

The *Multi-Feature Tagger of English* (MFTE): Rationale, description and evaluation

Elen Le Foll^a – Muhammad Shakir^b
University of Cologne^a / Germany
University of Münster^b / Germany

Abstract – The *Multi-Feature Tagger of English* (MFTE) provides a transparent and easily adaptable open-source tool for multivariable analyses of English corpora. Designed to contribute to the greater reproducibility, transparency, and accessibility of multivariable corpus studies, it comes with a simple GUI and is available both as a richly annotated *Python* script and as an executable file. In this article, we detail its features and how they are operationalised. The default tagset comprises 74 lexico-grammatical features, ranging from attributive adjectives and progressives to tag questions and emoticons. An optional extended tagset covers more than 70 additional features, including many semantic features, such as human nouns and verbs of causation. We evaluate the accuracy of the MFTE on a sample of 60 texts from the BNC2014 and COCA, and report precision and recall metrics for all the features of the simple tagset. We outline how that the use of a well-documented, open-source tool can contribute to improving the reproducibility and replicability of multivariable studies of English.

Keywords – software; multivariable analysis; multivariate analysis; open source; corpus linguistics; corpus tool; multi-dimensional analysis; *Python*

1. INTRODUCTION

The addition of multivariable¹ analysis methods to the linguist’s toolbox has proven indispensable to shed light on the intricate interplay between diverse linguistic features and the situational/contextual factors that shape them. One of the earliest such methods is multi-dimensional (MD) analysis—a framework pioneered by Douglas Biber in the 1980s and first applied to the study of register variation (see Biber 1984; 1988). What MD analysis and most other multivariable methods used in linguistics have in common

¹ In quantitative linguistics, the terms ‘multivariable’ and ‘multivariate’ are frequently equated. Statisticians, however, differentiate between ‘multivariable’ methods in which several independent variables (predictors) are used to explain or predict a single outcome variable and ‘multivariate’ ones in which there are two or more dependent (or outcome) variables (Hidalgo and Goodman 2013). As all multivariate methods are, by definition, also multivariable, we use the term ‘multivariable’ throughout this article as an overarching term encompassing methods such as multivariable linear and logistic regression, factor analysis, cluster analysis, and machine learning methods.



is that they rely on software capable of automatically identifying a large number of linguistic features in many texts. Since the co-occurrences of these linguistic features need to be identified and counted across hundreds or thousands of texts for such analyses to be feasible,² automatic feature taggers are needed. Indeed, they can be said to constitute the backbone of such analysis frameworks. In the context of MD analysis, Biber (2019: 14) stresses that “[a]lthough its importance is not widely recognized, the computer program used for grammatical tagging provides the foundation for MD studies.” This failure to recognise the crucial importance of the tools underlying most quantitative corpus-linguistic analyses is detrimental to the reproducibility of quantitative corpus-linguistic studies. Even when the corpus data are publicly available, if the tools used to process the data are not freely accessible, the results cannot be independently verified. It is also very difficult to test their robustness and generalisability on new data.

This is particularly problematic in the context of the ‘replication crisis’ (first named as such in Pashler and Wagenmakers 2012), which having been first exposed in social psychology has, over the past decade, spared almost no discipline, highlighting the pressing need to improve both the reproducibility of scientific studies and the replicability of their findings.³ Among the many causes of the replication crisis/crises, the lack of published (or otherwise freely accessible) research data, code and software ranks high. It undoubtedly constitutes a major barrier to both the reproducibility and replicability of published research (see, e.g. John *et al.* 2012; Baker 2016; Gewin 2016). (Corpus) linguists, too, are becoming increasingly aware of the implications of non-reproducible methods, as evidenced, for example, by a special issue of *Linguistics* devoted to the replication crisis (Sønning and Werner 2021) and several recent articles and monographs that tackle the issue head-on (e.g., Porte and McManus 2018; Wallis 2021; In’nami *et al.* 2022; McManus 2024).

² Synthesising data from 161 MD studies, Goulart and Wood (2021: 119) report a mean size of corpora used in MD analyses of 5.5 million words (with a very large range from under 10,000 words to over 206 million words). In a meta-analysis of 23 MD studies, Egbert and Staples (2019: 132) report that the mean number of variables entered in MD analyses is around 60.

³ Although the terms are sometimes used interchangeably, we use the term ‘reproducibility’ to refer to the ability to obtain the same results as an original study using the authors’ data and code, whilst we understand ‘replicability’ to be about obtaining compatible results with (more or less) the same method(s) but different data (see, e.g., Berez-Kroeker *et al.* 2018).

2. LITERATURE REVIEW

In the present paper, we introduce a substantially revised and extended version of the *Multi-Feature Tagger of English Perl*: the *MFTE Python* (hereafter MFTE), designed to facilitate the reproducible, multivariable analysis of large English corpora. Although our tool can be used for any type of quantitative corpus analysis, we developed the MFTE primarily with multivariable analyses in mind as it is often possible to manually examine the accuracy of a tagger for one or just a few feature(s). Such a procedure, however, becomes entirely unfeasible when a large(r) number of linguistic variables are entered in an analysis. The MFTE was designed to facilitate large-scale multivariable analyses such as MD studies by aiming to largely eliminate the need for the manual correction of feature identification ('fix-tagging', see Section 2.1.3)

In the following, we explain our rationale for the development of a new open-source tagger written in a modern, object-oriented programming language. The rationale in Section 2.1 is followed by the specifications of the tagger in Section 2.2. In Section 2.3, we list the linguistic features it tags and counts and explain how they are operationalised in Section 2.4. In Section 2.5, we explain the tagger's usage and outputs and then report on a detailed evaluation of the tagger's output, before discussing its strengths and limitations and concluding with an outlook on future possible uses and developments.

2.1. Rationale for a new multi-feature tagger of English

Several factors motivated the development of the MFTE. They are best summarised as concerns about the reproducibility, transparency, and accessibility of the corpus tools currently most frequently used in multivariable analyses of English. It should be mentioned that these same reasons had originally motivated the development of the *MFTE Perl*, an earlier but considerably less powerful version of the MFTE.⁴ Its documentation (Le Foll 2021) outlined the methodological decisions involved in the selection of its features, details of their operationalisations, and the rationale behind the use of different normalisation units for the feature frequencies. However, no research article about the *MFTE Perl* was submitted as it quickly became apparent that an entirely reworked version of it, ported to *Python* and considerably extended, would fulfil better

⁴ <https://github.com/elenlefol/MultiFeatureTaggerEnglish> (accessed 8 March 2024)

the three-fold objectives of reproducibility, transparency, and accessibility. It is this new *Python* version that we present and evaluate in the present paper.

2.1.1. Reproducibility

Using proprietary software for research constitutes a barrier to reproducibility as only a limited number of researchers (i.e., those who have personal contacts with the developer(s) or the (institutional) means to pay for a licence) can attempt to reproduce the results of published studies. In theory, publishing detailed information about the inner workings of a piece of software is an alternative to making software accessible to all for free and/or publishing its source code. This is essentially what was done in what remains the most cited MD analysis study to date, namely Biber (1988: Appendix II), which includes detailed descriptions of the algorithms of the tagger now widely known as the ‘*Biber tagger*’ (Gray 2019). It was used to identify the features entered in Biber’s (1988) seminal MD analysis. As a result, even though the *Biber tagger* is not available to the wider research community, it is possible to reconstruct it based on this list of algorithms. There is no doubt, however, that such a ‘reconstruction’ is a time-consuming process that requires advanced programming skills, which not all corpus linguists possess.

Demonstrating that reconstruction is possible based on Biber’s (1988: Appendix II) list of algorithms, Nini (2014; 2019) successfully reproduced the functions of the 1988 version of the *Biber tagger* with only very minor differences. The resulting corpus tool, labelled the ‘*Multidimensional Analysis Tagger*’ (MAT), was originally released as freeware in 2013 and subsequently made available under an open-source licence on GitHub in 2020.⁵ MAT allows researchers to conduct reproducible analyses using the linguistic features described in Biber (1988: Appendix II).⁶ More recent MD studies that rely on the *Biber tagger*, however, are not reproducible as the *Biber tagger* has considerably evolved since the 1988 publication (Biber and Egbert 2018: 22; Gray 2019:

⁵ <https://github.com/andreanini/multidimensionalanalysistagger> (accessed 14 December 2023).

⁶ It is worth noting that MAT does more than just tag and count the features used in Biber’s (1988) analysis: it also calculates dimension scores on Biber’s (1988) dimensions of *General Written and Spoken English*, outputs plots with mean values of the tagged texts/corpus against these dimensions, making it ideal for conducting additive MD analyses (Berber Sardinha *et al.* 2019) that rely on Biber’s (1988) *Model of General Spoken and Written English* as a comparison baseline. MAT also assigns each text to one of the eight text types proposed in Biber’s (1989) *Typology of English Texts*. Moreover, the GUI version (for *Windows* only) features a tool for visualising the features of an input text that load on a particular dimension.

