# A flexible and transparent feature processing pipeline for multivariate analysis of texts in context

**Florian Frenken**
RWTH Aachen University
florian.frenken@ia.rwth-aachen.de

**Tatiana Serbina**
Saint Louis University Madrid
tatiana.serbina@slu.edu

**Stella Neumann**
RWTH Aachen University
stella.neumann@ia.rwth-aachen.de

**Stephanie Evert**
FAU Erlangen-Nürnberg
stephanie.evert@fau.de

## Abstract

This paper presents a feature processing pipeline employed as the basis for multivariate linguistic analyses. At the core of this pipeline is a feature extraction script using sophisticated corpus queries geared towards broad coverage of linguistic features that reflect three dimensions of situational context. Queries are formulated in a linguistically transparent way and represented in a markdown document that allows the resesarcher to adapt them to their requirements. An analysis of precision and recall for individual features shows the overall good quality of feature extraction. We argue that the pipeline can be employed for various types of multivariate studies of texts in context.

## 1 Introduction

Corpora can provide solid empirical evidence for a variety of linguistic research questions, especially in large-scale quantitative analyses. The number and choice of features included in such studies is a crucial consideration for the interpretability of subsequent findings. As data size increases, improving the efficiency of extraction while maintaining its reliability and informativeness becomes essential to avoid errors introduced in automatic annotation that can compromise all later analyses and to obtain rich features comparable to those obtained from more costly manual methods. Script-based extraction using complex yet well-defined queries promises to strike a balance between informativeness and efficiency, while often yielding more consistent identification compared to a manual approach.

The goal of the present paper is to introduce and evaluate feature extraction for multivariate analyses of linguistic variation. One potential application of the feature processing pipeline is analysis of linguistic variation according to situational context, i.e. register variation (Biber, 1995). At the same time, the presented approach is also of interest for various other types of linguistic variation. The defining qualities of the introduced feature extraction approach are its linguistic motivation, transparency and flexibility. As Biber (2012) demonstrates, register represents a pervasive factor in language that influences the realization of all kinds of linguistic features, even in studies of other sources of variation (e.g. Kruger and van Rooy, 2018). Deriving features from comprehensive register constructs is, therefore, decisive to ensure broad applicability of the pipeline. The features included in our script are operationalizations of indicators for the three major dimensions of register variation – field, tenor, and mode –, which comprehensively capture the typical domains, interpersonal roles and the way language is organized in types of situations (Halliday and Hasan, 1989, 12). These largely represent queryable versions of operationalizations presented in Neumann (2014). Since such queries are necessarily approximations rather than one-to-one correspondencies to abstract features, one of the goals of the present paper is to evaluate the feature extraction script.

Our approach pursues a different goal from those often used in NLP that are more specialized (e.g. Lee and Lee, 2023; Maurer, 2026). In contrast to NLP features that often rely on shallow operationalizations of task-specific text properties, e.g. for the detection of translationese (Volansky et al., 2015), the proposed pipeline is designed as a task-neutral set of linguistically meaningful features that is applicable to different multivariate linguistic studies. Additionally, a maximalist approach that includes collinear features risks overemphasizing the effects of certain patterns. Our theory-guided approach avoids this by leaving out one option from all features with mutually exclusive choices. As such, taking mood as an example, we only cover interrogatives and imperatives, ignoring declaratives as the default variant. In terms of other broad coverage tools, the tagger used in Biber's (1988) seminal multidimensional analysis has been quite popular.

Unfortunately, it is not publicly accessible but has been recreated in various forms, most recently as the Multi-Feature Tagger of English (MFTE) by Le Foll and Shakir (2024), which also significantly improves and expands on the original feature set. A general problem with these approaches is that rule-based extraction employing regular expressions is rather brittle and difficult to read/modify compared to a pipeline using corpus queries (cf. Section 3).

Section 2 introduces the pipeline and its components. The feature extraction script is then described in Section 3, while Section 4 reports on the evaluation method. Section 5 presents results, comparing our performance with the MFTE tagger where applicable. This is followed by a discussion reflecting on its broader relevance for multivariate linguistic analyses in Section 6.

