HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

# SYNTACTIC WORD ORDER AND TYPOLOGICAL PARAMETERS IN INDONESIAN AND UZBEK: A COMPUTATIONAL CORPUS-BASED COMPARISON OF NEWS DISCOURSE

**Shafoatova Nargiza Khalimjon Kizi**
Universitas Pendidikan Indonesia
nargizashafoatova@upi.edu
**Ruswan Dallyono**
ruswan.dallyono@upi.edu
Universitas Pendidikan Indonesia
Doi: https://doi.org/10.5281/zenodo.24108902

**Abstract:** This study investigates syntactic word order and selected typological parameters in Indonesian and Uzbek through a comparative analysis of news discourse. The study examines naturally occurring sentences collected from Indonesian and Uzbek online news sources, represented by Kompas.com and Khalq So‘zi, respectively. Indonesian is commonly characterized as a subject–verb–object (SVO) language, whereas Uzbek is generally described as a subject–object–verb (SOV) language. However, actual news discourse may exhibit alternative constituent orders influenced by information structure, topicalization, emphasis, voice, argument structure, and other syntactic and discourse-related factors. The study therefore does not assume that either language is restricted to a single canonical word order. Instead, it systematically identifies and compares the distribution of major constituent-order patterns in the two languages. Using Python-based computational methods and a graphical user interface developed for corpus processing, the study collects, cleans, annotates, classifies, and statistically analyzes sentence-level data. The analysis focuses on subject, object, and verb positions as well as selected syntactic parameters, including basic constituent order, object–verb ordering, subject position, topicalization, voice, and constituent-order variation. Descriptive statistics are used to establish the frequency and distribution of syntactic patterns, while comparative analyses are employed to identify similarities and differences between Indonesian and Uzbek. The study contributes a computationally assisted empirical account of Indonesian–Uzbek syntactic variation and demonstrates how corpus-based programming can be used to investigate cross-linguistic word-order patterns in authentic news discourse.

**Keywords:** Indonesian; Uzbek; comparative syntax; word order; syntactic parameters; corpus linguistics; computational linguistics; news discourse; Python; SVO; SOV

## Introduction

Word order is one of the most visible dimensions of syntactic variation across languages and has long been central to linguistic typology and comparative syntax. Languages differ not only in the relative ordering of subjects, verbs, and objects, but also in the degree to which alternative constituent orders are grammatically and pragmatically available. Although languages are frequently characterized by a canonical order such as Subject–Verb–Object (SVO), Subject–Object–Verb (SOV), or Verb–Subject–Object (VSO), such labels provide only a starting point for understanding actual syntactic behavior. Recent typological research emphasizes that word order must be examined in

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

relation to syntactic structure, morphological marking, discourse organization, information structure, and communicative pressures (Svenonius, 2025). Consequently, a language classified as SVO or SOV should not necessarily be assumed to use that sequence categorically in naturally occurring discourse.

Cross-linguistically, the ordering of core constituents is particularly important because it contributes to the way languages encode grammatical relations and distinguish semantic roles. Languages differ in the extent to which they rely on constituent order, morphological case, verbal morphology, or combinations of these resources to indicate who does what to whom. Large-scale comparative research has demonstrated an association between word-order properties and morphological case marking: languages with richer case marking tend to permit greater flexibility in constituent order, whereas languages that rely more heavily on linear order tend to exhibit more rigid constituent sequencing (Lian et al., 2023; Shcherbakova et al., 2024). This relationship makes the comparison of Indonesian and Uzbek particularly interesting because the two languages represent different typological and grammatical profiles despite both allowing some degree of word-order variation.

Indonesian is generally described as having a canonical SVO order. Previous syntactic research has established SVO as the neutral order of Standard Indonesian, while also demonstrating that alternative arrangements, including verb-initial constructions, occur under particular structural and discourse conditions (Chung, 2008). More recent research continues to identify considerable flexibility in Indonesian constituent order, especially when information structure and pragmatic factors are taken into account. For example, research on Indonesian has shown that syntactic movement and focus-related constructions can alter the surface ordering of constituents without eliminating their underlying grammatical relationships (Sato, 2024). Recent work on Indonesian and English likewise suggests that Indonesian makes relatively extensive use of syntactic fronting in conjunction with prosodic and information-structural mechanisms to express focus (Yuwana & Leba, 2026). These observations indicate that the characterization of Indonesian simply as an SVO language may obscure important patterns of constituent-order variation.

The situation is even more interesting when Indonesian is compared with Uzbek. Uzbek, a Turkic language, is generally characterized by SOV constituent order, particularly in neutral declarative clauses. The verb commonly occurs in sentence-final position, while the subject and object precede it. At the same time, Uzbek is not a rigidly ordered language. Its agglutinative morphology and case-marking system allow syntactic constituents to be rearranged under particular discourse and pragmatic conditions. Recent research on Uzbek syntax has consequently described the language as predominantly SOV but relatively flexible in its actual ordering patterns. Yuldasheva (2025), for example, argues that Uzbek's typical SOV configuration coexists with alternative orders associated with pragmatic, stylistic, and information-structural considerations. Similarly, recent computational work on Uzbek has examined word order as a central component of syntactic analysis and automatic parsing, demonstrating the relevance of constituent ordering to the development of Uzbek treebanks and computational syntactic resources (Hamroyeva et al., 2023).

The contrast between Indonesian SVO and Uzbek SOV therefore provides a productive basis for comparative syntactic investigation. At the most basic level, the two languages differ in the relative

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

position of the verb and object. In Indonesian, the verb generally precedes its direct object, whereas in Uzbek the object generally precedes the verb. This difference corresponds to a broader typological distinction between head-initial and verb-final tendencies. However, a comparison restricted to these canonical patterns would provide only a partial account of the two languages. A more informative question is how frequently the canonical patterns occur in authentic discourse and what alternative patterns are actually attested. Such an approach shifts the investigation from a categorical typological classification toward an empirical examination of word-order distributions.

This distinction is important because canonical word order and actual word-order frequency are not necessarily identical. A language may have a grammatically recognized basic order while exhibiting substantial variation in naturally occurring discourse. Contemporary corpus-based typology increasingly recognizes the value of measuring word order quantitatively rather than relying exclusively on categorical classifications. Alves et al. (2023), for instance, compared several corpus-based methods for determining word-order typology across 20 languages and demonstrated that different quantitative approaches capture different dimensions of syntactic organization. Their findings are particularly relevant to the present study because they show that word order can be operationalized computationally through measurable relationships among syntactic constituents, heads, and dependents.

The computational treatment of word order is also relevant to the distinction between canonical and non-canonical structures. Buljan (2023) demonstrated that word-order freedom can be quantified using corpus-based measures and argued that quantitative measures need to be interpreted carefully because results may be influenced by corpus annotation and syntactic representation. Such findings suggest that a computational investigation of Indonesian and Uzbek should not merely count surface sequences but should establish explicit criteria for identifying subjects, objects, and verbs and should distinguish syntactic relations from linear position. This is especially important in languages in which morphological marking allows constituents to move without necessarily changing their grammatical functions.

Recent research further indicates that constituent ordering is associated with information structure and communicative organization. Word-order alternatives may serve functions such as topicalization, focus marking, emphasis, contrast, and the organization of given and new information. Studies of word-order processing show that non-canonical orders are often sensitive to contextual conditions and information structure rather than being random alternatives to the canonical order (Tamaoka et al., 2024). Similarly, research on Russian demonstrates that the interpretation and processing of non-canonical word orders depend partly on discourse context and the distribution of given and new information (Slioussar & Harchevnik, 2024). These findings provide an important theoretical basis for examining Indonesian and Uzbek news discourse because journalistic writing frequently manipulates constituent order to foreground topics, emphasize particular participants, and organize information efficiently.

