هارد درایو 7200 دور در دقیقه (RPM) بوده و از رابط پرسرعت $SATAIII$ با سرعت 6 گیگابیت بر ثانیه، برای انتقال اطلاعات بهره می برد -Memory: 2 TB -Size: 3.5 inch -Connection: SATA 3.0	OTHLI

Table 2: Tags Examples.

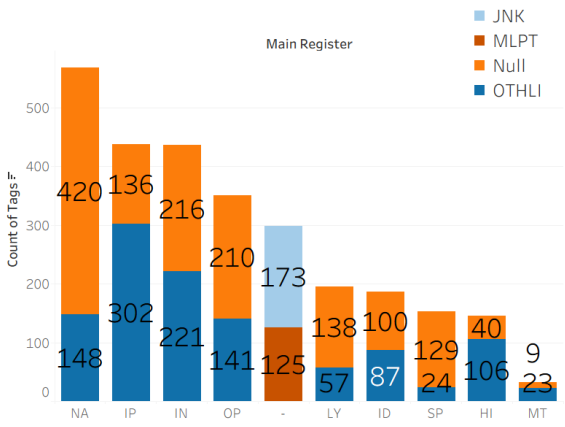

Figure 1: Distribution of Tags on registers

tem, where letters may be concatenated or written separately depending on orthographic conventions, certain elements (e.g., the present tense marker "می") may appear as independent tokens, and a single word may be split across multiple tokens. For each token, both document frequency and term frequency were calculated. Tokens were then ranked according to their log likelihood (G^2) scores to identify most distinctive vocabulary of each register.

4.2 Register labeling

Since web documents frequently exhibit characteristics of multiple registers, we formulate register identification as a multi-label classification task.

Automatic register labeling was performed using XLM-R Large (Conneau et al., 2020) and BGE-M3 (Chen et al., 2024), with maximum input sequence lengths of 512 and 1024 tokens, respectively. These models were selected for two main reasons. First, XLM-R Large has been shown to achieve state-of-the-art performance in multilingual web register identification, consistently outperforming monolingual and smaller multilingual

models across a wide range of languages and experimental settings (Henriksson et al., 2024). Second, XLM-R Large and BGE-M3 provide strong multilingual representations. In addition, BGE-M3 was included for its longer context window (up to 8,192 tokens), which is particularly important for register classification, as register cues may be distributed across extended document spans rather than concentrated locally.

We evaluated the models under four training and evaluation settings:

1. **Monolingual**: trained and evaluated on Persian.

2. **Multilingual-1**: trained on a mixture of Persian and Multilingual CORE, evaluated on Persian.

3. **Multilingual-2**: trained and evaluated on a mixture of Persian and Multilingual CORE.

4. **Zero-Shot**: trained on Multilingual CORE and evaluated on Persian.

The data was split for training and evaluation. In the Persian-only setting, 70% of the data was used for training, while the remaining 30% was reserved for evaluation purposes and model tuning. In the zero-shot setting, 80% of the data was allocated for training. For the Multilingual-1 setting (see Section 4.2), the full multi-CORE corpus was combined with 50% of the Persian data for training, with the remaining Persian data used for evaluation. In the Multilingual-2 setting, 90% of the multi-CORE corpus was combined with 50% of the Persian data for training, and the remaining data was held out for evaluation.

Model optimization was carried out using Bayesian hyperparameter optimization with the Optuna default optimizer, which is the Tree-structured Parzen Estimator (TPE) sampler. Continuous hyperparameters were sampled from predefined ranges, including the learning rate $[5 \times 10^{-6}, 3 \times 10^{-5}]$ on a logarithmic scale, weight decay $[0.0, 0.1]$, and warmup ratio $[0.0, 0.15]$. Discrete hyperparameters included batch size $4, 8$ and gradient accumulation steps $4, 8$. The maximum number of training epochs was set to 10.

