SEMANTIC AND SYNTACTIC COMPLEXITY OF INTERNATIONAL, NATIONAL AND LOCAL NEWSPAPERS

INDEX

Abstract

This study investigates whether local, national and international news articles differ in syntactic and lexical complexity, and whether such differences can be detected using a basic spaCy-based pipeline. A corpus of 156 news articles was compiled, organised into 52 triplets of local, national and international coverage of comparable events during October and November 2025. Each article was processed with spaCy for tokenisation, sentence segmentation and part-of-speech tagging. On this basis, Python scripts computed length, lexical diversity (STTR), POS distributions, a nominal style index, and a Zipf-based lexical rarity scores.

The results show many similarities across registers. All three use similar proportions of major word classes, comparable Zipf-based vocabularies and similar syntactic proxies. Local articles are shorter, slightly more diverse lexically and richer in proper nouns; while National articles are longer and more repetitive; and International articles generally fall in between. Overall, the study suggests that, in this controlled environment, local, national and international news are set in a common register family, and that the use of simple and transparent NLP tools are effective for mapping such subtle tendencies.

Keywords: journalistic style, syntactic complexity, lexical rarity, corpus linguistics, spaCy, register variation

INTRODUCTION

News now circulates almost instantaneously across digital platforms, yet the language of news reporting still varies with context, audience and institutional role. Local, national and international articles target different readers, cover different scales of events and operate under distinct constraints, which makes them a useful testing ground for studying how syntactic and lexical complexity change across registers (Biber, 1995; Biber & Conrad, 2009).

A possible expectation would be that “international news are more complex”: global agencies would rely on abstract and geopolitical prose; local outlets would produce shorter, and more concrete stories centred on specific communities and proper names; and national newspapers would fall in between. Rather than taking these assumptions for granted, this study treats them as empirical questions. It asks how far a set of reproducible corpus measures can detect register-level differences in complexity when applied to authentic English news writing. In that sense, the research functions both as an analysis of news discourse and as a methodological exercise in using a small and structured corpora combined with off-the-shelf NLP tools.

The empirical basis for the study is a 156-article corpus of English-language news, organised into 52 triplets of Local, National and International texts reporting on the same event per group (meaning that the three articles of each triplet talk about the same event from their respective points of address). Each article is automatically annotated using spaCy, which provides tokenisation, sentence segmentation and part-of-speech information (spaCy, 2016-2025). On top of this annotation, the analysis derives a set of complementary indicators: length (tokens and types), lexical diversity (STTR), POS-based distributions of content words, and external Zipf-based index of lexical rarity among others (Lu, 2010; McCarthy & Jarvis, 2010; Brysbaert & New, 2009).

These syntactic indicators might be related to formal discussions of finite recursive syntax and the ability of formal systems to generate unbounded hierarchical structures.

Subordination and the Coordination–Subordination Ratio reflect how natural language relies on recursive expansion (such as the recursive syntax of propositional logic and first-order logic) to capture more expressive structural relations. Although the study investigates these patterns

through corpus statistics rather than through formal proofs, it is motivated by the same questions about how far languages, and the registers that exploit them, push the resources of structural recursion and lexical choice.

The essay is organised as follows: the Literature Review situates the project within work on register variation, news discourse, readability and linguistic complexity. The Methodology section details the corpus design, spaCy-based annotation pipeline and the operationalisation of the complexity measures. The Results then present the distribution of these measures across Local, National and International articles. And the Discussion and Conclusion reflect on what these findings imply for both register theory and formal perspectives on linguistic complexity.

LITERATURE REVIEW

In recent years, extensive research has been conducted on register variation, which provides the theoretical foundation for analysing the linguistic differences between types of news writing. For example, Biber's *Dimensions of Register Variation: A Cross-Linguistic Comparison* (1995, p. 133-36) established that genres and sub-registers can be distinguished by measurable linguistic features such as clause density, nominalization, modality, and tense-aspect distributions. Biber's multi-dimensional (MD) analysis demonstrated that functional differences in communicative purposes reflect variations among registers (ranging from spoken conversation to formal academic writing). This framework might legitimize the assumption that "local", "national" and "international" news may themselves constitute distinct sub-registers within the broader news genre, each characterized by different stylistic and structural tendencies.

