Introducing the Multi-Feature Tagger of English (MFTE)

Version 3.0

Elen Le Foll

Please cite as:

Le Foll, Elen. 2021. Introducing the Multi-Feature Tagger of English (MFTE) version 3.0. https://github.com/elenlefoll/MultiFeatureTaggerEnglish.

Published on: 31 December 2021

Table of Contents

[N]ichtreproduzierbare Einzelereignisse sind, wie wir schon mehrfach erwähnt haben, für die Wissenschaft bedeutungslos [...].
– Karl Popper, 1935

1 Introduction

This document is intended as documentation for the Multi-Feature Tagger of English (hereafter: the MFTE) – a new automatic tagger for the analysis of situational variation in standard written and spoken general English. The MFTE was originally developed for use in multi-feature/multi-dimensional analysis (Biber 1984; 1988; 1995; Conrad & Biber 2013) (hereafter: MDA), a widely used framework first developed by Douglas Biber in the late 1980s. In short, MDA is based on the theoretical assumption that register-based variation can be observed as differences in patterns of co-occurring lexico-grammatical features, which result from texts having register-specific communicative goals and contexts of use (Biber & Conrad 2001; cf. Hymes 1984).

Since the co-occurrences of very many different linguistic features need to be counted across hundreds or thousands of texts for MDA to be feasible, automatic feature taggers can be said to constitute the backbone of the MDA framework. Most MDA studies focus on grammatical and lexico-grammatical features that, as opposed to surface lexical features such as those captured by n-grams, require the texts of the corpus to be part-of-speech (POS) tagged[1] (cf., however, Crossley et al. 2007; Bohmann 2017). To date, the majority of English MDA studies based on (lexico-)grammatical features have relied on the Biber Tagger,[2] although, in principle, any other (POS-)tagger could be used. In a synthesis of MDA studies published in English or Portuguese between 1984 and April 2020, Goulart & Wood (2021) report that 90 out of the 210 studies they surveyed reported using the Biber Tagger. However, at the time of writing, the key limitation of the Biber Tagger lies in the fact that it is not (freely) available to the wider research community. The fact that its source code is not available makes it very difficult for researchers not affiliated with Douglas Biber to replicate the findings of MDAs conducted on the basis of this tagger.

As a result, Andrea Nini (2014) attempted to replicate the 1988 version of the Biber Tagger and released the Multidimensional Analysis Tagger (MAT) as freeware. It was subsequently made available under an open-source license on GitHub[3] in August 2020. The availability of this software (which features a GUI for Windows) coincides with a rise in the use of MDA outside of the Biber Tagger's home institution, Northern Arizona University. Although Goulart & Wood's (2021) research synthesis only uncovered 25 studies (12%) that reported using the MAT, this number can be expected to rise thanks to the recent publication of the code on GitHub, which no longer means that the MAT functions as a black box, and the recent publication of a paper on the tool (Nini 2019).

The MFTE was partially built on the basis of Nini's MAT. Like the MAT, it is available under a GPL-3.0 License and is available on GitHub.[4] The present document outlines the steps involved in the development of the MFTE. Section 2.1 outlines its specifications, which were drawn up on the basis of the features needed to carry out MDA and taking account of the advantages and limitations of existing taggers (see Le Foll 2021a: chap. 3). The following sections explain the methodological decisions involved in the selection of the features to be

[1] POS-tagging is also referred to as "grammatical tagging" or "morpho-syntactic annotation" (cf. McEnery et al. 2006, 34).
[2] Though most publications refer to "*the* Biber Tagger", the tagger has in fact been continuously improved since it was first developed by Douglas Biber in the late 1980s, hence the denomination "Biber Tagger" refers to a whole family of related software and the results of MD analyses based on "*the* Biber Tagger" are therefore not always guaranteed to be comparable.
[3] https://github.com/andreanini/multidimensionalanalysistagger
[4] https://github.com/elenlefoll/MultiFeatureTaggerEnglish

identified by the MFTE (2.2), the details of the regular expressions used to identify these features (2.3) and the procedure for normalising the feature counts (2.4). Section 2.5 describes the outputs of the tagger. Chapter 3 then goes on to evaluate the accuracy of the MFTE. It reports the results of comparisons of the tags assigned by the MFTE and by two human annotators to calculate precision and recall rates for each linguistic feature across a range of contrasting text registers.

"Although its importance is not widely recognized, the computer program used for grammatical tagging provides the foundation for MD studies."
- Douglas Biber, 2019

2 Development

2.1 Specifications of the MFTE

The aim was to develop a new automatic tagger, which can be used to identify a broad range of grammatical, lexical, and semantic features for the multivariable analysis of linguistic variation in English. Given that modern corpora are often very large and that, for within the MDA framework, many different linguistic features are to be identified and counted, time and resource constraints mean that quantitative multivariable analyses such as MDA usually rely on large sets of linguistic features that can be relatively reliably retrieved using automated queries. The fact that some of Biber's (e.g., 1988 and 2006) features require (semi-)manual annotation has been highlighted as a weakness in the past:

> While Biber's multidimensional approach is doubtlessly the most sophisticated approach [to register analysis] to date, it also involves much supervised, semi-manual, if not completely manual, annotation, which makes it difficult to apply to large or constantly changing corpora (Gries et al. 2011: n.p.).

Studies, e.g., Le Foll (2021b), have indeed shown that this constitutes one of the limitations to applying additive MDA with the MAT. Given the huge number of texts analysed in most MDA studies, it is simply not feasible to manually correct the output of the tagger, even for just a few features. Thus, in developing this new tagger, the selection of linguistic features was also constrained by the feasibility of automatically extracting these features to a satisfactorily high degree of accuracy. An added difficulty consisted in the need to ensure that this automatic extraction process would work equally well for a broad range of registers, including transcriptions of spoken English and internet registers such as social media posts, chats, and forums.

