Castledown

OPEN ACCESS

Technology in Language Teaching & Learning
ISSN 2652-1687
https://www.castledown.com/journals/tltl/

Technology in Language Teaching & Learning, 8, 104641 (2026)
https://doi.org/10.29140/tltl.2026.104641

Meet MATT: A Human-Auditable Tool for Sentence-Sampled Analysis of English Article Use

ATSUSHI MIZUMOTO[a]
MATT LUCAS[b]

[a]*Kansai University, Japan*
atsushi@mizumot.com

[b]*Kansai University, Japan*
mlucas@kansai-u.ac.jp

Abstract

English articles (*a/an*, *the*, Ø) are central to second-language (L2) accuracy research but remain difficult to analyze because appropriate use depends on semantic, syntactic, and discourse-level interpretation. This paper introduces MATT (Model for Article Tracking & Typology), a web-based tool designed to produce sentence-sampled, human-auditable annotations of article use in learner writing. MATT accepts pasted text or uploaded files, segments the text into sentences, and evaluates one article-sensitive noun phrase per sentence. For each evaluated sentence, it generates a structured record that includes the target noun phrase, observed and expected article types, a broad error class aligned with omission, overuse, and misuse, a suggested minimal correction, a heuristic review-priority score, and a short rationale. Because MATT does not exhaustively annotate all noun phrases, its outputs should be interpreted as a sampled annotation layer rather than a complete corpus annotation. Accordingly, any accuracy values derived from MATT output, including MATT-derived Suppliance in Obligatory Contexts (MATT-SOC) and MATT-derived Target-Like Use (MATT-TLU), should be understood as sampled indicators based on the evaluated noun phrases. The interface provides correction views, summary tables, visualizations, and downloadable CSV files so that researchers can inspect individual decisions, recompute summary values, and document the scope of analysis transparently. The paper explains the tool workflow, interpretation of MATT-derived outputs, and recommended reporting practices for exploratory article-use analysis.

Keywords: English article use; learner writing; language learning technology; automated annotation; human-auditable feedback.

Copyright: *© 2026 Atsushi Mizumoto & Matt Lucas. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.* ***Data Availability Statement:*** *All relevant data are within this paper.*

Introduction

English articles (*a/an*, *the*, Ø) lie at the interface of syntax, semantics, and pragmatics, posing a well-documented challenge for second-language (L2) learners (Crosthwaite, 2016) and constituting one of the most difficult areas of English grammar (Chrabaszcz & Jiang, 2014). Appropriate article choice depends on multiple interacting factors, including noun countability and number (*a raindrop* vs. Ø *raindrops*), definiteness and identifiability (*the clouds in the sky*), genericity (Ø *dogs are loyal*), discourse status (first mention vs. subsequent mention), and, in some acquisition accounts, specificity (a particular item vs. any item). Specificity is relevant to some accounts of L2 article acquisition, but it should not be treated as the sole determinant of article choice. From a research-methods perspective, this interaction among semantic, syntactic, and discourse-level factors makes article use difficult to analyze consistently, because the same surface form may be judged differently depending on contextual interpretation.

These challenges are especially salient for learners whose first language (L1) lacks an article system, such as Japanese, Korean, and Russian (Ionin & Montrul, 2010). In such cases, learners must acquire form–meaning mappings that have no direct L1 equivalent (Shintani et al., 2014). Learners from article-less L1 backgrounds may therefore continue to experience difficulty with English articles even at advanced levels (Umeda et al., 2019). These learning difficulties also create methodological challenges: researchers must identify article-relevant noun phrases, infer the expected article in context, and decide whether a learner's production is target-like, non-target-like, or one of several possible acceptable alternatives.

Manual annotation remains a central method for analyzing article use, but it is labour-intensive and interpretively variable. Human annotators must integrate sentence-internal grammar, discourse context, referent identifiability, and possible alternative corrections. In learner writing, more than one correction may sometimes be acceptable, so disagreement among annotators does not necessarily indicate simple error or noise (Bryant & Ng, 2015). Comprehensive double-coding is also difficult to sustain for larger collections of learner texts. These constraints point to the value of tools that can support preliminary annotation, structured review, and transparent reporting, while still leaving final interpretive responsibility with the researcher.

Existing digital approaches offer useful support for learner-language analysis, but each addresses a somewhat different workflow. General grammar checkers can flag article-related issues, but their decision logic and category labels are not usually available as auditable, exportable research data. Earlier work on automated grammatical-error detection and article-error correction also demonstrates that articles and prepositions have long been treated as important targets for learner-language technologies (Chodorow et al., 2010; Leacock et al., 2010). Grammatical-error correction (GEC) systems can generate corrected versions of learner sentences. In turn, ERRANT, an automatic error-annotation and evaluation toolkit for GEC, can compare original and corrected sentences to extract edit types, including article- and determiner-related edits (Bryant et al., 2017). This makes a GEC-plus-ERRANT workflow an important reference point for article-error analysis. However, such a workflow requires a corrected sentence as input, whether produced by a human annotator or by another correction system. Learner corpora and annotation schemes also provide important resources for studying grammatical development. Large resources such as the Cambridge Learner Corpus (CLC) (Hawkins & Buttery, 2010), the EF-Cambridge Open Language Database (EFCAMDAT) (Alexopoulou et al., 2017), and the NUS Corpus of Learner English (NUCLE) (Dahlmeier et al., 2013) have supported influential work on learner errors, while recent work on error annotation for automated written feedback has proposed taxonomies relevant to article-related feedback and classification (Coyne et al., 2025).

More recently, language educators have begun to examine the roles of generative AI in language teaching and learning more broadly (Stockwell, 2024) and applied-linguistic work has begun to explore generative AI as a structured aid for analyzing learner production (Mizumoto, 2025). These approaches are valuable (Alexopoulou et al., 2017; Bryant et al., 2017; Coyne et al., 2025; Dahlmeier et al., 2013; Hawkins & Buttery, 2010; Mizumoto, 2025), but to our knowledge none of them provides a lightweight, web-based, article-specific annotation layer for new raw learner texts that researchers or teachers can inspect, export, and reanalyze directly.