46).⁷ Our reproducibility motivation for the development of the MFTE was therefore to allow researchers to conduct reproducible analyses involving a larger number of English lexico-grammatical and semantic features that go beyond those of the 1988-version of the *Biber tagger*, and which users can flexibly and transparently adapt and amend according to their needs.

2.1.2. Transparency

Ensuring that the concrete operationalisations of the features of the MFTE would be as transparent as possible was a further motivating factor for the development of the MFTE. Reproducibility and transparency of research processes are inextricably linked. Not only does the lack of access to the source code of a tagger mean that results cannot be reproduced, but it also means that researchers conducting and evaluating studies that rely on such tools have few means of understanding exactly how the features entered into these analyses were identified. While this is true for many tools used in corpus studies, it is particularly problematic in the context of multivariable studies such as MD analyses, which rely on counting large numbers of linguistic features across many texts, for which manual spot-checking of counts is simply not feasible.

A further aim of the MFTE was therefore to make available both detailed textual explanations of its feature operationalisations, as well as easily accessible source code to be able to examine their concrete operationalisation and, if need be, adapt them to specific language varieties and/or registers, linguistic theories,⁸ and research questions. Whilst it is possible to scrutinise the exact operationalisations of each linguistic feature of MAT, its code structure is relatively complex due to its graphical interface and many additional functions (see footnote 6), which means that only linguists with a strong programming background are likely to be able to edit the source code to introduce customised features/feature operationalisations. In developing the MFTE, we opted for the simplest

⁷ During the second round of revisions for this paper, an anonymous reviewer helpfully pointed out that a new tagger for MD analysis is now available in beta stage. According to the documentation, it is being developed by Kristopher Kyle and colleagues at the Linguistics Department of the University of Oregon in consultation with Douglas Biber's lab in Northern Arizona University: <https://github.com/kristopherkyle/LxGrTgr> (9 September 2024).

⁸ It is worth remembering that POS tagging is fundamentally a method of analysing grammar and morphology. As a consequence, the process inevitably implicitly reflects a specific approach or theory of grammar (McEnery *et al.* 2006; Lindquist 2009: 44–45; Gray 2019: 34)

structure possible: it consists of a single script that can be easily edited by linguists familiar with regular expressions. This brings us to our final, major motivation for developing the *MFTE Python*, which is to contribute to the better accessibility of taggers that can readily be used to conduct multivariable analyses of English.

2.1.3. Accessibility

For some features, both the *Biber tagger* and MAT require (semi-)manual ‘fix-tagging’ procedures to reach high levels of tagging accuracy (Gray 2019: 59–61). In fact, it is recommended that the *Biber tagger* be used in combination with an interactive tag-checker, see Biber and Gray (2013) for details. Although this process allows for the inclusion of linguistic features that cannot reliably be annotated automatically, it requires trained annotators to perform time-consuming manual checks and corrections.⁹ For a tool to be accessible for research projects with little or no funding, we therefore believe that the need for fix-tagging should be reduced to a minimum. Given that most multivariable linguistic methods, including MD analysis, require many different linguistic features to be identified across large corpora, we aimed for high tagging accuracy without the need for human intervention. Moreover, we concluded that the tagger documentation should include a detailed evaluation of the accuracy of the tagging procedure on a representative sample of texts. These evaluation results should be transparently reported for researchers to be able to decide which feature operationalisations are accurate enough for their specific research objectives.

Most standard POS taggers require an additional script to count the number of occurrences of each tag in each text and to normalise these counts if the texts of the corpora are of different lengths. This process adds an extra step in the preprocessing of tagged corpus data for multivariable analyses. In contrast, MAT and the *MFTE Perl* are more accessible in that they output tables of normalised frequencies that can readily be used as input for statistical tools and functions.

Whilst the *MFTE Perl* was designed to be used without fix-tagging, its outputs include normalised frequencies of each feature per text, and its documentation includes

⁹ For instance, for the TOEFL iBT project (with a corpus of 3,839 texts totalling 543,000 words), Gray and Biber (2013: 18) report having recruited and trained ten fix-taggers in addition to two independent coders and a project research assistant for the manual corrections of problematic tags.

detailed results of a thorough evaluation, it nonetheless failed to adequately meet our third aim of accessibility. This is because we believe that accessibility also entails ease of use. Regrettably, the *MFTE Perl* requires a separate installation of the *Stanford POS tagger* (Toutanova *et al.* 2003) which, even with detailed installation instructions, is likely to constitute a barrier for some linguists. This is not the case with the present version of the MFTE (see Section 2.5). Furthermore, the present version of the MFTE was written in *Python*, a more accessible programming language, with a large user base and hence many beginner-friendly tutorials and helper tools (e.g., *Anaconda* and *Anaconda Mini* for installation).

Finally, unlike MAT for *Windows*, the *MFTE Perl* also lacks a graphical user interface (GUI). We consider this to be an important aspect of making open-source tools accessible to the wider research community and argue that a genuinely accessible tagger ought to include a GUI for all major operating systems.

2.2. Specifications of the MFTE

Based on our triple motivation to improve the reproducibility, transparency, and accessibility of multivariable English corpus studies and our survey of the strengths and weaknesses of existing tools in these regards (see Section 2.1), we elaborated the tagger specifications for the MFTE by updating the specifications originally specified for the *MFTE Perl* (Le Foll 2021). These specifications are shown in Table 1.

1. Identify a broad range of lexico-grammatical and semantic features of English
a. that can each be meaningfully interpreted
b. to a satisfactorily high degree of accuracy (with precision and recall rates of > 90%)
c. without the need for human intervention
d. in a broad range of English registers
e. with standard American or British orthography.
2. Output
a. the full tagged texts in plain text format for qualitative analyses of the tagger’s accuracy
b. delimiter-separated values (DSV) files containing both raw and normalised feature counts per text.
3. Be available
a. as source code under a GNU licence for researchers with programming skills to scrutinise, adapt, improve and re-use and
b. as a GUI with adequate documentation for researchers with basic computer skills to be able to run the programme in all major operating systems.

Table 1: Tagger specifications for the MFTE

In what follows, we outline how we set out to meet these specifications. In Section 2.3, we list the linguistic features tagged by the MFTE and, in Section 2.4 we motivate their operationalisations, before explaining how to use the tagger and understand its outputs (Section 2.5).

2.3. Tagset

In line with the MFTE’s intended application for descriptive linguistic analyses (such as multivariable analyses of register variation) as opposed to classification tasks, we explicitly focused on the identification of linguistic features that can be “meaningfully interpretable” (specification criterion 1a). By this, we mean that the “scale and values [of each feature] represent a real-world language phenomenon that can be understood and explained” (Egbert *et al.* 2020: 24).

When designing the feature portfolio of the *MFTE Perl*, Le Foll (2021) examined simplified Hallidayan system networks grammars (see, e.g., Bartlett and O’Grady 2017) in an attempt to minimise researcher bias in the selection of linguistic features. The *MFTE Perl* identifies and counts 75 linguistic features covering lexical density and diversity, fine-grained POS classes, verb tense and aspect, various frequent lexico-grammatical constructions, and a selection of verb semantic categories. The *MFTE Python* builds on this original set of features from the *MFTE Perl*, of which 74 were retained to form the ‘simple tagset’ of the new MFTE. These are listed in Table 2. In addition, the present version of the MFTE features an ‘extended tagset’ with more than 70 additional features, mostly semantic features inspired from Biber *et al.* (1999) and Biber (2006), operationalised on the basis of the features from the simple tagset. A full list of all features including examples and a description of their operationalisation can be found on the MFTE’s [GitHub repository](#). Note that the latest features can be found in the developmental branch of the repository.