Though the tagger was foremost developed with the MDA framework in mind, its output could be used as the basis for any multivariable method. Crucially, however, its intended application

The tagger ought to:

1. Identify a broad, as comprehensive as possible, range of lexico-grammatical features of English
    a. that can each be meaningfully interpreted
    b. to a satisfactorily high degree of accuracy (precision and recall rates of > 90%)
    c. without the need for human intervention
    d. in a broad range of English registers with standard American or British orthography
    e. to examine register variation using multivariable methods in a broad range of general English registers.
2. Output both raw and normalised counts per text in a standard format.
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. in an accessible format with adequate documentation for researchers with basic computer skills to be able to run the programme.

for multi-dimensional analysis as opposed to classification tasks (see means that it aims to identify linguistic features that can be meaningfully interpreted and that are therefore linguistically motivated. Bearing in mind the advantages and limitations of the currently most widely used taggers for MDA (see Le Foll 2021a: chap. 3), the aim was to develop a tagger that meets the following specifications:

The rest of this chapter outlines and justifies the multiple decisions made to arrive at a new tagger that attempts to fulfil these specifications.

2.2 Feature selection

As with all unsupervised approaches based on text similarities across sets of quantitative features, the results of MDAs depend foremost on the choice (i.e., validity) and operational reliability of the features entered in the factor analysis. This was already pointed out in one of the first reviews of Biber's 1988 publication:

> It is obvious that a method that hinges on statistically determined patterns of co-occurring features will be very sensitive to the selection and identification of these features. If the features are ill-defined, functionally heterogeneous, stylistically skewed, etc., this is likely to have an immediate effect on the results (Altenberg 1989: 171).

Altenberg (1989) and others after him have pointed out that some of Biber's features and, in particular, operationalisations (at least, as outlined in the 1988 publication), are potentially problematic. For example, Le Foll (2021b) showed that the features relying on punctuation cannot be reliably applied to transcriptions of spoken English. As Picoral et al. (2019) have demonstrated: the operationalisations of linguistic features are, however, crucial to yield valid and reliable results. In many respects, this is now much easier than it was in the 1980s: computer power has expanded exponentially and POS-taggers are now considerably more reliable. However, it should also be stressed that linguistic features need not necessarily be captured by rule-based algorithms in the form of regular expressions that (partly) rely on POS tags as in the Biber Tagger and its replications. Crossley (2007), Bohmann (2017) and others have demonstrated that MDAs can successfully be carried out with sets of features that do not rely on POS-tagging. Such an approach can be advantageous because it is less computationally expensive, less error-prone, more portable, and therefore potentially easier to reproduce. In addition, POS-taggers are by no means 'linguistically agnostic', so that approaches that do not rely on them can also be argued to be more objective in their choice and/or operationalisations of linguistic features.

Biber (1988) and most other scholars applying MDA after him selected their sets of linguistic features from the results of previous studies and their intuition as to what may be potentially relevant to their research questions. Diwersy et al. (2014: 176) rightly warn of a risk of circularity in such top-down feature selection procedures:

> This is especially problematic if features are derived from theoretical assumptions about patterns of interest (subverting the inductive aims of the research) or if a researcher attempts to select features that steer the analysis towards an expected result (fearing that it might otherwise be hidden by more prominent variational patterns).

In a similar vein, McEnery & Hardie (2011: 114) suggest "that the MD methodology could be more solidly founded if based on a selection of features which is both *principled* and *exhaustive*". Whilst they provide a fragment of a feature tree that aims to exemplify this approach (see Fig. 1, which, incidentally, only lists features that Biber included in his original model), it is unclear how it would lead to a more principled and exhaustive list of features since the selection of features itself remains entirely "researcher-driven" (Taylor 2010: 223–224); it can merely be claimed that the feature tree representation allows for a more systematic organisation of the features. Taking a more pragmatic approach in discussing this issue, Szmrecsanyi (2013: 434) concludes:

Because there is typically a large to infinite number of linguistic features that could be used to characterize a language or language variety, often a choice has to be made when defining a feature portfolio. A certain degree of subjectivity is, therefore, alas, inevitable.

'Meaning'

- Actions (verbs)
    - Auxiliary forms
        - Aspect
        - Voice
            - Passives with by-phrase
            - Passives without by-phrase
        - Modality
    - Verb forms
        - Past
        - Present
    - Verb modification
- Entities (nouns)
    - Noun forms
        - Proper nouns
        - Pronouns
        - Derived nouns (nominalisations)
        - Plurals
    - Noun modification
        - Adjectives
        - Articles
        - …

Fig. 1: A fragment of a feature tree for English (as suggested in McEnery & Hardie 2011: 114)

Partly as a reaction to criticism of researcher-driven feature selection processes, several more recent, in particular machine learning, approaches to register analysis and/or classification have opted for data-driven, bottom-up feature selections. Although such methods can be argued to mitigate researcher bias in the feature selection process, ultimately, however, any multivariable approach will inevitably involve the researcher(s) making some choices about the set of variables to be examined. In the Biberian approach, the selection of linguistic features is both theoretically, linguistically motivated on the basis of previous research and strives to be as comprehensive as possible so as to have the potential to capture as-of-yet undiscovered salient aspects of linguistic variation (cf. Conrad & Biber 2013: 15; Egbert & Staples 2019: 127). Whilst inductive, emergent data-driven approaches laudably aim to minimise researcher bias, it can be argued that all research nevertheless remains first and foremost 'researcher-driven'. For example, in a case study demonstrating such a data-driven approach to the multivariable analysis of linguistic variation, Diwersy et al. (2014: case study 2) chose to focus on syntagmatic patterns involving nouns and specified a frequency threshold of a minimum of 500 occurrences for the noun colligations to be included in their model. Not only is 500 an arbitrary cut-off point, the choice of the frequencies of "noun lemma and syntactic functional label" pairs (Diwersy, Evert & Neumann 2014: 180) as a feature set is inherently researcher-driven, too. As hinted at above, proponents of rigorously data-driven approaches also argue that the use of NLP tools such as lemmatisers, POS-taggers and syntactic parsers also add layers of biases in that they assume the validity of existing (and well-established) theoretical models of language analysis (cf. Sinclair 1992: 385–390).