This gap motivates the development of a tool that starts from raw learner writing, focuses specifically on article use, and presents its decisions in a format that can be checked by humans. Such a tool should not be expected to replace exhaustive corpus annotation, human judgement, or GEC-based workflows. Rather, its value lies in providing a structured first-pass annotation that makes article-related decisions visible and reviewable. This is especially important for article analysis because article choice often depends on discourse assumptions that cannot always be resolved mechanically.

In response to this need, we introduce MATT (Model for Article Tracking & Typology), a web-based tool that produces sentence-sampled, human-auditable annotations of English article use in learner writing. MATT was developed in August, 2025 and is freely available at https://langtech.jp/matt/. MATT evaluates one article-sensitive noun phrase per sentence and records the evaluated noun phrase, observed article type, expected article type, broad error class, suggested minimal correction, heuristic review-priority score, and short rationale. The resulting summary values, including MATT-derived Suppliance in Obligatory Contexts (SOC) and MATT-derived Target-Like Use (TLU), are therefore intended to summarize the evaluated sample, not all obligatory contexts in the text. The present paper pursues three aims. The first aim is to document the design, workflow, and outputs of MATT so that researchers can understand exactly what the tool does and does not annotate. The second aim is to report a preliminary validation of MATT's output against human judgements, guided by the following research question: To what extent do MATT's article-use classifications agree with those of trained human annotators? The third aim is to propose reporting practices that allow MATT-based analyses to be documented transparently and audited by other researchers. Together, these aims position the paper as a tool-focused methodological contribution to exploratory analysis of English article use.

Background

Article Choice and L2 Difficulty

English article choice requires the coordination of morphosyntactic, semantic, and discourse-level information. Countability and number constrain whether *a/an* can occur with a singular countable noun and whether bare plurals or mass nouns are possible; definiteness and discourse identifiability constrain the use of *the*; and genericity affects whether plural and mass nouns appear with Ø or with an overt article (Hawkins, 1978; Snape & Yusa, 2013). Specificity has played an important role in L2 article-acquisition research, particularly in accounts of learners from article-less L1 backgrounds, but it should not be treated as a direct rule for article choice in all contexts (Ionin et al., 2004). Article use is therefore difficult to analyze because the expected form often depends not only on the noun phrase itself but also on discourse assumptions and the intended reference.

For learners whose L1 lacks an article system, English articles require the acquisition of obligatory form–meaning mappings that may have no direct counterpart in the L1. The Fluctuation Hypothesis, for example, proposes that learners from article-less L1 backgrounds may initially fluctuate between definiteness-based and specificity-based interpretations of English articles (Ionin et al., 2004). Such accounts help explain why learners may overuse *the* in indefinite but specific contexts

or omit articles in contexts where English requires an overt determiner. Corpus-based studies have also shown that L1 background affects article accuracy and the acquisition of English grammatical morphemes more generally (Murakami & Alexopoulou, 2016). Recent large-scale work on L2 English article acquisition further suggests that article accuracy is shaped not only by whether articles are available in the L1 but also by broader typological relations between the L1 and English (Öksüz et al., 2026). These findings support the need for article-analysis procedures that are sensitive to form, meaning, and learner background.

Article Error Categories and Interpretive Alternatives

Applied-linguistic studies of article use have often grouped article errors into three broad categories: omission, overuse, and misuse. Omission refers to non-suppliance of an article in an obligatory context, as in **She adopted Ø cat*. Overuse refers to the insertion of an article where no article is expected, as in **I like the music* when a general meaning is intended. Misuse refers to the selection of an inappropriate article form, as in **It was the beautiful day today* where an indefinite article is expected. These broad categories are useful because they distinguish failure to supply an article, unnecessary insertion, and incorrect article choice.

At the same time, article annotation is not always reducible to a single mechanical correction. Learner sentences may permit more than one acceptable correction, especially when discourse context is underspecified or when a sentence can be interpreted under more than one referential assumption. In grammatical-error correction research, this issue is often discussed as the problem of valid alternative corrections rather than simple annotator inconsistency (Bryant & Ng, 2015). For article analysis, this means that an annotation workflow should preserve enough information for human review, including the target noun phrase, expected article type, suggested correction, and rationale.

Article Cues and the Scope of Cue-Level Analysis

A more fine-grained way to understand article difficulty is provided by usage-based accounts of article cues. Zhao and MacWhinney (2018) examined form–function mappings in the English article system and characterized article cues in terms of their availability and reliability. This perspective is useful because article contexts differ in frequency, regularity, and lexical or discourse constraints. High-frequency cues may be learned earlier, whereas less frequent or more context-dependent cues may remain difficult even for advanced learners.

In the present paper, this cue-based account is used as theoretical motivation for why article use is difficult to learn and difficult to annotate. However, the current version of MATT does not output a full cue-level classification, nor does it assign each case to one of the 86 cues or to a structured top-ten cue category. Its structured output is limited to the evaluated noun phrase, observed article type, expected article type, broad error class, suggested correction, review-priority score, and rationale. Therefore, claims about MATT should be limited to broad article-type and error-class annotation unless future versions add explicit cue-level labels.

Sentence Sampling and Corpus Methodology

Sampling has long been recognized as a central methodological issue in corpus-based research. Corpus design involves principled decisions about which texts, text portions, and linguistic units to include, and the representativeness of the resulting sample determines what generalizations the analysis can support (Biber, 1990, 1993). Seen from this perspective, MATT's one-noun-phrase-per-sentence

policy is a form of systematic linguistic sampling: rather than exhaustively annotating every article context, the tool selects one analytically informative context per sentence under a consistent selection policy. The methodological value of such sampling lies in comparability. When the same selection policy is applied across texts, groups, or time points, the resulting sampled indicators can be compared meaningfully, even though they do not estimate the full population of obligatory contexts in each text.