Feature	Tag	Example
<i>BE able to</i>	ABLE	<i>It should be able to speak back to you. Would you be able to?</i>
Amplifiers	AMP	<i>I am very tired. They were both thoroughly frightened.</i>
Average word length	AWL	<i>It's a shame that you'd have to pay to get that quality. [AWL = 42/12 = 3.5]</i>
<i>BE as main verb</i>	BEMA	<i>It was nice to just be at home. She's irreplaceable. It's best.</i>
Coordinators	CC	<i>Instead of listening to us, he also told John and Jill but at least his parents don't know yet.</i>
Numbers	CD	<i>That's her number one secret. It happened on 7 February 2019.</i>
Concessive conj.	CONC	<i>Even though the antigens are normally hidden...</i>
Conditional conj.	COND	<i>If I were you... Even if the treatment works...</i>
Verbal contractions	CONT	<i>I don't know. It isn't my problem. You'll have to deal.</i>
Causal conjunctions	CUZ	<i>He was scared because of the costume. Yeah, coz he hated it.</i>
Demonstrative pronouns and articles	DEMO	<i>What are you doing this weekend? Whoever did that should admit it.</i>
Discourse/pragmatic markers	DMA	<i>Well, no they didn't say actually. Okay I guess we'll see.</i>
<i>DO auxiliary</i>	DOAUX	<i>Should take longer than it does. Ah you did. Didn't you?</i>
Determiners	DT	<i>Is that a new top? Are they both Spice Girls? On either side.</i>
Downtoners	DWNT	<i>These tickets were only 45 pounds. It's almost time to go.</i>
Elaborating conjunctions	ELAB	<i>Similarly, you may, for example, write bullet points.</i>
Emoji and emoticons	EMO	 :-(:DD XD :)
Emphatics	EMPH	<i>I do wish I hadn't drunk quite so much. Oh really? I just can't get my head around it.</i>
Existential <i>there</i>	EX	<i>There are students. And there is now a scholarship scheme.</i>
Filled pauses and interjections	FPUH	<i>Oh noooooo, Tiger's furious! Wow! Hey Tom! Er I know.</i>
Frequency references	FREQ	<i>We should always wear a mask. He had found his voice again.</i>
<i>Going to</i> constructions	GTO	<i>I'm not gonna go. You're going to absolutely love it there!</i>
Hedges	HDG	<i>There seemed to be no sort of chance of getting out. She's maybe gonna do it.</i>
<i>HAVE got</i> constructions	HGOT	<i>He's got one. Has she got any?</i>
Hashtags	HST	<i>#AcWri #Buy1Get1Free</i>
Prepositions	IN	<i>The Great Wall of China is the longest wall in the world. There are towers along the wall.</i>
Attributive adjectives	JJAT	<i>I've got a fantastic idea! Cheap, quick and easy fix!</i>
Predicative adjectives	JJPR	<i>That's right. One of the main advantages of being famous...</i>
Lexical density	LDE	<i>It's a shame that you'd have to pay to get that quality. [LDE = 3/14 = 0.21]</i>
<i>Like</i>	LIKE	<i>Sounds like me. And just like his father. I wasn't gonna like do it.</i>
Modal <i>can</i>	MDCA	<i>Can I give him a hint? You cannot. I can't believe it!</i>
Modal <i>could</i>	MDCO	<i>Well, that could be the problem. Could you do it by Friday?</i>
Modals <i>may</i> and <i>might</i>	MDMM	<i>May I have a word with you? But it might not be enough.</i>
Necessity modals	MDNE	<i>I really must go. Shouldn't you be going now? You need not have worried.</i>
Modal <i>would</i>	MDWO	<i>Wouldn't you like to know? I'd like to think it works.</i>
<i>Will</i> and <i>shall</i> modals	MDWS	<i>It won't do. Yes, it will. Shall we see?</i>
Noun compounds	NCOMP	<i>the dungeon entrance; this rare winter phenomenon</i>
Total nouns (including proper nouns)	NN	<i>on Monday 6 Aug, the U.S., on the High Street, comprehension</i>

Table 2: Features of the MFTE's simple tagset

Feature	Tag	Example
@mentions	NN	<i>@gretathunberg @MSF_france</i>
BE-passives	PASS	<i>He must have been burgled. They need to be informed.</i>
Perfect aspect	PEAS	<i>Have you been on a student exchange? He has been told.</i>
GET-passives	PGET	<i>He's gonna get sacked. She'll get me executed.</i>
It pronoun reference	PIT	<i>It fell and broke. I implemented it. Its impact is unproven.</i>
Place references	PLACE	<i>It's not far to go. I'll get it from upstairs. It's downhill.</i>
Politeness markers	POLITE	<i>Can you open the window, please? Would you mind giving me a hand?</i>
s-genitives	POS	<i>the world's two most populous country; my parents' house</i>
Reference to the speaker/writer and other(s)	PP1P	<i>We were told to deal with it ourselves. It's not ours either.</i>
Reference to the speaker/writer	PP1S	<i>I don't know. It isn't my problem. Nor is it mine.</i>
Reference to addressee(s)	PP2	<i>If your model was good enough, you'd be able to work it out.</i>
Single, female third person	PP3f	<i>She does tend to keep to herself, doesn't she?</i>
Single, male third person	PP3m	<i>He is beginning to form his own opinions. I trust him.</i>
Other personal pronouns	PPother	<i>One would hardly suppose that your eye was as steady as ever.</i>
Progressive aspect	PROG	<i>He wasn't paying attention. I'm going to the market.</i>
Quantifiers	QUAN	<i>Such a good time in half an hour. She's got all these ideas. It happens every time.</i>
Quantifying pronouns	QUPR	<i>She said addressing nobody in particular. Somebody will.</i>
Question tags	QUTAG	<i>Do they? Were you? It's just it's repetitive, isn't it?</i>
Other adverbs	RB	<i>Unfortunately, that's the case. Exactly two weeks.</i>
Particles	RP	<i>I'll look it up. It's coming down. When will you come over?</i>
So	SO	<i>She had spent so many summers there. So, there you go.</i>
Split auxiliaries/infinitives	SPLIT	<i>I would actually drive. You can just so tell. I can't imagine it.</i>
Stranded prepositions	STPR	<i>We've got more than can be accounted for. Open the door and let them in.</i>
Subordinator <i>that</i> omission	THATD	<i>I mean [THATD] you'll do everything. I thought [THATD] he just meant our side.</i>
<i>That</i> relative clauses	THRC	<i>I'll just run a cable that goes from here to there.</i>
<i>That</i> subordinate clauses	THSC	<i>Did you know that the calendar we use today was started by Julius Caesar?</i>
Time references	TIME	<i>It will soon be possible. Now is the time. I haven't it yet.</i>
Reference to more than one non-interactant and single <i>they</i> reference	TPP3t	<i>The text allows readers to grapple with their own conclusions. Do you see them?</i>
Lexical diversity	TTR	<i>It's a shame that you'd have to pay to get that quality. [TTR = 12/14 = 0.85]</i>
URL and e-mail addresses	URL	<i>www.faz.net https://twitter.com smith@gmail.au</i>
Past tense	VBD	<i>It fell and broke. I implemented it. If I were rich.</i>
Non-finite verb <i>-ing</i> forms	VBG	<i>He texted me saying no. He just started laughing.</i>
Non-finite <i>-ed</i> verb forms	VBN	<i>These include cancers caused by viruses. Have you read any of the books mentioned there?</i>
Imperatives	VIMP	<i>Let me know! In groups, share your opinion and take notes.</i>
Present tense	VPRT	<i>It's ours. Who doesn't love it? I know.</i>
Direct WH-questions	WHQU	<i>What's happening? Why don't we call the game off? How?</i>
WH subordinate clauses	WHSC	<i>I'm thinking of someone who is not here today. Do you know whether the banks are open?</i>
Negation	XX0	<i>Why don't you believe me? There is no way. Nor am I.</i>
Yes/no questions	YNQU	<i>Have you thought about giving up? Do you mind?</i>

Table 2: Features of the MFTE's simple tagset (Continuation)

The MFTE performs feature extraction in several steps. First, each text is tagged with basic POS labels using the POS tagger from the *stanza Python* library (Qi *et al.* 2020). Then, several loops of rule-based algorithms are run to refine some of the analyses of the POS tagger and to identify further linguistic features on the basis of syntactic patterns defined by a combination of regular expressions and dictionary lists. The lists are based on Biber (1988), Biber *et al.* (1999), Biber (2006), and the COBUILD (Sinclair *et al.* 1990).

Whilst many of the features of the simple tagset may look superficially like the 1988 version of the *Biber tagger* (and hence also to the MAT), their operationalisations often differ substantially. In particular, the algorithms used to capture verb tense, aspect, and voice are fundamentally different. For example, rather than tag the perfect aspect onto the auxiliary HAVE as in MAT, the MFTE assigns the ‘perfect aspect tag’ (PEAS) to the past participle form of the verb, whilst the auxiliary is marked for tense with either the VBD or the VPRT tag. Similarly, the ‘passive voice tag’ (PASS) is assigned to the past participle form rather than to the verb BE. These new operationalisations make possible the creation of a distinct, linguistically meaningful, VBN variable, which only includes non-finite uses of past participle forms. As the MFTE also identifies verbs in the progressive aspect (PROG) and the *going-to* construction (GTO), non-finite uses of present participle forms can also be accounted for in a separate variable (VBG). Moreover, these verb feature operationalisations are optimised for a range of syntactic patterns, e.g., by allowing for various combinations of intervening adverbs, negation, and paralinguistic sounds in verb phrases involving auxiliaries. Other lexico-grammatical features from the MFTE’s simple tagset that are not typically found in lexico-grammatical taggers include question tags (QUTAG), imperatives (VIMP), and emojis or emoticons (EMO).