## 2 The feature processing pipeline

The input for the pipeline is a part-of-speech tagged corpus. For English, the feature extraction requires annotations using the CLAWS 7 tagset (Garside and Smith, 1997) to take advantage of its more granular information compared to other taggers. Then, the corpus needs to be indexed in the IMS Open Corpus Workbench (CWB) (Evert and Hardie, 2011). The language used by its corpus query processor (CQP) enables highly specific and flexible searches that intuitively read as linguistic descriptions due to being explicitly designed for this purpose. Drawing on CQP's "macro substitution facility and support for matching against word lists" (Neumann and Evert, 2021), more complex queries can be easily subdivided into smaller, transparent units. The feature extraction script is represented as a markdown notebook that includes documentation of each operationalization, allowing users to interactively inspect the query code and adapt it to the needs of the respective study.

Compared to the MFTE (Le Foll and Shakir, 2024), which uses Python to run regular expressions on plain-text corpora annotated with a fixed tagset, CWB employs its own UTF-8 compatible tabular data model that supports diverse annotations and speeds up queries with the help of indexed look-ups as opposed to performing repeated slow linear searches. The system is designed to handle large corpora of up to 2.1 billion tokens and further benefits in terms of speed from mostly being written in optimized C code (Evert and Hardie, 2011, 3-4). While the encoding of the corpus incurs an additional upfront cost, the required input format – verticalized text (one token per line) – is readily produced from XML.

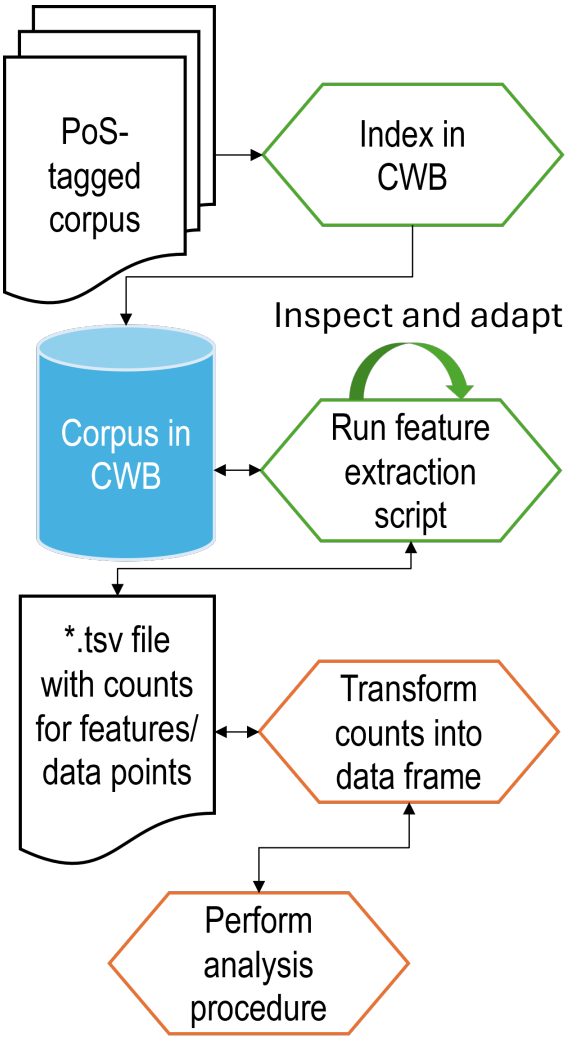


Figure 1: Graphical representation of the feature processing pipeline. Green components indicate parts of the feature extraction with the CWB, whereas orange represents subsequent post-processing and analysis steps.

Complete reproducibility is ensured by full automation of the feature extraction as a pipeline of interconnected components, with dependencies indicated by bidirectional arrows (see Figure 1). However, individual components do not automatically trigger subsequent steps. Reflecting its linguistically-motivated, broad, and flexible design, the pipeline thus facilitates intermediate inspections of the data and potential adaptations. After encoding the corpus, a helper program communicates with the CQP child process via standard input/output streams and collects instances per unit of analysis, which can be selected from any structural attribute present in the corpus (e.g., texts or

sentences), into a table as a .tsv file with one row for each item and one column for each feature. A separate script then aggregates features where desired (such as all future constructions) and, depending on the specifics of the study, transforms counts into a data frame including relative frequencies that serve as input for the linguistic analysis. If multiple corpora are used, the script further merges the results into one common .tsv file.