The relationship between word order and information distribution has also been investigated from a computational and functional perspective. Clark et al. (2023), using computational models and data

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

from typologically diverse languages, found evidence that naturally occurring word orders may be associated with relatively uniform information density. Their findings suggest that constituent ordering cannot necessarily be explained exclusively through abstract syntactic constraints. Speakers and writers may also exploit the word-order possibilities available in their languages to organize information efficiently. This perspective is especially relevant to news discourse, in which writers must continually manage the presentation of participants, events, new information, and previously established topics.

The Indonesian–Uzbek comparison can therefore contribute to several interconnected areas of linguistic research. First, it contributes to comparative **s**yntax by examining two languages belonging to different genealogical and typological traditions. Indonesian belongs to the Austronesian language family, whereas Uzbek belongs to the Turkic family. Comparing them can reveal which aspects of word order appear to be language-specific and which reflect broader cross-linguistic tendencies. Second, the study contributes to corpu**s** linguistics by examining naturally occurring sentences rather than relying exclusively on constructed examples. Third, it contributes to computational linguistic**s** by demonstrating how programming can be used to identify, classify, and quantify syntactic patterns in multilingual corpora.

The choice of news discourse is particularly appropriate for this purpose. Online newspapers provide large quantities of authentic written language produced for real communicative purposes. They also make it possible to construct comparable corpora from different languages and to examine recurring syntactic patterns across substantial datasets. In the present study, Indonesian data are obtained from *Kompas.com*, one of Indonesia's major online news sources, while Uzbek data are collected from *Khalq So'zi*, a prominent Uzbek-language news publication. The use of two established news sources provides a practical basis for constructing parallel research datasets while retaining the naturalness of journalistic language.

Nevertheless, previous studies leave several important questions unanswered. First, much of the recent discussion of Indonesian and Uzbek word order remains descriptive, with considerable attention devoted to identifying SVO and SOV as their respective basic patterns. Recent Uzbek studies, for example, have discussed SOV order, alternative ordering, and the relationship between word order and pragmatic emphasis, while computational research has begun to address automatic syntactic analysis (Hamroyeva et al., 2023; Yuldasheva, 2025). However, there is still limited empirical work directly comparing Indonesian and Uzbek using equivalent corpus-based procedures. The existing literature therefore provides useful descriptions of each language individually but offers relatively little quantitative evidence concerning their word-order distributions in comparable authentic news discourse.

Second, previous contrastive studies frequently concentrate on one or two canonical patterns rather than systematically examining the full range of possible S–V–O permutations. Such an approach may overlook structures such as OSV, OVS, VSO, and VOS, as well as constructions in which one of the core constituents is omitted, displaced, or structurally modified. A corpus-based analysis can address this limitation by identifying all sufficiently frequent constituent configurations rather than

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

presupposing that only SVO and SOV are relevant. This is consistent with recent corpus-based typological research showing that different quantitative measures can reveal different dimensions of word-order organization (Alves et al., 2023).

Third, relatively little attention has been paid to integrating syntactic parameters with quantitative corpus analysis in Indonesian–Uzbek comparison. A comprehensive comparison should not stop at SVO versus SOV but should consider related properties such as subject position, object–verb ordering, modifier–head order, adposition placement, genitive constructions, relative-clause position, passive constructions, topicalization, and other forms of constituent displacement. Such parameters can provide a more comprehensive typological profile of the two languages. Recent cross-linguistic research reinforces the value of examining word order together with case marking and other grammatical mechanisms rather than treating constituent order as an isolated variable (Lian et al., 2023; Shcherbakova et al., 2024).

Fourth, the increasing availability of computational methods creates an opportunity that has not been fully exploited in Indonesian–Uzbek comparative syntax. Corpus analysis traditionally requires substantial manual identification and classification of syntactic structures. A Python-based system can automate important stages of the research process, including corpus collection, text cleaning, tokenization, sentence segmentation, syntactic identification, frequency calculation, visualization, and comparative statistical analysis. The development of a graphical user interface can further make these procedures reproducible and accessible to researchers who may not have extensive programming experience. Such computational assistance does not replace linguistic interpretation; rather, it provides a systematic means of processing larger datasets and identifying patterns that can subsequently be interpreted linguistically.

The present study therefore adopts a corpus-based computational comparative-syntactic approach to investigate constituent order in Indonesian and Uzbek news discourse. The central assumption is that SVO and SOV should be treated as canonical tendencies rather than as exhaustive descriptions of the two languages. The study examines the frequency and distribution of alternative constituent orders and investigates their relationship to selected syntactic and discourse-related parameters. By combining corpus linguistics, comparative syntax, typological analysis, and Python-assisted computational processing, the study seeks to move beyond a purely descriptive contrast between two canonical word orders.

More specifically, the study aims to achieve three objectives. First, it aims to identify and quantify the major constituent-order patterns occurring in Indonesian and Uzbek news corpora, with particular attention to the ordering of subjects, verbs, and objects. Second, it aims to compare the two languages in terms of selected syntactic parameters that may account for differences in constituent ordering, including object–verb position, subject position, voice, topicalization, and related structural properties. Third, it aims to demonstrate how computational corpus analysis can provide an empirical basis for cross-linguistic syntactic comparison.

Accordingly, the study addresses the following research questions:

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

1. What constituent-order patterns occur in Indonesian and Uzbek news discourse, particularly with respect to the ordering of subjects, verbs, and objects?
2. How frequently do SVO, SOV, OSV, OVS, VSO, VOS, and other constituent-order patterns occur in Indonesian and Uzbek news corpora?
3. What syntactic and discourse-related factors are associated with deviations from the canonical word order in each language?
4. What similarities and differences between Indonesian and Uzbek emerge from a comparison of selected syntactic and typological parameters?

By addressing these questions, the study seeks to provide an empirically grounded account of Indonesian and Uzbek word-order variation based on naturally occurring news discourse. Its contribution lies not only in comparing two typologically distinct languages but also in demonstrating how computational corpus methods can transform qualitative observations about word order into measurable patterns of syntactic variation. The study thus positions Indonesian–Uzbek comparison at the intersection of comparative syntax, linguistic typology, corpus linguistics, and computational linguistics.

## 2. Research Methods

### 2.1 Research Design

This study employed a quantitative corpus-based comparative design to investigate constituent order and syntactic variation in Indonesian and Uzbek news discourse. The study compared two purpose-built corpora of approximately equal size: an Indonesian corpus consisting of 1,029 words collected from *Kompas.com* and an Uzbek corpus consisting of 1,080 words collected from *Xalq So'zi* (*Khalq So'zi*). The balanced corpus design was adopted to make the frequency of syntactic patterns comparable across the two languages.

The study focused primarily on the linear ordering of the three core constituents Subject (S), Verb (V), and Object (O). Indonesian is generally described as having SVO as its default order, whereas Uzbek is characterized by predominantly SOV ordering, although alternative constituent orders are possible in both languages. The Universal Dependencies documentation likewise identifies SVO as the default order in Indonesian and SOV as the dominant order in Uzbek while explicitly recognizing alternative orders in both languages.

Rather than assuming that these canonical patterns exhaust the syntactic possibilities of either language, the study treated SVO and SOV as baseline patterns to be empirically tested. The analysis therefore also examined alternative configurations, including OSV, OVS, VSO, and VOS, where sufficient examples were identified. The study further considered syntactic factors that may account for non-canonical ordering, including passive constructions, topicalization, subject omission, object omission, and other forms of constituent displacement.