In this sense, the multidimensional approach can still be related to the research hypothesis. Local news, typically focused on concrete events within a specific community, tends to favour shorter texts, a relatively higher density of proper nouns and fact-oriented description, while national and international outlets often handle events at broader scales and with more institutional actors. Rather than predicting grammatical contrasts, the MD framework here provides a way of anticipating subtle shifts in lexical diversity, referential style

and information packaging across registers, offering a rationale for measurable linguistic differences among local, national and international reporting.

In second place, several journalism and computational linguistics studies have examined how news language varies according to outlet type, audience, and topic. Large-scale automated analyses, such as those described by Flaounas et al. (2013), have demonstrated that digital tools can assess stylistic complexity, sentiment, and topical framing across millions of articles. Their work illustrates how automatic linguistic features, such as readability indices, syntactic counts, and lexical statistics, produce an understanding of journalistic practice, even if they admit that "readability cannot be entirely reduced to a set of linguistic properties, but they do provide a useful framework in which certain properties (…) are likely to be associated with higher levels of readability." (p. 104)

Parallel research (Shulman et al., 2024) suggests that online readers prefer simpler headlines and lighter syntactic constructions, what affects the interaction between linguistic complexity and audience behaviour, implying that editorial choices in syntax and vocabulary may be shaped by the demands of their readers.

So far, these studies indicate that local, national and international news differ not only in terms of topic and audience but also in measurable linguistic behaviour, what motivates a systematic investigation of syntactic and lexical complexity as register-level phenomena.

Syntactic complexity has been used as an indicator of linguistic sophistication, particularly in studies on second-language writing and genre analysis. Lu (2010) provided one of the most influential operational frameworks for quantifying syntactic complexity using the *L2 Syntactic Complexity Analyzer* (L2SCA). He validated several indices, like the mean length of sentence, clauses per sentence and dependent clause ratio that reliably capture structural elaboration. Although developed for L2 (Second Language Acquisition) research, these measures have also been adopted in corpus-based analyses of native language registers.

However, not all the methods to represent syntactic complexity are equally valid. In their diachronic study, Plavén-Sigray et al. (2017) showed a steady decrease in readability, which they interpreted as evidence of increasing lexical and syntactic difficulty over time, as well as "an increase in general scientific jargon over years". Their analysis stresses that

readability metrics are a mix of lexical and syntactic influences and should not replace specialized syntactic measures.

In third place, lexical sophistication can be theorised along two axes: word rarity (frequency-based) and lexical diversity (variety-based). The SUBTLEX frequency norms developed by Brysbaert and New (2012) can constitute a foundation for word rarity estimation. These norms are derived from film subtitles and naturalistic language data, providing Zipf-scale frequency approximations that better predict lexical processing than traditional frequency lists. In corpus analyses, lower mean Zipf scores correspond to rarer, less frequent vocabulary, allowing a precise quantification of lexical rarity across corpora.

Finally, in journalism studies, the automated analysis of news corpora has become increasingly central. Both Flaounas et al. (2013) and Horne et al. (2018) demonstrated that it is possible to evaluate the complexity of a news article according to its scale and its features. These studies combined lexical, syntactic, and metadata features to differentiate outlets and map stylistic tendencies across genres. Such computational methodologies not only expand the analysis scale but also enhance replicability.

Nevertheless, despite extensive research on readability and stylistic differences across media, few studies have directly compared syntactic and lexical complexity among *Local, National* and *International* news. The gap lies in combining both dimensions (syntax and vocabulary sophistication) within a unified framework. Moreover, few studies pay attention to the article length when comparing outlets, which are factors known to influence both syntactic and lexical measures.

METHODOLOGY

Corpus design and sampling

The corpus consists of 156 news articles written in English, divided into three registers: Local, National and International. The texts are organised into 52 triplets, each triplet containing a local, national and international article covering the same (or similar) event, which allows register comparisons while partially controlling for topic.