## 3 Feature extraction script

The Corpus Query Language supports queries for lexico-grammatical patterns beyond string searches with regular expressions. While some queries simply target individual tags such as prepositions, specific adverbs or pronouns, others rely on more complex sequences of PoS and lexical constraints or word lists, which are obtained from the relevant literature (e.g. Halliday and Matthiessen, 2014, Table 3-4). For example, the query below finds instances of *going to* future by searching for a form of the verb *be* (note that the CLAWS tagger also labels contracted forms of this feature such as *gonna*), followed optionally by a negator, zero or more adverbs, and then the expression *going to* plus the infinitive of a verb within one sentence:

```
[pos="VB.+"] [pos="XX"]? [pos="RR.*"]*
    [word="going" | pos="VVGK"] [pos="TO"]
[pos="V.I"] within s;
```

The CQP script automatically executes each query and writes the output in the defined form (individual hit or count, see Section 2) in a table. It comprises 81 linguistic features, including word and sentence counts. This number is somewhat inflated by the fact that larger feature categories (e.g. pronouns) are captured by multiple queries for each type so that they can be aggregated or (partially) left out depending on the needs of the study. Thanks to its notebook format, more sophisticated queries are documented with explanations. The script also comes with several useful macros, such as set operations, which help avoid duplicate counts. The word lists used by some queries are fully editable as regular text files.

All queries already underwent an interactive process of refinement, but their output should still be reviewed and the script, if necessary, expanded or adapted to the specifics of each new corpus when using it. However, this type of testing cannot substitute a more comprehensive evaluation of the output. Evaluation measures allow researchers to determine the quality of the retrieved data and the error analysis facilitates further refinement of queries. Since our script concentrates on interpretable linguistic features, it is necessarily language-dependent. This paper reports on the evaluation of feature extraction for English.

## 4 Evaluation method

The evaluation is based on 9 components from the International Corpus of English (ICE), covering the following varieties: Canada, Great Britain, Hong Kong, India, Ireland, Jamaica, New Zealand, The Philippines, and Singapore (Lehmann and Schneider, 2012). We have extracted a random subset of 576 sentences (2 sentences from each of the 32 text categories for all 9 components) using rejection sampling, comprising 30,099 tokens (27,222 words) that roughly equal 0.3% of the 10,095,536 tokens (9,276,664 words) in the corpus. The sentences therefore cover a small but varied sample that accounts for potential differences in accuracy depending on the variety and text type.

This study does not evaluate queries targeting individual PoS tags (e.g. common nouns) as their accuracy depends entirely on the tagger and not our operationalizations. Garside and Smith (1997) report a tagging accuracy of 97% for CLAWS on the British National Corpus. This may not hold true for other varieties of English, but this is a separate task not addressed here. Instead, the present study concentrates on 22 more complex features (see Table 1), such as passives or imperatives.

To create a gold standard for comparison, the sentences were uploaded to *Segmatator* (Weber and Heinrich, 2026), a tool for collaborative annotation. Each sentence was annotated separately by two student assistants trained on annotation guidelines based on grammatical definitions in Quirk et al. (1985), Martin et al. (1997), and Thompson (2014). Cases of disagreement were resolved through adjudication by a third annotator. Inter-annotator agreement was calculated using the Jaccard coefficient, defined as the size of the intersection of all annotated spans divided by the size of their union (see Appendix A). We do not use Cohen's $\kappa$ because chance agreement for token-level annotation would be too low to give meaningful results.

The evaluation metrics used to measure the accuracy of the CQP feature extraction script are precision and recall (Manning and Schütze, 1999), calculated based on the proportion of cases correctly extracted from the corpus (true positives),

falsely extracted (false positives), or failed to be extracted (false negatives). Additionally, we calculate the support (true positives + false negatives) alongside binominal 95% confidence intervals to provide a range of plausible values for the true probability of success in the feature classification process given the observed data, despite different distributions of features in the sample. Due to the linguistic research interest behind multivariate analyses, maximizing precision was a guiding principle for developing the feature extraction script, even if this meant compromising on recall. For all metrics, any positional overlap between two annotations with the same label was considered to match. The calculations were performed in Python.

## 5 Results

Overall, our feature extraction script achieves good results, with a precision between 0.85 and 0.95 for most features (see Table 2). In line with our goal of prioritizing precision, recall is often about 0.1 points lower. The accuracy scores are generally in line with inter-annotator agreement (see Table 1), indicating that many errors result from genuine ambiguity and not only shortcomings of the automatic extraction. Considering that the evaluation focuses on features that are more difficult to identify, these values therefore indicate decent agreement between the script and gold standard despite this complexity. In the following, we will assess the quality of individual queries by discussing their accuracy in terms of common error patterns.