Whilst the idea of a "non-biased variable set" (cf. Ruette et al. 2014: 211, who, however, claim that "generating a truly unbiased variable set is probably impossible") seems superficially attractive, another major drawback of genuinely data-driven features sets is that they cannot pretend to capture much beyond form, thus ignoring the thorny issue of polysemy. This is easily illustrated with one of the oldest and easiest to implement set of features: the frequency of some of the most common words of the English language: function words. As the most frequent words in any text, they bear the advantage of occurring largely independently of a text's thematic focus. Whilst such methods, which remain widely used in stylometry and other fields, have also proven to be relatively effective in register classification tasks (e.g., Biber 1991, Nowson 2006, Argamon & Levitan 2005, Herring & Paolillo 2006), they do not lend themselves well to meaningful linguistic interpretations of the classification results. As an example, consider the function word *that*: differences in the frequencies of occurrence of *that* cannot be meaningfully interpreted in a functional-linguistic sense across different registers. In academic writing, for instance, high frequencies of *that* are likely to point to complex sentence structures such as verb complementation and relative clauses, e.g.:

(1) The use of quantitative analyses means **that** we can see the extent of variation in texts and analyze the complex interactions among linguistic features, while at the same time the more qualitative interpretations ensure **that** we understand the communicative functions **that** the linguistic features are serving. <from Conrad 1996: 301>

In contrast, high occurrences of *that* in spontaneous conversation are more likely to be the result of the interlocutors taking advantage of their shared knowledge and environment. This is illustrated in the transcripts of two snippets of conversations in (2) and (3). As context-less transcripts, these are largely unintelligible to the outsider as a result of their very high proportions of demonstrative *that* pronouns used to refer to various aspects of the interlocutors' shared knowledge and environment:

(2) **that** one see I think like **that**'s
lighter
that's darker than **that** isn't it?
that one's slightly darker <Extract from Spoken BNC2014: S38V.txt[5]>

(3) I thought **that**'s what he was
oh **that**'s oh I've never heard of **that** but **that** could work
you can use **that**
I could use **that**
yeah **that**'s quite funny <Extract from the Spoken BNC2014: S5DJ.txt>

Even without comparing "extreme" registers, such as academic writing and informal conversation as above, *that* fulfils too many different functions to be meaningfully incorporated in linguistic analysis using simple counts of its surface form – especially, since, as illustrated in (4), these different functions often co-occur within a single text (see also Gray 2019: 52):

(4) I expect **that** not only can the hon. Gentleman remember **that**, but **that** he will recall **that** we warned at **that** time **that** the reforms were illusory and would not work. <BNC1994: HHX.txt>

This issue of polysemy is even more conspicuous when it comes to relying on features generated bottom-up, as these features often transcend quite fundamental (though certainly not unproblematic) linguistic units: such as syllables, words, utterances, and sentences. Character n-grams are a case in point. Their attractivity is undeniable: they are computationally inexpensive and, unlike most approaches, can be argued to be genuinely data-driven (at least for written language and in the rare cases when they include whitespace and punctuation). As a result, they are frequently employed in machine learning approaches to text classification. In many respects, they yield similar results to the long-established tradition of using function words because frequencies of character n-grams are inevitably highly correlated with those of function words. However, drawing meaningful conclusions from character n-grams results is even more difficult than with function words. What's more, in many cases, their high accuracy scores can be traced back to rather trivial reasons – for instance, they have been found to be very powerful predictors of translated as opposed to non-translated texts: one reason for this is that proper nouns with rare letter combinations are more likely to be found in translated texts simply because these more often refer to people,

[5] All text examples from here onwards are drawn from the BNC2014 Baby+ corpus (see 3.1) unless otherwise indicated. Emphasis in bold was added by the author unless otherwise stated.

places, or institutions with foreign names featuring “foreign” letter combinations (e.g., references to the city of *Put**eaux*** or the surnames *Dub**ois*** or *R**oux*** in English translations of French source texts) (cf. Baroni & Bernardini 2005: 264; Popescu 2011: 638; Volansky, Ordan & Wintner 2015: 111). Thus, while relative frequencies of character n-grams have been shown to be remarkably effective in a range of register and genre classification tasks (e.g., Amasyalı & Diri 2006; Kanaris & Stamatatos 2007), their lack of “direct linguistic motivation or interpretation” (Argamon 2019: 111) makes them unsuitable for the aims of the present tagger.

Most other “bottom-up” or “data-driven” attempts to generate sets of features, e.g., word forms, n-grams, lemmas, POS tags, shallow parsing chunk types, grammatical relations, and combinations thereof, suffer from the same issues though to different degrees. In addition to their inherent reliance on theory-dependent tools for their extraction and their limited interpretability, these sets of features also frequently inadvertently capture text topic and domain information. Consequently, careful experimental controls are needed to ensure that models based on such features genuinely represent register variation in the sense of structural, linguistic differences as opposed to “accidentally correlated topics” (Argamon 2019: 111; cf. Volansky, Ordan & Wintner 2015: 100). As a result, such radical bottom-up approaches tend to present serious scalability issues (cf. Luyckx 2010). Berber Sardinha’s (2017) MDA of register variation in the Corpus of American English (COCA) using collocations provides fascinating insights into lexis-based register variation but the results themselves are inherently tied to the corpus data they are based on and therefore cannot readily be generalised to other texts in (American) English. To counter this, Crossley & Louwerse (2007) only entered bigrams that were shared across different (sub)corpora in their MDA. However, this resulted in very sparse correlation matrices. Their method, which in choosing to focus on bigrams bears the advantage of combining lexical, semantic, syntactic, and discursive aspects of register, successfully distinguished between broad registers; however only with large amounts of data that was, due to data sparsity reasons, analysed at the corpus/subcorpus level only, as opposed to the arguably linguistically more valid text-level. The authors concluded that “[w]hile this analysis works well at distinguishing disparate registers, it does not seem to discriminate between similar registers” (Crossley & Louwerse 2007: 475). The aim, here, however is to propose an English tagger that may be used across a broad range of different registers and topics, as well as for the analysis of register within specific domains (e.g., Textbook English; Le Foll in preparation). This is why such an approach, although arguably more “exhaustive and principled”, was not pursued as part of the feature selection process.