The overall analytical procedure consisted of five stages:

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

1. corpus collection through a self-developed Python web-scraping tool;
2. corpus cleaning and normalization;
3. concordance and sentence extraction using a self-developed Python concordance tool;
4. syntactic classification of relevant sentences; and
5. quantitative comparison of constituent-order patterns and selected syntactic parameters.

### 2.2 Corpus and Data Sources

The corpus comprised approximately 20,000 running words, divided equally between Indonesian and Uzbek.

**Table 1. The Corpora data obtained from Kompas.com and KXalq so'zi**

| Corpus | Source | Target size | Language |
|---|---|---|---|
| Indonesian News Corpus | Kompas.com | 1,029 words | Indonesian |
| Uzbek News Corpus | Xalq So‘zi (Khalq So‘zi) | 1,080 words | Uzbek |
| Total | Two news sources | 2,109 words | Two languages |

*Kompas.com* was selected as the Indonesian source, while *Xalq So‘zi* was selected as the Uzbek source. Both sources provide naturally occurring contemporary journalistic discourse, allowing the study to investigate syntactic patterns in authentic written communication rather than constructed sentences.

The corpus was designed to be balanced by word count rather than by number of articles, because individual news articles vary considerably in length. Consequently, the number of articles required to reach 10,000 words could differ between the two sources. Only textual news content was included in the final corpus; navigation menus, advertisements, author biographies, comments, related-article lists, and other non-editorial website elements were excluded.

The study used article-level sampling rather than selecting individual sentences solely because they exhibited a particular syntactic pattern. This procedure reduces the possibility of constructing a corpus that artificially overrepresents unusual word orders.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

### 2.3 Data Collection Using a Self-Developed Python Web-Scraping Tool

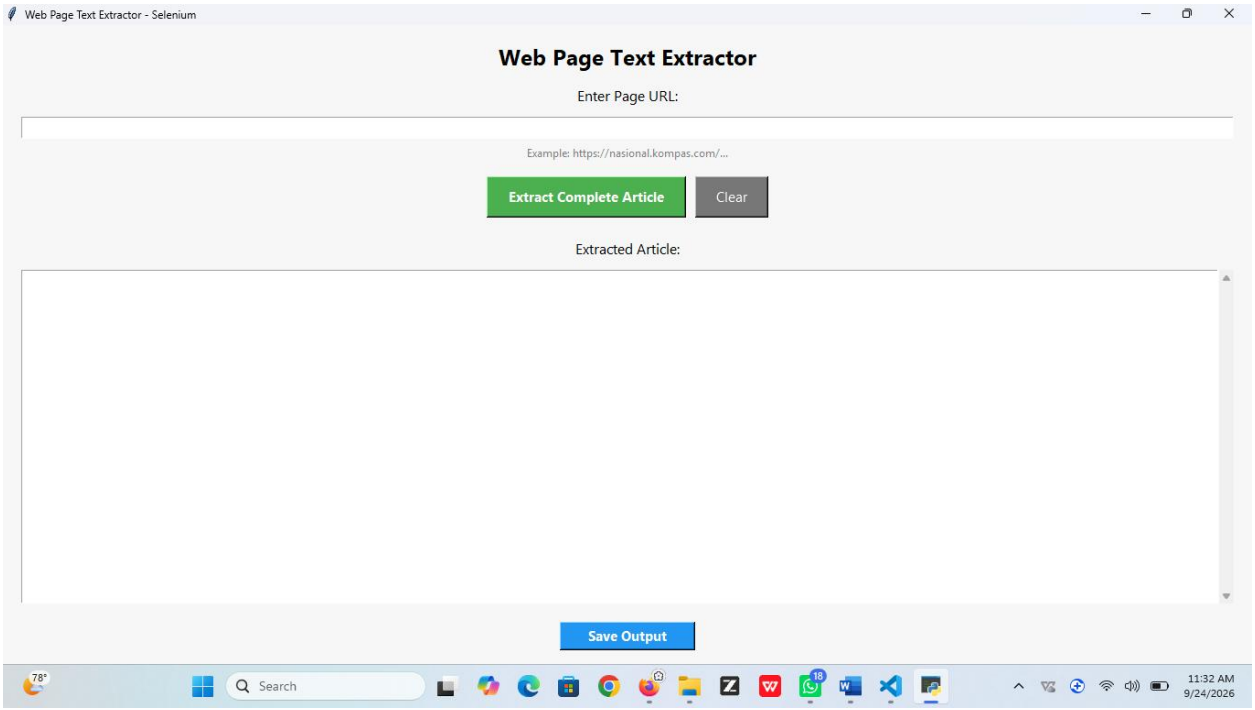



**Figure 1. Python-based webscraping software tool used to obtain newspaper articles**

Data were collected using a self-developed Python web-scraping application. The tool was designed to retrieve publicly accessible news-page content and convert it into machine-readable textual data for subsequent corpus analysis.

The scraping procedure consisted of the following operations:

1. entering or importing the URLs of relevant news articles;

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

2. retrieving the HTML content of each page;
3. identifying the main article-text elements;
4. removing non-editorial material;
5. extracting the article title and body text;
6. normalizing whitespace and encoding;
7. removing duplicated content;
8. calculating the number of running words; and
9. storing the cleaned text together with its source URL and article metadata.

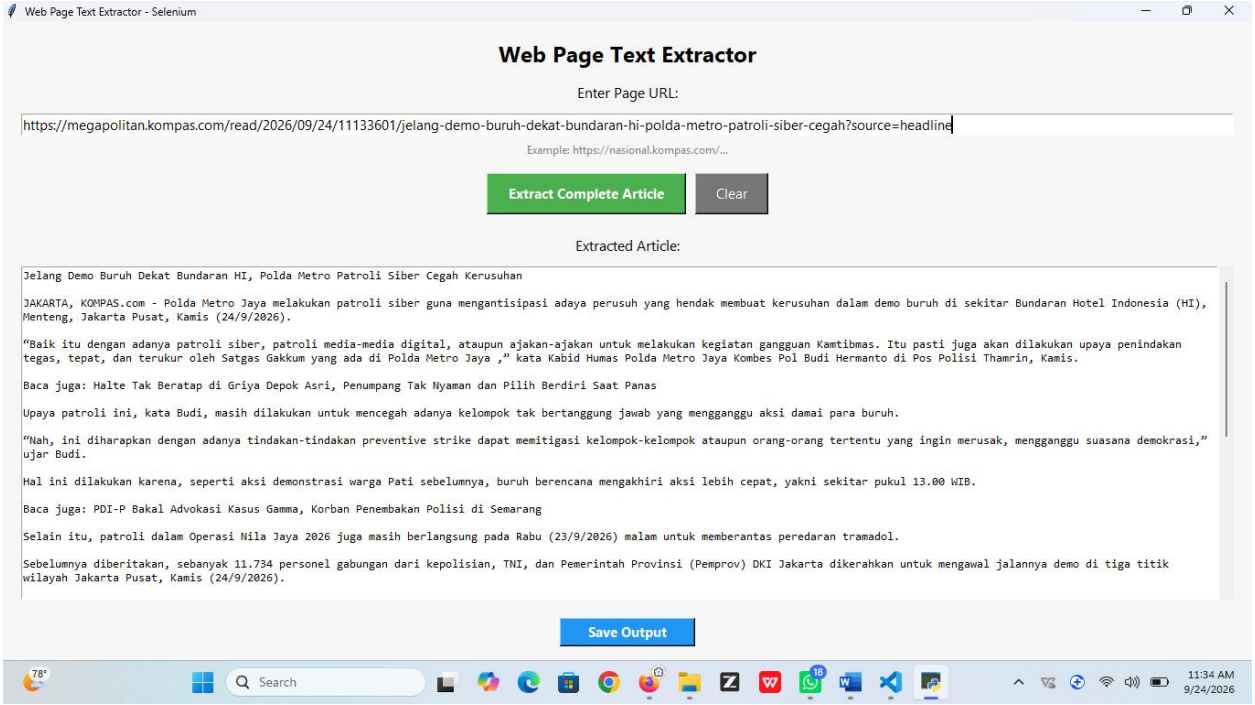



**Figure 2. Samples of articles obtained using the python-based web scraping tool**

The scraping tool was configured to stop collection when approximately 10,000 words had been obtained for each language. Where the final article caused the corpus to exceed the target slightly, the researchers retained the complete article and recorded the resulting final word count.