Starting with future references, which have perfect precision (95% CI $[0.89, 1.00]$), false negatives are cases of present tense and *would* used in the past referring to the future (e.g. *their lives would never be forgotten*). These require contextual knowledge, so the recall of 0.82 ($[0.67, 0.93]$) likely cannot be improved much. Le Foll and Shakir (2024) only include *will* and *going to* futures, thus reporting better recall but slightly worse precision.

Attributive adjectives have only moderate precision and recall (both 0.87, $[0.84, 0.89]$). False negatives can mainly be attributed to tagging errors (e.g. *public opinion* being tagged as two nouns) and fixed expressions like *New Zealand*, which were considered adjective-noun combinations by CLAWS. Since predicative adjectives are defined by exclusion of attributive adjectives, they naturally perform worse, achieving a precision of only 0.63 ($[0.54, 0.72]$). By comparison, the MFTE reaches 0.88 for predicative and 0.96 for attributive adjectives using the same strategy. Partially due to these same issues, passives also do not quite reach satisfactory accuracy rates (P = 0.80, $[0.72, 0.87]$, R = 0.71, $[0.62, 0.79]$) as verbs are often confused for adjectives or nouns. The comparatively low annotator agreement (0.74) suggests that there are indeed many ambiguous cases of this. Le Foll and Shakir (2024), in contrast, reach a precision of 0.93 and 1 for normal and *get* passives respectively. However, their operationalization does not include reduced relative clauses as in *evidence put before you*.

One particular advantage of our script is that it also accounts for theme, covering different patterns of discourse organization (Halliday and Matthiessen, 2014, 89), using different types of sentence-initial elements as a proxy. This is necessarily reductive because theme is a clause-level feature, but the script lacks dependency information. Themes reflecting the organization of the text such as conjunctions (P = 0.95, $[0.87, 0.99]$, R = 0.96, $[0.88, 0.99]$) and conjunctive adjuncts (both 0.88, $[0.62, 0.98]$) perform well, with false negatives largely attributable to gaps in the word lists. This is because textual themes tend to be less ambiguous in their forms. Themes linked to the interpersonal relationship between interactants yield slightly lower accuracy. For example, the script generally fails to differentiate *now* as a continuative from the adverb of time (P = 0.80, $[0.65, 0.93]$, R = 0.75, $[0.58, 0.88]$). Similarly, initial modal adjuncts (0.73, $[0.39, 0.94]$ and finite verbs (0.85, $[0.69, 0.95]$) have only moderate precision, although the latter is mostly due to tagging errors.

Similar discrepancies in accuracy emerge for the queries of process types, five theoretically motivated semantic verb classes (Halliday and Matthiessen, 2014, 213–214), an important indicator of differences in field of discourse. MFTE uses a larger set of verb classes, whose organization is not theoretically motivated and which are not evaluated. Like MFTE, most of our operationalizations of these verb classes rely on word lists. The precision of verbal (0.69, $[0.56, 0.79]$), material (0.72, $[0.66, 0.78]$), and mental (0.82, $[0.73, 0.89]$) processes is among the lowest ones of all features. Existential processes perform much better (0.91, $[0.72, 0.99]$) because the existential *there* serves as a reliable constraint. The same would be true for relational processes, which are typically realized by *be* or *have*; the issue lies in including those that are

not and excluding cases where *be* and *have* are not the main verb (further complicated by the predicative adjectives above) or express another process. That's why relational processes are roughly on par with material processes in terms of precision but have better recall (0.70, $[0.65, 0.76]$).

Neo-classical compounds, i.e. words that contain an element of Greek or Latin origin, have very low precision (0.18, $[0, 28, 0, 71]$) but great recall (0.94, $[0.71, 1.00]$). Looking at the false positives, we find many instances that should have been annotated in the gold standard (e.g. *ultraviolet*) but also overgeneralizations (e.g. *subtleties* due to the prefix *sub-*). That annotators overlooked obvious instances of such compounds suggests that the neoclassical elements are cognitively not salient and that the compounds are perceived as lexicalized wholes. This illustrates the trade-off in using an automatic matching strategy, which helps avoid human error (also indicated by the relatively low annotator agreement of 0.11), at the cost of flexibility. While the negative list can be continuously expanded to improve precision, it is impossible to account for all exceptions. A similar yet not as pronounced pattern emerges for nominalizations.