Various strategies were employed to deal with multifunctional/polysemous lexical items: in many cases, they were included in the semantic categories corresponding to their most frequent functions (e.g., *only* is assigned to the category of downtoners rather than hedges). When this proved too error-prone, infrequent items were excluded. Moreover, for two highly frequent multifunctional words (*so* and *like*) separate feature categories were created to capture all potentially problematic cases, i.e., all occurrences of *so* and all occurrences of *like* tagged as a preposition or adjective by the stanza POS tagger. Both words are frequently used as discourse markers and fillers in conversation. These uses are

very difficult to automatically disambiguate. Consequently, if they are not excluded, they run the risk of severely distorting the frequencies of adverbs and prepositions in conversational texts. More recent versions of the *Biber tagger* make use of different algorithms and probabilities depending on the mode/register of the texts to be tagged to (partially) circumvent this issue (Gray 2019: 46). We believe that such a procedure bears the risk of predetermining patterns of occurrences and have therefore opted to relegate these items to separate feature categories for *like* (when not tagged as a verb) and *so*, instead. While this is by no means a perfect solution, it allows users of the MFTE to decide—depending on the resources available to them—whether to perform manual fix-tagging to assign the correct functional tag to each occurrence of *so* and *like*, or to exclude these tokens by not entering the counts of the SO and LIKE variables in their analyses.

When using POS taggers, Brezina (2018: 192) suggests removing paralinguistic sounds (e.g., *um*, *er*, *mhm*, *mm*) from the spoken data as these are frequently misidentified as nouns, thus potentially vastly overestimating the noun count in natural conversation corpora in which these sounds have been transcribed. Rather than remove these tokens as part of the pre-processing of the corpus texts, we opted to retain them as they constitute a defining characteristic of natural spoken language. The MFTE counts these as a self-contained linguistic variable (FPUH), which, like all the linguistic features counted by the MFTE, the user is free to either include or exclude from their analyses, depending on their texts (considering, for instance, how reliably these were transcribed across of the texts under study) and research questions.

A final major departure from the feature extraction principle applied by some lexico-grammatical taggers that is worth mentioning concerns the treatment of multiword items. The 1988 version of the *Biber tagger* and MAT (Nini 2014; 2019), for example, tag the first token of the multiword *on the other hand* as a conjunct and assigns NULL tags to the remaining tokens, as illustrated in (1) below. This means that, when tagged with these taggers, the string *on the other hand* counts as just one conjunct. In contrast, the MFTE assigns the conjunct tag (CONJ) in addition to the usual preposition, determiner, adjective, and noun tags, as shown in (2). NULL tags were only retained for a select few multiword units that largely resist compositional analysis: *of course* (tagged as a discourse marker, DMA), *all right* (DMA) and *no one* (tagged as a quantifying pronoun, QUPR).

- (1) CONJ the_NULL other_NULL hand_NULL
- (2) on_IN CONJ the_DT other_JJ hand_NN

The three semantic verb categories employed in Biber (1988) are only used in the MFTE for the identification of the *that*-omissions in subordinate clauses (THATD). Instead, the MFTE's extended tagset adopts, with some minor corrections, the semantic categories described in Biber *et al.* (1999) and Biber (2006). As a result, in addition to ten semantic verb features, the extended tagset also includes fine-grained semantic categories for adverbs, nouns, and adjectives, as well as combined grammatico-semantic features such as 'to clauses preceded by verbs of desire' (ToVDSR) or 'stance nouns without prepositions' (NSTNCothers). The extended tagset also includes comparative and superlative constructions (see [GitHub repository](#) for details).

The user should be aware that the tables of counts generated when using the extended tagset include some composite features. These consist of aggregates of individual feature counts, e.g., the variable PASSall contains the sum of the counts of *BE*-passives (PASS) and *GET*-passives (PGET), two features from the simple tagset. It goes without saying that, to avoid redundant correlations, these aggregate counts should not be entered in MD analyses together with their respective individual feature counts. In proposing this large number of lexical, grammatical, and semantic features, we are not suggesting that it makes sense to enter them all in multi-feature analyses of English, but rather that, with the help of our detailed documentation, users can make informed, linguistically motivated choices as to which features are relevant and reliable enough (see evaluation results in Section 4.3) for their use cases.

2.4. Usage

Since the alpha version was first released in April 2023, the MFTE has been accessible under an open-source GPL-3.0 licence. It can be downloaded from a dedicated GitHub repository at <https://github.com/mshakirDr/MFTE>. Detailed installation and usage instructions can be found on the main page of the [GitHub repository](#).

The MFTE relies on commonly utilised *Python* dependencies whose installation are not particularly involved. To expand its accessibility to non-programming linguists, we have also released the programme as an executable file for *Windows*. Once this executable file has been downloaded, the GUI can be run without installing *Python* or any

dependencies and is therefore the most user-friendly version of the MFTE (currently only for *Windows*; see Figure 1).

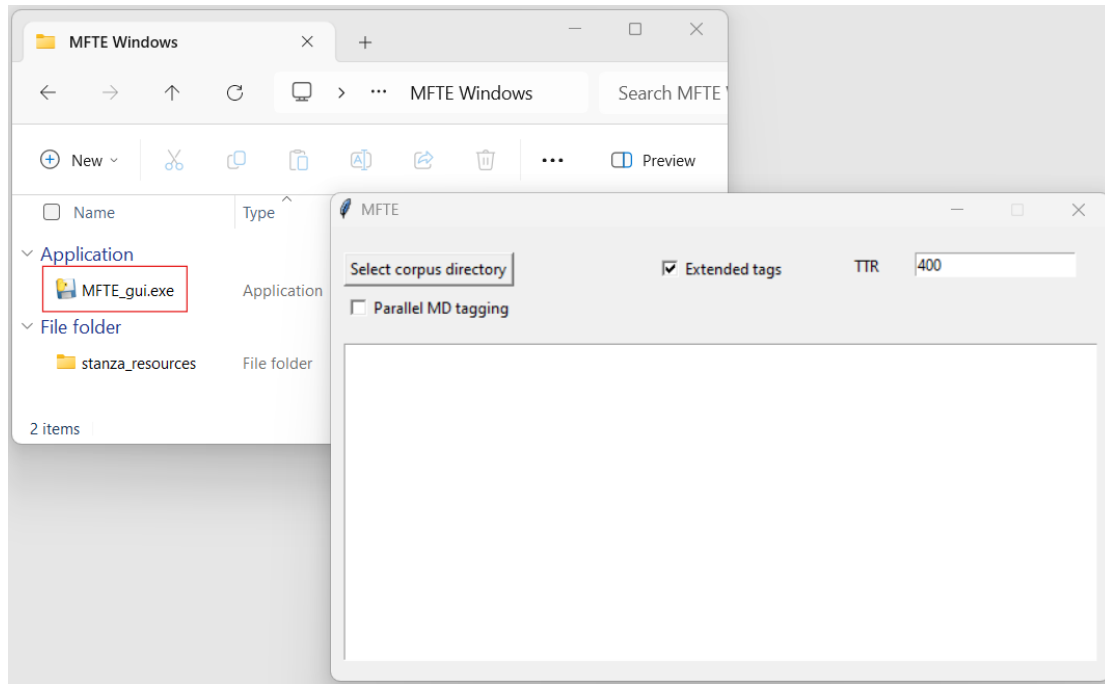


Figure1: Screenshot of the downloaded MFTE GUI executable for *Windows* (MFTE_gui.exe), alongside the running MFTE GUI application

For *Mac OS*, *Linux*, and *Windows*, both the command-line version and the GUI version can be run from a terminal. The command-line version is straightforward (see Figure 2), requiring only four arguments:

1. the path to the folder containing the text files to be tagged;
2. the number of words used to calculate the type-token ratio (TTR)—as in the MAT, the default is 400 but this should be set lower if any texts are shorter than 400 words;
3. an option to use the extended tagset (the default is True);
4. an option to tag multiple files in parallel (this reduces running time for large corpora significantly but may also cause high CPU usage, hence the default setting is False).

Each CSV file consists of a data matrix in which each row corresponds to a text file from the tagged corpus and each column to a linguistic feature. In all three tables, the first five columns of the count tables are identical: the first column lists the filenames, the second the total number of words in each text as used for word-based normalisation (excluding fillers, see online appendix). The remaining columns correspond to the linguistic features listed in Table 2 for the simple tagset followed by those of the extended tagset (see online appendix), if this option was selected.