At the same time, this sampling policy is not random, and its selection criteria have directional consequences that researchers should understand. Because the policy prioritizes noun phrases whose article choice is likely to be analytically informative, such as singular countable common nouns and discourse-affected noun phrases, the evaluated sample may overrepresent contexts in which learner errors are more likely and underrepresent unproblematic contexts, which would tend to lower sampled accuracy values relative to exhaustive annotation. In addition, syntactically dense sentences contribute only one evaluated noun phrase regardless of how many article contexts they contain, so writers who produce more complex sentences will have a smaller proportion of their article contexts evaluated. These properties do not invalidate sampled analysis, but they mean that MATT-derived accuracy values should be compared only under the same sampling policy and should not be treated as unbiased estimates of overall article accuracy. These points are taken up again in the Limitations and Future Directions section.

Corpus-Based and GEC-Based Annotation Workflows

Existing resources and tools provide important reference points for article-error analysis. Large learner corpora such as the Cambridge Learner Corpus and EFCAMDAT have supported research on L1 effects, proficiency development, and grammatical accuracy in learner writing (Alexopoulou et al., 2017; Hawkins & Buttery, 2010; Murakami & Alexopoulou, 2016). NUCLE also provides a major annotated corpus of learner English for grammatical-error correction research (Dahlmeier et al., 2013). These resources are valuable for corpus-based research, but their annotations are not designed specifically as a lightweight article-focused review layer for new raw texts submitted by individual researchers or teachers.

A second important workflow comes from grammatical-error correction. GEC systems can generate corrected versions of learner sentences, and ERRANT can then compare original and corrected sentences to extract edits and assign error types, including article- and determiner-related categories (Bryant et al., 2017). This makes ERRANT an important baseline and point of comparison for any new tool that claims to support article-error analysis. However, an ERRANT-based workflow requires a corrected sentence as input. The corrected sentence may come from a human annotator, a GEC system, or a correction procedure based on a large language model (LLM). By contrast, MATT is designed to produce an article-focused annotation table directly from raw learner text, while keeping the individual decision visible for human inspection.

Recent work on automated written feedback also shows the value of annotation schemes that connect learner errors to feedback-relevant categories. Coyne et al. (2025), for example, propose an annotation framework that models error type and generalizability in order to support feedback generation. This work is relevant to MATT not because it demonstrates the absence of error types in commercial grammar checkers, but because it shows how error annotation can be designed around pedagogically meaningful categories. MATT differs in scope by focusing narrowly on English articles and by prioritizing exportable, sentence-level records for researcher review. Table 1 summarizes how MATT is positioned relative to these existing annotation workflows.

Table 1 *Positioning MATT Relative to Existing Annotation Workflows*

Approach	Required input	Main output	Relation to MATT
Manual article annotation	Raw learner text	Exhaustive article judgements and corrections	Most complete option, but time-intensive and interpretively variable
Learner corpora such as the Cambridge Learner Corpus, the EF-Cambridge Open Language Database, and the NUS Corpus of Learner English	Pre-existing corpus data	Corpus annotations, error tags, or learner metadata	Useful reference resources but not a lightweight tool for new raw texts
GEC plus ERRANT	Original and corrected sentence pairs	Edit spans and error types	Strong baseline for edit-based analysis, but requires corrected text
Coyne et al.-style feedback annotation	Learner errors and feedback labels	Error type, generalizability, and feedback categories	Relevant to pedagogical error taxonomies
MATT	Raw learner text	One sentence-sampled article annotation per sentence	Lightweight, article-focused, human-auditable review layer

Quantifying Article Accuracy: SOC, TLU, and MATT-Derived Metrics

Two established measures in morpheme and article-acquisition research are SOC and TLU. SOC captures the proportion of obligatory contexts in which the learner supplies the target form (Dulay & Burt, 1973):

$$SOC = \frac{\text{Number of correct articles in obligatory contexts}}{\text{Number of obligatory contexts}} \times 100\%$$

TLU extends this logic by also penalizing overuse in non-obligatory contexts (Pica, 1983):

$$TLU = \frac{\text{Number of correct articles in obligatory contexts}}{\text{Number of obligatory contexts} + \text{Number of unnecessary articles used in non} - \text{obligatory contexts}} \times 100\%$$

In fully annotated datasets, these measures depend on identifying all relevant obligatory and non-obligatory contexts in the target sample. This point is crucial for the present paper because MATT does not exhaustively annotate all noun phrases. For this reason, the metrics reported by MATT should be labeled MATT-SOC and MATT-TLU. These values summarize article accuracy only over the evaluated noun phrases, not over all article-bearing noun phrases or all obligatory contexts in the text.

MATT-SOC and MATT-TLU should therefore be interpreted as sampled indicators rather than direct equivalents of conventional SOC and TLU; researchers who require exhaustive counts of obligatory contexts should use manual annotation or a GEC-plus-ERRANT workflow.

Implications for the Present Tool

The literature reviewed above suggests that a useful article-analysis tool should make article-related decisions inspectable rather than simply outputting a final correction. It should record what was

evaluated, what article was observed, what article was expected, how the error was broadly classified, and why the decision was made. It should also make clear whether the output is exhaustive or sampled. On this basis, MATT is positioned as a sentence-sampled, human-auditable annotation aid for exploratory article-use analysis. Its contribution lies in providing a structured first-pass review layer that researchers can inspect, export, recompute, and report transparently.

