Genre Variation in Dependency Types: A Two-Level Genre Analysis Using the Czech National Corpus

Xinying Chen
University of Ostrava
Ostrava, Czech Republic
cici13306@gmail.com

Miroslav Kubát
University of Ostrava
Ostrava, Czech Republic
miroslav.kubat@gmail.com

Abstract

This paper examines how dependency type distributions vary across genres in the Czech National Corpus (SYN2020). Using a two-level genre classification, broad categories and fine-grained subgenres, we identify genre-sensitive syntactic patterns through relative frequency analysis. The results show that some dependency types (e.g. Atr 'attribute') vary consistently across genres, while others (e.g. ExD 'part of discourse ellipsis') show sensitivity only at the subgenre level. Our dependency-based approach extends common multidimensional analyses based on lexical-grammatical co-occurrences, directly capturing syntactic evidence and improving interpretability. Our findings also highlight the importance of fine-grained genre distinctions in revealing syntactic variation.

1 Introduction

Syntactic structure plays a central role in how information is organized and interpreted across different communicative contexts. One of the most important contextual factors that influence language use is the genre or style of the text (Biber and Conrad, 2009). While it is well recognized that genres impose different communicative goals and stylistic conventions, most of the existing work in stylometry studies focuses primarily on lexical features, such as word frequencies, stylistic markers, or vocabulary richness (Stamatatos, 2009; Kestemont, 2014), leaving syntactic variation relatively underexplored. However syntax, particularly as represented through dependency relations, provides valuable insights into how information is structured and presented differently across genres (Nivre, 2005; Roland et al., 2007; Webber, 2009).

This study contributes to the relatively unexplored area by investigating how the usage of dependency types varies across genres in the Czech National Corpus (SYN2020) (Křen et al., 2020; Jelínek et al., 2021; Křivan and Šindlerová, 2022). Specifically, we take advantage of the corpus's hierarchical genre structure. It categorizes texts into three broader groups: fiction, non-fiction, newspapers and magazines. At the same time, it also provides more fine-grained subcategories, such as novels, short stories, scientific literature, professional literature, newspapers, leisure magazines, etc. This hierarchical genre organization provides a unique opportunity to explore genre sensitivity in syntactic structures both across and within broader genre categories.

Our primary objective is to identify which dependency types remain stable across genres and which ones display genre-sensitive variation. We examine the relative frequencies of all dependency types in both broader and fine-grained genre categories, using descriptive statistics such as maximum-minimum differences and standard deviation to quantify variability. This two-level analysis allows us to highlight dependency types that are structurally central and consistent, as well as those that are more genre-dependent. By comparing genre sensitivity at two levels of granularity, we provide a more nuanced understanding of how syntactic preferences are shaped by genre.

While previous influential genre analyses, notably Biber and Conrad (2009), utilized multidimensional analysis primarily focusing on English and lexical-grammatical feature co-occurrences, our study explicitly employed dependency tag, providing direct, transparent syntactic evidence across hierarchical genre distinctions. This approach enables deeper cross-linguistic and syntactic insights that complement and extend their foundational work.

Therefore, our results contribute to the growing body of research that integrates syntactic analysis into genre studies (Biber and Conrad, 2009; Oostdijk, 1998; Kubát et al., 2021, 2024; Chen and Kubát, 2024) and highlight the importance of us-

Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025), pages 84–92

ing hierarchical genre structures in corpus-based syntactic research.

2 Data and Methodology

This study is based on data from the Czech National Corpus, specifically the SYN2020 subcorpus (Křen et al., 2020; Jelínek et al., 2021; Křivan and Šindlerová, 2022), a representative collection of contemporary written Czech containing approximately 100 million words mainly from 2015 to 2019. SYN2020 is annotated with morphosyntactic and syntactic information, including dependency relations. The syntactic annotation follows the principles outlined in the Prague Dependency Treebank framework (Hajič et al., 2020) and is performed automatically using a neural network-based parser from the NeuroNLP toolkit (Ma et al., 2018). The parser was trained on data from the analytical layer of the Prague Dependency Treebank and the syntactically annotated FicTree fiction corpus (Jelínek, 2017). The automatic syntactic annotation achieves a labeled attachment score (LAS) of 88.73% and an unlabeled attachment score (UAS) of 92.39% on test data[1]. While not manually verified, SYN2020 represents a significant improvement over previous corpus versions and provides consistent annotation across the entire 100-million-word corpus, ensuring reliable frequency comparisons across genres.