Register	Translation	Keyword	Keyness
NA_ne	Report	گزارش	109.884
NA_ne	Added	افزود	103.847
NA_ne	He/She	وی	82.510
NA_ne	Deputy	معاون	81.432
NA_ne	Reporter	خبرنگار	79.825
NA_ne	Head of/Boss	رئیس	69.856
NA_ne	Specify/State	تصریح	66.140
NA_ne	Minister	وزیر	64.012
NA_ne	(X) Council	شورای	56.164
NA_ne	Ministry	وزارت	53.543
IP_ds	Product	محصول	76.842
IP_ds	Comments/Opinions	دیدگاهی	50.065
IP_ds	Products	محصولات	40.0852
IP_ds	Specifications	مشخصات	37.653
IP_ds	Capability/Capable	قابلیت	33.982
IP_ds	Product	کالا	29.970
IP_ds	Dimensions	ابعاد	29.370
IP_ds	Material	جنس	29.200
IP_ds	Frame/Body	بدنه	28.821
IP_ds	Warranty	ضمانت	28.375
IP_oe	Download	دانلد	138.783
IP_oe	(Down)Load	لوڈ	138.783
IP_oe	Download	داونلود	138.783
IP_oe	Download	ddl	138.783
IP_oe	Down(load)	ڈاؤن	138.783
IP_oe	Download	donload	138.783
IP_oe	usnet	usnet	138.783
IP_oe	Download	nhkgn	138.783
IP_oe	Download	danload	138.783
IP_oe	uznet	یوزنت	138.783
IN_dtp	Qajar	قاجار	20.471
IN_dtp	1080p	1080p	19.778
IN_dtp	Herbert	هربرت	19.778
IN_dtp	Jordan	اردن	17.191
IN_dtp	It has been	شدهاست	15.560
IN_dtp	Architecture	معماری	14.793
IN_dtp	Fame	شهرت	13.412
IN_dtp	2020	2020	13.363
IN_dtp	(X) Studio	استودیوی	13.363
IN_dtp	Small stick (Gillidanda)	پیل	13.199
OP_av	Benefits	فواید	39.407
OP_av	Hormonal	هورمونی	33.001
OP_av	Don't	نکنید	31.440
OP_av	Avoid	اجتناب	30.996
OP_av	Fat	چربی	30.0467
OP_av	Diabetes	دیابت	29.656
OP_av	Body	بدن	29.259
OP_av	Do/Give	دهید	27.824
OP_av	Inflammatory	التهابی	24.363
OP_av	You can	بتوانید	23.874

Table 3: 10 top keywords from the five highest-percentage registers

5 Evaluation

5.1 Keyness

Table 3 presents the top ten keywords for the most frequent registers in our data: News, Description with Intent to Sell, Other Informational Persuasion (which broadly persuades but lacks characteristics of any more specific register), Description of a Thing or Person, and Advice. These registers were selected for analysis based on their frequency; the

top 20 registers are presented in Table 4, showing that hybrids are less common. The full distribution of registers is available in the Appendix, Table 8. The results exhibit a long-tail distribution, in which primary registers account for most documents, while diverse hybrid combinations occur less frequently.

Register	Percentage
-	14.95%
NA_ne	10.46%
IP_ds	5.80%
IP_oe	4.15%
IN_dtp	3.70%
OP_av	3.05%
LY	3.00%
NA_on	2.90%
SP_os	2.30%
SP_it	1.65%
MT	1.55%
IN_ra	1.35%
NA_nb	1.35%
ID	1.30%
NA_sr	1.20%
LY+ID	1.10%
IN_dtp+OP_av	1.00%
IP_ds+IN_fi	0.90%
NA_on+LY	0.85
IN_oi	0.85%

Table 4: Distribution of registers.

In the "News" register, most keywords are nouns related to administration and governance, a pattern also observed for Turkish (Erten, 2023). In contrast to findings for Turkish and French (Hellström et al., 2025), our analysis did not identify past-tense verbs as salient keywords. This absence may be due to tokenization issues in Persian, where verb constructions can be split into multiple tokens by spaces or half-spaces. For example, the noun "گزارش" ('report') becomes a verb in the construction "گزارش کرد" ('reported'), which may prevent such forms from being captured as single verbal units in keyword analysis. Additionally, the pronoun "وی", equivalent to "he" or "she", is characteristic of narrative usage, where it is commonly employed in recounting events. The same characteristic of the Narrative register occurs in English, concerning recounting events (Biber and Egbert, 2018).

In the "Description with the intention to sell" class, a product or a service is described, with the purpose of selling and overt marketing. The keywords include nouns related to products or their features, such as "جنس" {'Material'}, "ابعاد" {'Dimensions' }, and "مشخصات"" {'specifications'}. Unlike English product descriptions, which tend to make extensive use of adjectives and evaluative language (Biber and Egbert, 2018), Persian tends toward noun repetition, despite the grammatical possibility of noun-adjective modification.