MATT Tool Description and Workflow

Implementation

MATT uses the Groq application programming interface (API) to access an LLM for article annotation. The version of MATT reported in this paper is version 2.1, released in July 2026 and available at https://langtech.jp/matt/. The implementation reported in this paper uses OpenAI's GPT-OSS 120B open-weight model (openai/gpt-oss-120b), accessed through the Groq infrastructure. An earlier version of MATT used Meta's Llama 3.3 70B Versatile model (llama-3.3-70b-versatile); because that model is scheduled for deprecation by the API provider, it was replaced with GPT-OSS 120B, and all validation results reported in this paper were obtained with the current model. The following inference parameters were fixed for the analyses reported here: temperature = 0.0, maximum completion length = 3,000 tokens (a cap that accommodates the internal reasoning tokens consumed by GPT-OSS models), and request timeout = 60 seconds. These settings were chosen to reduce output variability and keep web-based response times manageable. They do not, however, guarantee exact reproducibility over time, because cloud-hosted models and provider-side infrastructure may change.

The annotation prompt consists of three components: a system prompt that defines the annotation task, the classification schema (CORRECT, OMISSION, OVERUSE, MISUSE, UNMATCHED), an explicit decision procedure that applies three ordered rules (predicate-nominal classification, coreferent anaphora, and bridging anaphora), a hard grammatical constraint that blocks the zero article for singular countable common nouns, a selection rule that treats cardinal-number plural noun phrases as evaluable, and the required output format; fourteen few-shot examples illustrating the expected input-output pattern across article-use cases; and a user message containing the target sentence together with preceding discourse context. In addition, an application-level consistency check verifies that the categorical labels in each returned record are mutually consistent and, where necessary, corrects them, flagging each correction in the rationale field so that these automatic adjustments remain auditable. The full system prompt, few-shot examples, and output schema are provided in the online supplementary materials (https://osf.io/6b9hq/). These materials allow researchers to inspect the prompt-level instructions that guide MATT's output, although they should not be interpreted as revealing the internal decision process of the LLM.

Because MATT processes submitted texts through the authors' web server and an external LLM API, users should not upload identifiable or sensitive learner writing unless appropriate consent, institutional approval, and data-handling procedures are in place. For research use, learner texts should be anonymized before submission, and researchers should report that the data were processed through remote services. MATT outputs should also be treated as annotations for researcher review rather than as final judgements for high-stakes assessment, placement, grading, or learner evaluation.

Web Interface and Accepted Inputs

MATT is implemented as a lightweight Flask application with a single-page workflow. Users can paste learner writing into the 'Text to Analyze' field or upload a file through the 'Upload Text File (Optional)' control (Figure 1). The upload function accepts .txt and .docx files. When both pasted text and an uploaded file are provided, the uploaded file is used as the analysis input.

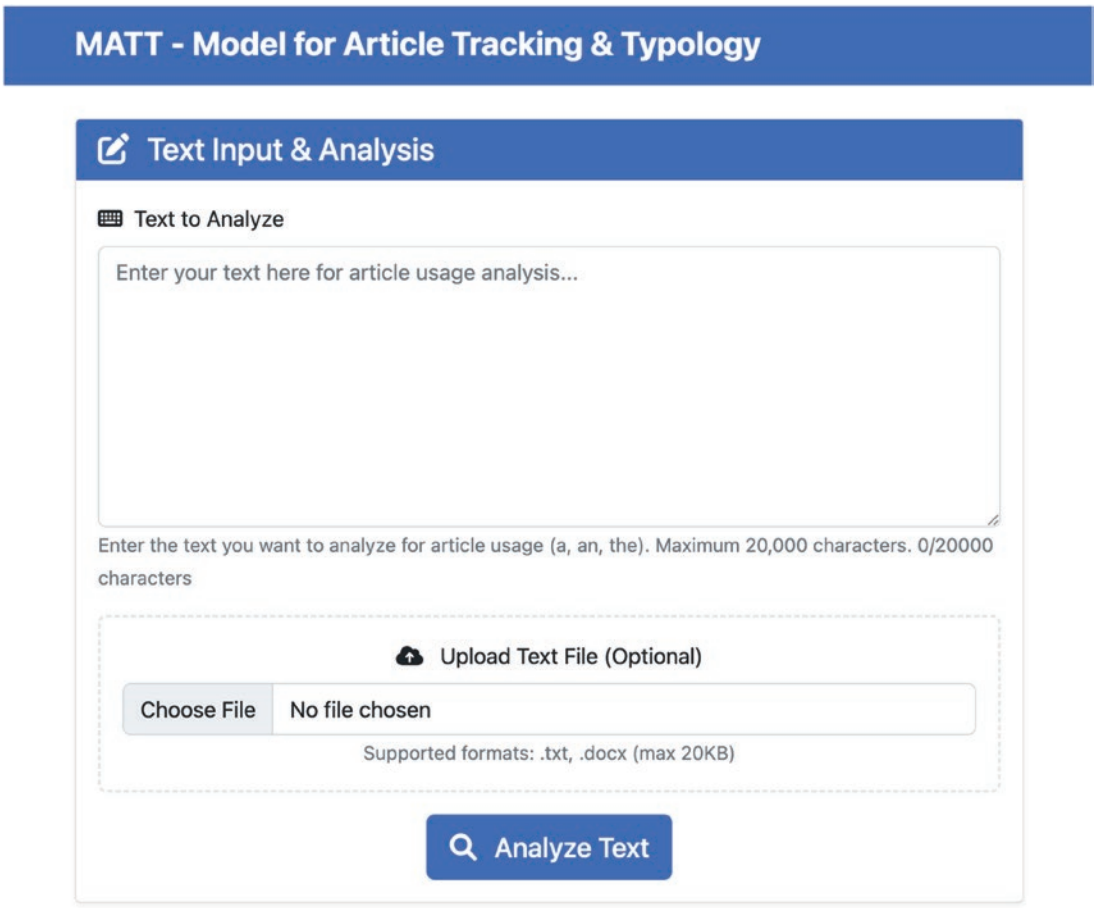

Figure 1 *MATT Landing Page*

The current interface processes one submitted text at a time. For corpus projects, researchers should therefore preserve document-level metadata outside MATT, such as learner ID, document ID, task, date, proficiency level, and L1 background, and add these fields when combining exported files. Researchers should not concatenate multiple learner scripts into a single input unless document-level distinctions are irrelevant to the analysis.