The difference between the three tables that the MFTE outputs is that the first contains feature frequencies normalised per 100 words. In the second table, ‘complex’ normalised feature frequencies are calculated based on the three normalisation baselines listed in the fifth column of the table in the appendix. This means that, in this table, the present tense variable (VPRT) represents the percentage of finite verbs in the present tense. As such, it can range from zero, i.e., texts in which no single verb is in the present tense to 100, i.e., texts that are exclusively in the present tense, and therefore does not violate the assumptions of the binomial distribution required for many statistical methods commonly used in corpus linguistics research (Wallis 2020: 56). For details on the normalisation units used for each feature and the rationale behind these choices, see Le Foll (2021: 20–23; 2024: 120–124). Lastly, the MFTE outputs a table of raw counts. Unless the text samples of the examined corpus are all the same length, this table should not be used *as is* for any statistical analyses; however, it may be used by researchers who wish to implement their own normalisation baseline(s). We have also found it very useful for test purposes, i.e., to check how the tagger deals with certain strings.

3. EVALUATING TAGGER ACCURACY

Given that taggers provide “the foundation for MD studies” (Biber 2019: 14) and other multivariable studies, their accuracy is crucial for drawing reliable conclusions that can contribute to building cumulative knowledge. In a comparison of the accuracy of just four linguistic features as identified by three different taggers, Picoral *et al.* (2021) reported differences large enough to lead to significantly different conclusions. Despite this, Goulart and Wood (2021: 123) concluded that tagger accuracy is underreported in the majority of published MD studies. This is not surprising, as the large number of linguistic features typically entered in such analyses makes comprehensive evaluations of tagger

accuracy an arduous task. In the context of most small- to mid-scale projects, it is simply not feasible. This is why we believe that, to meet our aim for providing the research community with an accessible tagger, it is necessary to conduct and publish the results of a comprehensive evaluation of the accuracy of the MFTE.

When evaluating the performance of a tagger, several accuracy measures can be used, as shown in Table 3 below. On the one hand, the number of correct tags, i.e., true positives, can be counted and, on the other hand, the number of incorrect tags, i.e., false positives. The simplest measure of accuracy is ‘precision’: the ratio of true positives to all tags assigned by the tagger. ‘Recall’, by contrast, is the ratio of true positives to all instances of a given tag in the data. It takes into account both true positives and false negatives (instances where a particular tag should have been assigned by the tagger but was not). In other words, recall provides a measure of the tagger's ability to correctly identify and classify all instances of a particular feature. While both precision and recall provide valuable information about a tagger's performance, they only give a partial picture of its accuracy. If precision is high but recall is low, the tagger may be too conservative, tagging only those instances about which it is particularly confident. Conversely, a tagger with low precision but high recall will assign some tags too liberally.

Term	Definition
True positive	Feature correctly identified as X
True negative	Feature correctly identified as not X
False positive	Feature incorrectly identified as X
False negative	Feature incorrectly identified as not X
Precision	$\text{True positive count} / (\text{true positive count} + \text{false positive count})$
Recall	$\text{True positive count} / (\text{true positive count} + \text{false negative count})$
F1 score	$2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$

Table 3: Summary of the terminology frequently used in tagger performance evaluations

In an ideal world, a tagger would have 100 per cent precision (i.e., all assigned tags are correct) and 100 per cent recall (i.e., all features are labelled with all the correct tags). In practice, however, attempts to increase precision on any one feature will usually result in lower recall rate for that particular feature (and many others) and vice versa. An important aspect of tagger development therefore consists in striking the appropriate balance between precision and recall. If it is feasible to complement automatic tagging with a manual fix-tagging phase, then it makes sense to prioritise recall. If, however, a tagger is to be used without any manual intervention, both precision and recall are important. This

is why a third accuracy measure is often calculated: the F1 score, which combines precision and recall (see Table 3). It ranges from 0 (entirely inaccurate) to 1 (perfectly accurate) and provides a single number that can be used to compare the performance of different taggers.

A common strategy in POS tagger evaluations is to report one overall measure of per-token accuracy across all tags (typically slightly over 97%, e.g., Manning 2011). This is common practice in NLP but invariably leads to misleadingly inflated results. Indeed, some of the most frequent tokens, in particular punctuation markers and determiners, are both highly frequent and extremely easy to ‘get right’ (e.g., *the* and *I*). By contrast, per-sentence accuracy rates of POS taggers tend to be considerably more modest (hovering around 50–57%) and considerably lower rates for non-standard varieties and registers for which there is little training data (Manning 2011).

Although it is crucial for users of a tagger to be aware of its per-feature accuracy to gauge the reliability of its annotation, few taggers have documentation that includes detailed results of a comprehensive accuracy evaluation study. The *Biber tagger* constitutes an exception. We are aware of two comprehensive evaluations of its tagging accuracy: Biber and Gray (2013: 16–18) and Gray (2015). The first, however, only reports post-fix-tagging precision and recall rates (Biber and Gray 2013: Appendices C and D). In other words, users of the *Biber tagger* who do not have access to the fix-tagging scripts used as part of this project cannot expect similar accuracy levels when tagging their own texts. Moreover, in Biber and Gray (2013: 18) the same 5 per cent sample of texts that were originally manually checked for tagging errors was analysed to calculate the reported precision and recall rates after fix-tagging. This procedure entails the risk of inflating the tagger’s accuracy metrics as the fix-tagging scripts may have been constructed based on errors specific to the sample. By contrast, Gray (2015) reports both “initial reliability rates” of the uncorrected *Biber tagger* output and “final reliability rates after [fix-tagging] scripts” (Gray 2015: Appendix B) on excerpts of a random sample of 15 research articles across different disciplines and registers. On close inspection of the initial reliability rates, it transpires that not all of the features of the *Biber tagger* (or rather the version used in this particular project) seem suitable for use without some manual or semi-automatic ‘fix-tagging’ procedure, as recall and precision rates for some features are well below 90 per cent.

An anonymous reviewer drew our attention to the fact that several recent PhD projects conducted at Northern Arizona University (the ‘home’ of the *Biber tagger*) include small-scale evaluations of the accuracy of more recent versions of the *Biber tagger* for selected features. Goulart (2022: Appendix A) reports recall and precision rates for 29 features in L1 and L2 academic writing. Wood (2023: 59–63) does so for 28 features in a sample of statutory texts, while Dixon (2022: 71–72) evaluated the tagger’s accuracy for 20 features in 17 texts representing gaming language. Before fix-tagging, Goulart (2022: Appendix A) reports precision and recall rates well above 90 per cent for all the selected features except nominalisations. In Wood’s (2023) corpus of statutory language, on the other hand, accuracy is much lower for many of the selected features (e.g., noun and adjective complement clauses, agentless passives, clausal and phrasal coordinated conjunctions). These results confirm that the accuracy of taggers varies widely across different linguistic features and depends on the type of texts being tagged. While it is inevitable that such labour-intensive evaluations cannot realistically be carried out on large samples, it is problematic that most of these evaluations do not report the number of tags on which recall and precision were calculated. This means that confidence intervals around the reported accuracy metrics cannot be computed (see Section 4.2). Dixon’s (2022: 72) evaluation constitutes an exception: the figures reveal that some of the reported accuracy metrics were calculated on as few as three or four occurrences, thus confirming that the results of such small-scale, project-specific evaluations cannot meaningfully be used in other projects.

The documentation of the *MFTE Perl* (Le Foll 2021) includes a detailed evaluation of the tagger on a stratified random sample from the *British National Corpus 2014* (BNC2014; Brezina *et al.* 2021). Samples of 24 texts covering written and spoken British English from a range of registers were tagged using the *MFTE Perl*. The resulting 31,311 tags were manually annotated by two linguists. For each tag, the annotators had three options: 1) mark it as correct, 2) mark it as incorrect and assign the correct tag, or 3) mark it as an unclear or ambiguous context/token for which no tag could reliably be assigned. Most of the 75 features of the *MFTE Perl* achieved a “satisfactorily high degree of accuracy” (Le Foll 2021: 14) with high rates of both recall and precision. F1 scores below 90 per cent were reported for 10 features. These include two rare features for which Le Foll (2021: 45) warns that accuracy rates are unreliable because they are based on very few occurrences.

4. EVALUATION

In the following, we report on the accuracy of the simple tagset of the MFTE on British and American English texts from a range of registers. Given how misleading overall accuracy rates may be (see Section 3), we report recall, precision, and the combined F1 measure for each feature of the simple tagset, thus allowing users to make an informed decision as to which set of features they wish to include in their analyses and which they would like to exclude depending on their research questions and the characteristics of their data. Whenever possible, we also report 95 per cent confidence intervals around these accuracy metrics.

4.1. Data

In line with the specifications of the MFTE (see Section 2.2), for the evaluation of the *MFTE v.1.0*, we chose a stratified random sample of 30 texts from the BNC2014 and 30 texts from the *Corpus of Contemporary American English* (COCA; Davies 1990) representing diverse registers, as shown in Table 4. To maximise the potential for text/topic-specific tagging errors, we analysed samples of roughly 1,000 tokens from the longer texts in our evaluation sample. This was motivated by an observation from the evaluation of the *MFTE Perl* in which most tagging errors were found to be text/topic-specific and therefore highly clustered (see Le Foll 2021: 25–43). The sampled texts from the spoken subcorpus of the BNC2014 were pre-processed to remove the anonymisation tags and metadata following the procedure documented in Le Foll (2021: 28).