The corpus is organized into three primary groups: fiction(FIC), non-fiction(NFC), newspapers and magazines (NMG). Each of these groups contains multiple subcategories, including novels (NOV), short stories (COL), poetry (VER), drama or screenplays (SCR), scientific literature (SCI), professional (PRO) and popular writing (POP), memoirs and autobiographies (MEM), administrative documents (ADM), newspapers (NEW), and leisure magazines (LEI). Table 1 presents the structure of the corpus, and additional details are available in SYN2020 website.

For our analysis, we extracted frequency data for all dependency types occurring in the corpus. We first gathered absolute counts of each dependency type in all broad genre categories and subcategories. These counts were converted into relative frequencies within each genre, enabling fair comparisons across genres of differing sizes. This normalization was particularly important for the subgenre-level analysis, where category sizes varied considerably, see Table 1.

To assess genre sensitivity, we employed two descriptive statistical measures:

- Maximum-Minimum Difference (Max-Min): This value captures the absolute difference between the highest and lowest relative frequency of a dependency type across genres. A high Max-Min value suggests that a dependency type is used very differently depending on the genre.
- Standard Deviation (SD): This measure reflects the overall spread of the relative frequency of each dependency type across the fine-grained subgenres. Unlike Max-Min, SD accounts for all intermediate values and provides a more balanced view of variability. Due to the limited number of broader genre categories (only three), SD was not computed for the broader genre level.

We calculated these measures separately for the broader genre categories and the more fine-grained subgenres. To address possible internal heterogeneity within broad genre categories, we also examined variation across subgenres using standard deviation for each dependency type. This allowed us to quantify the extent to which syntactic usage diverges within genre groups, especially in categories like non-fiction where textual functions vary widely. It also enables us to compare how certain dependency types behave in both coarse and more fine-grained genre classifications, providing a systematic view of syntactic variability shaped by genre.

3 Analysis of Broader Genre Categories

In the first stage of analysis, we examined the relative frequency distributions of the dependency types in the three broader genre categories: FIC, NFC, and NMG. The metric used to assess variation was the Max-Min in relative frequency across the three genres. Dependency types with higher Max-Min values were interpreted as more genre-sensitive, while those with smaller values were considered more stable across genres. Table 2 presents the observed dependency variation (Max-Min > 0.5) across 3 broad genres, where Atr represents attributive relations (nominal modifiers), Obj marks direct objects, AuxK indicates sentence-ending punctuation, and ExD captures elliptical

[1] For detailed documentation of the automatic annotation pipeline, including morphological tagging and syntactic parsing procedures, see https://wiki.korpus.cz/doku.php/cnk:syn2020:automaticka_anotace.

Text-Type/Genre Group	Text-Type/Genre	Number of Word Tokens	Number of Sentences
FIC: fiction	NOV: novels	26,059,743	2,360,348
	COL: short stories	5,350,850	442,758
	VER: poetry	1,002,449	178,844
	SCR: drama, screenplays	1,003,033	156,750
NFC: non-fiction	SCI: scientific literature	9,284,751	459,801
	PRO: professional literature	7,013,611	405,546
	POP: popular literature	13,431,550	801,639
	MEM: memoirs and autobiographies	4,030,874	270,245
	ADM: administrative	348,920	24,795
NMG: newspapers and magazines	NEW: newspapers	20,393,309	1,402,548
	LEI: leisure magazines	13,601,340	1,049,767

Table 1: Text-type structure of SYN2020.

constructions. For definitions of the dependency type abbreviations, please refer to the Table 3.

Dependency	FIC	NFC	NMG	Max-Min
Atr	12.55	24.58	22.74	12.04
Obj	9.77	6.74	7.83	3.02
AuxK	7.72	4.70	5.84	3.02
Adv	12.72	9.91	11.04	2.81
AuxT	2.78	1.41	0.16	2.62
Pred_Co	5.10	2.90	3.26	2.20
Atr_Co	0.92	3.00	2.02	2.08
AuxP	7.83	9.36	9.86	2.02
ExD	2.72	1.41	1.55	1.31
Pred	4.35	3.05	3.96	1.30
AuxG	3.43	3.08	2.29	1.14
AuxC	2.47	1.58	1.57	0.90
AuxX	6.02	5.41	5.21	0.82
Sb	6.04	5.62	6.43	0.80
AuxV	1.72	1.16	1.17	0.55

Table 2: Dependency variation across 3 broad genres.

The most genre-sensitive dependency type was Atr (attribute), where non-fiction and news texts show markedly higher frequencies compared to fiction (Max-Min = 12.04). This pattern reflects the tendency of non-fiction and journalistic writing to make extensive use of noun modifiers in order to express specific, technical, or formal information. This result aligns with previous findings in English that associate high syntactic density and nominal modification with informational density in informational genres (Biber, 1988; Biber and Conrad, 2009). However, by explicitly analyzing syntactic dependencies rather than lexical co-occurrence patterns, our analysis provides direct syntactic evidence and a more precise characterization of these genre distinctions.

