Word Order Variation in Spoken and Written Corpora: A Cross-Linguistic Study of SVO and Alternative Orders

Nives Hüll
University of Ljubljana
nives.hull@gmail.com

Kaja Dobrovoljc
University of Ljubljana
Jozef Stefan Institute, Ljubljana, Slovenia
kaja.dobrovoljc@ff.uni-lj.si

Abstract

This study investigates word order variation in spoken and written corpora across five Indo-European languages: English, French, Norwegian (Nynorsk), Slovenian, and Spanish. Using Universal Dependencies treebanks, we analyze the distribution of six canonical word orders (SVO, SOV, VSO, VOS, OSV, OVS). Our results reveal that spoken language consistently exhibits greater word order flexibility than written language. This increased flexibility manifests as a decrease in the dominant SVO pattern and a rise in alternative orders, though the extent of this variation differs across languages. Morphologically rich languages such as Slovenian and Spanish show the most pronounced shifts, while English remains syntactically rigid across modalities. These findings support the claim that modality significantly affects syntactic realizations and highlight the need for typological studies to account for spoken data.

1 Introduction

Word order is a fundamental parameter in linguistic typology and syntactic theory. It plays a central role in language classification and shapes our understanding of cross-linguistic variation and syntactic universals. Typological databases such as WALS (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023) document dominant word order patterns like subject–verb–object (SVO), but these generalizations are typically based on written, formal sources and often fail to capture variation across genres or modalities.

Recent corpus-based work (e.g., Naranjo and Becker, 2018; Levshina, 2019; Baylor et al., 2024) has challenged this categorical view. These studies advocate for a gradient, usage-based approach to word order typology, emphasizing observed token frequencies in syntactically annotated corpora. This shift has enabled a more nuanced classification of languages and has revealed that word order is not solely a matter of structural constraints, but also reflects contextual factors such as genre, domain, and modality (Levshina et al., 2023; Baylor et al., 2023).

Despite this growing awareness of contextual variation, there remains a lack of systematic, cross-linguistic studies that focus specifically on modality, understood here as the distinction between spoken and written language. While modality is often acknowledged, most existing work incorporates it only indirectly or treats it as a secondary factor within broader genre-based analyses. As a result, cross-linguistic studies that systematically examine the influence of modality on constituent order remain scarce. The extent to which spoken and written language diverge in word order—especially across typologically diverse languages—has yet to be addressed in a comparative framework.

This study offers an exploratory contribution to this gap by examining cross-modal variation in constituent order. We analyze a sample of five Indo-European languages—English, French, Norwegian (Nynorsk), Slovenian, and Spanish—using both spoken and written corpora within the Universal Dependencies framework. Focusing on clause-level syntax, we investigate the distribution of six canonical word order permutations—SVO, SOV, VSO, VOS, OSV, and OVS—as a typological baseline for comparing modalities. While the tendency for spoken language to show more variation is often assumed, we argue that systematically capturing how this plays out across languages adds empirical weight to such claims and exposes patterns missed in writing-based investigations.

Our goal is to assess whether and how word order differs between speech and writing, and whether these differences follow consistent cross-linguistic patterns. The analysis is guided by two central research questions: **(1)** Does word order differ between spoken and written language? and **(2)** If so, how does it differ?

Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025), pages 150–155
August 27-28, 2025 ©2025 Association for Computational Linguistics

The remainder of the paper outlines our data and methods (Section 2), presents the results (Section 3), and discusses cross-linguistic trends (Section 4), followed by concluding remarks (Section 5).

2 Data and Methods

2.1 Corpus Selection and Preparation

The analysis includes five Indo-European languages for which both spoken and written data are available in Universal Dependencies v2.15 (Zeman et al., 2024). We focused on languages where spoken and written data are clearly separated, either across different treebanks or within a single treebank. To keep the work manageable and grounded in languages we are familiar with, we limited the study to five treebank pairs listed below.