Corpus	Subcorpus	Number of texts	Number of tags
BNC2014	Academic writing	3	2,617
BNC2014	E-Language: Blogs	3	2,092
BNC2014	E-Language: E-Mails	2	1,964
BNC2014	E-Language: Forums	2	2,822
BNC2014	E-Language: Reviews	2	3,291
BNC2014	E-Language: Social Media Posts	3	3,263
BNC2014	E-Language: Text Messages	1	396
BNC2014	Fiction	3	4,144
BNC2014	News: Magazines	2	2,363
BNC2014	News: Newspapers	6	5,036
BNC2014	Spoken: Conversation	3	5,312
COCA	Academic writing	3	2,877
COCA	E-Language: Blogs	3	3,337
COCA	E-Language: Web Pages	3	2,416
COCA	Fiction	3	3,129
COCA	News: Magazines	3	2,202
COCA	News: Newspaper Articles	4	2,342
COCA	News: Newspaper Opinion Pieces	4	2,993
COCA	Spoken: Conversation	4	6,015
COCA	Spoken: TV/Movies	3	2,541
Totals		60	61,154

Table 4: Evaluation data

4.2. Methodology

The 60 files tallied in Table 4 were tagged using the *MFTE Python v.1.0* with its simple tagset. The resulting tagged text files were then converted to a spreadsheet format for manual evaluation.¹⁰ Each tag was marked as either correct, unclear, or incorrect. In the case of an incorrect tag, a corrected version of the tag was added to the corresponding column. In addition, tags were added where they were missing.¹¹ These 60 spreadsheet files were subsequently processed and merged using custom *R* functions.

¹⁰ The evaluation was performed by the first author and her research assistant, Tatjana Winter, whom we thank for her meticulous work.

¹¹ For details of the procedure, see Le Foll (2021: 29–33) and Le Foll and Shakir (2023).

Statistics and data visualisations were computed in *R* and *Python* (see [GitHub repository](#) for data and code). Bootstrap simulation was used to calculate 95 per cent confidence intervals (CI) for the precision, recall and F1 measures of each feature based on the results of 1,000 bootstrapped samples (in a procedure inspired by Picoral *et al.* 2021). We performed this task in *Python* and applied it to all tags for which there were more than 100 occurrences in the 60 evaluation files (see Table 4).

4.3. Results

Of the 61,154 manually reviewed tags, 294 (0.48%; 95% CI [0.43–0.54%]) were deemed by the human annotators to be ‘unclear’ due to misspellings, text processing errors, ambiguous contexts, or fragmented sentences (particularly in the COCA data where 5% of each text was removed for copyright reasons). A further 8,506 tags were excluded from the evaluation metrics because they are not tallied in the tables of frequencies generated by the MFTE. As these include all the tags denoting punctuation marks, as well as foreign words, symbols, and other non-word tokens, they would also have considerably inflated the overall accuracy metrics. Excluding these, the number of correctly assigned tags across the 74 linguistic features of the MFTE simple tagset was 51,139. This corresponds to an overall precision of 97.13 per cent (95% CI: 96.99–97.23%) across the 60 evaluation files. Figure 3 shows the recall, precision and F1 measures for each feature of the simple tagset with at least 100 occurrences in the evaluation corpus. The error bars correspond to bootstrapped 95 per cent CI. The colours indicate how many times each tag occurred across the 60 evaluation files (note that the colour scale is logarithmic).

Twelve features from the simple tagset did not meet the precision and/or recall rates of at least 90 per cent stipulated in the tagger specifications (see Section 2.2), although five of these do have F1 scores above 90 per cent. The least accurate features are stranded prepositions (STPR), verbs in the imperative (VIMP), *that* omission (THATD), WH-questions (WHQU), non-finite past participle forms (VBN), and *GET*-passives (PGET).

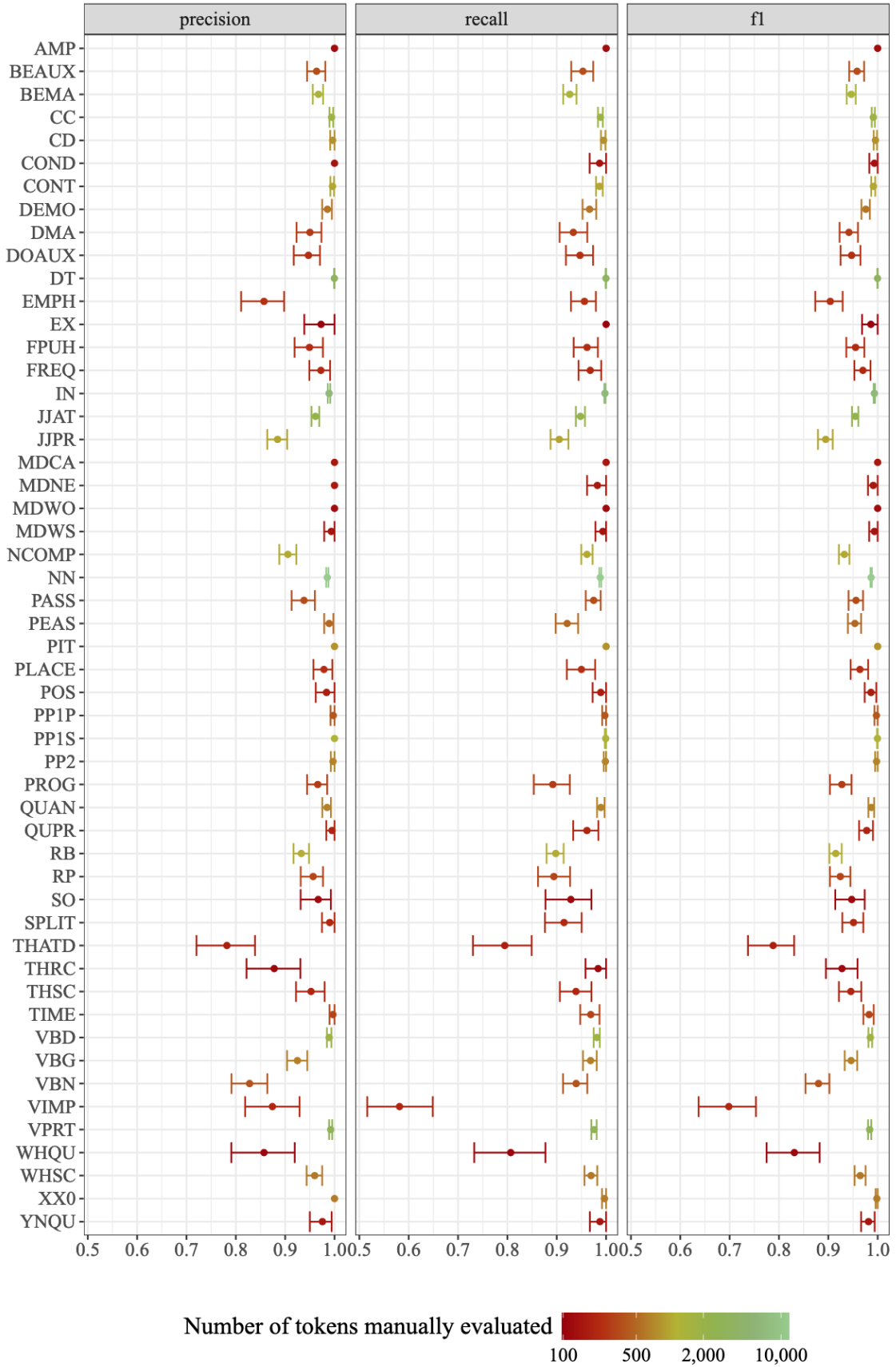


Figure 3: Precision, recall, and F1 score with bootstrapped 95 per cent CI for features with ≥ 100 occurrences across the 60 evaluation files

Figure 4 displays the most frequent tagger errors as clusters of red points. The y -axis lists the tags as assigned by the MFTE, whilst the x -axis corresponds to those assigned by the two human annotators. For the sake of readability, only the most frequent tags are included in Figure 4. The code provided in the repository may be used to compute and examine the full matrix. The figure shows that notable clusters of errors involve the confusion of infinitives (VB) vs. imperatives (VIMP), attributive (JJAT) vs. predicative (JJPR) adjectives, non-finite past participles (VBN) vs. finite verbs in the perfect aspect (PEAS), and WH-questions (WHQU) vs. WH-subordinate clauses (WHSC). We can also see that several clusters of red points involve the tag 'NONE'. NONE is not part of the tagset. We used this tag as a placeholder to indicate that the MFTE assigned an unwarranted second or third-order tag to this token (e.g., when two adjacent nouns were incorrectly identified as a noun compound by the tagger), or to indicate that a necessary tag was omitted (e.g., if *am* in *I am happy* was assigned the present tense tag, but not the one for *BE* as a main verb).