Dependency	Definition
Atr	Attribute (adjective)
Obj	Object
AuxK	Sentence-ending punctuation
Adv	Adverbial (adverbial determination)
AuxT	Reflexive particle 'se' in inherently reflexive verbs
Pred	Predicate
AuxP	Preposition
ExD	Part of discourse ellipsis
AuxG	Other graphic symbols that do not end a sentence
AuxC	Subordinating conjunction
AuxX	Comma
Sb	Subject
AuxV	Auxiliary verb být (to be)
Coord	Coordination node
AuxZ	Emphatic word
AuxY	Adverbs and particles that cannot be classified elsewhere
Pnom	Nominal part of a verbonominal predicate
Apos	Apposition (main node)

Table 3: Definitions of dependency type abbreviations in Czech syntactic analysis. Dependency ending _Co is for tokens that are coordinated and ending _pa is for part of parentheses. For example, coordinated attributes are assigned the function Atr_Co. For more information, please check the introduction page of the Czech National Corpus and Prague Dependency Treebank Annotation Manual.

Obj (object) and AuxK (sentence-ending punctuation) also demonstrated notable genre sensitivity. Although their Max-Min values were smaller than Atr, this does not imply insignificant variation. Instead, it suggests subtler stylistic variation across genres. For instance, the relatively higher frequency of Obj in fictional texts (Max-Min = 3.02) reflects their narrative-driven syntax, emphasizing events and actions. Regarding AuxK (Max-Min = 3.02), its frequency directly corresponds to the number of sentences in the corpus. Narrative texts (FIC) typically contain shorter sentences and frequent dialogues, resulting in a higher number of sentences and thus increasing the relative fre-

quency of sentence-ending punctuation. In contrast, informational texts (NFC and NMG) often feature longer sentences designed to convey complex ideas, leading to fewer sentences overall and consequently reducing the frequency of AuxK. Therefore, the observed genre variation in AuxK primarily represents differences in sentence segmentation, sentence count, and syntactic complexity across genre categories.

The majority of dependency types such as Sb (subject), AuxV (Auxiliary verb *být* 'to be'), and many other dependency types (Max-Min ≤ 0.5) demonstrated relatively stable distributions across all three genre categories. These dependency types exhibit consistent usage patterns across genres, suggesting they fulfill fundamental syntactic roles in Czech that are relatively unaffected by stylistic variation.

The analysis of the broader genre categories reveals both structural constants and genre-sensitive syntactic choices. Types like Atr, Obj, and AuxK demonstrate meaningful variation across genres and can thus serve as indicators of broader stylistic tendencies in written Czech.

4 Analysis of Fine-Grained Genre Subcategories

While the analysis of broader genre categories revealed general patterns of syntactic variation, it is at the subgenre level that genre sensitivity becomes more evident and interesting. Our subgenre analysis reveals dramatically increased variation, with ExD frequencies ranging from 0.20 in leisure magazines to 14.03 in poetry, a 70-fold difference invisible at the broader genre level.

To account for variation in the size of these subcategories, we used normalized relative frequencies and applied both Max-Min and sample standard deviation (SD) as metrics of variability. Whereas Max-Min highlighted extreme contrasts in usage across subgenres, SD allowed us to capture more distributed forms of variation. The results of analysis (Max-Min > 0.5) are presented in Table 4.

Several dependency types emerged as highly sensitive at this level of analysis. Atr once again topped the list. The consistently high frequency of Atr in ADM and SCI genres highlights their requirement for syntactic density and precision. Administrative texts, characterized by formality and specificity, rely heavily on noun modifiers to express precise legal or bureaucratic concepts. Similarly, scientific literature employs dense noun phrases extensively to contain technical details and methodological precision clearly and concisely, aligning with the informational focus of these genres. This reinforces the role of Atr as a marker of dense, information-heavy discourse (Biber and Conrad, 2009).

In contrast, ExD (part of discourse ellipsis) showed marked sensitivity at the subgenre level, contrasting its relative stability at the broader genre level. Its distribution was notably uneven, peaking in literary texts, especially poetry and drama. ExD refers specifically to elements omitted from sentences because they can be inferred from the context. This high variation likely reflects genre-specific stylistic conventions related to brevity, informality, and implied meaning. For example, poetry frequently utilizes elliptical constructions to create ambiguity, enhance rhythmic conciseness, or engage readers in interpreting implicit meanings. Similarly, dramatic texts commonly feature discourse ellipsis to simulate natural speech patterns, spontaneous dialogues, or emotional intensity by omitting linguistic elements clearly understood from the conversational context. The relatively low frequency of ExD in more formal or informational genres, such as scientific literature and administrative texts, aligns with their explicitness and precision, which discourage reliance on contextual inference. This pattern aligns with previous findings highlighting genre-specific syntactic phenomena distinguishing narrative and expressive texts from expository and formal writing (Biber and Conrad, 2009).