Regarding mood, imperatives have better precision (0.91, $[0.72, 0.99]$) than interrogatives (0.85, $[0.62, 0.97]$) but worse recall (0.57, $[0.39, 0.73]$ vs. 0.61, $[0.49, 0.87]$). This is different from the MFTE where interrogatives (reported separately for *wh-* and *yes/no*-questions) are more accurate. Overall, the approaches are fairly similar, with our script showing better performance for imperatives and the MFTE for interrogatives. It should be noted, however, that we cannot rely on punctuation in the spoken parts of our sample for the latter, unlike Le Foll and Shakir (2024). The most common error type for both mood types are cases preceded by other material, e.g. *Either that or be repared...*, which is difficult to account for without parsing.

## 6 Discussion

We have presented in this paper a theoretically informed pipeline for extracting a set of meaningful linguistic features that captures three dimensions of contextual variation. Its usefulness lies in its interpretability and flexibility. The feature extraction script transparently documents the search strategy and, along with the aggregation script, can be adapted to the researcher's specific requirements with respect to the grammatical characterization and dedicated wordlists. Its quality obviously hinges on the accuracy of the queries. As shown in the previous section, the overall accuracy is at least satisfactory or even good. The results generally achieve high precision, with recall at somewhat lower levels in line with our priorization for the sake of linguistic rigor. In many cases, the accuracy is similar to that of the MFTE, while offering additional features that round off the range of contextual variation. Interestingly, in some cases, automatic feature extraction outperforms the consistency of human annotation. Together, this suggests that our feature extraction pipeline offers a meaningful addition to the toolbox for multivariate quantitative studies.

The feature extraction pipeline is openly accessible via Github[1]. Thanks to its comprehensive coverage of meaningful lexico-grammatical features, it can be employed for a wide range of quantitative studies in NLP and many areas of linguistics such as learner corpus research and studies of translationese requiring the quantification of a comprehensive, rather than specialized feature set. Crucially, it allows the researcher to inspect its accuracy on the chosen dataset and adapt it to the specific requirements with each study, which is made easy thanks to the well-documented notebook format and rigorous theoretical grounding of the features.

## Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 528467412

## References

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

Douglas Biber. 1995. *Dimensions of register variation*. Cambridge University Press, Cambridge.

Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, pages 1–21. Birmingham: University of Birmingham.

Roger Garside and Nicholas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In Roger Garside,

[1] https://github.com/quantor-project/featex

Geoffrey Leech, and Anthony McEnery, editors, *Corpus Annotation: Linguistic information from computer text corpora*, pages 102—-121. Longman, London.

M.A.K. Halliday and Christian M.I.M. Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*. Routledge, London.

Michael A. K. Halliday and Ruqaiya Hasan. 1989. *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford University Press, Oxford.

Haidee Kruger and Bertus van Rooy. 2018. Register variation in written contact varieties of English: A multidimensional analysis. *English World-Wide: A Journal of Varieties of English*, 39.

Elen Le Foll and Muhammad Shakir. 2024. The Multi-Feature Tagger of English (MFTE): Rationale, description and evaluation. *Research in Corpus Linguistics*, 13(2):63–93.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. ACL.

Hans Martin Lehmann and Gerold Schneider. 2012. BNC Dependency Bank 1.0. In Signe Oksefjell, Jarle Ebeling, and Hilde Hasselgard, editors, *Aspects of Corpus Linguistics: Compilation, Annotation, Analysis*. Research Unit for Variation, Contacts, and Change in English, Helsinki.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge.

J. R. Martin, Christian M. I. M. Matthiessen, and Clare Painter. 1997. *Working with functional grammar*. Arnold, London.

Maximilian Maurer. 2026. elfen: A python package for efficient linguistic feature extraction for natural language datasets. In *Proceedings of the 19th Conference of the EACL Conference (Vol. 3: System Demonstrations)*, pages 61–74, Rabat, Marocco. ACL.

Stella Neumann. 2014. *Contrastive register variation: A quantitative approach to the comparison of English and German*. de Gruyter, Berlin.

Stella Neumann and Stefan Evert. 2021. A register variation perspective on varieties of English. In Elena Seoane and Douglas Biber, editors, *Corpus based approaches to register variation*. de Gruyter, Berlin.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Geoff Thompson. 2014. *Introducing Functional Grammar*. Routledge.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30:98–118.

Timm Weber and Philipp Heinrich. 2026. Segmatator. `https://github.com/fau-klue/segmatator`.