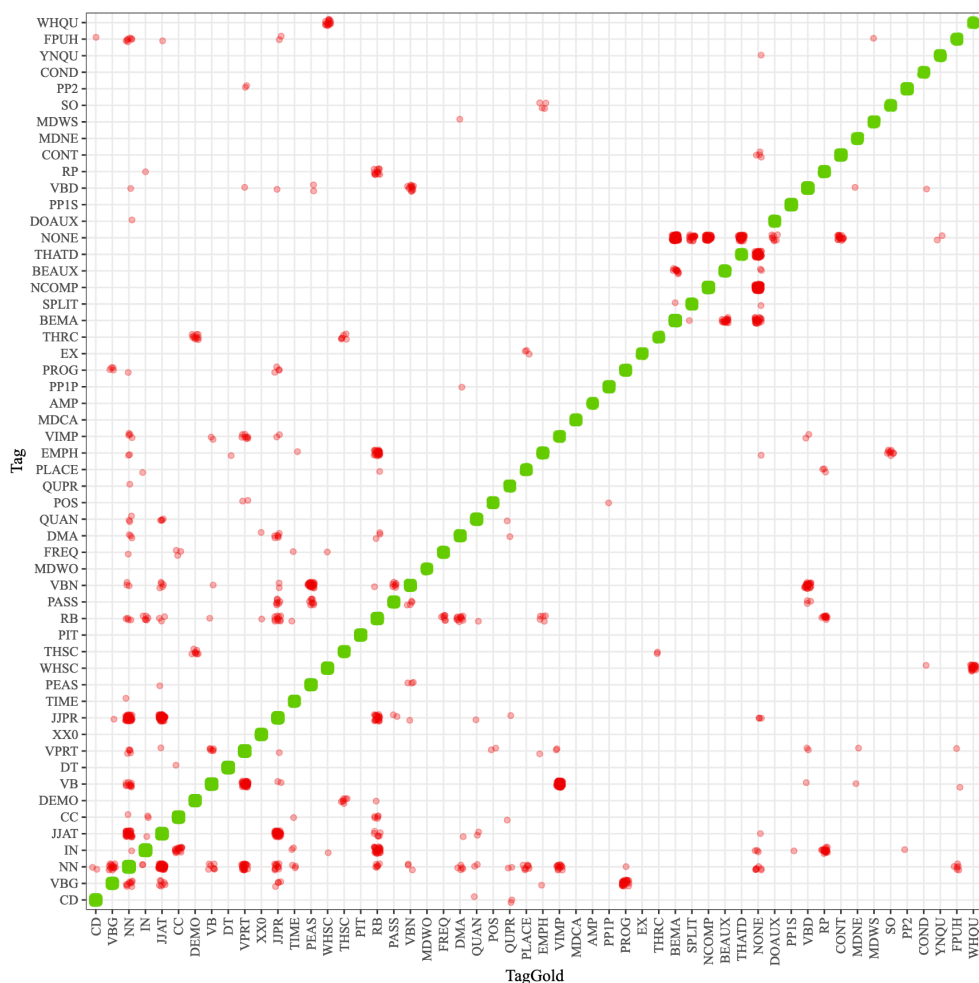


Figure 4: Confusion matrix showing mismatches between the MFTE output (Tag) and the human annotators' assessments (TagGold) in red

As was to be expected, the MFTE performs slightly less well on internet and spontaneously spoken registers than on professionally written and edited texts such as those typically found in academic writing and fiction (see Table 5). For most purposes, however, these differences are unlikely to be significant. The code provided in the repository also allows for the calculation of per-feature accuracy measures for each of the six broad register categories of the BNC2014 and COCA.

Register	Accuracy	Lower CI	Upper CI
Academic writing	97.83	97.37	98.23
E-language	96.56	96.27	96.83
Fiction	97.89	97.49	98.24
News writing	97.77	97.50	98.02
Spoken	96.62	96.28	96.94

Table 5: Overall precision of the MFTE across the broad register categories

5. DISCUSSION

Based on the results of the evaluation, we now examine the extent to which the MFTE meets the tagger specifications elaborated in Section 2.2. We further discuss the limitations of both the evaluation and the tagger itself.

With a total of 74 ‘simple’ and over 70 additional ‘extended’ features covering fine-grained part-of-speech classes, tenses and aspects, frequent lexico-grammatical constructions, and numerous semantic categories, the MFTE’s feature portfolio meets the first criterion of identifying a broad range of lexico-grammatical and semantic features of English (see Table 1 in Section 2.2). Aside from the SO and LIKE features (see Section 2.4) they can readily be meaningfully interpreted (criterion 1a). Of the features in the simple tagset, all but twelve features reached “a satisfactorily high degree of accuracy” (see Table 1), with both precision and recall rates of > 90 per cent (criterion 1b). That said, it is worth remembering that some of the per-feature precision and recall rates presented in this study are based on relatively very few data points (see section 4.3).

Some of the most frequently mistagged features are among the problematic features for which some users of even the more recent versions of the *Biber tagger* have reported performing (semi-)automatic fix-tagging (e.g., features involving various uses of *that* and non-finite *-ed* and *-ing* clauses, see Goulart 2022 or Wood 2023). Low accuracy metrics

for some of these features may be the price of meeting specification criterion 1c: feature identification “without the need for human intervention” (see Section 2.2). For some features, we consider some of relatively low accuracy rates to be necessary to meet criterion 1a: if linguistic features are to be “meaningful” (in the sense of functionally and linguistically interpretable), it is crucial to distinguish, for example, non-finite past participle forms (VBN) from finite verbs in the perfect aspect (PEAS), and imperatives (VIMP) from infinitives (VB). Many existing taggers do not do this and therefore achieve higher accuracy rates overall, but the features are less readily interpretable. In sum, the results of the evaluation suggest that, at least for the simple tagset, most feature counts can be entered into multivariable analyses “without the need for human intervention” (1c).

While the MFTE struggled with e-language most (especially text messages and social media posts), the results of our evaluation nevertheless confirm that the MFTE performs sufficiently well across “a broad range of English registers” (thus meeting criterion 1d) with texts in “standard American or British orthography” (thus meeting criterion 1e). We recommend that researchers carry out additional evaluation tests if they intend to use the MFTE for other registers and/or varieties of English. Particular care should be taken when applying the extended tagset of the MFTE, as the accuracy of these features has not been subject to such a systematic evaluation.

The outputs of the MFTE also conforms to the tagger specification. Not only does the MFTE produce a table of raw counts (thus allowing researchers to apply their own normalisations) and two tables of normalised counts (criterion 2b), but it also saves the tagged texts for detailed examination of the texts themselves (criterion 2a). This is important to ensure full transparency of the tagging process and to verify the accuracy of the tagger’s output.

To meet the final requirement of the tagger specification, the MFTE source code and all the additional evaluation materials are available under a GPL-3.0 licence for use and scrutiny by the research community (criterion 3a). Criterion 3b is also satisfied with the publication of step-by-step instructions on how to install and run the MFTE on the landing page of its [GitHub repository](#), together with the present article describing the development and evaluation of the tagger.

6. CONCLUSION

In this article, we have presented the *Multi-Feature Tagger of English* (MFTE), an open-source tool designed for multivariable analyses of English. Characterised by transparency, adaptability, and accessibility, the MFTE offers promising avenues for future research endeavours in line with the principles of Open Science. Unlike ‘standard’ POS taggers such as the *Stanford tagger* (Toutanova *et al.* 2003) or CLAWS (Leech *et al.* 1994; Rayson and Garside 1998), its output does not require any additional processing (i.e., it outputs tables of counts with different normalisation options) and aims to tag only linguistically meaningful, functionally relevant features. As a free tool, the MFTE can contribute to making multivariable analysis of English more accessible to researchers and students from institutions with fewer resources.

While MAT and the initial version of the MFTE are in *Perl*, the new MFTE runs in *Python*, a programming language increasingly familiar to linguists. We hope that both the use of an accessible, object-based language and of an open-source licence will encourage colleagues not only to make use of the tagger, but also to contribute improvements. Our hope is that, as the tool gains traction within the research community, collaborative efforts will lead to further enhancements, expanding the tagset and refining the tagger’s performance across varieties and text types. Compared to the *MFTE Perl*, this new *Python* version benefits from being written in an actively developed language with a large user base, thus facilitating updates and improvements from developers (e.g., the integration of additional features using native parser libraries in *Python*).

In addition, the transparency and accessibility of the MFTE may also inspire linguists to develop similar taggers for languages other than English. Indeed, although MD analysis is, in theory, applicable to any language, in practice, most MD studies to date have examined varieties of English, which we believe is in part attributable to the lack of open-source taggers of lexico-grammatical features for languages other than English.