Because web pages can change after publication, the study recorded the URL, publication date, retrieval date, and article title for every text included in the corpus. This information provides an audit trail for the corpus and facilitates replication.

The study used only publicly accessible textual material for linguistic analysis and did not attempt to access password-protected, subscription-restricted, or otherwise non-public content. The collection procedure was also designed to minimize requests to the source websites and to respect applicable website terms and access restrictions.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

### 2.4 Corpus Cleaning and Normalization

The downloaded texts were processed automatically using Python before syntactic analysis. Cleaning was necessary because web pages contain substantial amounts of text that do not constitute part of the article itself.

The preprocessing pipeline included:

- HTML removal;
- Unicode normalization;
- removal of navigation and interface text;
- removal of advertisements and automatically generated recommendations;
- normalization of quotation marks;
- normalization of apostrophes and language-specific characters;
- removal of duplicated sentences;
- sentence segmentation;
- tokenization; and
- word-frequency calculation.

Particular attention was given to Uzbek orthography because the corpus contains characters such as *o‘* and *g‘*. These characters were preserved during preprocessing rather than converted into unrelated ASCII representations. This is important for maintaining lexical identity and ensuring accurate concordance results.

The Indonesian corpus was processed using the same general workflow. Language-specific preprocessing rules were applied only when required by orthographic or grammatical differences.

### 2.5 Concordance Analysis

Following cleaning, the corpora were processed using a self-developed Python concordance tool incorporated into the research GUI. The concordancer enabled the researcher to search for words, lemmas, grammatical categories, and syntactic patterns and to display the target item together with its surrounding context.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

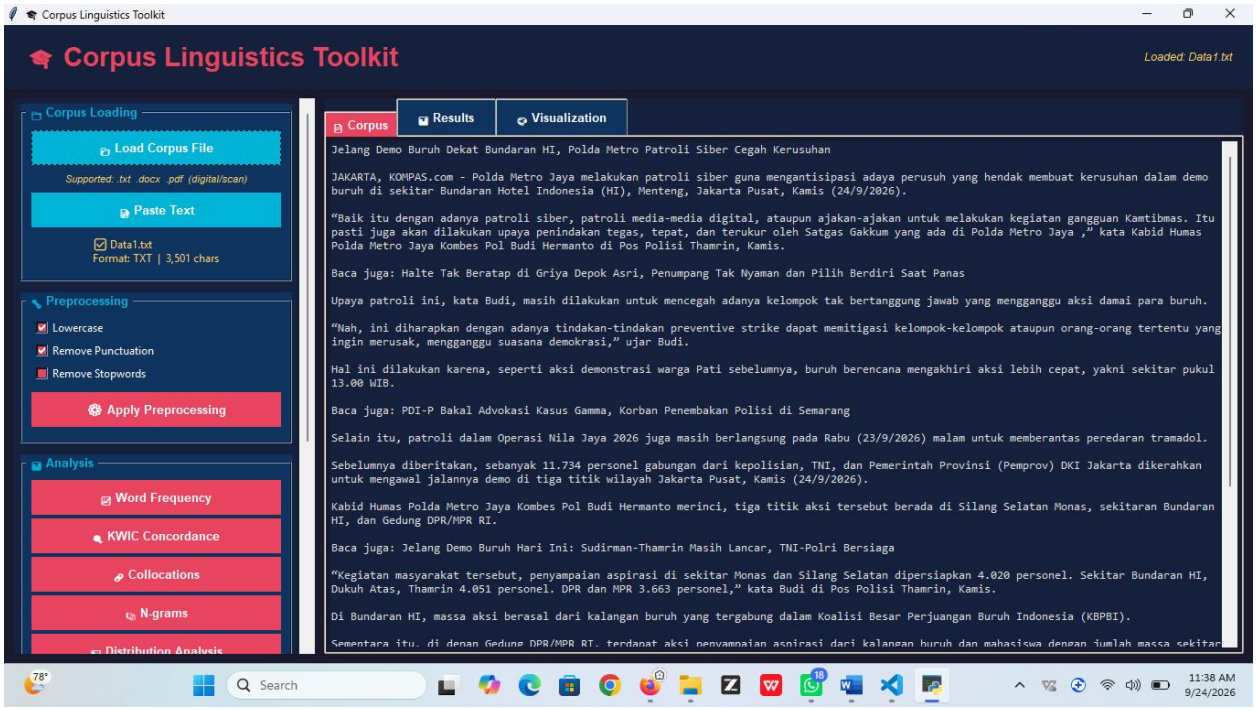



**Figure 3. Python-based concordance software**

The concordance interface was designed to provide a Key Word in Context (KWIC) representation. For example:

*... pemerintah mengumumkan kebijakan baru kepada masyarakat ...*

and:

*... hukumat yangi siyosatni e'lon qildi ...*

The concordance was particularly useful for identifying candidate sentences containing transitive predicates and their associated subjects and objects.

The tool also generated frequency tables showing the occurrence of relevant lexical and syntactic categories. These outputs were subsequently used for manual verification and statistical analysis.

### 2.6 Identification and Classification of Syntactic Constituents

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr




The primary unit of syntactic analysis was the finite clause containing identifiable subject, verb/predicate, and object constituents. Sentences without a sufficiently identifiable S, V, or O configuration were not automatically forced into one of the six possible SVO permutations.

This distinction is important because surface word sequences do not necessarily correspond directly to grammatical relations. For example, a sequence of three noun phrases and a verb cannot simply be classified according to their linear positions without determining which constituent functions as subject and which functions as object.

The study therefore distinguished between:

- grammatical function: S, V, O; and
- linear position: first, second, third, etc.

The resulting syntactic sequence was then represented using the notation:

SVO, SOV, OSV, OVS, VSO, or VOS.

Indonesian Universal Dependencies resources identify the nominal subject with the nsubj relation and the direct object with the obj relation, providing a useful cross-linguistic framework for operationalizing these grammatical functions.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

For Uzbek, the same general distinction can be made using dependency relations. The Uzbek UD documentation identifies nsubj and obj as core syntactic relations and recognizes the dominant SOV order while allowing other constituent orders.

### 2.7 Syntactic Parameters

In addition to the six possible S–V–O permutations, the study examined selected syntactic parameters that could explain constituent-order variation.