The *Rhapsodie* (spoken) and *GSD* (written) corpora were used for French (Gerdes et al., 2012; Guillaume et al., 2019); *NynorskLIA* (spoken) and *Nynorsk* (written) for Norwegian (Nynorsk) (Øvrelid et al., 2018; Solberg et al., 2014); *SST* (spoken) and *SSJ* (written) for Slovenian (Dobrovoljc and Nivre, 2016; Dobrovoljc et al., 2017); *COSER* (spoken) and *GSD* (written) for Spanish (Fernández-Ordóñez, 2005–present; Ballesteros et al., 2024); and the *GUM* treebank for English (Zeldes, 2017), which was manually divided into spoken and written subsets based on genre metadata.[1]

2.2 Data Extraction

We conducted a quantitative analysis using the STARK tool, designed for querying syntactic patterns in UD-formatted dependency trees (Krsnik and Dobrovoljc, 2025).[2] For each language and modality pair, we extracted all instances in which a finite verb governs both a nominal subject (`nsubj`) and a direct object (`obj`), regardless of clause type—this includes main and subordinate, declarative and interrogative clauses alike. Each sentence was then classified based on the linear order of the subject, verb, and object into one of six canonical word orders: SVO, SOV, VSO, VOS, OSV, or OVS. This procedure yielded a dataset containing word order distributions for each corpus, which we compared across modalities. To illustrate the six possible orders, Table 1 provides examples in Slovenian—a language that permits all six permutations—along with their English translations.

With this approach, our analysis aligns with inclusive, usage-based studies (e.g., Gerdes et al., 2019; Östling, 2015; Naranjo and Becker, 2018; Baylor et al., 2024), which aim to capture naturally occurring syntactic variation across clause types. Rather than limiting the analysis to main declarative transitive clauses only (e.g., Levshina, 2019; Dryer, 2013), we include all instances of subject–verb–object structures, regardless of clause type. This enables us to more fully capture modality-sensitive variation, while keeping the analysis straightforward and the results easily interpretable.

Order	Slovenian	Gloss
SVO	*Mama kupuje jabolka*	mother buys apples
SOV	*Mama jabolka kupuje*	mother apples buys
VSO	*Kupuje mama jabolka*	buys mother apples
VOS	*Kupuje jabolka mama*	buys apples mother
OSV	*Jabolka mama kupuje*	apples mother buys
OVS	*Jabolka kupuje mama*	apples buys mother

Table 1: Canonical word order examples in Slovenian. All sentences translate as 'Mother buys apples'.

3 Results

3.1 General Observations

Figure 1 summarizes the distribution of six canonical word orders across written and spoken corpora for each language. It shows the relative frequency of SVO, SOV, VSO, VOS, OSV, and OVS, allowing for a direct comparison between modalities.

The results confirm that word order in spoken language differs from written language across all examined languages. In every case, speech exhibits greater variation than writing, with the dominant SVO pattern decreasing in spoken data. Additionally, the degree of flexibility in word order varies across languages, with some showing more pronounced shifts than others. These findings are consistent across the sample.

3.2 Language-Specific Findings

English shows the least variation. SVO remains dominant, dropping only slightly from 97.4% in written to 93.0% in spoken data. OSV rises modestly from 2.6% to 6.9%, while other patterns remain marginal. This limited shift may suggest that English maintains a relatively high degree of syntactic rigidity even in spontaneous speech. The

[1] The spoken subset includes interviews, conversations, podcasts, vlogs, courtroom transcripts, and speeches; the written subset includes news articles, academic texts, fiction, how-to guides, biographies, essays, letters, textbooks, and travel guides.