Other types such as Pred_Co (coordinated predicates) and AuxP (preposition) also ranked highly in variability. Pred_Co exhibited notable differences across genres. This variation reflects stylistic preferences for predicate coordination, which are more frequent in conversational or literary subgenres probably due to their use of compound predicates that facilitate narrative flow or rhythmic expression. Conversely, scientific and administrative texts tend toward simpler predicate structures to enhance precision and clarity, thus explaining their lower frequencies of Pred_Co. AuxP variation highlights genre differences in prepositional phrase complexity and density. Technical genres like scientific literature and administrative documents often exhibit a higher frequency of AuxP due to their reliance on prepositional phrases to precisely convey complex information, whereas literary genres typically use

Dependency	NOV	COL	VER	SCR	SCI	PRO	POP	MEM	ADM	NEW	LEI	Max-Min	SD
Atr	12.73	18.21	2.83	9.44	28.66	30.07	23.94	18.78	34.02	23.84	22.01	31.19	9.32
ExD	2.57	2.98	14.03	6.51	1.76	1.36	1.43	1.15	3.37	1.28	0.20	13.83	3.92
Obj	10.31	1.21	13.60	9.61	6.19	6.30	7.40	0.98	5.85	7.59	8.56	12.62	3.69
Adv	13.26	1.71	2.06	9.93	9.49	9.42	10.70	13.90	7.14	10.68	12.11	12.19	4.04
AuxK	7.93	9.55	1.31	11.84	4.42	4.92	4.88	6.63	4.68	5.58	6.52	10.52	2.82
AuxP	8.07	10.74	1.70	5.70	10.19	10.64	9.22	10.01	12.09	10.16	9.82	10.39	2.91
Sb	6.22	0.81	10.44	6.41	5.79	5.95	5.95	6.28	4.78	6.76	6.18	9.63	2.22
Pred	4.46	5.10	12.23	5.63	2.88	3.55	3.10	3.84	3.49	4.01	4.07	9.34	2.61
AuxX	6.33	8.33	0.67	4.16	5.91	4.89	5.69	6.87	0.45	5.04	5.70	7.88	2.41
Coord	4.37	6.01	8.33	4.44	4.73	0.45	4.35	4.77	4.72	3.86	4.90	7.88	1.84
Pred_Co	5.27	6.96	7.82	4.95	0.25	0.24	3.23	4.80	0.18	2.79	4.17	7.64	2.65
ExD_Co	1.22	1.87	5.87	3.80	1.68	1.05	1.13	1.12	2.09	0.99	1.72	4.88	1.50
AuxG	3.69	4.05	0.27	3.63	4.43	2.94	2.80	2.35	0.07	2.24	0.25	4.36	1.58
Atr_Co	0.09	1.55	0.15	0.47	0.42	3.72	2.62	1.74	4.24	2.02	2.11	4.15	1.41
AuxC	2.61	3.14	3.63	1.98	0.14	1.18	0.18	2.55	0.77	1.47	0.18	3.48	1.25
AuxT	2.94	3.50	3.82	2.20	1.31	1.08	1.54	2.46	0.64	1.48	1.84	3.19	1.02
AuxV	1.80	2.23	0.22	0.18	0.10	1.08	1.04	2.70	1.09	1.16	1.24	2.60	0.83
Obj_Co	0.11	1.59	1.26	0.68	1.47	1.50	1.39	0.14	1.87	1.32	1.61	1.76	0.60
AuxZ	0.13	1.84	0.27	1.04	1.65	1.72	1.71	1.66	0.86	1.77	0.19	1.71	0.70
AuxY	0.78	1.13	1.23	0.77	0.94	0.72	0.87	0.91	0.04	0.70	0.83	1.19	0.30
Adv_Co	0.69	1.11	0.10	0.38	1.08	0.94	0.94	0.94	1.27	0.72	0.96	1.17	0.34
Coord_Co	0.51	0.83	0.96	0.60	0.05	0.36	0.44	0.55	0.40	0.34	0.50	0.91	0.24
Pnom	1.30	1.72	2.18	1.72	1.63	1.70	1.56	1.57	1.34	1.35	1.61	0.88	0.24
Sb_Co	0.42	0.65	0.65	0.33	1.13	1.02	0.84	0.70	0.92	0.80	0.85	0.79	0.24
ExD_Pa	0.31	0.38	0.61	0.93	0.45	0.22	0.20	0.17	0.35	0.15	0.19	0.78	0.23
Apos	0.18	0.29	0.39	0.47	0.07	0.55	0.45	0.35	0.63	0.36	0.04	0.59	0.19

Table 4: Dependency variation across fine-grained genres.

fewer dense prepositional constructions.