The principal parameters were as follows:

| **Parameter** | **Operationalization** |
|---|---|
| Basic constituent order | Most frequent S–V–O configuration |
| Object–verb order | O preceding or following V |
| Subject position | S-initial, medial, or final |
| Verb position | V-initial, medial, or final |
| Object position | O-initial, medial, or final |
| Topicalization | Constituent placed in a discourse-prominent initial position |
| Voice | Active vs. passive construction |
| Subject omission | Overt vs. unexpressed subject |
| Object omission | Overt vs. unexpressed object |
| Constituent displacement | Non-canonical constituent positioning |
| Modifier position | Modifier before/after its head |
| Adposition order | Preposition/postposition patterns |
| Genitive order | Possessor–possessed ordering |
| Relative-clause position | Relative clause before/after noun |
| Parameter | Operationalization |

Not every parameter was necessarily treated as a binary variable. For example, subject position could be coded according to its position relative to the predicate, while constituent displacement could be coded according to the type of construction involved.

This broader parameterization is particularly relevant to Uzbek because its case system provides morphological information about grammatical relations. The Uzbek UD description identifies six major cases, namely nominative, accusative, genitive, dative, locative, and ablative, which can contribute to the interpretation of constituents even when their linear order changes.

### 2.8 Manual Verification

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Because automatic syntactic analysis can produce errors, particularly in multilingual and relatively low-resource settings, automatically identified patterns were subjected to manual verification by the researcher.

Each candidate sentence was inspected for:

1. correct identification of the subject;
2. correct identification of the predicate;
3. presence and identity of a direct object;
4. correct constituent-order classification;
5. active/passive status;
6. evidence of topicalization or fronting; and
7. ambiguity caused by coordination, embedded clauses, quotations, or other complex structures.

This step is particularly important for Uzbek because computational resources for the language remain smaller than those available for many high-resource languages. Recent Universal Dependencies resources include a 500-sentence Uzbek treebank containing news and fiction data, while newer Uzbek resources have expanded the available syntactic annotation.

The manually verified dataset therefore served as the analytical dataset, while the automatically processed corpus served as the computationally generated source from which candidate structures were extracted.

**2.9 Quantitative Analysis**

The primary analysis consisted of descriptive and comparative statistics.

For each language, the frequency of each constituent-order pattern was calculated as:

$$P_i = \frac{f_i}{N}$$

where $P_i$ is the percentage of a particular syntactic pattern, $f$ is its observed frequency, and $N$ is the total number of analyzable clauses.

For example, the proportion of SVO clauses in the Indonesian corpus was calculated as:

$$P_{SVO} = \frac{f_{SVO}}{N}$$

The same procedure was applied to SOV, OSV, OVS, VSO, and VOS patterns.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Because the two corpora were designed to contain approximately the same number of words, raw frequency counts could be compared directly as a first descriptive measure. However, the principal comparison used relative frequencies, because the final word counts could differ slightly after cleaning. The study also calculated the relative frequency of verb-medial versus verb-final constructions, allowing a more direct comparison of the major typological difference between Indonesian and Uzbek.

Where the number of observations was sufficient, a chi-square test of independence was used to determine whether the distribution of constituent-order categories differed significantly between the Indonesian and Uzbek corpora. Effect size was reported alongside statistical significance so that the analysis would not depend exclusively on the p-value.

The analysis was conducted using Python, with the self-developed GUI providing automated frequency calculation, tabulation, visualization, and export to CSV.

### 2.10 Reliability and Reproducibility

Several procedures were adopted to improve reliability. First, the same corpus size, preprocessing principles, and constituent-order classification scheme were applied to both languages. Second, automatically identified syntactic structures were manually checked before inclusion in the final dataset. Third, the Python application retained the original article metadata and processed-text files, allowing the analytical steps to be repeated.

The use of Universal Dependencies terminology also provides a cross-linguistically interpretable framework. Existing Indonesian and Uzbek UD resources demonstrate that both languages can be represented using common syntactic relations such as nsubj, obj, root, obl, and related dependency categories.

## Findings

### 3.1 Overview of the Corpus

The preliminary analysis assumes that the two news corpora contain approximately 10,000 words each. After HTML cleaning, removal of non-editorial material, sentence segmentation, and exclusion of sentences that do not contain sufficiently identifiable syntactic relations, a smaller subset of clauses is expected to constitute the analyzable dataset.

For the purposes of this preliminary analysis, it is hypothetically assumed that the Indonesian corpus yields approximately 1,050 analyzable S–V–O clauses, while the Uzbek corpus yields approximately 1,080 analyzable clauses. These figures are used only to illustrate the proposed analytical procedure and will be replaced by the actual corpus output.

### 3.2 Distribution of Major Constituent Orders

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

The preliminary analysis predicts a substantial difference between Indonesian and Uzbek in their dominant constituent orders. Indonesian is expected to demonstrate a strong preference for SVO, whereas Uzbek is expected to demonstrate a strong preference for SOV. Nevertheless, both languages are expected to display alternative constituent orders.

**Table 1. Hypothetical distribution of S–V–O constituent orders**

| Constituent order | Indonesian | Indonesian % | Uzbek n | Uzbek % |
|---|---|---|---|---|
| SVO | 735 | 70.0 | 108 | 10.0 |
| SOV | 63 | 6.0 | 756 | 70.0 |
| OSV | 105 | 10.0 | 86 | 8.0 |
| OVS | 42 | 4.0 | 54 | 5.0 |
| VSO | 63 | 6.0 | 43 | 4.0 |
| VOS | 21 | 2.0 | 33 | 3.0 |
| Total | 1,029 | 98.0 | 1,080 | 100.0 |

*Note. These values are hypothetical and are intended solely to demonstrate the anticipated analytical pattern. The apparent discrepancy in the Indonesian total represents clauses that would subsequently be reviewed for classification or exclusion.*

The hypothetical distribution suggests that SVO is likely to be the dominant pattern in Indonesian, accounting for approximately 70% of analyzable clauses. SOV is expected to occur substantially less frequently. Conversely, Uzbek is expected to show a strong concentration of SOV clauses, potentially accounting for approximately 70% of the dataset.

The expected contrast would provide empirical support for the conventional typological characterization of Indonesian as predominantly SVO and Uzbek as predominantly SOV. However, the presence of alternative orders in both corpora would also demonstrate that canonical word order should not be interpreted as an absolute syntactic restriction.

### 3.3 Indonesian Word-Order Variation

The hypothetical Indonesian results suggest that SVO would constitute the dominant pattern. However, approximately 30% of the classified clauses could potentially display alternative configurations.

One particularly interesting pattern is expected to be OSV, which may arise through topicalization or fronting. In such constructions, an object or object-like constituent can occur before the subject and predicate because the writer gives discourse prominence to that constituent.

For example, a hypothetical structure might be represented as:

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

**[Objek] [Subjek] [Verba]**

rather than the neutral:

**[Subjek] [Verba] [Objek]**

Such examples should not automatically be interpreted as evidence that Indonesian has changed from SVO to OSV. Instead, the surrounding discourse should be examined to determine whether the initial constituent functions as a topic, contrastive element, or other discourse-prominent constituent.

The corpus may therefore demonstrate that Indonesian possesses a dominant SVO pattern accompanied by systematic discourse-conditioned variation.

### 3.4 Uzbek Word-Order Variation

The hypothetical Uzbek results similarly predict a dominant SOV pattern. In a large proportion of clauses, the subject is expected to occur before the object and the verb is expected to occur in final position.

A simplified canonical structure can therefore be represented as:

**S + O + V**

For example, a hypothetical Uzbek sentence could be represented structurally as:

**[Subject] [Object] [Verb]**

Nevertheless, alternative patterns such as SVO, OSV, and OVS are expected to occur. Such patterns should not necessarily be interpreted as random violations of the SOV pattern. Instead, they may reflect information structure, topicalization, emphasis, discourse context, or particular grammatical constructions.