[2] https://github.com/clarinsi/STARK

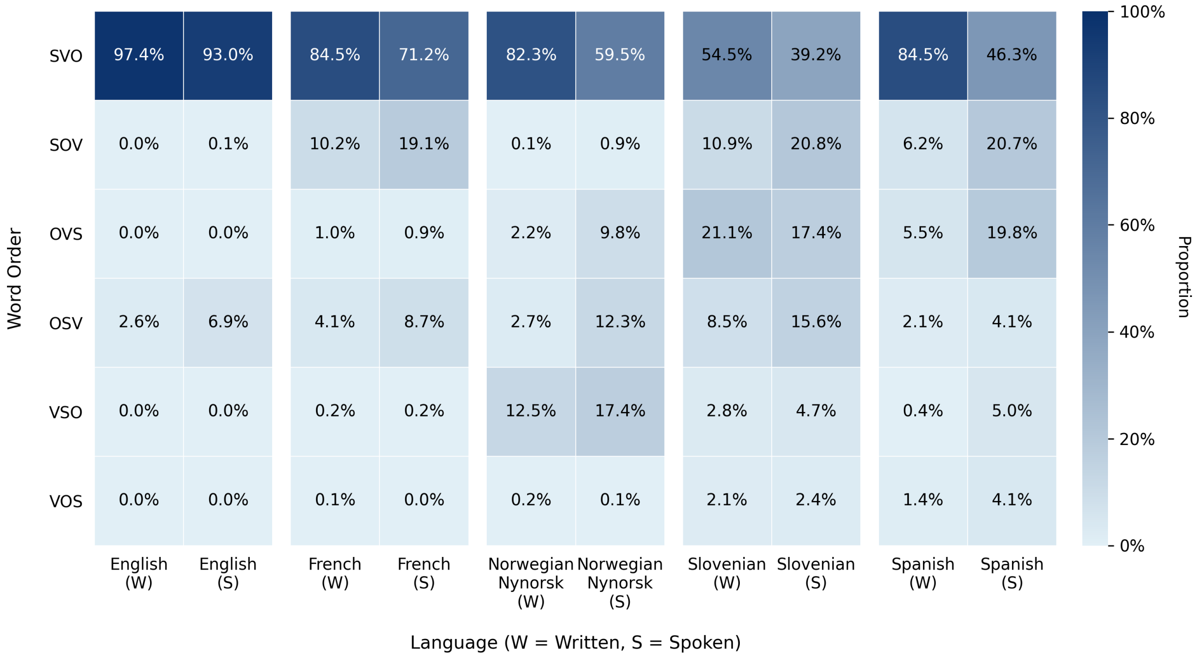

Figure 1: Word order frequencies in written and spoken language across five Indo-European languages.

slight rise in OSV might be due to marked topicalizations or interrogative constructions. For example, *That we did this summer* and *That I give you* illustrate how speakers may foreground the object for emphasis, while questions such as *What did he say?* could also contribute to this pattern.

French exhibits a more noticeable change: SVO usage drops from 84.5% to 71.2%, accompanied by increases in SOV (10.2% → 19.1%) and OSV (4.1% → 8.7%). These surface patterns may be influenced by the frequent use of object clitic pronouns and dislocation structures in spontaneous speech. For example, *on l'a mise à l'épreuve* (*we/one her put to the test*) and *je le mets également à l'intérieur* (*I it put also inside*) contain preverbal object clitics, which may lead to an apparent rise in non-SVO orders without indicating a change in underlying syntax.

Norwegian Nynorsk displays moderate flexibility. SVO decreases from 82.3% to 59.5%, while OSV and OVS rise significantly (from 2.7% to 12.3% and from 2.2% to 9.8%, respectively). The increase in OSV patterns may partly result from interrogative structures where question elements and pronouns are fronted, as in *Kva den kallast den fóra?* (*What that is called, that feed?*). OVS constructions, such as *Det veit eg ikkje* (*That I don't know*) are common in spoken discourse and may reflect object-fronting for emphasis or information structure.

Slovenian is the most flexible language in the sample. SVO accounts for only 39.2% of spoken clauses, with SOV (20.8%), OSV (15.6%), and OVS (17.4%) forming near-equal shares. This distribution may reflect the language's high degree of pragmatic word order variation. SOV patterns, such as *To mi je šlo zelo na živce* (*That really annoyed me*) and *Jaz ti zdaj pomagam* (*I now help you*) may result from object fronting, emphasis, or prosodic rhythm in spontaneous speech. OSV examples like *To jaz nisem* (*This I am not*) are frequently used for contrastive focus, especially in expressive or corrective contexts.