The "Other informational persuasion" register is characterized by frequent occurrences of the word "download", either as a single token or split according to orthographic conventions, and contains the highest proportion of non-Persian tokens. The token "nhkgn" represents the Persian word for "download" typed with a Persian keyboard while the input language remains English, a practice common on software distribution websites. Similarly, "usnet" and "uznet" are alternative transliterations of "یوزنت", a broadband service provider.

In the "Description of a thing or a person" register, proper names such as Qajar, Jordan, and Herbert appear among the keywords. Some items point to Persian-specific historical or cultural references. For example, "Qajar" refers to a formerly aristocratic Iranian dynasty that ruled Iran from 1789 to 1925, while "Jordan" reflects the long, complex historical relationship between Iran and Jordan, dating back to the Achaemenid Empire according to Wikipedia. The word "پیل" denotes the "Gilli" in the traditional game Gillidanda "پیل دسته", also known as "الک دولک".

Other tokens likely reflect sampling randomness rather than register-specific properties. For instance, "Herbert" appears in reference to multiple entities, including Herbert Le Porrier, Herbert Hoover, and the Herbert Berghof Studio. Similarly, tokens such as "1080p" and "2020" typically indicate video resolution or production year, reflecting broader media-description conventions rather than Persian-specific features.

In the "Advice" register, which is based on opinions that suggest actions to solve a particular problem, keywords are primarily health-related nouns, content-focused and topic-driven, rather than narrative. At the same time, the relatively high diversity of verbs reflects the directive nature of the advice register, where actions and recommendations are foregrounded, often through imperative forms (e.g., negative imperatives such as "don't"). Moreover, lexical borrowing with phonological adapta-

tion has happened to health-related nouns, for instance, like “Hormonal” and “Diabetes”, instead of writing them in English, and in contrast with the “Informational Persuasion” category.

It is essential to note that the symbol (X) in Table 3 marks the presence of the suffix ی, which allows a word to function as a head noun and take dependents such as adjectives or nouns. However, this only happens when a word ends with a vowel, such as “شورا” {“Council”}, while for other words like “وزارت” {“Ministry”} that end with a consonant, adding a {/e/} vowel at the end of the word will do the same. Moreover, in the Persian orthography of Iran, writing vowels {/æ/}, {/e/}, and {/o/} is omitted. Therefore, it cannot be generalized that using head nouns that take dependents is much frequent in a specific register by just looking at separate words.

Taken together, the analysis reveals both features that align with patterns reported for other languages and keywords that reflect language-specific characteristics of Persian. This highlights the need for language-specific approaches for modeling registers.

5.2 Register labeling

Model(max_lenght)	Setting	F1 Score(μ)
XLM-R (512)	Monolingual	0.72
XLM-R (512)	Multilingual-1	0.75
XLM-R (512)	Multilingual-2	0.75
XLM-R (512)	Zero-shot	0.76
bge-m3 (1024)	Monolingual	0.74
bge-m3 (1024)	Multilingual-1	0.75
bge-m3 (1024)	Multilingual-2	0.75
bge-m3 (1024)	Zero-shot	0.76

Table 5: Model performances.

We assess model performance using micro F1, computed by aggregating true positives, false positives, and false negatives across all labels before calculating precision and recall. In a multi-label setting with imbalanced class distributions, micro F1 is well-suited as a performance measure, as it weights labels proportionally to their frequency and reflects the model’s effectiveness across all individual label decisions. The results of the automatic register labeling are presented in Table 5. The multilingual and zero-shot settings achieved comparable performance, with micro F1 scores close to 0.75. The multilingual training improved model performance, with the monolingual Persian-only model achieving 0.72 micro F1.

In addition to the overall micro F1 score, the classification report in Table 6 shows that performance varies substantially across registers. Frequent and well-represented registers (e.g., NA, IP) achieve higher F1 scores, while sparse registers (e.g., ed, en, lt) show lower performance. Among the main registers, NA (F1 = 0.84), SP (F1 = 0.92), and IP (F1 = 0.74) show a relatively balanced precision–recall trade-off, indicating stable performance for these categories. In contrast, other main registers exhibit larger discrepancies between precision and recall, suggesting less consistent detection. General performance on subregisters is weaker and more variable. Subregister classes such as ds (F1 = 0.82) and ne (F1 = 0.85) achieve comparatively strong results with a balanced precision–recall trade-off. However, several low-resource subregisters show low F1 scores and large precision–recall imbalances. In several low-support classes, high precision but low recall suggests the model makes conservative predictions, reflecting the difficulty of learning reliable decision boundaries in a multi-label setting. Results for classes with very limited support are less reliable and further highlight the difficulty of fine-grained subregister classification under data-sparse conditions. Therefore, additional and cleaner data might be required to achieve better performance.