Instead, simplified Hallidayian system networks for various aspects of English lexicogrammar were examined in an attempt to arrive at a more systematically selected set of linguistic features. This approach is not novel. It was mentioned by Matthiesen (2019: 30) in response to McEnery & Hardie’s (2011: 114) suggestion that the feature selection for MDA ought to ideally be “both *principled* and *exhaustive*”. Prior to this, Whitelaw & Argamon (2004) had also attempted to resolve text classification tasks and carry out sentiment/appraisal analysis using a selection of system networks elaborated within the framework of Systemic Functional Linguistics (hereafter: SFL) (note that the code from these projects was never made public and sadly no longer exists, Whitelaw, personal communication 2021). Other scholars have explored specific aspects of Halliday’s (1993; cf. Halliday & Hasan 1991) tripartite model of register by extracting one or more lexico-grammatical features associated with either field, tenor or mode. Whilst these approaches explicitly make no attempt to be exhaustive, they can be said to be principled. For example, Teich et al. (2016) derived, for each of their two hypotheses on the specialisation and diversification of academic language as scientific fields develop, small, custom sets of features which they selected based on previous observations made by Halliday (1988) on smaller corpora (requests as to whether code for this paper was available or could be made available remained unanswered). For the diversification hypothesis, they chose the relative frequency of verb-argument classes for field, modal verbs for tenor, and theme type and conjunctive cohesive relations to account for mode (Teich et al.

2016: 1671). This approach is similar to Neumann's (2014; cf. Evert & Neumann 2017; Neumann & Evert 2021) whose selection of features is also motivated by Halliday's theory of register. Though she includes a much broader range of features than Teich et al. (2016), Neumann (2014), too, does not make any claims to include an exhaustive list of linguistic features. On the contrary, in addition to practical and computational constraints, Neumann's selection of features is tailored to her research aims.

By contrast to Teich et al. (2016) and Neumann's (2014) studies, the aim, here, is to develop an English tagger that may be a useful tool for researchers studying situational variation across different domains and (sub-)registers and potentially seeking to explore a range of different research questions. At the end of the day, whether it is by selecting which SFL system models are to be applied (e.g., the conjunction, comment, modality and pronominal determination systems in Whitelaw & Argamon 2004, all taken from Matthiessen 1995) or which features to implement to represent particular aspects of the SFL register theory (e.g., lexical chains for field of discourse in Neumann 2014: 52–53), it is clear that such approaches run the risk of having the researchers selecting features "derived from theoretical assumptions about patterns of interest" (Diwersy, Evert & Neumann 2014: 176). Nonetheless, considering that, from its very beginnings, one of the theoretical goals of SFL was to describe language *systematically*, the use of system networks to arrive at a plausibly exhaustive and principled set of linguistic features remains attractive. In SFL, system networks model "levels of the linguistic system [that] are characterised as resources by describing them as interlinked collections of alternations accompanied by specifiable distinctions in their associated linguistic forms" (Bateman 2017: 15). Note that although the feature selection procedure was inspired by such simplified system networks of SFL English grammars (many of which appear in Bartlett & O'Grady 2017), the present tagger is by no means an attempt to parse texts following all aspects of SFL formalism. As O'Donnell (2017) describes in a review of interactions between NLP and SFL, this task has been attempted in the past but has proven to be immensely complex and highly computationally expensive. Instead, as part of the feature selection process of the present tagger, system networks were examined to minimise researcher bias by ensuring that no major aspect of English lexicogrammar would be overlooked. Ultimately, however, the final choice of features was necessarily restricted by practical and computational constraints (cf. tagger specifications in 2.1).

In particular, it is important to mention that the features of most SFL system networks are organised in a highly complex manner with, if represented visually as trees as in Fig. 2, both conjunctive and disjunctive branches. This means that any one token or set of tokens may be assigned tags for several features. For example, in the modality system outlined in Fig. 2, any one token, say *probably*, will be identified for type (in this case, the 'modalisation' type: 'probability'), value ('median), orientation ('objective') and manifestation ('explicit').

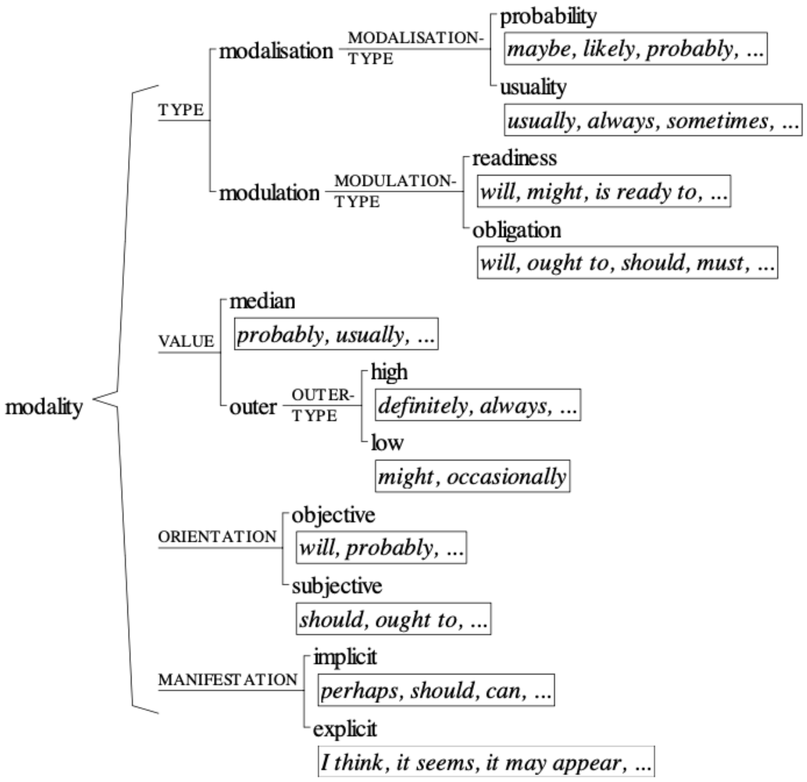


Fig. 2: The modality system (from Whitelaw & Argamon 2004: Fig. 4). Options in the curly brace are conjunctive (i.e., one option in every child must be chosen), all other options are disjunctive; examples of some lexical realisations for the leaves are given in italics.

From a theoretical point of view, this makes sense; however, it can cause issues in the context of tagging for multivariable register analysis. Take the number of occurrences of 'medium' value modal tokens: this value will by default be directly correlated with the total number of modal tokens. This can lead to "dishonest correlations" being entered into a model. As will be explained in more depth in 2.4, this issue can be (partially) resolved by normalising the frequencies of occurrences of each feature to the total number of features corresponding to the node immediately to the left of that feature, e.g., the number of 'median' value modals can be expressed as a percentage of all occurrences of modals, whereas the number of occurrences of 'high' value modals may be expressed as a percentage of all 'outer value' modals (see Fig. 2).