Spanish undergoes the strongest shift from a canonical pattern. SVO falls from 84.5% to 46.3%, while SOV (20.7%), OVS (19.8%), and VSO (5.0%) become more frequent. The increased presence of OV patterns may be partly attributed to clitic constructions and discourse-driven reordering. For instance, *las yuntas lo trillaban* (*the oxen it threshed*) shows preverbal clitic placement that results in an apparent SOV order, while *cómo lo hacía su padre* (*how it did his father*) illustrates an OVS structure that may arise in embedded or emphatic contexts.

4 Discussion

Our findings confirm that word order varies significantly between spoken and written modalities across the examined languages. Although all are classified as SVO-dominant in WALS, spoken data consistently exhibit greater flexibility, with higher frequencies of SOV, OSV, and OVS orders. We observe a cross-linguistic rise in postverbal subjects and object-initial configurations—patterns that are rarely captured in typological descriptions based on written sources. The extent of this variation differs by language: it is most pronounced in morphologically rich systems such as Slovenian and Spanish, and more limited in structurally rigid languages like English.

Several factors may account for the greater flexibility observed in speech. As expected, morphological richness plays a central role: languages with robust case marking, such as Slovenian and Spanish, can overtly signal grammatical roles, reducing the need for fixed word order and allowing more pragmatic or prosodically driven constituent placement.

Second, prosodic structure in speech—intonation, rhythm, and stress—can help disambiguate syntactic relations and guide listener interpretation, even in non-canonical orders (Levshina et al., 2023). In morphologically rich languages, prosody may interact with syntax and discourse to license constituent placement (Gerken, 1996).

Third, discourse-related considerations shape spoken word order. The distinction between given and new information often drives constructions like Left-Dislocation, which promote new or contrastive elements to the left periphery of the sentence (Prince, 1981, 1997; Gregory and Michaelis, 2001). This reflects how spoken syntax is sensitive to real-time communicative needs rather than fixed structural defaults.

Finally, cognitive and psycholinguistic constraints influence linearization. Speakers often place accessible or low-load elements earlier in the sentence to ease comprehension and gain time to plan semantically complex constituents (Schouwstra et al., 2022; Levshina et al., 2023). Features typical of spontaneous speech—such as repairs, hesitations, questions, and topic shifts—also encourage deviations from canonical order. These effects are particularly evident in Slovene dialectal discourse (Kumar, 2019), and more generally in languages where grammatical structure permits flexible sequencing (Levshina, 2019).

Taken together, ur findings support previous calls for more gradual, context-aware investigations of constituent order that move beyond dominant patterns and account for variation across modalities (e.g., Baylor et al., 2024; Levshina et al., 2023). In particular, they highlight speech as a crucial communicative context—shaped by a complex interplay of morphosyntactic, prosodic, cognitive, and discourse factors.

Future work should extend this approach to additional languages, including those outside the Indo-European family. With the growing availability of spoken UD treebanks (Dobrovoljc, 2022; Kahane et al., 2021), there is now concrete potential to uncover cross-linguistic patterns that have long remained underdocumented—not only in typological accounts, but in linguistic research more broadly.

5 Conclusion

This study highlights the significant impact of modality on SVO word order variation across five Indo-European languages. Spoken language consistently shows greater syntactic flexibility, especially in morphologically rich systems like Slovenian and Spanish. These findings challenge typological generalizations based primarily on written data and underscore the need for future studies to incorporate spoken corpora for a more accurate picture of constituent order variation.

Limitations

This study focuses on five Indo-European languages, limiting typological diversity. Only the Nynorsk variety of Norwegian was included to ensure consistent comparison between spoken and written data.