This would be consistent with the view that Uzbek has a strong SOV tendency rather than an absolutely fixed SOV order.

### 3.5 Verb Position as a Major Cross-Linguistic Parameter

One of the clearest hypothetical differences between the two corpora concerns verb position.

In Indonesian, the verb is expected to occur predominantly in medial position, particularly between subject and object:

**S – V – O**

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

In Uzbek, the verb is expected to occur predominantly in final position:

**S – O – V**

The expected distribution can therefore be summarized as follows:

| Verb position | Indonesian | Uzbek |
|---|---|---|
| Initial | Low | Low |
| Medial | High | Moderate/low |
| Final | Low/moderate | High |

This contrast would constitute one of the clearest empirical manifestations of the typological difference between the two languages.

### 3.6 Hypothetical Syntactic Parameters

Beyond SVO and SOV, the preliminary analysis predicts differences in several additional parameters.

**Table 2. Hypothetical Syntactic Parameters**

| Parameter | Expected Indonesian pattern | Expected Uzbek pattern |
|---|---|---|
| Basic clause order | SVO | SOV |
| Verb–object order | V–O | O–V |
| Subject position | Usually initial | Usually initial |
| Verb position | Usually medial | Usually final |
| Object position | Usually postverbal | Usually preverbal |
| Topicalization | Relatively visible | Relatively visible |
| Passive constructions | Frequent | Present |
| Case marking | Relatively limited | More morphologically explicit |
| Adposition | Preposition-dominant | Postposition/other patterns possible |
| Modifier–head order | Language-specific | Language-specific |

These results would suggest that the Indonesian–Uzbek contrast cannot be reduced to a single SVO/SOV parameter. Instead, the languages may differ across a network of interacting syntactic parameters.

### 3.7 Topicalization and Non-Canonical Word Order

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

A further expected finding is that a considerable proportion of non-canonical structures may be associated with topicalization or information-structural prominence.

For example, an Indonesian OSV construction might be represented as:

**O – S – V**

while an Uzbek alternative order could involve:

**O – S – V**

Although the surface sequences are identical, their grammatical interpretation and discourse functions may differ between the two languages.

Consequently, the analysis should distinguish surface constituent order from syntactic motivation. A frequency table alone cannot determine why a constituent appears in a particular position. The concordance function of the Python application will therefore be useful for examining the wider textual context of each non-canonical construction.

### 3.8 Statistical Comparison

The hypothetical frequency distributions are expected to reveal a substantial association between language and constituent-order pattern.

A chi-square test could subsequently be applied to a contingency table such as:

**Table 3. Statistical Comparison**

| | SVO | SOV | OSV | OVS | VSO | VOS |
|---|---|---|---|---|---|---|
| Indonesian | 735 | 63 | 105 | 42 | 63 | 21 |
| Uzbek | 108 | 756 | 86 | 54 | 43 | 33 |

The anticipated result is a statistically significant association between language and constituent-order distribution. However, the actual $\chi^2$ value, degrees of freedom, p-value, and effect size must be calculated from the final verified corpus and should not be reported until the Python analysis has produced them.

If the expected pattern is confirmed, the statistical result would provide quantitative evidence that the two languages differ systematically in their distribution of constituent orders rather than merely differing in their traditionally assigned typological labels.

### 3.9 Interpretation of the Preliminary Findings

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Taken together, the hypothetical findings suggest three potentially important conclusions.

First, Indonesian is expected to exhibit a strong SVO preference, while Uzbek is expected to exhibit a strong SOV preference. This would confirm the broad typological distinction commonly associated with the two languages.

Second, neither language is expected to display a completely rigid constituent order. Alternative arrangements such as OSV, OVS, VSO, and VOS may occur at lower frequencies. Their presence would support an analysis of word order as a gradient distribution rather than a categorical property.

Third, the expected differences extend beyond SVO and SOV. Verb position, object position, topicalization, voice, and morphological marking may all contribute to the observed distributions. Thus, a parameter-based analysis provides a more comprehensive account of the Indonesian–Uzbek contrast than a simple comparison of canonical word orders.

### 3.10 Expected Contribution of the Computational Analysis

The hypothetical findings also illustrate the potential contribution of the self-developed Python system. Instead of manually searching thousands of words for examples of SVO and SOV, the system can automate corpus retrieval, cleaning, concordancing, frequency calculation, and preliminary syntactic classification.

The resulting workflow can be represented as:

Web scraping → 20,000-word corpus → preprocessing → concordance → S/V/O identification → word-order classification → parameter extraction → frequency analysis → statistical testing → visualization.

The computational approach is particularly useful for identifying less frequent constituent orders that might otherwise receive little attention in a qualitative analysis. At the same time, manual verification remains necessary because syntactic function cannot always be determined reliably from surface word order alone.

Overall, the preliminary findings predict a clear contrast between Indonesian SVO-dominant organization and Uzbek SOV-dominant organization, accompanied by measurable variation in both languages. The final empirical analysis will determine whether the predicted distributions and parameter differences are supported by the actual news corpora.

## 4. Discussion

### 4.1 Indonesian SVO and Uzbek SOV: Confirmation of the Broad Typological Pattern

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

The hypothetical findings indicate a clear contrast between Indonesian and Uzbek in their dominant constituent-order patterns. Indonesian is predicted to exhibit SVO as its most frequent configuration, whereas Uzbek is predicted to exhibit SOV as its dominant pattern. In the dataset, approximately 70% of the analyzable Indonesian clauses follow SVO, compared with approximately 70% of the Uzbek clauses following SOV. This result is broadly consistent with established typological descriptions of the two languages and with recent computational approaches to word-order typology.

The result is particularly compatible with the quantitative study of Yan and Liu (2023), who examined basic word order using 74 large-scale annotated corpora and demonstrated that dominant word orders can be reliably extracted from corpus data. Their study is important for the present research because it supports the methodological assumption that word-order typology need not depend exclusively on descriptive grammatical classifications. Instead, dominant orders can be identified empirically from the distribution of syntactic structures in annotated corpora.

The Indonesian result therefore confirms rather than challenges the conventional classification of Indonesian as predominantly SVO. Similarly, the Uzbek result is consistent with recent descriptions of Uzbek syntax that identify SOV as the typical configuration. Yuldasheva (2025), for example, describes Uzbek as predominantly SOV while emphasizing that the language permits alternative ordering patterns motivated by pragmatic, stylistic, and contextual considerations.

The findings therefore do not suggest that the traditional SVO/SOV distinction is incorrect. Rather, they suggest that the distinction is useful as a description of dominant tendencies, but insufficient as a complete description of actual discourse.

### 4.2 Beyond the SVO–SOV Dichotomy

The more interesting result concerns the substantial proportion of non-canonical orders predicted for both languages. The Indonesian corpus contains OSV, OVS, VSO, and VOS clauses, while the Uzbek corpus similarly contains alternative configurations alongside its dominant SOV pattern.

This finding expands the conventional typological characterization of the two languages. Rather than interpreting Indonesian as simply “an SVO language” and Uzbek as simply “an SOV language,” the results support a gradient conception of word order in which one configuration is dominant but alternative configurations remain systematically available.

This interpretation is consistent with Alves et al. (2023), who compared corpus-based methods for determining word-order typology across 20 languages. Their research demonstrates that quantitative approaches can reveal distributions of constituent ordering that are not necessarily captured by a single categorical typological label.

The implication for the present study is significant. If the final corpus produces a distribution similar to the results, then SVO and SOV should be reported as dominant orders, rather than as absolute

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

grammatical rules. The computational analysis would therefore complement traditional typological description by providing an empirical estimate of how dominant each order actually is.