All articles were taken mostly from the online sources of the Waterford-news (Local), The Irish Time (National) and BBC News (International) and collected between October and November 2025 (because some online articles may disappear from the Internet after some time, what makes more difficult the gathering). This homogeneity is deliberate: the primary aim of the project is not to map differences between all news providers, but to test what simple and replicable tools and metrics can detect in a controlled setting where many external variables are held constant.

Only news reports written in English were included. Very short briefs and opinion pieces with overt first-person commentary were excluded to keep the corpus focused on comparable news reporting. All texts were downloaded in digital form, converted to plain .txt files and cleaned to remove navigation menus, links, captions and other non-editorial material.

Pre-processing and annotation with spaCy

All processing was carried out in Python using the spaCy natural language processing library. A contemporary English pipeline (en_core_web_sm) was loaded, and each article was processed as a separate document. The spaCy pipeline provided:

- tokenisation (segmenting raw text into tokens)
- sentence segmentation (doc.sents)
- part-of-speech (POS) tags (token.pos_ / token.tag_)
- basic morphological information

SpaCy's linguistic-features framework maps language-specific tags onto a small, universal POS inventory which was then used (i.e. NOUN, VERB, ADJ, ADV, PROPN, PRON, CCONJ, SCONJ) and that can be accessed via token.pos_. This universal tagset is particularly suited to cross-text comparisons of grammatical composition.

For all quantitative measures:

- Punctuation and spaces were excluded using "token.is_punct" and "token.is_space".
- Only tokens with "token.is_alpha == True" were treated as words when computing lexical measures and Zipf scores.
- All wordforms were lower-cased before type counting, so that *Government* and *government* were treated as the same type.

Custom Python scripts then looped over the tokens in each Doc file and, following spaCy's own usage examples for linguistic features, incremented counters whenever a token matched the relevant POS category. For example, the Adjective Rate is computed by increasing an "adj_count" each time "*token.pos_ == 'ADJ'*" and then dividing by the total number of word tokens in that article. The same pattern is used for other categories. This counting logic was used in order to discover what can be learned using minimal but well-documented scripts built on top of spaCy's annotation.

For interpretability, the universal POS tags were grouped into the following categories:

- Adjectives (ADJ): token.pos_ == "ADJ"
- Adverbs (ADV): token.pos_ == "ADV"
- Common nouns (NOUN): token.pos_ == "NOUN"
- Proper nouns (PROPN): token.pos_ == "PROPN"
- Verbs (VERB): token.pos_ == "VERB"
- Pronouns (PRON): token.pos_ == "PRON"
- Coordinating conjunctions (CCONJ): token.pos_ == "CCONJ"
- Subordinating conjunctions (SCONJ): token.pos_ == "SCONJ"

For each article, the scripts gathers raw counts and normalised percentages for these categories, as well as sentence counts (len(list(doc.sents))). All subsequent statistics (means, standard deviations, etc.) were computed from these article-level values.

Complexity measures

The choice of measures follows previous work on register variation and complexity that recommends transparent, multi-index approaches rather than single scores (Biber, 1995; Lu, 2010; McCarthy & Jarvis, 2010; Brysbaert & New, 2009).

Length and lexical diversity

For each article, the scripts computed:

- Total tokens: number of word tokens (excluding punctuation and spaces) produced by spaCy.
- Total types: number of distinct lower-cased wordforms.

Although Type-Token Ratio (TTR) would have been the most expected measure having these data, TTR is sensitive to sample length (McCarthy & Jarvis, 2010). Therefore, a better lexical-diversity indices was implemented: Segmented TTR (STTR): the token sequence is divided into contiguous windows of 200 tokens. For each full window, TTR is calculated and the mean of those window TTRs is taken as the STTR for that article. If the text is shorter than the window, a single TTR is returned. This approach reduces length dependence by comparing like-sized segments.