Another major practical restriction is that, although many system networks have been formalised to a great degree of precision to cover almost all aspects of the English language, to date, the more lexis-based system networks in particular (e.g., for conjunctions) have not yet been concretised with (relatively) exhaustive lists of lexical resources that (can) fulfil the categories of such systems. Furthermore, many of the features that form part of SFL system networks would require syntactic parsing to be automatically detectable (e.g., to disambiguate between the different types of clauses in the clause system, cf. Matthiesen 1995). Whilst using an open-source syntactic parser for English (e.g., NLTK, SpaCy, Stanford Parser) was originally considered, preliminary tests showed that their accuracy was too low to serve as a first layer of automatic tagging for the present purposes – in particular given the fact that the present tagger is to be used across different registers of English, including transcripts of conversations and pseudo-spoken internet registers (see tagger specifications in 2.1).

Although the use of syntactic parsing was dismissed for the present project due to low accuracy with non-edited text registers such as Internet language, the use of POS-tagging is clearly helpful or even necessary for most of the selected linguistic features. The CLAWS tagger spans a large set of lexico-grammatical features, which may have formed a very good basis for the present tagger. However, CLAWS is not freely available and therefore does not satisfy the specifications outlined in 2.1. As a result, another widely-used POS-tagger was chosen: the Stanford Tagger (Toutanova & Manning 2000; Toutanova et al. 2003). The Stanford Tagger is also used as the basis of Nini's (2014) MAT tagger.

The author is well aware of the controversies and limitations of relying on an off-the-shelf POS-tagger as the underlying linguistic annotation layer for the present tagger. The first usually revolves around the accuracy of automatic taggers. Here, it is argued that modern POS-taggers such as the Stanford Tagger are, overall, admirably accurate (e.g., Toutanova et al. 2003; Spoustová et al. 2009; Manning 2011) and that, crucially, the gain made through their linguistic annotation is much greater than the relatively small amount of noise they add where tags are erroneous. In the context of multivariable linguistic analysis such as MDA, the most pressing point of criticism, however, concerns the nature of the layer of linguistic analysis which POS-taggers add. Indeed, even major reference grammars do not converge on basic terminology and definitions, POS-tagging procedures inevitably reflect a specific theory of grammar (cf. Lindquist 2009; McEnery et al. 2006). Gray (2019: 45–46) illustrates how POS-tagging constitutes a form of linguistic analysis with the example of the word *tomorrow*. Whilst *tomorrow* is identified by CLAWS as a "quasi-nominal adverb of time", the Biber Tagger assigns it the tag: "singular adverbial noun". Thus, two taggers can arrive at very different solutions: not only because errors can, and do, occur, but also because they can reflect different linguistic interpretations. This is demonstrated with a brief comparison in Table 1 of the tags assigned by CLAWS7 (CLAWS using the C7 tagset) (via http://ucrel-api.lancaster.ac.uk/claws/free.html on 14 Sep. 2021), the Stanford Tagger (via http://nlp.stanford.edu:8080/parser/index.jsp on 14 Sep. 2021) and the Biber Tagger[6] (see Biber et al. 2004: Appendix A for a comprehensive list of the tag codes as of 2004) to the following short excerpt:

(5) At home, my husband and I put our garbage in paper bags. When one is full, we close it and staple it shut. Stapled is better since we store the bags for a while on the back porch and since I'm just compulsive enough that I don't want my business to blow away and become everybody's business.

Table 1: Comparison of three taggers' output for text excerpt (5). Incorrectly tagged tokens and those with disagreements between the taggers are highlighted.

CLAWS7 (vertical output)			Stanford Tagger	Biber Tagger
At	93	II	At/IN	At I++++++ II
home	97	NN1	home/NN	home N+CM+SING++++ NN1
,	03	,	,/,	, Y+COM+++++ ,
my	93	APPGE	my/PRP$	my D+PER+SING+1+V++ APPGE
husband	93	[NN1/100] VV0%/0	husband/NN	husband N+CM+SING++++ NN1
and	93	CC	and/CC	and C+CRD+++++ CC
I	93	[PPIS1/100] ZZ1%/0 MC1%/0	I/PRP	I P+PER+SING+1+S++ PPIS1
put	93	[VV0/52] VVD/48 VV N/0 NN1%/0	put/VBD	put VL+BF+++++ VV0
our	93	APPGE	our/PRP$	our D+PER+PLUR+1+V++ APPGE
garbage	93	NN1	garbage/NN	garbage N+CM+SING+NOM+++ NN1
in	93	[II/99] RP@/1	in/IN	in I++++++ II
paper	93	[NN1/100] VV0%/0	paper/NN	paper N+CM+SING+++PMOD+ NN1
bags	93	[NN2/99] VVZ%/1	bags/NNS	bags N+CM+PLUR++++ NN2
.	03	.	./.	._.

[6] Many thanks to Larissa Goulart for providing the version tagged with the Biber Tagger.