Previously, Henriksson et al. (2024) reported a best average micro F1 of 0.77 across all languages (0.79 on main labels) using XLM-R, XLMR-XL, and BGE-M3 (2048 tokens) in the multilingual setting. However, the multilingual CORE dataset is approximately ten times larger than the Persian corpus, and Persian may exhibit distinctive linguistic features. Moreover, even among the best models in Henriksson et al. (2024), performance varies substantially across languages, with micro F1 scores ranging from 0.72 to 0.81 (Table 3).

In addition, based on manual analysis, many instances in our Persian sample from HPLT3 contained noisy content, such as irrelevant page tags generated for search engine indexing, short clickable links to other pages, and brief advertising texts. This noise may have hindered model learning and introduced confusion, even when longer context windows were used.

class	precision	recall	f1-score	support
HI	0.59	0.83	0.69	35
ID	0.80	0.68	0.74	47
IN	0.68	0.62	0.65	110
IP	0.73	0.76	0.74	111
LY	0.87	0.94	0.90	50
MT	1.00	0.33	0.50	9
NA	0.84	0.85	0.84	139
OP	0.62	0.49	0.55	87
SP	0.92	0.92	0.92	38
av	0.65	0.54	0.59	37
ds	0.80	0.85	0.82	52
dtp	0.69	0.39	0.50	74
ed	0.50	0.27	0.35	11
en	0.00	0.00	0.00	3
fi	1.00	0.45	0.62	11
it	0.88	0.79	0.83	19
lt	0.50	0.33	0.40	3
nb	0.87	0.76	0.81	17
ne	0.86	0.85	0.85	78
ob	0.60	0.25	0.35	12
ra	0.83	0.62	0.71	8
re	0.71	1.00	0.83	5
rs	0.67	0.25	0.36	16
rv	1.00	0.62	0.77	8
sr	0.89	0.80	0.84	10

Table 6: Classification report

6 Conclusion

This study addresses a notable gap in register-based research by focusing on Persian, a digitally low-resource language that has so far been underrepresented in large-scale web register corpora and multilingual register identification efforts. By combining manual annotation with computational modeling, the work contributes both empirical data and methodological insights relevant to cross-linguistic register analysis. We presented ParsCORE, the first large-scale, human-annotated corpus of Persian web registers, and evaluated register identification models under three different settings as an initial step toward automatic identification. In addition, we conducted keyword analysis across major registers. The results show that model performance is comparable to that reported for digitally high-resource languages. However, additional data is required, and further investigation is needed to identify potential linguistic and cultural specificities of Persian registers. The keyness analysis provides insights into differences across registers and categories. The manually annotated dataset and the model optimization pipeline are publicly available on GitHub. Taken together, ParsCORE provides a foundation for future research on Persian web registers, including the development of more robust multi-label models, improved handling of morphologically complex constructions, and broader cross-lingual comparisons involving both high- and low-resource languages.

Limitations and further work

Regarding inter-annotator agreement, this study follows the taxonomy and annotation schema established in prior work to ensure methodological consistency. Rather than conducting full parallel annotation, only challenging or ambiguous instances were discussed with experts who had previously applied this framework to other languages. This approach was adopted due to the difficulty of identifying experts with specific experience in Persian register variation who are also available and willing to perform manual annotation. Consequently, formal inter-annotator agreement measures were not calculated. Another limitation is that only the beginnings of documents (512 or 1024 tokens) were used due to token constraints. Future research should explore the use of document endings, combined beginning–end segments, and alternative windowing approaches.

Future studies could apply clustering methods to the semantic embeddings of classified documents in order to elucidate register relationships and to qualitatively investigate both correctly classified and misclassified documents (Santini, 2005; Gries, 2021). Moreover, keyness analysis was limited to the top 10 keywords for the five most frequent registers. Future work could extend this analysis to include all registers and a larger set of keywords, providing a more comprehensive view of register-specific lexical patterns across the corpus.

Acknowledgments

Alireza Razzaghi received funding from the European Union's Horizon Europe research and innovation program under the 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. Veronika Laippala and Erik Henriksson received funding from the Research Council of Finland through "FIN-CLARIAH research infrastructure" (project 358720, which has also received funding from the European Union – NextGenera-

tionEU instrument), "Mechanisms of register variation in massively multilingual web-scale corpora" (project 362459), and "Green NLP - controlling the carbon footprint in sustainable language technology" (project 353167). Furthermore, we also wish to acknowledge CSC – IT Center for Science Ltd. for providing computational resources.