POS-based distributions and nominal style

Using the POS categories defined above, the following POS rates were obtained for each article: ADJ%, ADV%, NOUN%, VERB%, PRON%, PROPN%, CCONJ%, SCONJ%.

Each percentage is computed as:

$$\text{CATEGORY\%} = \frac{\text{CATEGORY tokens}}{\text{total tokens}} \times 100$$

These proportions characterise the grammatical profile of the three registers and follow the tradition of using word-class distributions to describe register variation and information packing (Biber, 1995; Biber & Conrad, 2009).

In addition, a nominal style index was calculated by combining common and proper nouns:

$$\text{Nominal index} = \frac{\text{NOUN tokens} + \text{PROPN tokens}}{\text{total tokens}} \times 100$$

Higher values indicate a denser nominal style, which is often associated with informational prose and facts.

Lexical rarity (Zipf-based frequency)

To estimate lexical rarity, the project uses an external Zipf frequency measure via the wordfreq Python library. wordfreq.zipf_frequency(word, "en") returns a Zipf score based on large subtitle-based frequency norms, following the approach advocated by Brysbaert and New (2009). In this frequency, the higher the value obtained for a word, the more common that word is.

For each article:

1. Every token.is_alpha word was lower-cased and passed to zipf_frequency.
2. The script averaged all Zipf scores in the article to obtain a mean Zipf value.

Syntactic complexity proxies

The study uses proxies for syntactic complexity, inspired by Lu's (2010) syntactic-complexity indices but adapted to what can be computed reliably from spaCy in a small corpus.

1. Mean Length of Sentence (MLS):

MLS reflects how much syntactic material is packed into each sentence (syntactic elaboration). By using spaCy's sentence segmentation (doc.sents), the script counts: number of sentences in the article and number of word tokens (excluding punctuation and spaces), and calculates:

$$\text{MLS} = \frac{\text{words}}{\text{sentences}}$$

2. Subordination and coordination rates (SUBORD%):

It is computed as:

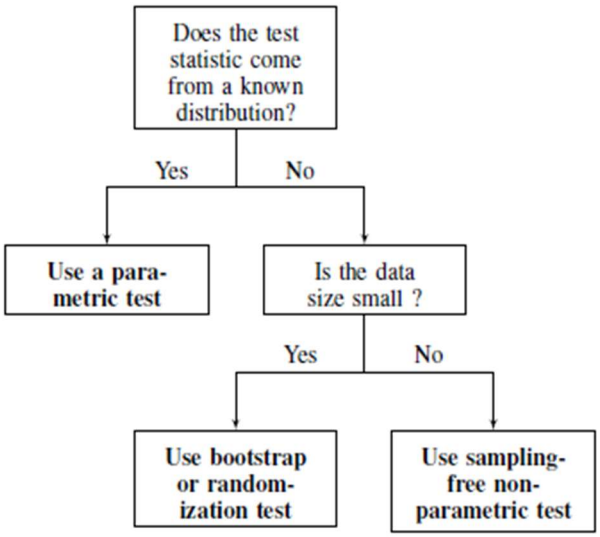
$$\text{SUBORD\%} = \frac{\text{SCONJ tokens}}{\text{total tokens}} \times 100$$

where SCONJ tokens represent subordinating conjunctions (*because, although, if…*). Coordination rate (COORD%) is computed the same way using CCONJ tokens (*and, but, or…*).

These rates capture how frequently each register marks a clause relations with explicit coordinators and subordinators, in line with the use of clause-combining features as indicators of syntactic complexity in previous studies (Biber & Conrad, 2009; Lu, 2010).

Statistical Analysis

The numerical information was obtained by using the programming language R (Mair, 2020; Team, 2025), and following the decision tree proposed on "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing" (Dror, Baumer, Shlomov, & Reichart, 2018):



RESULTS

For each article of the 156 articles, the spaCy-based pipeline produced tokenisation, sentence segmentation and POS tags, on top of which the custom Python scripts computed lexical and syntactic metrics. This section summarises the main patterns that emerge when those metrics are aggregated by register (Local, National, International).