---------------	--	------		
When	93	[CS/62] RRQ/38	When/WRB	When C+SRD+++++ CS
one	93	[PN1/97] MC1/3	one/PRP	one P+++INDEF+++ PN1
is	93	VBZ	is/VBZ	is VL+Z++COP+++ VBZ
full	93	[JJ/94] RR%/6	full/JJ	full J++PRDV++++ JJ
,	03	,	,/,	, Y+COM+++++ ,
we	93	PPIS2	we/PRP	we P+PER+PLUR+1+S++ PPIS2
close	93	[VV0/100] RR@/0 JJ /0 NN1@/0	close/VBP	close VL+BF+++++ VV0
it	93	PPH1	it/PRP	it P+IM++3+++ PPH1
and	93	CC	and/CC	and C+CRD+++++ CC
staple	06	[NN1/63] VV0/37	staple/NN	staple VL+BF+++++ VV0
it	93	PPH1	it/PRP	it P+IM++3+++ PPH1
shut	03	[VVD/92] JJ/4 VVN/3 VV0/1	shut/VBD	shut VL+ED+++++ VVD
.	03	.	./.	._.
---------------	--	------		
Stapled	06	[NP1@/96] JJ@/2 VV N@/2 VVD/0	Stapled/NNP	Stapled N+PR+SING++++ NP1
is	93	VBZ	is/VBZ	is VL+Z++COP+++ VBZ
better	97	JJR	better/JJR	better J++PRDV+ER+++ JJR
since	93	[CS/100] RR@/0 II@/0	since/IN	since C+SRD+++++ CS
we	93	PPIS2	we/PRP	we P+PER+PLUR+1+S++ PPIS2
store	93	[VV0@/100] NN1/0	store/VBP	store VL+BF+++++ VV0
the	93	AT	the/DT	the DET+AT+SING+DEF+++ AT
bags	93	[NN2/100] VVZ%/0	bags/NNS	bags N+CM+PLUR++++ NN2
for	93	[IF/100] CS%/0	for/IN	for I++++++ IF
a	93	AT1	a/DT	a D+AT+SING+INDEF+++ AT1
while	93	[NNT1@/100] CS/0 VV 0%/0	while/NN	while N+CM+SING+++TIME+ NNT1
on	93	[II/99] RP@/1	on/IN	on I++++++ II
the	93	AT	the/DT	the DET+AT+SING+DEF+++ AT
back	93	[NN1/72] JJ@/28 RP/0 VV0%/0	back/JJ	back N+CM+SING+++PMOD+ NN1
porch	93	NN1	porch/NN	porch N+CM+SING++++ NN1
and	93	CC	and/CC	and C+CRD+++++ CC
since	97	CS	since/IN	since C+SRD+++++ CS
I	93	[PPIS1/100] ZZ1%/0MC1%/0	I/PRP	I P+PER+SING+1+S++ PPIS1
’m	03	VBM	’m/VBP	m VL+AM++COP+CONT++ VBM
just	93	[RR/95] JJ@/5	just/RB	just R++++++ RR
compulsive	06	JJ	compulsive/JJ	compulsive J++PRDV++++ JJ
enough	93	[RR/99] DD/1	enough/RB	enough R++++++ RR
that	96	[CST/100] DD1%/0	that/IN	that C+SRD+THT+REL+++ CST
I	93	[PPIS1/100] ZZ1%/0MC1%/0	I/PRP	I P+PER+SING+1+S++ PPIS1
do	56	VD0	do/VBP	do VA+BF+++++ VD0
’nt	56	XX	n’t/RB	n’t XX++++SPLT++ XX
want	97	VVI	want/VBP	want VL+INF+++++ VVI

my	93	APPGE	my/PRP$	my D+PER+SING+1+V++ APPGE
business	93	NN1	business/NN	business N+CM+PLUR+NOM+++ NN1
to	97	TO	to/TO	to TO+++COMP+VERB++ TO
blow	97	VVI	blow/VB	blow VL+INF+++++ VVI
away	93	[RL/100] JJ%/0	away/RB	away R+++++LOC+ RL
and	93	CC	and/CC	and C+CRD+++++ CC
become	98	VVI	become/VB	become VL+INF++COP+++ VVI
everybody	03	PN1	everybody/NN	everybody P+++INDEF+++ PN1
’s	03	[GE/96] VBZ/3 VHZ@/0	’s/POS	’s GE++++++ GE
business	93	NN1	business/NN	business N+CM+PLUR+NOM+++ NN1
.	03	.	./.	._.

As part of the development of the MFTE, two practical solutions to mitigate the impact of these problems inherent to automatic POS-taggers were adopted. First, the tagging rules are made as transparent as possible. Such an approach calls for extensive documentation and the publication of the source code. In this context, it may also be argued that the use of a well-known and documented POS-tagger, such as the Stanford Tagger, can contribute to making the results more accessible to a wider usership. Second, as already stressed by McEnery et al. (2006: 31), the results of the analysis must be “recorded and encoded in the annotated corpus” such that it is “explicit and recoverable”, thus ensuring that humans can disambiguate between what are genuine tagging errors and what represents different approaches to linguistic analysis. In other words, a tagger for MDA should not simply produce large tables of (normalised) counts, but also supply the researchers with the fully annotated texts so that they may understand exactly which features contributed to which counts (see 2.5).

Whilst the examination of system networks allowed for a systematic approach to the selection of linguistic features, the final set of features only includes linguistic features for which automatic extraction using the Stanford Tagger as the first layer of linguistic analysis was not only possible but also to a satisfactorily high degree of accuracy. As 2.3 will show, this set of features shares many similarities with those identified by the Biber Tagger (in both its 1988 and 2006 versions) though, in most cases, their identification is operationalised rather differently. Appendix I (ListFullMDAFeatures_v3.0.xlsx) lists the features of the MFTE. This “feature portfolio” or “catalogue” was devised so as to arrive at a set of features in which each feature is meaningfully interpretable in the sense that “its scale and values represent a real-world language phenomenon that can be understood and explained” (Egbert, Larsson & Biber 2020: 24), whilst mitigating the risks of circularity and researcher bias that Altenberg (1989), Diwersy et al. (2014) and others have rightly warned about. As such, what may, at first sight, seem like an eclectic collection of linguistic features in fact reflects the tagger’s underlying aim to be able to describe registers of English in the most comprehensive way possible. The following section explains how the features of the MFTE feature portfolio were operationalised.

2.3 Feature operationalisation

This section describes how the custom-built perl script (Wall 1994) that constitutes the MFTE derives countable indicators from abstract linguistic concepts and outlines the various decision-making processes involved in arriving at the final set of features and feature operationalisations. Before proceeding, it should be noted that, although the processes of feature selection (2.2), operationalisation (2.3) and tagger accuracy evaluation (3.3) are described sequentially in the present document, for many features these processes were repeated in a cyclical manner over multiple iterations before arriving at the final feature portfolio and feature operationalisations described in the following and subsequently evaluated in Chapter 3.

The code of the MFTE was originally based on Andrea Nini's MAT (Nini 2014; 2019) and, although the vast majority of features are now operationalised differently, it nevertheless follows its structure.[7] Like the MAT, the MFTE performs feature extraction over several iterations over the texts of the corpus/corpora to be examined. First, each text is tagged with basic POS-labels using the Stanford Tagger (bidirectional version 3.9.2; Toutanova & Manning 2000; Toutanova et al. 2003). Next, rule-based algorithms are run to refine some of the analyses of the Stanford Tagger and to identify the selected linguistic features.