At the same time, several core syntactic functions, such as Sb (subject), Coord (coordination), and AuxX (auxiliary in coordinated constructions), continued to show low variability across subgenres. This consistency supports the notion that certain syntactic dependencies remain relatively unaffected by genre, functioning as part of the grammatical infrastructure of Czech syntax.

Our fine-grained analysis further extends beyond the level of detail achievable through lexical-grammatical multidimensional analyses, as employed by Biber and Conrad (2009). The dependency tag explicitly reveals subtle yet important stylistic differences and subgenre-specific syntactic variations, such as the distinctive high frequency ExD in literary subgenres. It provides syntactic insights that indirect co-occurrence analyses may not capture.

Together, these findings reveal that while some dependency types maintain stability across genre levels, others become more genre-sensitive when fine-grained distinctions are considered. The results underscore the importance of using hierarchical genre structures in syntactic analysis to avoid averaging out meaningful stylistic variation.

These findings also highlight internal genre heterogeneity. For instance, the non-fiction category includes both scientific texts and memoirs, which show sharply different syntactic profiles, Atr ranges from 18.78% in memoirs to over 34% in administrative texts. Similarly, genres like leisure magazines and newspapers differ in ExD usage despite both falling under NMG. These internal divergences show that traditional genre groupings may mask important syntactic variation, which supports the need for finer-grained or data-driven genre modeling.

5 Comparison and Interpretation

Having examined dependency type distributions across both broader genre categories and fine-grained subcategories, we now compare the findings to better understand how syntactic variation is shaped by different levels of genre granularity. This comparison offers insights into the types of dependencies that are consistently genre-sensitive, those that are genre-neutral, and those whose variability becomes more apparent at the subgenre level.

To facilitate this comparison, we ranked all dependency types according to their Max-Min values

in both levels of analysis.[2] Our results reveal significant granularity effects: Atr maintains rank 1 at both levels but with obvious value changes (Max-Min: 12.04 broad, 31.19 fine), while ExD shifts from rank 9 (Max-Min: 1.31) to rank 2 (Max-Min: 13.83).

To visualize these dynamics, we created a scatterplot comparing coarse-grained and fine-grained sensitivity ranks (Figure 1). Points along or close to the diagonal indicate consistent genre sensitivity across both classification levels. Points deviating more from the diagonal represent dependency types whose genre sensitivity becomes more apparent at finer granularity. For instance, ExD, positioned more far from the diagonal than Atr, demonstrates that its sensitivity to genre is not as evident at the broader level but becomes more pronounced within subgenre distinctions, highlighting the importance of analyzing detailed subgenre classifications to uncover syntactic patterns. This visual comparison demonstrates how coarse categorization can sometimes obscure important syntactic variability.

Figure 1 shows that while many dependency types exhibit stable genre sensitivity regardless of granularity, a smaller subset displays considerable divergence between the two levels. These divergences are particularly important, as they highlight constructions that are sensitive to more subtle communicative or stylistic demands found only in specific subgenres. Importantly, the most genre-sensitive types (i.e., those with the lowest ranks) cluster in the lower-left corner of the plot. Identifying these genre-sensitive dependencies such as Atr, ExD, and Obj has practical implications for computational linguistics applications, particularly automated genre classification. Dependency types sensitive to genre differences can improve the accuracy of classifiers by incorporating genre-specific syntactic features, thus enhancing linguistic modeling in computational frameworks.

In a second visualization, we plotted Max-Min values against SD for each dependency type in the fine-grained analysis, see Figure 2. The overall distribution in Figure 2 reveals a positive relationship between Max-Min and SD: dependency types that show stronger sensitivity to genre distinctions also tend to fluctuate more across subgenres. This confirms that genre-sensitive types are not only skewed toward specific contexts but also exhibit greater instability, further reinforcing their role as stylistically responsive constructions.