Length and lexical diversity

Across the corpus, National and International articles tend to be longer on average than Local ones, both in raw token counts and in the number of distinct types. However, once length is considered, the lexical diversity measures show a more nuanced picture.

STTR, calculated on fixed 200-token windows, reduces the effect of length and confirms that: Local articles generally show higher window-based TTRs, while National articles are somewhat more repetitive lexically, and International articles cluster between them.

Lexical rarity (Zipf-based frequencies)

Mean Zipf values are very similar across the three registers. All three show average Zipf scores within a tight range, indicating reliance on vocabulary that is, in frequency terms, broadly comparable. In the Zipf file, individual article scores range only between 5.35 and 5.92, with most Local, National and International texts clustered around 5.6–5.8, and differences between register means staying below 0.1 Zipf units.

There is no evidence that one register systematically uses rarer or more specialised words throughout its texts. Local news occasionally includes more specific place names and locally relevant entities, which may slightly lower Zipf scores, while International coverage sometimes introduces geopolitical or institutional terminology that is less frequent in general language. Nonetheless, these effects are localised: overall averages remain extremely close, and lexical rarity alone does not differentiate clearly between the three registers in this corpus.

POS distributions and nominal style

In all three categories, common nouns and verbs together account for an important portion of tokens, with adjectives and adverbs contributing a smaller but stable share. Pronouns are infrequent, and their proportions do not differ much between registers.

Proper nouns are the one POS category where clearer differences emerge. As expected, Local articles display a higher density of proper names, reflecting their focus on specific people, institutions and locations relevant to the immediate community. National and International articles also use proper names frequently, but they tend to spread them across a wider range of political, economic and international references. When common and proper nouns are combined into a nominal style index, all three registers look highly nominal, but Local texts often are at the top of the range.

These patterns are shown in the POS-rate metrics. For one representative content-word category, values across all triplets remain in a relatively narrow band, with individual article percentages ranging from about 7.25% to 15.67% and Local, National and International means all clustered around 11–13%. The three registers thus differ only by one or two percentage points on average, and their ranges overlap almost completely, reinforcing the idea that they share a common grammatical template with only small shifts in the relative prominence of particular word classes.

Syntactic complexity proxies

The syntactic proxies derived from spaCy's sentence segmentation and conjunction tags also point to subtle differences.

Mean Length of Sentence (MLS) shows that National and International articles generally have longer sentences on average than Local articles, which often rely on slightly

shorter, more segmented sentences. However, some Local texts contain long, multi-clause sentences, and some International texts are relatively short. MLS therefore indicates a small tendency for more syntactically elaborated sentences in higher-level coverage, but not a categorical distinction.

Subordination and coordination rates, based on SCONJ and CCONJ tokens, are also similar across registers. All three make comparable use of explicit subordinating conjunctions (*because, although, if…*) and coordinating conjunctions (*and, but, or…*). None of the registers clearly favours subordination over coordination (or vice versa) to a degree that would suggest a different clause-combining strategy.

The additional subordination index, which combines SCONJ use with other syntactic cues, reinforces this conclusion. While individual texts within each register can be embedded, the overall distributions across Local, National and International articles are closely aligned. International news, in particular, does not show a markedly higher level of subordination, despite often dealing with complex geopolitical topics.

Finally, a summary of the results (concerning whether there is a difference between two categories) is available in Appendix B.

DISCUSSION

This study was made to test whether local, national and international news differ in syntactic and lexical complexity and to explore how far a pipeline (spaCy) can detect such differences in a homogeneous environment. Using 156 news articles organised into 52 triplets of local, national and international coverage of comparable events, the analysis combined lexical diversity measures (STTR), Zipf-based lexical rarity, POS distributions, a nominal index, mean sentence length and subordination/coordination proxies. Overall, the results show many similarities rather than strong differences across registers, with only small and gradual shifts in emphasis.

From a descriptive point of view, the clearest contrasts involve text length, lexical diversity and proper noun density. Local articles are generally shorter but often show slightly higher lexical diversity and a higher proportion of proper nouns.