To support stable performance in a web setting, the interface applies an input-length cap and a maximum number of sentences for detailed article annotation in a single run. These limits should be reported when MATT is used for research. In the current implementation, the maximum input length is 25,000 characters, and the maximum number of sentences analyzed in one run is 25 sentences.

Input Processing, Sampling Policy, and Outputs

Before article annotation, MATT applies automatic cleaning to the submitted text. It removes extra spaces, normalizes selected typographic characters such as curly quotation marks, and removes characters outside the standard English alphabet and punctuation. This cleaning procedure helps maintain a consistent input format for typical English learner writing, but it can also alter multilingual text, names with diacritics, symbols, or non-standard punctuation. Researchers should therefore archive the original texts separately and report the preprocessing procedure when using MATT for research.

After cleaning, MATT segments the text into sentences using sentence-ending punctuation, including periods, exclamation marks, and question marks. Very short fragments are excluded from detailed

article annotation. Because learner writing often contains unconventional punctuation, sentence segmentation should be treated as part of the processing pipeline rather than as a perfect representation of the learner's intended sentence boundaries. Researchers should inspect the sentence-level output when sentence boundaries are analytically important.

MATT analyzes article use one sentence at a time. In each sentence, it evaluates one article-sensitive noun phrase. The selection policy prioritizes noun phrases where article choice is likely to be analytically informative, especially singular countable common nouns and noun phrases affected by discourse factors such as anaphoric reference, restrictive modification, or uniqueness. If the model does not identify an article-sensitive noun phrase, the sentence is marked as UNMATCHED.

This one-noun-phrase-per-sentence policy, discussed in the Background as a form of systematic sampling, is a design constraint rather than an exhaustive annotation strategy. It reduces the number of model calls and keeps the output table manageable for human review, but it necessarily omits additional article-sensitive noun phrases in the same sentence. MATT outputs should therefore be interpreted as sentence-sampled annotations. Researchers should report the number of sentences submitted, the number of evaluated noun phrases, and the number and proportion of UNMATCHED cases.

For each evaluated sentence, MATT produces a structured record containing the sentence, the evaluated noun phrase, the observed article type, the expected article type, a broad error class, a suggested minimal correction, a heuristic review-priority score, and a short rationale. The review-priority score should not be interpreted as a calibrated probability of correctness. It is included only to help researchers identify cases that may deserve closer human inspection. Likewise, the rationale field should be read as a model-generated explanation for review, not as direct evidence of the model's internal reasoning.

From these sentence-level records, MATT produces summary tables and visualizations (Figure 2). The accuracy values reported by the interface are the MATT-derived indicators MATT-SOC and MATT-TLU, introduced in the Background. These values are calculated over the evaluated noun phrases only. They should not be reported as conventional SOC or TLU values from fully annotated obligatory contexts. Additional tables summarize output by expected article type (*a/an*, *the*, or Ø) and by broad error class (omission, overuse, misuse, or other tool-specific labels).

In the 'Detailed Annotations' table (Figure 3), each sentence appears with its evaluated noun phrase and the observed and expected article types, allowing users to inspect individual decisions before using the summary values. The table also includes the suggested minimal correction. Because article corrections may depend on discourse assumptions and because more than one correction may be acceptable, users should review the correction field before treating it as final.

In addition to the annotation tables, MATT generates a sentence-correction view. For each sentence, the tool constructs a corrected version by applying the suggested minimal correction to the original sentence and visualizes differences using a word-level comparison. The interface supports a combined view and a side-by-side view. These displays are intended to support inspection of the annotation rather than to replace human correction.

MATT provides an export function that downloads the sentence-level annotation table as a comma-separated values (CSV) file. The exported file contains one row per evaluated sentence and includes the fields needed to recompute MATT-SOC, MATT-TLU, and error distributions over the evaluated sample. This supports workflows in which MATT provides a first-pass annotation layer and subsequent checking, filtering, and statistical analysis are conducted in R, Python, or spreadsheet software.

Analysis Results

Overall Accuracy (SOC/TLU)

SOC	SOC_denom	TLU	TLU_total
0.33	12	0.33	12

SOC: Suppliance in Obligatory Contexts (% correct when an article was required (a/an or the)

TLU: Target-Like Use (% correct in all contexts, including zero article)

Error Rates by Article Type

expected_type	Accuracy	ErrorRate	N
def	0.4	0.6	10
indef	0.0	1.0	2

Error Distribution

expected_type	error_class	count	total_by_type	prop
def	CORRECT	4	10	0.4
def	MISUSE	1	10	0.1
def	OMISSION	5	10	0.5
indef	MISUSE	1	2	0.5
indef	OMISSION	1	2	0.5

Figure 2 *Aggregated Accuracy Outputs*

Detailed Annotations (6 sentences)

Download CSV

Sentence	Target NP	Observed	Expected	Error Class	Correction	Confidence	Notes
Yesterday I visited a small bookstore near my university	small bookstore	indef	indef	CORRECT	a small bookstore	0.95	First mention of a singular countable noun; 'a' is correct for introducing a new, non-specific entity.
I bought book about language learning	book about language learning	zero	indef	OMISSION	a book about language learning	0.90	First mention of a singular countable noun requires 'a' or 'an'.
The book was useful for my paper	book	def	def	CORRECT	the book	0.95	Anaphoric reference to previously mentioned book; 'the' correctly used for specific, previously mentioned entity.
A clerk helped me find another title	clerk	indef	indef	CORRECT	a clerk	0.90	First mention of a singular countable noun; 'a' is correct for introducing a new entity.
I received the helpful suggestion	helpful suggestion	def	indef	MISUSE	a helpful suggestion	0.85	First mention of a singular countable; should be 'a' instead of 'the'.
At home, I realized that the practice is important for improving writing	practice	def	zero	OVERUSE	practice	0.81	Generic activity without prior specific mention; zero article expected.