In discussing the results of the extensive evaluation of the MFTE’s simple tagset, we have also acknowledged the limitations of the tagger. Its accuracy will undoubtedly vary with respect to registers, topic domains, and varieties of English not included in the evaluation corpus. The extended tagset has not yet been systematically evaluated, and its reliance on dictionary lists for the semantic features is clearly a limitation. Refining and

updating these lists will be essential for the continued accuracy of the tagger. Future studies could explore the generation of tailored test data using Large Language Models (LLMs) as a means of evaluating the precision and recall rates of infrequent linguistic features. Finally, although the results of the evaluation are considerably more detailed than those of most linguistic taggers, they should nonetheless be interpreted with an awareness of the inherent challenges of making objective, categorical judgments when interpreting complex and often ambiguous linguistic phenomena. It would be misleading to suggest that these judgments are theory-free. As Gray (2019: 45) points out in an article focusing on the tagging and counting of linguistic features for MD analysis, “conflicting POS categorisation reflects a different grammatical interpretation or theory of the nature of this word.”

With these considerations in mind, we also hope that the MFTE will not only make a significant contribution to multivariable corpus linguistics research, but also stimulate ongoing methodological discussions on the transparency, validity, and reliability of the tools and methods used in corpus linguistics research. Ultimately, we hope that, in the near future, making research materials, data, and code available alongside linguistics publications will no longer be the exception (Wieling *et al.* 2018; Bochynska *et al.* 2023), but the norm.

REFERENCES

- Baker, Monya. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533/7604: 452–454.
- Barlett, Tom and Gerard O’Grady eds. 2017. *The Routledge Handbook of Systemic Functional Linguistics*. London: Routledge.
- Berber Sardinha, Tony, Marcia Veirano Pinto, Cristina Mayer, Maria Carolina Zuppari and Carlos Henrique Kauffmann. 2019. Adding registers to a previous multi-dimensional analysis. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multidimensional Analysis: Research Methods and Current Issues*. New York: Bloomsbury, 165–188.
- Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton and Peter Pulsifer. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56/1: 1–18.
- Biber, Douglas. 1984. *A Model of Textual Relations within the Written and Spoken Modes*. California: University of Southern California dissertation.
- Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, Douglas. A typology of texts. *Linguistics* 27: 3–43.

- Biber, Douglas. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.
- Biber, Douglas. 2019. Multidimensional Analysis: A historical synopsis. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic, 11–26.
- Biber, Douglas and Jesse Egbert. 2018. *Register Variation Online*. Cambridge: Cambridge University Press.
- Biber, Douglas and Bethany Gray. 2013. Discourse characteristics of writing and speaking task types on the TOEFL IBT test: A lexico-grammatical analysis. ETS Research Report Series 2013/1. <https://doi.org/10.1002/j.2333-8504.2013.tb02311.x>.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Bochynska, Agata, Liam Keeble, Caitlin Halfacre, Joseph V. Casillas, Irys-Amélie Champagne, Kaidi Chen, Melanie Röthlisberger, Erin M. Buchanan and Timo B. Roettger. 2023. Reproducible research practices and transparency across linguistics. *Glossa Psycholinguistics* 2/1. <https://doi.org/10.5070/G6011239>.
- Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.
- Brezina, Vaclav, Abi Hawtin and Tony McEnery. 2021. The written British National Corpus 2014 – design and comparability. *Text & Talk* 41/5–6: 595–615.
- Davies, Mark. 1990. *Corpus of Contemporary American English (COCA)*. <https://www.english-corpora.org/coca/>.
- Dixon, Daniel Hobson. 2022. *The Language in Digital Games: Register Variation in Virtual and Real-World Contexts*. Flagstaff: Northern Arizona University dissertation.
- Egbert, Jesse, Tove Larsson and Douglas Biber. 2020. *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. Cambridge: Cambridge University Press.
- Egbert, Jesse and Shelley Staples. 2019. Doing multi-dimensional analysis in SPSS, SAS, and R. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic, 125–144.
- Gewin, Virginia. 2016. Data sharing: An open mind on open data. *Nature* 529/7584: 117–119.
- Goulart, Larissa. 2022. *Communicative Text Types in University Writing*. Flagstaff: Northern Arizona University dissertation.
- Goulart, Larissa and Margaret Wood. 2021. Methodological synthesis of research using multi-dimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science* 6/2: 107–137.
- Gray, Bethany. 2015. *Linguistic Variation in Research Articles: When Discipline Tells only Part of the Story*. Amsterdam: John Benjamins.
- Gray, Bethany. 2019. Tagging and counting linguistic features for multi-dimensional analysis. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic, 43–66.
- Gray, Bethany and Douglas Biber. 2013. Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics* 18/1: 109–135.
- Hidalgo, Bertha and Melody Goodman. 2013. Multivariate or multivariable regression? *American Journal of Public Health* 103/1: 39–40.

- In'nami, Yo, Atsushi Mizumoto, Luke Plonsky and Rie Koizumi. 2022. Promoting computationally reproducible research in applied linguistics: Recommended practices and considerations. *Research Methods in Applied Linguistics* 1/3. 100030. <https://doi.org/10.1016/j.rmal.2022.100030>.
- John, Leslie K., George Loewenstein and Drazen Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23/5: 524–532.
- Le Foll, Elen. 2021. *Introducing the Multi-Feature Tagger of English (MFTE)*. Perl. Osnabrück University. <https://github.com/elenlefol/MultiFeatureTaggerEnglish>.
- Le Foll, Elen. 2024. *Textbook English: A Multi-Dimensional Approach*. Studies in Corpus Linguistics 116. Amsterdam: John Benjamins.
- Le Foll, Elen and Muhammad Shakir. 2023. *Introducing a New Open-Source Corpus-Linguistic Tool: The Multi-Feature Tagger of English (MFTE)*. Paper presented at the 44th International Computer Archive of Modern and Medieval English Conference. NWU Vanderbijlpark: South Africa.
- Leech, Geoffrey, Roger Garside and Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational Linguistics*. Kyoto: Association for Computational Linguistics, 622–628.
- Lindquist, Hans. 2009. *Corpus Linguistics and the Description of English (Edinburgh Textbooks on the English Language – Advanced)*. Edinburgh: Edinburgh University Press.
- Manning, Christopher D. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In Alexander F. Gelbukh ed. *Computational Linguistics and Intelligent Text Processing*. Berlin: Springer, 171–189.
- McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Taylor & Francis.
- McManus, Kevin. 2024. Replication and open science in applied linguistics research. In Luke Plonsky ed. *Open Science in Applied Linguistics*. Applied Linguistic Press, 148–165.
- Nini, Andrea. 2014. *Multidimensional Analysis Tagger (MAT)*. <http://sites.google.com/site/multidimensionaltagger>.
- Nini, Andrea. 2019. The multi-dimensional analysis tagger. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multi-Dimensional Analysis: Research Methods and Current Issues*. New York: Bloomsbury, 67–96.
- Pashler, Harold and Eric-Jan Wagenmakers. 2012. Introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science* 7/6: 528–530.
- Picoral, Adriana, Shelley Staples and Randi Reppen. 2021. Automated annotation of learner English: An evaluation of software tools. *International Journal of Learner Corpus Research* 7/1: 17–52.
- Porte, Graeme and Kevin McManus. 2018. *Doing Replication Research in Applied Linguistics*. Milton Park: Routledge.
- Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv*. <https://doi.org/10.48550/arXiv.2003.07082>.
- Rayson, Paul and Roger Garside. 1998. The CLAWS web tagger. *ICAME Journal* 22: 121–123.
- Sinclair, John McH., Gwyneth Fox, Stephen Bullon, Ramesh Krishnamurthy, Elisabeth Manning and John Todd eds. 1990. *Collins Cobuild English grammar: Helping learners with real English*. Glasgow: Harper Collins.

- Sönning, Lukas and Valentin Werner. 2021. The replication crisis, scientific revolutions, and linguistics. *Linguistics* 59/5: 1179–1206.
- Toutanova, Kristina, Dan Klein, Christopher D. Manning and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Marti Hearst and Mari Ostendorf eds. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Edmonton: Association for Computational Linguistics, 173–180.
- Wallis, Sean. 2020. *Statistics in Corpus Linguistics Research: A New Approach*. London: Routledge.
- Wieling, Martijn, Josine Rawee and Gertjan van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics* 44/4: 641–649.
- Wood, Margaret. 2023. *Communicative Function and Linguistic Variation in State Statutory Law*. Flagstaff: Northern Arizona University dissertation.

Corresponding author

Elen Le Foll
University of Cologne
Department of Romance Studies
Universitätsstraße 22
50937 Cologne
Germany
E-mail: elefoll@uni-koeln.de

received: December 2023
accepted: September 2024