Appendix I provides an overview of the tagger's feature portfolio and descriptions of the operationalisation of each feature (see Table 2 for an extract). To preclude any misunderstandings, it should be noted that although the table is subdivided into broad linguistic categories, these merely serve organisational purposes and do not seek to represent any definite functional categorisation of these features. Indeed, the reader will notice that many features could equally be subsumed under a different category. The second column of the table in Appendix I (see also Table 2 for illustration purposes) provides a very brief (and, due to space limitations, often simplified) description of each linguistic feature. The third column corresponds to the tags assigned by the MFTE (note that these same abbreviations are also used in the tables and figures presented in Chapter 3). Examples of different language patterns exemplifying these features are found in the fourth column. Note that, except for the first four rows, the tokens highlighted in bold in the example column refer to the tokens to which the tags are assigned, e.g., for the BE-passive variable, the tag PASS is assigned to the past participle tokens (rather than to the BE tokens), but that it is also possible for one token to be assigned more than one tag, e.g., a PASS token may also be a verb of communication and therefore also assigned the tag COMM. The fifth, "operationalisation" column is essentially a (simplified) written-out explanation of the combinations of regular expressions used to identify these variables. For more details, the reader is referred to the full source code available on https://github.com/elenlefoll/MultiFeatureTaggerEnglish.

Table 2: Excerpt of Appendix I: Operationalisation of the DO auxiliary variable [DOAUX]

Category	Feature	Code	Examples	Operationalisation
Verb semantics	*DO auxiliary*	*DOAUX*	*Should take longer than it* ***does****.* *Ah you* ***did****.* *She needed that house,* ***didn't*** *she?* *You* ***don't*** *really pay much attention,* ***do*** *you?* *Who* ***did*** *not already love him.*	Assigned to *do, does* and *did* as verbs in the following patterns: (a) when the next but one token is a base form verb (VB) (e.g., *did it work?, didn't hurt?)*; (b) when the next but two token (+3) is a base form verb (VB) (e.g., *didn't it work*); (c) when it is immediately followed by an end-of-sentence punctuation mark (e.g., *you did?*); (d) when it is followed by a personal pronoun (PRP) or *not* or *n't* (XX0) and an end-of-sentence punctuation mark (e.g., *do you? He didn't!)*; (e) when it is followed by *not* or *n't* (XX0) and a personal pronoun (PRP) (e.g., *didn't you?*); (f) when it is followed by a personal pronoun followed by any token and then a question mark (e.g., *did you really? did you not?*); (g) when it is preceded by a WH question word. Additionally, all instances of *DO* immediately preceded by *to* as an infinitive marker (TO) are excluded from this tag.

[7] I am indebted to Andrea Nini for having made the source code of this MAT tagger available. Heartfelt thanks also go to Stephanie Evert for contributing a more efficient way to run the Stanford Tagger over the texts of a corpus to be tagged (see Appendix III for details).

Table 6 (excerpt from Appendix I) explains how DO auxiliaries are identified thanks to various combinations of POS tags and forms of the verb DO. The descriptions of the operationalisations also contain many cross-references to other tags. For instance, the operationalisation description of imperatives (VIMP) makes it clear that they can only be identified after DO auxiliaries have been tagged. This ensures that imperative forms of the verb DO can be disambiguated from auxiliary forms, in particular those included in *yes/no* questions where the *do/does/did* frequently occur after an end-of-sentence punctuation mark (see Appendix I for details).

Since the Stanford Tagger provides the first layer of linguistic annotation (tokenisation and POS tagging), the accuracy of the feature extraction is heavily dependent on the accuracy of the Stanford Tagger. Whilst it is a well-tested and robust model, it is by no means perfect (Toutanova et al. 2003; Spoustová et al. 2009; Manning 2011). As a result, some of the feature operationalisations outlined in Appendix I include more tags than would be necessary if the underlying Stanford-based POS-tagging process were failproof. For instance, since the Stanford Tagger was found to frequently fail to differentiate between past tense (VBD) and past participle forms (VBN), the algorithms designed to capture passives (PASS and PGET) and the perfect aspect (PEAS) include syntactic patterns with either the VBN or the VBD tag in order to improve recall rates whenever past participles have been erroneously tagged as VBD (cf. Nini 2014). Similarly, some of the distinctions made by the Stanford Tagger were not retained in the MFTE. A case in point concerns the tagging of nouns: various development-evaluation cycles showed that the Stanford Tagger failed to differentiate these to a satisfactorily high degree in texts with no or inconsistent capitalisation (see 3.3.3) so that the two categories were merged.

On the whole, however, there is no doubt that POS-taggers have considerably improved since Biber conducted his first set of MDAs in the late 1980s and, consequently, many of the present feature operationalisations rely much more on this first layer of linguistic annotation than Biber's Tagger could. This is notably the case for the identification of *that* relative clauses (tagged as WDT by the Stanford Tagger, though see also Appendices I and III for the adjustments that were made to account for frequent tagging errors). Whilst it is true that using a POS-tagger as the basis for the feature extraction process reduces the reproducibility of the method as different tagging software (and models/versions) will inevitably produce different results (cf. Bohmann 2017: 165), the gain in recall and precision is huge and many of the most meaningful linguistic features simply cannot be extracted without this initial annotation layer.

The MFTE involves too many iterations over the texts for a legible graphical representation of the code, but the table in Appendix II (Tagger_Example_PEAS_PASS.pdf) aims to give the reader an impression of the order in which tokens are tagged. For each example sentence, the first layer of annotation comes from the Stanford Tagger. Whenever nothing else appears below the tag assigned by the Stanford Tagger, that original tag was retained for the feature count (e.g., *the* is counted as a determiner, DT); in other cases, the tags are overwritten by the MFTE whenever one of the patterns described in Appendix I is found. This sometimes happens more than once, e.g., *done* in Example 6 is first identified as a past participle by the Stanford Tagger (VBN), second as a passive (PASS) by the MFTE on the basis that *'s* could be a contracted form of the verb BE; however, *done* is subsequently identified as a frequent verb more likely to occur in the perfect aspect than the passive voice leading to the PASS tag being overwritten to PEAS (perfect aspect) instead. The examples in Appendix II thus illustrate that the order of the tagging loops is not trivial: countless tags are used to identify other tags, either by inclusion or exclusion. An example of tagging by inclusion is the identification of split auxiliaries (SPLIT) which relies on the previous identification of DO auxiliaries (DOAUX). Tagging by exclusion is more common, e.g., the DOAUX tags are assigned before the semantic verb category tags so that any occurrences of DO not tagged as auxiliaries are subsequently tagged as activity verbs (ACT) (see Appendix I).