References

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*. Cambridge University Press, Cambridge.

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(1):1–22.

Douglas Biber, Jesse Egbert, and Mark Davies. 2015. Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora*, 10(1):11–45.

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

Marina Bondi. 2010. *Perspectives on keywords and keyness: An introduction*, pages 1–18. John Benjamins Publishing Company.

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.

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

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

Jonathan Culpeper and J. Demmen. 2015. *Keywords*, pages 90–105.

Jesse Egbert and Doug Biber. 2018. Incorporating text dispersion into keyword analyses. *Corpora*, 14(1):77–104. Publisher Copyright: © Edinburgh University Press.

Selcen Erten. 2023. Exploring register variation in turkish web corpus. *14–15 September 2023, University of Mannheim, Germany*, page 60.

Selcen Erten-Johansson, Valtteri Skantsi, Sampo Pyysalo, and Veronika Laippala. 2024. Linguistic variation beyond the Indo-European Web: Analyzing Turkish Web registers in TurCORE. *Register Studies*.

Stefan Gries. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9:1–33.

Saara Hellström, Valtteri Skantsi, Anna Salmela, and Veronika Laippala. 2025. From keywords to key embeddings – contrasting french and swedish web registers using multilingual deep learning. *Corpus Linguistics and Linguistic Theory*.

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

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

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO training dataset for web genre identification of documents out in the wild. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1584–1594, Marseille, France. European Language Resources Association.

Mikhail Lepekhin and Serge Sharoff. 2021. Experiments with adversarial attacks on text genres. *CoRR*, abs/2107.02246.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

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 *Second Conference on Language Modeling*.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the english web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*.

Marina Santini. 2005. Genres in formation? An exploratory study of web pages using cluster analysis: Proceedings of the 8th annual colloquium for the UK special interest group for computational linguistics (CLUK05). In *Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK05)*, Manchester.

Mike Scott. 1997. Pc analysis of key words — and key key words. *System*, 25(2):233–245.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

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

Lukas Sönning. 2023. Evaluation of keyness metrics: performance and reliability. *Corpus Linguistics and Linguistic Theory*, 20.

A Appendix

Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb
Other narrative	on
Opinion	OP
Review	rv
Opinion blog	ob
Advice	av
Denominational religious blog / sermon	rs
Other opinion	oo
Informational Description	IN
Description of a thing or a person	dtp
Research article	ra
Encyclopedia article	en
FAQ	fi
Legal terms and conditions	lt
Other Informational description	oi
Interactive Discussion	ID
How-to	HI
Recipe, Other how-to	
Informational Persuasion	IP
Description with intent to sell	ds
News & opinion blog or editorial	ed
Other informational persuasion	oe
Lyrical	LY
Spoken	SP
Interview	it
Other spoken	os
Machine Translated	MT
Tags	
Junk	JNK
Other language included	OTHLI
Other language	OTHL