Figure 3 *Detailed Annotations*

Example Annotation Walkthrough

To illustrate how MATT's output should be read and audited, this section walks through four sentences drawn from the learner texts in the validation data set, together with the complete structured record produced by the tool and the outcome of human review. Table 2 shows, for each sentence, the preceding context supplied to the model (sentence boundaries marked by slashes), the evaluated noun phrase, the observed and expected article types, the broad error class, the suggested minimal correction, and the model-generated rationale.

The first three examples illustrate the three broad error classes and show the typical case in which the reviewer accepts MATT's record without change. The fourth example illustrates why human audit matters. MATT classified *a waiter* as a correct first-mention indefinite. Both annotators, however, noted that the restaurant context makes the waiter identifiable and judged *the waiter* at least equally natural, while still accepting *a waiter* as a defensible first-mention alternative. This is an instance of the valid-alternative-correction problem discussed in grammatical-error correction research

Table 2 *Worked Examples of MATT Annotation and Human Review*

Field	Example 1	Example 2	Example 3	Example 4
Learner sentence	They waited for bus	He found an advice	A team leader helped the cat	A waiter served two teas
Preceding context	Three people met their friend yesterday	He searched information on the Internet/The information was helpful	The boy called a team of experts/ Experts came soon around wood to help the cat	After eating all, they wanted to eat an ice cream, but the restaurant didn't serve any deserts/Instead of eating an ice cream, they ordered a tea, matcha
Evaluated NP	bus	advice	team leader	waiter
Observed article	Ø	an	a	a
Expected article	a/an	Ø	the	a/an
Error class	OMISSION	OVERUSE	MISUSE	CORRECT
Suggested correction	a bus	Ø advice	the team leader	a waiter
Rationale (abridged)	First mention of singular countable noun 'bus' in object position requires the indefinite article.	'advice' is an uncountable mass noun; generic statements require zero article, so 'an advice' overuses an indefinite article.	Bridging from prior mention of 'a team of experts' makes the leader uniquely identifiable, so definite article is expected; indefinite 'a' is misuse.	First mention of singular countable noun 'waiter' in subject position requires the indefinite article.
Human review	Accepted	Accepted	Accepted	Alternative noted: both annotators judged *the waiter* preferable in the restaurant context but accepted *a waiter* as a valid first-mention alternative.

(Bryant & Ng, 2015). In such cases, the reviewer may accept the suggested correction, substitute an alternative, or record the case as permitting multiple corrections before recomputing summary values from the exported CSV file. The walkthrough thus shows the intended division of labor: MATT makes each decision explicit and exportable, and the human reviewer resolves the interpretive cases that sampling and automatic annotation cannot settle.

Known Failure Modes

MATT is designed as a first-pass review aid, and several failure modes should be anticipated. First, the tool may miss an article-sensitive noun phrase and return UNMATCHED even when a human annotator would identify an obligatory context. Second, because only one noun phrase is evaluated per sentence, the output can underrepresent article use in syntactically dense sentences that contain multiple article-bearing noun phrases. Third, article judgements that depend on world knowledge, brand-mediated reference, or implicit discourse relations may be unstable. For example, a noun phrase such as *the driver* may be interpreted differently depending on whether the preceding context establishes a taxi, ride-sharing service, or another transport situation. Fourth, suggested corrections may not always be the most minimal or the only acceptable correction. These cases are not exceptions to be hidden, but expected reasons for treating MATT output as human-auditable rather than fully automatic annotation.

Preliminary Validation

To address the research question stated in the Introduction—to what extent MATT's article-use classifications agree with those of trained human annotators—we conducted a binary acceptability check on 450 sentences sampled from learner texts. Two annotators, both experienced English instructors, independently annotated the sentences using the same broad classification schema as MATT. The annotators recorded the observed article type, expected article type, and broad error class. For the agreement analysis reported here, these annotations were collapsed into a binary outcome variable (1 = acceptable article use, 0 = article error).

Because the underlying language model was updated during revision (see the Implementation subsection), the 450 sentences were re-annotated with the current version of MATT, and the agreement analysis was anchored to the tool's actual output. For 74 sentences in which the updated model evaluated a different noun phrase from the one originally rated, both annotators independently re-rated the new target noun phrase using the same procedure (inter-rater agreement on this subset: $\kappa = 0.96$, percent agreement = 98.6%). For six sentences the model returned UNMATCHED; these were excluded from the agreement analysis and audited: the annotators jointly judged none of the six abstentions defensible (they split in four cases, and in two cases both identified an evaluable noun phrase). Finally, because one prompt revision was made and then frozen after a diagnostic run on these same 450 sentences, the values reported in this section should be read as post-revision development-set agreement rather than as performance on unseen data.

The binary analysis evaluates whether MATT's final correct/error judgement aligns with human judgement at a broad level. To provide a more differentiated picture, we additionally compared each field of MATT's structured output with the annotators' independent field-level judgements, reported below. Target noun phrase selection, suggested corrections, rationale quality, and review-priority scores were not validated separately and remain targets for future validation.

Inter-rater agreement between the two human annotators was substantial (Cohen's $\kappa = 0.72$, 95% CI [0.65, 0.80], percent agreement = 89.6%). Agreement between the human consensus judgement and

MATT's binary classification was $\kappa = 0.58$, 95% CI [0.50, 0.67], agreement rate = 82.7% (n = 444). These results provide preliminary support for MATT as a first-pass annotation aid, but they should not be interpreted as evidence that all structured fields are fully validated.