Using the mean values of Max-Min and SD as thresholds, we divided the space into four interpretive zones:

- Low Range / Low Spread (bottom-left): 113
- High Range / High Spread (top-right): 19
- High Range / Low Spread (bottom-right): 4
- Low Range / High Spread (top-left): 0

The vast majority of dependency types fall into the low range / low spread quadrant, indicating that they are largely genre-neutral and stable across subgenres. In contrast, a small but crucial set of types cluster in the high range / high spread zone. These include types such as Atr, ExD, and Sb, which exhibit both strong genre sensitivity and high variability. This identifies them as key indicators of subgenre-specific syntactic preferences.

Interestingly, the absence of types in the low range / high spread quadrant suggests that high variability almost never occurs without accompanying genre sensitivity. That is, wide fluctuations in usage typically correspond to meaningful genre-driven effects, rather than random variation.

This analysis further supports the importance of examining genre at multiple levels of resolution. While many dependency relations remain stable regardless of context, a focused view on fine-grained distinctions reveals important dimensions of syntactic variability that would otherwise remain hidden.

This comparative approach demonstrates explicitly how dependency-based syntactic analysis provides methodological depth and granularity beyond previous multidimensional analyses (Biber and Conrad, 2009). By directly mapping syntactic patterns onto genre distinctions at both coarse and fine-grained levels, our method explicitly identifies syntactic features sensitive to subtle genre differences, thus notably enriching the theoretical and methodological scope of genre analysis.

6 Discussion

The results of this study underscore the importance of incorporating genre structure into syntactic analysis. The comparison between broader and fine-grained genres revealed both the stability and variability of dependency types. While majority of syntactic relations showed consistent distributions regardless of genre, others such as Atr, ExD, and

[2]Tied values were given the same lowest possible rank.

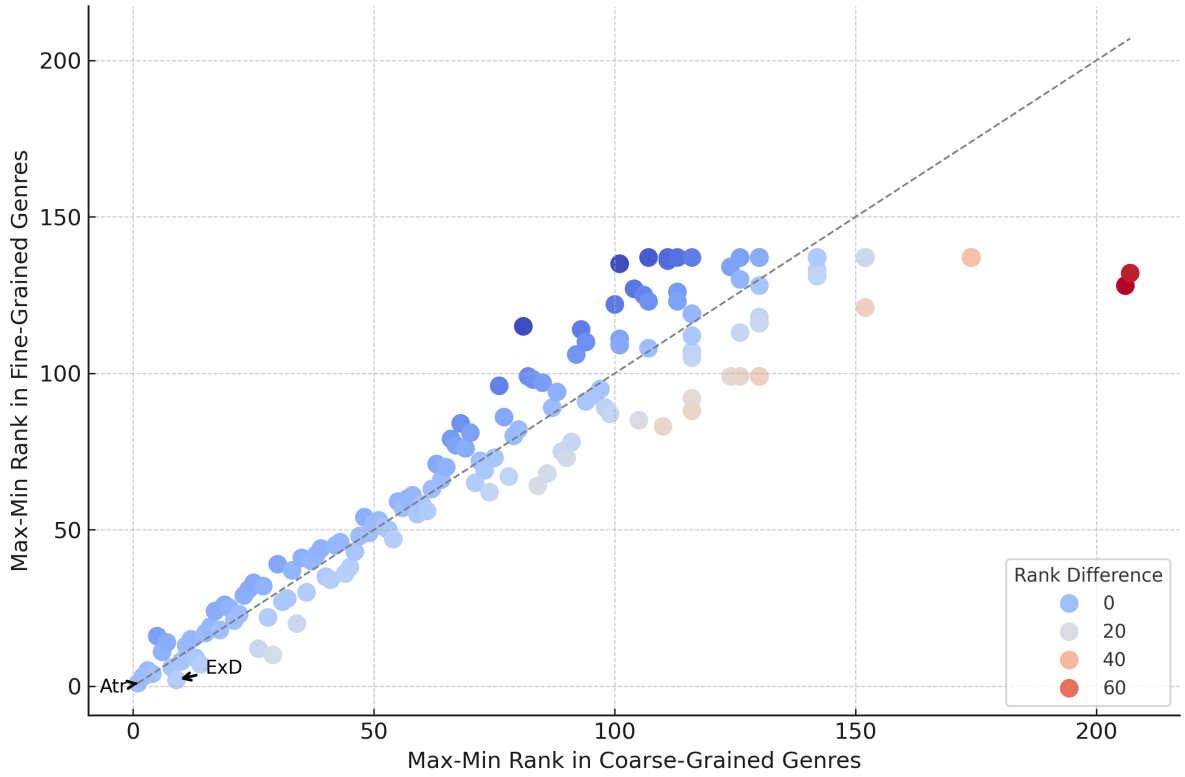

Figure 1: Comparison of dependency sensitivity ranks between coarse- and fine-grained genres.