Table 7: Registers abbreviation

Table 8: Distribution of registers

Register	Count	Percentage	Tag	Tag Count	Tag Percentage
-	173	8.65%	JNK	173	57.86%
-	125	6.25%	MLPT	125	41.81%
-	1	0.05%	OTHL	1	0.33%
NA_ne	209	10.46%	OTHLI	40	19.14%
IP_ds	116	5.80%	OTHLI	84	72.41%
IP_oe	83	4.15%	OTHLI	67	80.72%
IN_dtp	74	3.70%	OTHLI	43	58.11%
OP_av	61	3.05%	OTHLI	20	32.79%
LY	60	3.00%	OTHLI	18	30.00%
NA_on	58	2.90%	OTHLI	16	27.59%
SP_os	46	2.30%	OTHLI	3	6.52%
SP_it	33	1.65%	OTHLI	4	12.12%
MT	31	1.55%	OTHLI	22	70.97%
IN_ra	27	1.35%	OTHLI	13	48.15%
NA_nb	27	1.35%	OTHLI	1	3.70%
ID	26	1.30%	OTHLI	14	53.85%
NA_sr	24	1.20%	OTHLI	4	16.67%
LY+ID	22	1.10%	OTHLI	2	9.09%
IN_dtp+OP_av	20	1.00%	OTHLI	8	40.00%
IP_ds+IN_fi	18	0.90%	OTHLI	14	77.78%
NA_on+LY	17	0.85%	OTHLI	2	11.76%
IN_oi	17	0.85%	OTHLI	5	29.41%
OP_ob	16	0.80%	OTHLI	5	31.25%
IP_oe+LY	16	0.80%	OTHLI	15	93.75%
OP_oo	14	0.70%	OTHLI	2	14.29%
NA_ne+IN_dtp	14	0.70%	OTHLI	5	35.71%
IP_oe+HI_oh	14	0.70%	OTHLI	14	100.00%
IP_ds+ID	13	0.65%	OTHLI	9	69.23%
HI_oh	13	0.65%	OTHLI	12	92.31%
IN_dtp+HI_oh	13	0.65%	OTHLI	11	84.62%
IP_ed	13	0.65%	OTHLI	3	23.08%
IN_dtp+LY	13	0.65%	OTHLI	4	30.77%
IN_dtp+ID	12	0.60%	OTHLI	5	41.67%
HI_re	11	0.55%	OTHLI	1	9.09%
IP_oe+ID	11	0.55%	OTHLI	9	81.82%
IP_ds+HI_oh	10	0.50%	OTHLI	7	70.00%
NA_nb+LY	10	0.50%	OTHLI	2	20.00%
OP_rs	9	0.45%	OTHLI	3	33.33%
LY+IN_dtp	9	0.45%	OTHLI	1	11.11%
IP_ds+IN_dtp	9	0.45%	OTHLI	6	66.67%
OP_av+ID	9	0.45%	OTHLI	4	44.44%
HI_oh+ID	9	0.45%	OTHLI	8	88.89%
NA_ne+IN_oi	8	0.40%	OTHLI	3	37.50%
NA_ne+OP_rs	8	0.40%	OTHLI	2	25.00%
NA_on+OP_oo	8	0.40%	OTHLI	4	50.00%
OP_rv	8	0.40%	OTHLI	4	50.00%
IN_dtp+IP_ds	7	0.35%	OTHLI	6	85.71%
IN_dtp+OP_rs	7	0.35%	OTHLI	3	42.86%
IP_oe+IN_dtp	7	0.35%	OTHLI	3	42.86%

Register	Count	Percentage	Tag	Tag Count	Tag Percentage
SP_os+OP_rs	7	0.35%	OTHLI	2	28.57%
IN_en	7	0.35%	OTHLI	5	71.43%
NA_ne+IP_ed	6	0.30%	OTHLI	3	50.00%
NA_on+OP_av	6	0.30%	OTHLI	4	66.67%
NA_on+IN_dtp	6	0.30%	OTHLI	4	66.67%
IN_dtp+NA_on	6	0.30%	OTHLI	1	16.67%
OP_ob+ID	6	0.30%	OTHLI	2	33.33%
IN_dtp+IP_oe	6	0.30%	OTHLI	5	83.33%
NA_ne+OP_av	5	0.25%	OTHLI	2	40.00%
IN_dtp+OP_rv	5	0.25%	OTHLI	4	80.00%
NA_nb+NA_on	5	0.25%	OTHLI	1	20.00%
NA_ne+SP_it	5	0.25%	OTHLI	2	40.00%
NA_ne+OP_oo	5	0.25%	OTHLI	2	40.00%
IP_ed+SP_it	5	0.25%	OTHLI	1	20.00%
IN_dtp+OP_ob	4	0.20%	OTHLI	3	75.00%
NA_on+SP_it	4	0.20%	OTHLI	2	50.00%
LY+OP_rs	4	0.20%	OTHLI	2	50.00%
OP_av+IP_oe	4	0.20%	OTHLI	2	50.00%
NA_nb+OP_rs	4	0.20%	OTHLI	1	25.00%
OP_av+HI_oh	4	0.20%	OTHLI	2	50.00%
NA_on+OP_rs	4	0.20%	OTHLI	2	50.00%
HI_oh+IN_dtp	4	0.20%	OTHLI	3	75.00%
IP_oe+NA_sr	4	0.20%	OTHLI	4	100.00%
IN_oi+OP_av	3	0.15%	OTHLI	2	66.67%
NA_nb+OP_oo	3	0.15%	OTHLI	1	33.33%
LY+HI_oh	3	0.15%	OTHLI	3	100.00%
IP_oe+SP_it	3	0.15%	OTHLI	2	66.67%
IP_ed+ID	3	0.15%	OTHLI	1	33.33%
IN_oi+ID	3	0.15%	OTHLI	1	33.33%
IN_dtp+HI_oh+ID	3	0.15%	OTHLI	3	100.00%
OP_oo+ID	3	0.15%	OTHLI	1	33.33%
NA_nb+ID	3	0.15%	OTHLI	1	33.33%
HI_oh+IN_oi	3	0.15%	OTHLI	1	33.33%
IN_dtp+SP_os	3	0.15%	OTHLI	2	66.67%
IN_dtp+HI_oh+OP_av	3	0.15%	OTHLI	3	100.00%
OP_av+HI_re	3	0.15%	OTHLI	3	100.00%
IP_oe+NA_ne	3	0.15%	OTHLI	2	66.67%
IP_ds+NA_on	2	0.10%	OTHLI	1	50.00%
IN_dtp+IN_lt	2	0.10%	OTHLI	1	50.00%
ID+IN_dtp+OP_av	2	0.10%	OTHLI	1	50.00%
IN_oi+IP_oe	2	0.10%	OTHLI	1	50.00%
OP_rv+NA_on	2	0.10%	OTHLI	1	50.00%
NA_on+IP_oe	2	0.10%	OTHLI	2	100.00%
IP_ds+OP_av	2	0.10%	OTHLI	1	50.00%
OP_rs+LY	2	0.10%	OTHLI	1	50.00%
IP_ds+LY	2	0.10%	OTHLI	1	50.00%
IP_ds+OP_rv+ID	2	0.10%	OTHLI	1	50.00%
IN_fi	2	0.10%	OTHLI	1	50.00%
NA_on+ID	2	0.10%	OTHLI	1	50.00%
IP_oe+IN_oi+HI_oh	2	0.10%	OTHLI	1	50.00%
IP_ds+IN_oi	2	0.10%	OTHLI	1	50.00%