The measures of lexical rarity and POS distributions reinforce this view of a common core. Mean Zipf scores are extremely similar for Local, National and International articles, indicating that all three rely on vocabulary of comparable frequency in general language. None of the registers systematically uses rarer or more specialised words across entire articles. Differences in perceived difficulty are therefore unlikely to stem from global word-frequency differences alone; instead, they may arise from localised jargon in specific domains (for example, finance or international law) or from discourse-level factors such as how information is ordered and contextualised.

POS-based indicators show the same kind of gentle variation within a stable pattern. All three registers have a similar number of common nouns, verbs, adjectives and adverbs, but not in the use of pronouns. The nominal index, which combines common and proper nouns, is high in all categories, thus showing the informational nature of news prose oriented towards facts. The slightly higher nominal index in local articles suggests a stronger emphasis on entities and places, but the grammatical architecture remains similar across the three registers.

The syntactic complexity representations derived from spaCy's sentence segmentation and POS tags tell a compatible story. Mean sentence length tends to be higher in National and International coverage, hinting at a somewhat greater degree of syntactic elaboration, but again there is considerable overlap across registers. Local texts are not systematically "simple": many contain long, multi-clause sentences; on the other hand, some International articles use short and punchy sentences. This stability suggests that, at least in this sample, Local, National and International coverage rely on very similar clause-combining strategies, even when dealing with topics that differ in geopolitical scale.

These findings have several implications for the methodological aims of the project. First, they show that by using a relatively small set of transparent metrics from spaCy, which were implemented in short Python scripts that loop over tokens, check token.pos_ and count categories, can capture interpretable register tendencies, such as the higher density of proper names and slightly richer lexical diversity in local reporting, even if these are small. From a formal perspective, the subordination and coordination indices also provide a window on how often articles exploit the recursive resources of syntax to build complex propositions.

Second, the results show the limits of what such simple measures can reveal in a deliberately homogeneous environment. When most texts come from the same outlet, over a short period (October–November 2025), with similar editorial norms, the basic metrics are naturally similar.

At the same time, the study comes with clear limitations. The corpus is restricted to one outlet, one time frame and a specific regional focus, which is a strength for experimental control but a limitation for generalisation. For a future study, other newspapers with different house styles or political orientations might give different profiles. Likewise, the reliance on spaCy's sentence segmentation and POS tags means that the syntactic measures are only proxies for deeper structure: labels such as SCONJ and CCONJ do not distinguish between different types of subordinate clauses, nor do they capture the full depth of embedding or the presence of non-finite structures. More fine-grained syntax-based indices, such as dependency-based complexity measures or clause-level metrics in the style of Lu (2010), would be needed to explore those dimensions.

Finally, the project suggests several directions for future work that follow directly from its methodological orientation. One obvious extension would be to apply the same spaCy-based pipeline to a longer diachronic sample, for example a ten-year archive of local, national and international articles, to see whether the same stability holds over time or whether shifts in journalistic practice alter the complexity profile. Another is to broaden the range of outlets to test how robust the patterns are beyond a single institutional setting. A third avenue would be to combine the current metrics with readability indices (such as Flesch scores) or with more advanced syntactic measures, to relate surface POS-based complexity to reader-oriented difficulty and to more formal accounts of recursion and expressive power.

In summary, within this corpus, local, national and international news are closely related variants of a single register family than as strongly differentiated systems. The tools implemented here are reinforces the value of using simple, replicable NLP pipelines as a starting point for investigating formal and functional aspects of complexity in real-world texts, while also signalling the need for richer syntactic and discourse-level tools in future research.

CONCLUSION

This project set out to investigate whether local, national and international news differ in syntactic and lexical complexity, and to do so using a minimal, transparent spaCy-based pipeline rather than heavy, opaque tooling. A corpus of 156 BBC News articles, organised into 52 triplets of local, national and international coverage of comparable events, was annotated with spaCy and processed using short Python scripts that count tokens, POS categories and sentence-based metrics, and that compute lexical diversity and Zipf-based rarity indices.