Obj were markedly sensitive to the communicative and stylistic demands of specific genres.

Importantly, the internal heterogeneity observed within broad genre groups is not a limitation of genre-based analysis but an opportunity for refinement. Our two-level analysis illustrates that genre categories are often composed of syntactically distinct subtypes. Rather than assuming genre homogeneity, our approach enables empirical evaluation of genre cohesion and reveals when fine-grained distinctions are warranted. This supports a more dynamic, corpus-driven model of genre.

Crucially, these findings question the adequacy of relying exclusively on broad genre categories for syntactic analyses, as such coarse classifications may level up subtle yet important stylistic features. Our results align with Biber and Conrad (2009), emphasizing that communicative, stylistic, and functional differences in language frequently manifest at fine-grained levels of genre variation. For instance, dependency types like ExD, clearly more sensitive at the subgenre level, illustrate precisely the kind of stylistic phenomenon that broader classifications may obscure. Thus, incorporating multiple granularity levels is essential not only theoretically but also practically for linguistic analyses that aim for accuracy and depth.

The use of both Max-Min and SD measures further allowed us to differentiate between types that exhibit obvious shifts and those not. This dual perspective provided a richer view of how syntactic preferences are shaped across the genre spectrum. Moreover, the visual analyses confirmed that variability is not uniformly distributed; some types are tightly linked to fundamental syntactic roles of language, while others are more responsive to genre-specific stylistic conventions. These results offer practical implications for areas such as genre-aware syntactic parsing, authorship attribution, and language modeling, where understanding genre-specific syntactic tendencies can improve performance and interpretability.

7 Conclusion

This study has presented a two-level genre analysis of dependency type distributions in the Czech National Corpus. By examining both broader genre categories and fine-grained subgenres, we identified which dependency relations are structurally stable and which are sensitive to genre distinctions. The analysis demonstrated that certain types, such as Atr, show strong and consistent variation across