Agreement was highest for overuse errors in zero-article contexts (81.2%), lower for omission errors (77.5%), and lowest for misuse errors (63.0%). This pattern is descriptively useful, but it should be interpreted cautiously because the main agreement analysis was binary.

To complement the binary analysis, we compared the individual fields of MATT's structured output with the annotators' independent judgements of the same noun phrases over all 444 evaluated sentences. For each record, the annotators had recorded, without seeing MATT's output, the observed article type, the expected article type, and the broad error class; a field was scored as agreeing when MATT's value matched the value shared by both annotators, and disagreements between the annotators were counted conservatively as non-agreement. Agreement was 97.3% for the observed article type, 76.1% for the expected article type, and 79.1% for the broad error class. In addition, in seven cases at least one annotator explicitly accepted a valid alternative realization, recording an error class while judging the learner's form acceptable, which provides a lower bound on the number of contexts permitting more than one correction. These field-level results are descriptive, but they show that surface detection of the observed article is highly reliable, whereas inference of the contextually expected article is the interpretively demanding component and the appropriate focus of human review effort.

Recommended Research and Reporting Workflow

As emphasized throughout this paper, MATT provides a sentence-sampled annotation layer that requires human audit. For research use, researchers can run the tool on individual learner texts or controlled excerpts, export the CSV files, and combine them into a dataset for analysis. When combining files, researchers should add document-level and learner-level metadata outside MATT, including document ID, learner ID, task, proficiency level, L1 background, and any sampling decisions relevant to the study.

Researchers should inspect the sentence-level annotations before reporting summary values. At minimum, they should review low review-priority-score cases, UNMATCHED cases, and cases where the suggested correction depends on discourse assumptions. For projects requiring higher reliability, a stratified sample of MATT annotations should be checked by human annotators, and any changes to the exported annotations should be documented.

When reporting MATT-based analyses, researchers should specify the operational unit of analysis. They should report whether complete scripts or excerpts were analyzed, how long texts were handled, how many sentences were submitted, how many noun phrases were evaluated, how many sentences were marked UNMATCHED, and how MATT-SOC and MATT-TLU denominators were defined. Because MATT evaluates one noun phrase per sentence, researchers should avoid presenting MATT-SOC and MATT-TLU as exhaustive article-accuracy measures.

Reproducibility also depends on documenting the annotation configuration. Researchers should archive the exported CSV files and report the tool version, model name, analysis date, prompt version, inference parameters, input-length limit, and sentence limit. The summary calculations can be recomputed exactly from archived CSV files, but as noted in the Implementation subsection, exact reproduction of the original model outputs cannot be guaranteed over time. For this reason, MATT should be reported as a transparent, auditable workflow rather than as a fully reproducible automatic annotator.

Pedagogical Use Cases

Although MATT is not intended to provide final judgements for grading, placement, or high-stakes assessment, it can support low-stakes pedagogical review when used with teacher oversight. For example, teachers can use the sentence-level output to identify article contexts that may be useful for classroom discussion, such as first-mention singular countable nouns, subsequent-mention definite noun phrases, or zero-article contexts involving plural and mass nouns. Because each annotation includes the evaluated noun phrase, expected article type, suggested correction, and rationale, the output can help teachers select examples for focused feedback or learner reflection.

In classroom-oriented use, MATT output should be treated as a source of candidate examples rather than as automatic corrective feedback; recent classroom studies that automate corrective written feedback with large language models (Hirschel & Horai, 2025) pursue a complementary but distinct goal. Teachers can ask learners to compare the original sentence, the suggested correction, and the rationale, and then discuss whether the correction is appropriate in context. This use is consistent with the human-auditable design of the tool: the value of the output lies not in replacing teacher judgement, but in making article-related decisions visible enough to support review, discussion, and selective feedback.

The use cases described in this section are anticipated applications that follow from the design of the tool; they have not yet been evaluated empirically with practicing teachers. A systematic usability evaluation, such as a cognitive walkthrough in which trained teachers use MATT with their own learners' writing and report their experiences, difficulties, usability barriers, and feature requests, is an important next step and is identified as future work in the Limitations and Future Directions section.

Limitations and Future Directions

Several limitations follow from the current implementation. The one-noun-phrase-per-sentence policy may underrepresent article use in syntactically dense sentences and, as discussed in the Background, may overrepresent error-prone contexts relative to exhaustive annotation. The UNMATCHED label may include cases where a human annotator would identify an article-sensitive noun phrase; in the abstention audit reported in the Preliminary Validation section, the annotators jointly endorsed none of the six abstentions. Suggested corrections may not always be the only acceptable correction, especially when discourse context is underspecified. The current preprocessing procedure removes non-ASCII characters, which may be unsuitable for multilingual corpora, proper names with diacritics, or texts with non-standard punctuation. Sentence segmentation is punctuation-based and may require inspection when learner writing contains unconventional punctuation. Finally, because expected article type is inferred through a cloud-hosted large language model, exact reproduction of outputs cannot be guaranteed across model or provider updates.

The validation reported in this paper covers binary correct/error agreement and field-level agreement over the evaluated noun phrases, and it reflects one disclosed prompt revision made after a diagnostic run on the same sentences, so the reported values are development-set agreement rather than performance on unseen data; target noun phrase selection, suggested corrections, rationale quality, and review-priority scores remain to be validated. Future work should include (a) validation of these remaining components, (b) a systematic comparison with grammatical-error-correction plus ERRANT workflows on the same texts, (c) a cognitive-walkthrough-style usability study in which five or six trained teachers use MATT in their L2 classrooms and report their experiences, difficulties, usability barriers, and feature requests, and (d) technical extensions including multiple noun phrase targets per sentence, improved handling of document-level metadata, more robust sentence segmentation, privacy-sensitive deployment options, and evaluation of smaller or locally deployable models.