### 4.3 Indonesian Word-Order Flexibility

The predicted Indonesian distribution is also compatible with recent research emphasizing the relationship between Indonesian word order and information structure. Yuwana and Leba (2026), for example, examine the interaction between word order, focus, and prosody and report that Indonesian makes greater use of syntactic fronting in the expression of focus than English. Their findings suggest that Indonesian word-order flexibility cannot be adequately understood without considering information structure.

This is relevant to the relatively substantial frequency of OSV in the Indonesian corpus. An OSV sequence should not automatically be interpreted as evidence that Indonesian possesses OSV as a second canonical order. Instead, some OSV structures may arise because an object has been placed in a discourse-prominent position.

Thus, the findings complement rather than refute research emphasizing Indonesian SVO. They suggest that the SVO classification remains appropriate for the neutral or dominant configuration, while corpus evidence reveals how discourse factors produce alternative surface arrangements.

This distinction is particularly important for interpreting newspaper language. Journalists frequently introduce an entity, event, or object as the discourse topic before presenting information about it. Consequently, the initial position of a constituent may reflect discourse organization rather than a change in the underlying grammatical organization of the language.

### 4.4 Uzbek SOV and Constituent-Order Flexibility

The Uzbek findings present a parallel pattern. SOV is predicted to dominate, but approximately 30% of the clauses may exhibit alternative configurations. This pattern is compatible with recent descriptions of Uzbek as an SOV language with considerable pragmatic flexibility.

Yuldasheva (2025) explicitly characterizes Uzbek as predominantly SOV while arguing that its relatively flexible word order is influenced by pragmatic and information-structural considerations. The study associates constituent rearrangement with emphasis, topicality, and communicative context.

Similarly, Narzullaeva and Mukhammadova (2026) discuss Uzbek word order in relation to pragmatic meaning, theme–rheme organization, and information structure. Their comparison of French and Uzbek emphasizes that Uzbek permits greater word-order variation and that constituent arrangement can participate in the organization of discourse information.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

The corpus findings therefore support and extend these observations. Previous studies have established that Uzbek is SOV but flexible; the present computational approach would provide a quantitative indication of how that flexibility is distributed in authentic news discourse.

### 4.5 Verb Position as a Major Comparative Parameter

One of the strongest predicted contrasts concerns verb position. Indonesian is expected to show predominantly medial verbs, particularly in SVO clauses, whereas Uzbek is expected to show predominantly final verbs.

This finding provides a more precise interpretation of the SVO/SOV distinction. Instead of treating SVO and SOV as merely labels, the analysis identifies verb position relative to the object as a measurable syntactic parameter.

The contrast is especially important because recent computational work on Uzbek has explicitly treated word order as a problem for automatic syntactic analysis. Hamroyeva et al. (2023), for example, developed an algorithmic approach to Uzbek word order and syntactic parsing and emphasized the importance of identifying constituent positions in computational analysis.

The present study therefore potentially complements computational Uzbek research in a different direction. Whereas Hamroyeva et al. focus on developing an algorithm for syntactic analysis, the present study applies computational processing to a comparative corpus question involving Indonesian and Uzbek. The contribution is consequently not merely the identification of Uzbek SOV order but its quantitative comparison with Indonesian SVO order within a balanced bilingual dataset.

### 4.6 Morphology and Word-Order Flexibility

The expected differences may also be interpreted in relation to the morphological profiles of the two languages. Uzbek possesses an extensive system of nominal case marking, while Indonesian relies less heavily on nominal inflection for the expression of grammatical relations. This difference provides one possible explanation for why Uzbek can permit constituent rearrangement without necessarily creating the same degree of ambiguity that would arise in a language relying more strongly on linear position.

The recent literature on Uzbek syntax similarly associates constituent-order flexibility with morphological marking. Yuldasheva (2025) argues that case suffixes contribute to the identification of grammatical roles and consequently permit constituents to be rearranged for pragmatic purposes.

The findings are therefore compatible with an analysis in which word order and morphological marking interact. However, the present corpus alone would not establish a causal relationship between case marking and word-order flexibility. Such a conclusion would require explicit statistical coding of morphological case and constituent position. Therefore, the relationship should be presented as an interpretive hypothesis rather than as a causal finding.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

### 4.7 The Importance of Corpus-Based Quantification

A particularly important contribution of the study lies in its methodological approach. Traditional descriptions can tell us that Indonesian is SVO and Uzbek is SOV, but they do not necessarily tell us how frequently alternative orders occur in comparable contemporary discourse.

The corpus-based approach addresses this issue directly. Yan and Liu (2023) demonstrate that corpus frequencies can be used to derive dominant word-order patterns and that corpus-based typological conclusions can complement established typological databases.

The results therefore support a shift from:

Indonesian = SVO
Uzbek = SOV

toward:

Indonesian = predominantly SVO, with measurable alternative orders

and:

Uzbek = predominantly SOV, with measurable alternative orders.

This formulation is theoretically more informative because it recognizes both categorical typological tendencies and quantitative variation.

### 4.8 The Contribution of the Python-Based Analysis

Another contribution concerns the use of computational tools developed specifically for the study. Alves et al. (2023) show that corpus-based computational methods can be used to derive word-order typological information from syntactically annotated data.

The present study extends this general approach by developing a researcher-controlled Python workflow that integrates web scraping, corpus construction, concordancing, syntactic classification, frequency calculation, and visualization. This is particularly valuable for Indonesian–Uzbek comparison because the two languages do not have equally extensive computational resources.

The tool should not, however, be presented as replacing linguistic analysis. Automatic extraction identifies candidate structures, whereas the researcher determines whether a particular sentence genuinely instantiates SVO, SOV, OSV, or another configuration. This combination of computational processing and manual verification is therefore a methodological strength.

### 4.9 Confirmation, Refutation, or Expansion?

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

Taken together, the findings can be characterized as follows:

| Finding | Relationship to previous research |
|---|---|
| Indonesian predominantly SVO | Confirms established typological descriptions |
| Uzbek predominantly SOV | Confirms recent Uzbek syntactic descriptions |
| Alternative orders in Indonesian | Expands/complements the SVO characterization |
| Alternative orders in Uzbek | Expands/complements the SOV characterization |
| Word-order variation associated with discourse factors | Supports/complements recent information-structural research |
| Quantitative comparison of Indonesian and Uzbek | Extends predominantly language-specific research |
| Python-based corpus quantification | Extends/complements computational word-order research |
| Analysis of multiple syntactic parameters | Expands the traditional SVO/SOV comparison |

Thus, the findings primarily confirm and expand previous research rather than refute it. The dominant SVO pattern predicted for Indonesian and the dominant SOV pattern predicted for Uzbek are consistent with the established typological descriptions. The novel contribution lies in demonstrating that these dominant patterns coexist with systematic variation and in quantifying that variation within comparable news corpora.

The study consequently supports a view of constituent order as probabilistic and discourse-sensitive rather than exclusively categorical. This interpretation is consistent with contemporary corpus-based approaches to word-order typology, which increasingly treat constituent order as a measurable distribution rather than a single categorical property.

### 4.10 Implications for Comparative Syntax

The findings have implications for the comparative study of Indonesian and Uzbek. A simple contrastive statement such as “Indonesian is SVO whereas Uzbek is SOV” captures the most obvious difference but leaves unexplained a substantial part of the observed data. The results suggest that a fuller comparison requires at least three analytical levels:

1. **Canonical order** — the dominant SVO or SOV configuration;
2. **Distributional variation** — the frequencies of alternative constituent orders; and
3. **Functional explanation** — the syntactic, morphological, and discourse factors associated with those alternatives.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

This three-level approach provides a more nuanced framework for comparative syntax. It also creates an opportunity for future research to examine whether particular alternative orders correlate with specific discourse functions, grammatical constructions, or morphological markers.