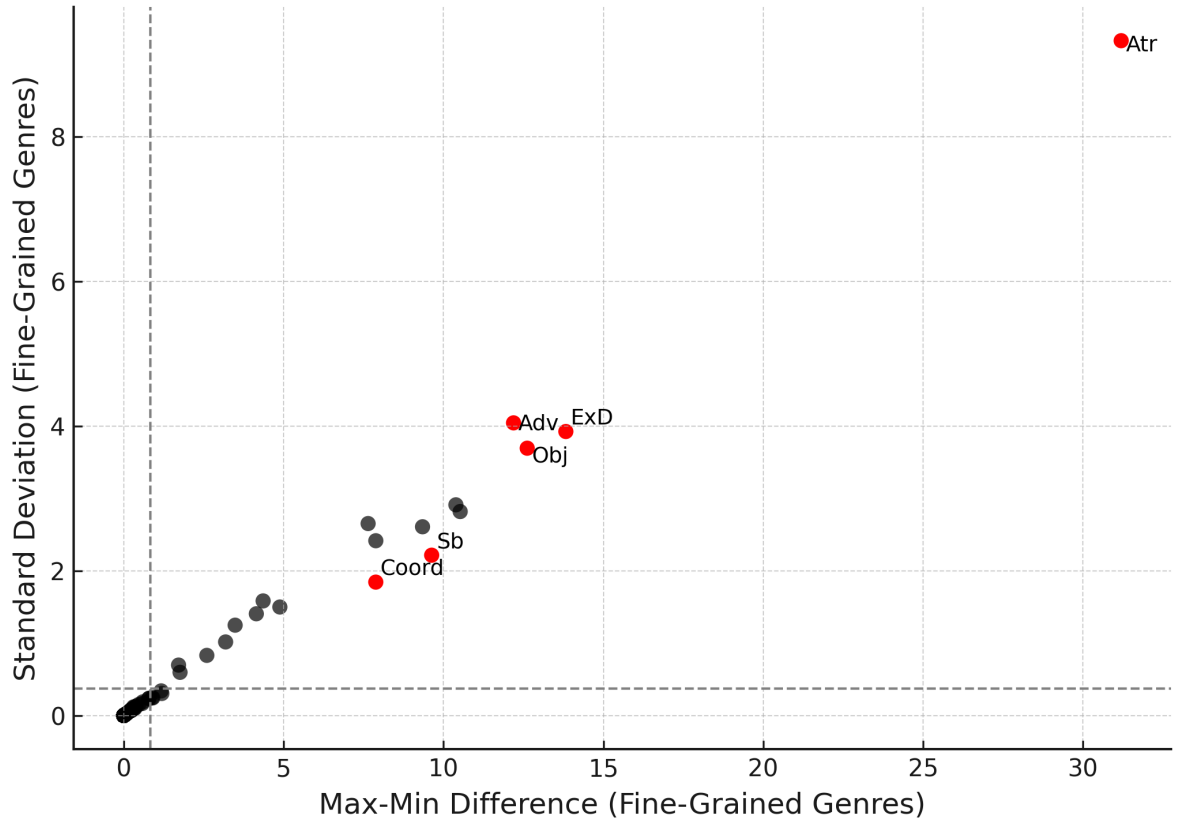

Figure 2: Scatter plot of dependency types based on Max-Min and SD across fine-grained genres. Red points are selected samples for illustration. Dashed lines indicate the mean values for each axis, dividing the plot into four interpretive regions.

both levels, serving as indicators of informational density in formal genres. Others, like ExD, revealed their stylistic specificity more at the subgenre level, highlighting the expressive features of narrative and literary texts.

Overall, the findings confirm that genre plays a critical role in shaping syntactic preferences and that this role can only be fully appreciated by analyzing data at multiple levels of granularity. By explicitly examining dependency relations across multiple genre levels, our study substantially complements foundational multidimensional analyses (Biber and Conrad, 2009), offering direct syntactic evidence and enhanced theoretical insights that deepen our understanding of genre-specific linguistic variability.

Future research could beneficially extend these analyses cross-linguistically or explore computational approaches to utilize genre-sensitive syntactic patterns in natural language processing applications. Investigating the consistency of these findings across different languages and genre classification frameworks would further clarify the relationship between syntactic variation and communicative context.

Limitations

Despite the insights provided, several limitations should be acknowledged. First, this study exclusively relies on the SYN2020 from the Czech National Corpus, which covers texts mainly from 2015–2019. Consequently, the findings may not generalize to other time periods or linguistic contexts, as language use can evolve considerably even within relatively short spans.

Second, while SYN2020 offers a robust genre classification scheme, the hierarchical categorization used in this analysis may still obscure more complex stylistic variations. Certain genres could contain internal heterogeneity, and additional subclassifications might yield further insights.

Last but not the least, dependency frequency measures alone do not capture the complexity of syntactic variation fully. Including additional linguistic features such as dependency distance might enhance the understanding of genre variations.

Acknowledgments

This work was supported by the Czech Science Foundation (GAČR), project No. 22-20632S.

References

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

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press, Cambridge, UK.

Xinying Chen and Miroslav Kubát. 2024. Quantifying syntactic complexity in czech texts: An analysis of mean dependency distance and average sentence length across genres. *Journal of Quantitative Linguistics*, 31(3):260–273.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague dependency treebank - consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Tomás Jelínek. 2017. Fictree: A manually annotated treebank of czech fiction. In *ITAT 2017 Proceedings*, pages 181–185.

Tomáš Jelínek, Jan Křivan, Vladimír Petkevič, Hana Skoumalová, and Jana Šindlerová. 2021. SYN2020: A new corpus of Czech with an innovated annotation. In *Text, Speech, and Dialogue. TSD 2021*, volume 12848 of *Lecture Notes in Computer Science*, pages 48–59, Cham. Springer.

Mike Kestemont. 2014. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66.

Michal Křen, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Dominika Kováříková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal. 2020. *SYN2020: reprezentativní korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha.

Jan Křivan and Jana Šindlerová. 2022. Změny v morfologické anotaci korpusů řady SYN: nové možnosti zkoumání české gramatiky a lexikonu. *Slovo a slovesnost*, 83(2):122–145.

Miroslav Kubát, Ján Mačutek, Radek Čech, and Michaela Nogolová. 2024. Automatic genre classification of czech texts based on syntactic functions. In G. Giordano and M. Misuraca, editors, *New Frontiers in Textual Data Analysis*, pages 163–172. Springer.

Miroslav Kubát, Radek Čech, and Xinying Chen. 2021. Attributivity and subjectivity in contemporary written czech. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 58–64. Association for Computational Linguistics.

Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, page 1403–1414, Melbourne, Australia. Association for Computational Linguistics.

Joakim Nivre. 2005. Dependency grammar and dependency parsing. Technical Report MSI report 05133, Växjö University, School of Mathematics and Systems Engineering.

Nelleke Oostdijk. 1998. A corpus-based model of syntactic variation with applications for english. *Literary and Linguistic Computing*, 13(3):147–153.

Douglas Roland, Frederic Dick, and Jeffrey L. Elman. 2007. Frequency of basic english grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57(3):348–379.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682, Suntec, Singapore. Association for Computational Linguistics.