Conclusion

This paper introduced MATT as a web-based tool for sentence-sampled, human-auditable analysis of English article use in learner writing. The tool is designed to provide a structured first-pass annotation layer rather than a complete replacement for manual annotation, learner-corpus annotation, or grammatical-error-correction workflows. For each evaluated sentence, MATT records the selected noun phrase, observed and expected article types, a broad error class, a suggested correction, a review-priority score, and a short rationale. These outputs are intended to make article-related decisions visible, inspectable, and exportable for subsequent researcher review.

A central point in interpreting MATT output is that the tool evaluates one article-sensitive noun phrase per sentence. This design keeps the output manageable and supports sentence-level inspection, but it also means that MATT does not provide an exhaustive inventory of all article-bearing noun phrases or all obligatory article contexts in a text. Consequently, MATT-SOC and MATT-TLU should be interpreted as sampled indicators, and researchers should report the number of submitted sentences, evaluated noun phrases, UNMATCHED cases, and the denominators used.

MATT's contribution is therefore best understood as practical and methodological in a limited sense: it offers a focused, article-specific workflow for generating reviewable annotations from raw learner writing. When its sampling policy, validation scope, and model-based nature are clearly reported, MATT can support exploratory analysis of article use while preserving the need for human judgement in final interpretation.

Acknowledgment

In preparing this manuscript, we used ChatGPT (GPT-5.5) to improve the clarity and coherence of the writing and to ensure compliance with scholarly journal standards. ChatGPT was used solely for language editing and did not contribute to the development of the study's concepts, methods, or conclusions.

References

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing Natural Language Processing techniques. *Language Learning, 67*, 180–209. https://doi.org/10.1111/lang.12232

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing, 5*(4), 257–269. https://doi.org/10.1093/llc/5.4.257

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Bryant, C., & Ng, H. T. (2015). How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (pp. 697–707). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1068

Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 793–805). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1074

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing, 27*(3), 419–436. https://doi.org/10.1177/0265532210364391

Chrabaszcz, A., & Jiang, N. (2014). The role of the native language in the use of the English nongeneric definite article by L2 learners: A cross-linguistic comparison. *Second Language Research, 30*(3), 351–379. https://doi.org/10.1177/0267658313493432

Coyne, S., Galvan-Sosa, D., Spring, R., Guerraoui, C., Zock, M., Sakaguchi, K., & Inui, K. (2025). Annotating errors in English learners' written language production: Advancing automated written feedback systems. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education: AIED 2025* (Lecture Notes in Computer Science, Vol. 15880, pp. 292–306). Springer. https://doi.org/10.1007/978-3-031-98459-4_21

Crosthwaite, P. (2016). L2 English article use by L1 speakers of article-less languages: A learner corpus study. *International Journal of Learner Corpus Research, 2*(1), 68–100. https://doi.org/10.1075/ijlcr.2.1.03cro

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 22–31). Association for Computational Linguistics.

Dulay, H. C., & Burt, M. K. (1973). Should we teach children syntax? *Language Learning, 23*(2), 245–258. https://doi.org/10.1111/j.1467-1770.1973.tb00659.x

Hawkins, J. A. (1978). *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. Routledge. https://doi.org/10.4324/9781315687919

Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal, 1*, e5. https://doi.org/10.1017/S2041536210000103

Hirschel, R., & Horai, K. (2025). Exploring AI to automate EFL corrective written feedback in the first language. *Technology in Language Teaching & Learning, 7*(1), 2208. https://doi.org/10.29140/tltl.v7n1.2208

Ionin, T., Ko, H., & Wexler, K. (2004). Article semantics in L2 acquisition: The role of specificity. *Language Acquisition, 12*(1), 3–69. https://doi.org/10.1207/s15327817la1201_2

Ionin, T., & Montrul, S. (2010). The role of L1 transfer in the interpretation of articles with definite plurals in L2 English. *Language Learning, 60*(4), 877–925. https://doi.org/10.1111/j.1467-9922.2010.00577.x

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Article and preposition errors. In C. Leacock, M. Chodorow, M. Gamon, & J. Tetreault (Eds.), *Automated grammatical error detection for language learners* (pp. 47–61). Springer. https://doi.org/10.1007/978-3-031-02137-4_6

Mizumoto, A. (2025). Automated analysis of common errors in L2 learner production: Prototype web application development. *Studies in Second Language Acquisition*, *47*, 867–884. https://doi.org/10.1017/S0272263125100934

Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition, 38*(3), 365–401. https://doi.org/10.1017/S0272263115000352

Öksüz, D., Alexopoulou, T., Derkach, K., & Tsimpli, I. M. (2026). The influence of L1 typology on the acquisition of the L2 English article: A large-scale corpus study. *Second Language Research, 42*(2), 281–311. https://doi.org/10.1177/02676583251395876

Pica, T. (1983). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition, 6*(1), 69–78. https://doi.org/10.1017/S0272263100000309

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language Learning, 64*(1), 103–131. https://doi.org/10.1111/lang.12029

Snape, N., & Yusa, N. (2013). Explicit article instruction in definiteness, specificity, genericity and perception. In M. Whong, K. H. Gil, & H. Marsden (Eds.), *Universal grammar and the second language classroom* (pp. 161–183). Springer. https://doi.org/10.1007/978-94-007-6362-3_9

Stockwell, G. (2024). ChatGPT in language teaching and learning: Exploring the road we're travelling. *Technology in Language Teaching & Learning, 6*(1), 2273. https://doi.org/10.29140/tltl.v6n1.2273

Umeda, M., Snape, N., Yusa, N., & Wiltshier, J. (2019). The long-term effect of explicit instruction on learners' knowledge on English articles. *Language Teaching Research, 23*(2), 179–199. https://doi.org/10.1177/1362168817739648

Zhao, H., & MacWhinney, B. (2018). The instructed learning of form–function mappings in the English article system. *The Modern Language Journal, 102*(1), 99–119. https://doi.org/10.1111/modl.12449