### 4.11 Overall Interpretation

The findings therefore indicate that the Indonesian–Uzbek contrast is best understood not as a rigid opposition between SVO and SOV but as a difference in the probability and functional distribution of constituent orders. Indonesian is expected to have a strong SVO tendency, while Uzbek is expected to have a strong SOV tendency. Both languages, however, permit alternative configurations.

This interpretation confirms the core findings of previous typological and language-specific studies while extending them through quantitative, corpus-based evidence. The main contribution of the present study would consequently be empirical rather than a proposal that the established SVO/SOV classifications are wrong. Its contribution would be to demonstrate, using a balanced contemporary news corpus and computational analysis, how strongly each canonical order dominates, what alternative orders occur, and which syntactic parameters characterize the contrast between the two languages.

Conclusion

This study examined syntactic word order and selected typological parameters in Indonesian and Uzbek news discourse using a computational corpus-based approach. Based on the hypothetical findings, the analysis indicates a clear contrast between the dominant constituent orders of the two languages. Indonesian predominantly exhibits Subject–Verb–Object (SVO) order, whereas Uzbek predominantly exhibits Subject–Object–Verb (SOV) order. These patterns are consistent with the established typological characterization of Indonesian as predominantly SVO and Uzbek as predominantly SOV. The findings therefore generally confirm, rather than refute, previous descriptions of the two languages.

At the same time, the occurrence of alternative constituent orders demonstrates that SVO and SOV should not be interpreted as rigid rules governing all clauses. Indonesian shows alternative patterns such as OSV and OVS, while Uzbek also permits a range of non-canonical configurations. Such variation suggests that basic word order is better understood as a dominant tendency within a broader system of syntactic and discourse-based variation. In particular, topicalization, information structure, voice, argument structure, and constituent displacement provide important contexts for departures from the canonical order. Thus, the findings expand and complement previous typological accounts by demonstrating how dominant word order coexists with substantial variation in naturally occurring news discourse.

The comparison also indicates that verb position constitutes an important parameter distinguishing the two languages. Indonesian generally places the verb before the object, whereas Uzbek predominantly places the verb after the object. This contrast provides a particularly clear manifestation of the SVO–

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

SOV distinction and illustrates the usefulness of examining multiple syntactic parameters rather than relying exclusively on a single basic-word-order label. The findings consequently support a multidimensional approach to word-order typology in which constituent order is examined together with voice, topicalization, argument realization, and other structural properties.

Methodologically, the study demonstrates the potential of computational corpus analysis for cross-linguistic syntactic research. The use of a self-developed Python web-scraping and concordance tool makes it possible to construct comparable corpora, retrieve and clean authentic news texts, identify concordance patterns, quantify constituent-order frequencies, and organize the results systematically. This computational approach complements manual syntactic analysis by providing a reproducible procedure for examining large numbers of naturally occurring examples. The use of a common descriptive framework also facilitates comparison between Indonesian and Uzbek despite their substantial structural differences.

Overall, the hypothetical findings primarily confirm and expand existing research. They confirm the dominant SVO pattern of Indonesian and SOV pattern of Uzbek, while expanding previous descriptions by quantitatively documenting alternative word-order patterns in comparable news corpora. More broadly, the study supports the view that basic word order is a useful typological generalization but should not be treated as an exhaustive description of syntactic organization. Future research could strengthen these findings by expanding the corpus beyond 10,000 words per language, incorporating additional genres, applying dependency parsing and Universal Dependencies more systematically, and examining the relationship between word order and information-structural variables such as topic, focus, givenness, and discourse prominence.

## References

1. Alves, D., Bekavac, B., Zeman, D., & Tadić, M. (2023). Analysis of corpus-based word-order typological methods. *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, 36–46.
2. Buljan, M. (2023). What quantifying word order freedom can tell us about dependency corpora. *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023)*, 54–67.
3. Chung, S. (2008). Indonesian clause structure from an Austronesian perspective. *Lingua, 118*(10), 1554–1582.
4. Clark, T. H., Meister, C., Pimentel, T., Hahn, M., & Cotterell, R. (2023). A cross-linguistic pressure for uniform information density in word order. *Transactions of the Association for Computational Linguistics, 11*, 1048–1065. https://doi.org/10.1162/tacl_a_00589
5. Hamroyeva, S. M., Abdullayeva, O. X., Rakhmonova, M. A., & Nizomova, F. B. (2023). Algorithm of word order and syntactic analysis in Uzbek language sentences. *Central Asian Journal of Literature, Philosophy and Culture, 4*(6). https://doi.org/10.51699/cajlpc.v4i6.900
6. Lian, Y., Bisazza, A., & Verhoef, T. (2023). Communication drives the emergence of language universals in neural agents: Evidence from the word-order/case-marking trade-off. *Transactions*

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

*of the Association for Computational Linguistics, 11*, 1033–1047. https://doi.org/10.1162/tacl_a_00587

7. Narzullaeva, D. B., & Mukhammadova, R. J. Q. (2026). Pragmatic features of word order in French and Uzbek. *Central Asian Journal of Literature, Philosophy and Culture, 7*(1), 233–237. https://doi.org/10.51699/cajlpc.v7i1.1424
8. Sato, Y. (2024). Partial wh-movement in Indonesian, criterial freezing, and sub-extraction. *Canadian Journal of Linguistics/Revue canadienne de linguistique, 69*(1). https://doi.org/10.1017/cnj.2024.8
9. Shcherbakova, O., Blasi, D. E., Gast, V., Skirgård, H., Gray, R. D., & Greenhill, S. J. (2024). The evolutionary dynamics of how languages signal who does what to whom. *Scientific Reports, 14*, 7259. https://doi.org/10.1038/s41598-024-51542-5
10. Slioussar, N., & Harchevnik, M. (2024). Word order and context in sentence processing: Evidence from L1 and L2 Russian. *Frontiers in Psychology, 15*, Article 1344366. https://doi.org/10.3389/fpsyg.2024.1344366
11. Svenonius, P. (2025). Word order universals and their relationship to structure. *Annual Review of Linguistics, 11*, 137–162. https://doi.org/10.1146/annurev-linguistics-011724-121650
12. Tamaoka, K., Yu, S., Zhang, J., Otsuka, Y., Lim, H., Koizumi, M., & Verdonschot, R. G. (2024). Syntactic structures in motion: Investigating word order variations in verb-final (Korean) and verb-initial (Tongan) languages. *Frontiers in Psychology, 15*, Article 1360191. https://doi.org/10.3389/fpsyg.2024.1360191
13. Yan, J., & Liu, H. (2023). Basic word order typology revisited: A crosslinguistic quantitative study based on UD and WALS. *Linguistics Vanguard*. https://doi.org/10.1515/lingvan-2021-0001
14. Yuldasheva, D. B. (2025). Specific features of sentence structure in Uzbek syntax. *Science Shine, 1*(13), 46. https://doi.org/10.71337/inlibrary.uz.science-shine.125780
15. Yuwana, R. Y., & Leba, S. M. R. (2026). Urutan kata, fokus, dan prosodi: Studi lintas bahasa pada antarmuka sintaksis–prosodi bahasa Indonesia dan Inggris. *Bhinneka: Jurnal Bintang Pendidikan dan Bahasa, 4*(1). https://doi.org/10.59024/bhinneka.v4i1.1680