The corpora vary in size, balance, and genre coverage, particularly between spoken and written modalities, which may influence pattern distribution. Only clauses with overt nominal subjects and objects were included, following WALS criteria, excluding constructions common in morphologically rich languages where arguments are omitted.

We also restricted the analysis to verbal predicates, excluding nominal and adjectival constructions, and did not distinguish between declarative and interrogative clauses. Our findings are based on quantitative distributions, with no in-depth qualitative analysis.

Lastly, while Universal Dependencies aims for consistency, differences in annotation guidelines or treebank practices may affect comparability.

Acknowledgment

This work was financially supported by the Slovenian Research and Innovation Agency through the research project Treebank-Driven Approach to the Study of Spoken Slovenian (Z6-4617), and the research program Language Resources and Technologies for Slovene (P6-0411). The authors also made selective use of generative AI tools during the writing phase, but all proposals were independently reviewed and finalized by the authors.

References

Miguel Ballesteros, Héctor Martínez Alonso, Ryan McDonald, Elena Pascual, Natalia Silveira, Daniel Zeman, and Joakim Nivre. 2024. Universal dependencies 2.15: Spanish gsd. `https://github.com/UniversalDependencies/UD_Spanish-GSD`. Accessed: 2025-04-23.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2023. The past, present, and future of typological databases in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1163–1169, Singapore. Association for Computational Linguistics.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual gradient word-order typology from Universal Dependencies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–49, St. Julian's, Malta. Association for Computational Linguistics.

Kaja Dobrovoljc. 2022. Spoken language treebanks in universal dependencies: An overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017. The universal dependencies treebank for slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, pages 33–38, Valencia.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1566–1573. European Language Resources Association (ELRA).

Matthew S. Dryer. 2013. Order of subject, object and verb (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

I. Fernández-Ordóñez. 2005–present. Corpus oral y sonoro del español rural. urlhttp://www.corpusrural.es/. Retrieved April 15, 2022.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2019. Rediscovering greenberg's word order universals in UD. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 124–131, Paris, France. Association for Computational Linguistics.

Kim Gerdes, Sylvain Kahane, Anne Lacheret, Arthur Truong, and Paola Pietrandrea. 2012. Intonosyntactic data structures: The Rhapsodie treebank of spoken french. In *Proceedings of the Sixth Linguistic Annotation Workshop (LAW VI)*, Jeju, South Korea. Held in conjunction with ACL-2012.

LouAnn Gerken. 1996. Prosody's role in language acquisition and adult parsing. *Journal of Psycholinguistic Research*, 25(3):345–356.

Michelle L. Gregory and Laura A. Michaelis. 2001. Topicalization and left-dislocation: A functional opposition revisited. *Journal of Pragmatics*, 33(11):1665–1706.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en universal dependencies. *Traitement Automatique des Langues*, 60(2):71–95.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Luka Krsnik and Kaja Dobrovoljc. 2025. STARK: A Toolkit for Dependency (Sub)Tree Extraction and Analysis. In *Proceedings of the SyntaxFest 2025*. To appear.

Danila Zuljan Kumar. 2019. Word order in slovene dialectal discourse. *Slovenski jezik*, 12:53–74. URN:NBN:SI:DOC-VC42NQN7.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Matías Guzmán Naranjo and Laura Becker. 2018. Quantitative word order typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 91–104. Linköping University Electronic Press.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. The LIA treebank of spoken Norwegian dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. *Radical Pragmatics*, pages 223–256.

Ellen F. Prince. 1997. On the functions of left-dislocation in english discourse. In Akio Kamio, editor, *Discourse and Functional Linguistics*, pages 117–144. John Benjamins.

Marieke Schouwstra, Danielle Naegeli, and Simon Kirby. 2022. Investigating word order emergence: Constraints from cognition and communication. *Frontiers in Psychology*, 13.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, and 85 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16).

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The norwegian dependency treebank. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA).

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noémi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, and ... 2024. Universal dependencies 2.15. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.