All three types of news have similar proportions of major word classes, comparable Zipf-based lexical rarity, and broadly similar syntactic proxies such as mean sentence length. Within this shared template, Local articles tend to be shorter, slightly more lexically diverse and richer in proper nouns, reflecting their focus on concrete, place-bound events and named entities. National articles are longer and somewhat more repetitive lexically. And International articles usually sit between the two, but the distributions overlap heavily. Substantively, the findings suggest that in a controlled environment, the differences between local, national and international news are best understood as small shifts in emphasis within a common register family, rather than as qualitatively distinct systems.

Methodologically, the project demonstrates that a lightweight pipeline built on spaCy can provide a robust framework for exploring register-related complexity in real news data. The approach is fully reproducible, easy to extend, and directly compatible with more advanced metrics.

Future work can apply the same pipeline to a longer historical archive to test temporal stability; integrating readability indices and other syntactic complexity measures; and linking these empirical patterns more tightly to formal theories of recursion and expressive power. In that sense, the present study should be read as a demonstration that carefully controlled corpora, combined with inspectable NLP tools, can illuminate both the possibilities and the limits of empirical work on linguistic complexity in contemporary news.

BIBLIOGRAPHY

Biber, D. (1995). Chapter 2., The comprehensive analysis of register variation. En D. Biber, *Dimensions of register variation. A cross-linguistic comparison* (pages. 34-37). Cambridge: Cambridge University Press.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *The Psychonomic Society, Inc*, 977-990. Seen on October 17th 2025, from https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus/brysbaertnew.pdf

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)* (pages. 1383-1392). Melbourne (Australia): Association for Computational Linguistics.

Flaounas, I., Ali, O., Lansdall-Welfare, T., De Bie, T., Mosdell, N., Lewis, J., & Cristianini, N. (2013). RESEARCH METHODS IN THE AGE OF DIGITAL JOURNALISM: Massive-scale automated analysis of news-content—topics, style and gender. *Digital Journalism, 1(1)*, 102-116. Seen on October 6th, 2025, from https://doi.org/10.1080/21670811.2012.714928

Horne, B. D., Khedr, S., & Adalı, S. (2018). Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape. *Association for the Advancement of Artificial Intelligence*, 518-527.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 474-496.

Mair, P. & Wilcox R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*, 464–488. https://doi.org/10.3758/s13428-019-01246-w

McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *The Psychonomic Society, Inc*, 381-392. Seen on October 10th, 2025 from:

https://www.academia.edu/5913160/MTLD_vocd_D_and_HDD_A_validation_study_of_sophisticated_approaches_to_lexical_diversity_assessment

Team, R. C. (2025). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. , Vienna, Austria, URL https://www.R-project.org/

Shulman, H. C., Markowitz, D. M., & Rogers, T. (5 de June de 2024). *Science Advances*. Seen on October 10th, 2025, from Science Advances: https://www.science.org/doi/10.1126/sciadv.adn2555

spaCy. (2016-2025). *Linguistic Features*. Obtained from spaCy: https://spacy.io/usage/linguistic-features

APPENDIX A: Links to the articles and metrics

Available on GitHub: https://github.com/pedrogl2002/News-complexity-analysis

APPENDIX B: Tables with results

Metric	Type of analysis	Loc-Nat	Loc-Int	Nat-Int
ADJ Rate	Parametric	NO	NO	NO
ADV Rate	Robust	NO	NO	NO
Coord. Rate	Parametric	NO	NO	NO
Nominal Rate	Robust	NO	NO	NO
Noun Rate	Parametric	NO	NO	NO
Pronoun Rate	Parametric	NO	NO	NO
STTR	Parametric	NO	NO	NO
Subord. Rate	Parametric	NO	NO	NO
Syntactic Cplx - MLS	Parametric	NO	YES	NO
Verb Rate	Parametric	NO	NO	NO
Zipf Score	Robust	NO	NO	NO

A value of 'NO' denotes no significant differences between the distributions; 'YES' denotes the presence of significant differences.