## A Inter-annotator agreement

Table 1 shows inter-annotator agreement for each feature (Jaccard coefficient).

## B Detailed evaluation results

Table 2 shows detailed evaluation results for the feature extraction pipeline. Precision and recall are computed for each query together with binomial confidence intervals at 95%.

| Feature | Jaccard Index |
|---|---|
| Greetings | 1.00 |
| InitialContinuatives | 0.89 |
| FutureReferences | 0.85 |
| AttributiveAdjectives | 0.85 |
| MaterialProcesses | 0.83 |
| InitialConjunctions | 0.82 |
| ExistentialProcesses | 0.76 |
| Passives | 0.74 |
| RelationalProcesses | 0.72 |
| InitialConjunctiveAdjuncts | 0.68 |
| Imperatives | 0.63 |
| PredicativeAdjectives | 0.61 |
| MentalProcesses | 0.61 |
| Nominalizations | 0.60 |
| VerbalProcesses | 0.58 |
| InitialFiniteVerbs | 0.54 |
| InitialModalAdjuncts | 0.54 |
| Interrogatives | 0.53 |
| Titles | 0.46 |
| InitialNumbers | 0.44 |
| PostpositiveAdjectives | 0.11 |
| NeoclassicalCompounds | 0.11 |

Table 1: Inter-annotator agreement for all features (Jaccard index)

| | Precision | | | Recall | | | |
|---|---|---|---|---|---|---|---|
| Feature | Estimate | LL | UL | Estimate | LL | UL | Support |
| FutureReferences | 1.00 | 0.89 | 1.00 | 0.82 | 0.67 | 0.93 | 40 |
| Greetings | 1.00 | 0.29 | 1.00 | 0.50 | 0.12 | 0.88 | 6 |
| InitialNumbers | 1.00 | 0.48 | 1.00 | 0.56 | 0.21 | 0.86 | 9 |
| Titles | 0.95 | 0.75 | 1.00 | 0.45 | 0.30 | 0.61 | 42 |
| InitialConjunctions | 0.95 | 0.87 | 0.99 | 0.96 | 0.88 | 0.99 | 73 |
| ExistentialProcesses | 0.91 | 0.72 | 0.99 | 0.91 | 0.72 | 0.99 | 23 |
| Imperatives | 0.91 | 0.72 | 0.99 | 0.57 | 0.39 | 0.73 | 37 |
| AttributiveAdjectives | 0.88 | 0.85 | 0.90 | 0.87 | 0.84 | 0.89 | 590 |
| InitialConjunctiveAdjuncts | 0.88 | 0.62 | 0.98 | 0.88 | 0.62 | 0.98 | 16 |
| InitialFiniteVerbs | 0.85 | 0.69 | 0.95 | 0.85 | 0.69 | 0.95 | 34 |
| Interrogatives | 0.85 | 0.62 | 0.97 | 0.71 | 0.49 | 0.87 | 24 |
| MentalProcesses | 0.82 | 0.73 | 0.89 | 0.68 | 0.58 | 0.76 | 115 |
| InitialContinuatives | 0.82 | 0.65 | 0.93 | 0.75 | 0.58 | 0.88 | 36 |
| Passives | 0.80 | 0.72 | 0.87 | 0.71 | 0.62 | 0.79 | 131 |
| InitialModalAdjuncts | 0.73 | 0.39 | 0.94 | 0.73 | 0.39 | 0.94 | 11 |
| MaterialProcesses | 0.72 | 0.66 | 0.78 | 0.30 | 0.26 | 0.34 | 565 |
| RelationalProcesses | 0.71 | 0.66 | 0.76 | 0.70 | 0.65 | 0.76 | 297 |
| VerbalProcesses | 0.69 | 0.56 | 0.79 | 0.55 | 0.44 | 0.66 | 83 |
| PredicativeAdjectives | 0.63 | 0.54 | 0.72 | 0.68 | 0.59 | 0.76 | 119 |
| Nominalizations | 0.32 | 0.28 | 0.37 | 0.77 | 0.71 | 0.83 | 212 |
| NeoclassicalCompounds | 0.19 | 0.11 | 0.28 | 0.94 | 0.71 | 1.00 | 17 |
| PostpositiveAdjectives | 0.00 | 0.00 | 0.60 | 0.00 | 0.00 | 0.31 | 10 |

Table 2: Precision and recall for all features including lower (LL) and upper limit (UL) of 95% binomial confidence intervals (CI)