Register	Count	Percentage	Tag	Tag Count	Tag Percentage
IP_ds+HI_re	2	0.10%	OTHLI	2	100.00%
IP_ds+HI_oh+ID	2	0.10%	OTHLI	2	100.00%
OP_oo+HI_oh	2	0.10%	OTHLI	2	100.00%
IP_oe+HI_oh+ID	2	0.10%	OTHLI	2	100.00%
IN_fi+IP_ds	2	0.10%	OTHLI	2	100.00%
IP_ds+IN_en	2	0.10%	OTHLI	1	50.00%
IN_dtp+NA_ne	2	0.10%	OTHLI	1	50.00%
IP_ds+IP_oe	2	0.10%	OTHLI	1	50.00%
OP_av+IN_dtp	2	0.10%	OTHLI	1	50.00%
OP_ob+NA_ne	2	0.10%	OTHLI	1	50.00%
NA_sr+IP_oe	2	0.10%	OTHLI	1	50.00%
NA_ne+IP_oe	2	0.10%	OTHLI	2	100.00%
IN_ra+IP_ds	1	0.05%	OTHLI	1	100.00%
OP_av+OP_rs	1	0.05%	OTHLI	1	100.00%
SP_os+OP_rs+IN_dtp	1	0.05%	OTHLI	1	100.00%
IN_dtp+IN_fi+IP_ds	1	0.05%	OTHLI	1	100.00%
NA_ne+IN_lt	1	0.05%	OTHLI	1	100.00%
SP_it+OP_oo	1	0.05%	OTHLI	1	100.00%
NA_on+IN_dtp+OP_oo	1	0.05%	OTHLI	1	100.00%
NA_on+LY+NA_nb	1	0.05%	OTHLI	1	100.00%
OP_av+OP_oo	1	0.05%	OTHLI	1	100.00%
NA_nb+NA_on+OP_ob	1	0.05%	OTHLI	1	100.00%
OP_oo+NA_on	1	0.05%	OTHLI	1	100.00%
LY+HI_oh+IP_oe	1	0.05%	OTHLI	1	100.00%
OP_rv+LY+ID	1	0.05%	OTHLI	1	100.00%
HI_oh+IP_oe+LY	1	0.05%	OTHLI	1	100.00%
IN_dtp+IP_ds+LY	1	0.05%	OTHLI	1	100.00%
IN_en+ID	1	0.05%	OTHLI	1	100.00%
IN_oi+ID+HI_oh	1	0.05%	OTHLI	1	100.00%
IN_dtp+ID+OP_rv	1	0.05%	OTHLI	1	100.00%
NA_nb+HI_oh+ID	1	0.05%	OTHLI	1	100.00%
IP_oe+IN_dtp+ID	1	0.05%	OTHLI	1	100.00%
ID+OP_rs	1	0.05%	OTHLI	1	100.00%
NA_nb+OP_oo+ID	1	0.05%	OTHLI	1	100.00%
OP_rs+IN_oi+ID	1	0.05%	OTHLI	1	100.00%
ID+IN_oi+OP_oo	1	0.05%	OTHLI	1	100.00%
IP_ds+ID+OP_rv	1	0.05%	OTHLI	1	100.00%
HI_oh+OP_av	1	0.05%	OTHLI	1	100.00%
NA_ne+SP_os+OP_oo	1	0.05%	OTHLI	1	100.00%
OP_av+NA_ne	1	0.05%	OTHLI	1	100.00%
NA_nb+SP_it+IN_dtp	1	0.05%	OTHLI	1	100.00%
OP_rs+SP_os	1	0.05%	OTHLI	1	100.00%
SP_os+IP_ds	1	0.05%	OTHLI	1	100.00%
HI_oh+ID+IN_dtp	1	0.05%	OTHLI	1	100.00%
IN_oi+HI_oh	1	0.05%	OTHLI	1	100.00%
IN_fi+HI_oh+IP_oe+NA_ne	1	0.05%	OTHLI	1	100.00%
IN_fi+HI_oh	1	0.05%	OTHLI	1	100.00%
IP_oe+HI_oh+IN_oi	1	0.05%	OTHLI	1	100.00%
IN_fi+HI_oh+ID	1	0.05%	OTHLI	1	100.00%
IN_dtp+IP_oe+HI_oh	1	0.05%	OTHLI	1	100.00%
NA_on+HI_oh	1	0.05%	OTHLI	1	100.00%

Register	Count	Percentage	Tag	Tag Count	Tag Percentage
HI_re+IN_dtp	1	0.05%	OTHLI	1	100.00%
IP_ds+IN_dtp+ID	1	0.05%	OTHLI	1	100.00%
IN_dtp+OP_av+HI_re	1	0.05%	OTHLI	1	100.00%
HI_oh+IN_dtp+IN_fi	1	0.05%	OTHLI	1	100.00%
NA_nb+HI_re+IP_oe	1	0.05%	OTHLI	1	100.00%
IN_dtp+IN_fi+HI_oh+ID	1	0.05%	OTHLI	1	100.00%
NA_on+IN_oi+HI_oh+ID	1	0.05%	OTHLI	1	100.00%
HI_oh+IN_fi+ID	1	0.05%	OTHLI	1	100.00%
OP_ob+HI_oh	1	0.05%	OTHLI	1	100.00%
IN_dtp+HI_oh+IP_ds	1	0.05%	OTHLI	1	100.00%
OP_av+IN_dtp+LY	1	0.05%	OTHLI	1	100.00%
NA_ne+IN_dtp+ID	1	0.05%	OTHLI	1	100.00%
IN_dtp+IN_lt+NA_ne	1	0.05%	OTHLI	1	100.00%
NA_ne+IN_oi+OP_oo	1	0.05%	OTHLI	1	100.00%
NA_ne+IN_dtp+OP_oo	1	0.05%	OTHLI	1	100.00%
NA_ne+OP_ob	1	0.05%	OTHLI	1	100.00%
IN_ra+IP_oe	1	0.05%	OTHLI	1	100.00%
IP_ds+OP_rs	1	0.05%	OTHLI	1	100.00%
IN_ra+HI_oh	1	0.05%	OTHLI	1	100.00%
OP_rs+IN_dtp	1	0.05%	OTHLI	1	100.00%
OP_rs+NA_nb	1	0.05%	OTHLI	1	100.00%
OP_rv+ID	1	0.05%	OTHLI	1	100.00%
OP_av+NA_on	1	0.05%	OTHLI	1	100.00%
IP_oe+OP_rv	1	0.05%	OTHLI	1	100.00%
ID+HI_oh	1	0.05%	OTHLI	1	100.00%
NA_ne+OP_rs+HI_oh	1	0.05%	OTHLI	1	100.00%
OP_av+HI_oh+IP_oe	1	0.05%	OTHLI	1	100.00%
NA_nb+OP_av	1	0.05%	OTHLI	1	100.00%
NA_on+IP_ds+IN_dtp	1	0.05%	OTHLI	1	100.00%
IP_ds+MT	1	0.05%	OTHLI	1	100.00%
IP_ed+OP_rs	1	0.05%	OTHLI	1	100.00%
OP_rs+OP_ob	1	0.05%	OTHLI	1	100.00%
IP_ds+OP_rv	1	0.05%	OTHLI	1	100.00%