Although many of the features may look superficially similar, many improvements were made to Biber's (1988) selection and operationalisations of features capturing verb tense, aspect and voice (see Appendix I). For example, rather than tag the perfect aspect onto the auxiliary HAVE, the PEAS tag becomes an attribute of the past participle form and overwrites the corresponding VBN tag assigned by the Stanford Tagger. Similarly, the passive voice (PASS) also becomes an attribute of the past participle rather than the verb BE. These new operationalisations make possible the creation of a new, linguistically meaningful VBN variable, which only includes non-finite uses of past participle forms (e.g., *These include cancers **caused** by viruses*). Similarly, the addition of two new variables for the progressive aspect (PROG) and the *going*-to construction (GTO) also led to the creation of a new VBG category for non-finite uses of present participle forms. Moreover, these new operationalisations capture more complex syntactic patterns, e.g., allowing for various combinations of intervening adverbs and negation in verb patterns involving auxiliaries. Efforts were also made to disambiguate *'s + past participle* constructions which can either be present in passive or present perfect forms. Manual tag checking as part of the code development phase revealed that many of these constructions are instances of the HAVE-*got* construction. Due to its idiom-like character and restricted usage (see, e.g., Huddleston & Pullum 2016: 111–113), it was decided to assign this construction a separate tag (HGOT). For the remaining instances of *'s* + past participle structures, various options were considered. In the end, the most efficient and accurate method was to disambiguate between passive and perfect forms on the basis of the lemma of the past participle form. Hence, verbs that are known to never or rarely occur in the passive voice are assigned the perfect tag (PEAS) (e.g., BE, HAVE, DO and stative verbs see Appendix I for more details) whilst the rest are assigned the passive tag (PASS).

In addition to the new verb tense, aspect and voice features mentioned above, the MFTE also tags for WH-questions (WHQU), question tags (QUTAG) and imperatives (VIMP), all of which were not accounted for in Biber's (1988) model, but which feature in even the most simplified of Hallydian system networks and are undeniably relevant to the exploration of register variation in English. In addition, the creation of an imperative category (VIMP) allows for the disambiguation of non-finite infinitives from imperatives (both of which are tagged by the Stanford Tagger as VB).

At the lexical level, too, Altenberg (1989) and others have pointed out that inclusion of some highly multifunctional items in some of Biber's (1988) categories is problematic, e.g., *just, most* and *really*, which are all classified as belonging to the emphatics category, regardless of their actual contextual use. Since linguistic features closer to the lexical end of the lexico-grammatical continuum have also been shown to also be of high relevance for the register variation analysis (Biber et al. 2004; e.g., Crossley & Louwerse 2007; Berber Sardinha 2017), the lexical features of the Biber Tagger (in its 1988 version since this is the version with the most detailed descriptions of the feature operationalisations) were closely examined and, whenever deemed necessary, features were either operationalised differently or entirely removed. As documented in the sixth column of the table in Appendix I, this resulted in just one of Biber's (1988) original variables being retained without any changes (phrasal coordination; PHC)[8]. Five more were kept with Nini's (2014) operationalisations, which include minor, but very meaningful, corrections to increase extraction accuracy when using the Stanford Tagger. Relatively minor changes were made to a further twenty features originally included in Biber (1988) – the details of which are laid out in Appendix I. Changes made include rectifying some of the inconsistencies pointed out as early as 1989 by Altenberg (1989: 172), e.g., adding *perhaps* to the category of hedges, correcting the fact that *almost* was listed as both an approximator and a downtoner, and finding solutions to the classification of multifunctional items such as *a lot, just, really*. Across all lexis-based features, various

[8] Though this feature was later merged with the general coordinating conjunction variable (CC) due to low precision and recall (see 3.4).

strategies were employed to deal with multifunctional/polysemous words and phrases: in many cases, they were included in the categories that correspond to their most frequent functions. When this proved too error-prone, items were excluded or, if they are highly frequent, separate categories were created: this led to the creation of the SO and LIKE categories. Both of these are frequently used as discourse markers and fillers in everyday conversation, but these uses are very difficult to automatically disambiguate from other uses; thus, if they are not excluded, they run the risk of skewing the frequencies of the adverb and preposition variables, respectively. Depending on the research questions under study, users of the tagger may decide to include or exclude these variables. One specific use of *like* which can be relatively reliably extracted using regular expressions is in the quotative construction *BE* + *like*, for which a separate category was therefore created (QLIKE). In practice, it is often not frequent enough to reach text/corpus coverage threshold levels and users of the MFTE may therefore also decide to exclude it and/or merge the counts with the LIKE variable – provided that the latter is also entered in the analysis (see 3.4).

The functional nature of Biber's (1988: 241-242) modal verbs categories[9] was also deemed to be unsatisfactory due to modal verbs' diverse meanings and contextual uses. Counting each central modal as a distinct variable proved, however, not to be feasible either because some have relatively low relative frequencies, e.g., *ought* or *might*, which would likely result in several modals having to be discarded in the factor analysis. At the same time, however, considering all modals as a single linguistic feature is also inadequate since register variation can reasonably be expected to influence the distributions of modals (as indeed demonstrated in many past MDA studies, including Biber 1988). As summarised in Appendix I, the following compromise was therefore made: *can, could* and *would* are counted as standalone features (MDCA, MDCO, MDWO), whilst *may* and *might*, and *will* and *shall* that share more functional similarities are grouped together in two categories (MDMM and MDWS respectively). The necessity modal variable (MDNE) is the only functional modal category that was retained from Biber's (1988) set of features. However, in addition to *must*, *ought* and *should*, the semi-modal NEED, when identified as a modal by the Stanford Tagger, or when followed by *to* and an infinitive, was also included in this category.

The three semantic verb categories employed in Biber's (1988) original MDA (public, private, and suasive verbs) are only used in the present analysis for the identification of the subordinator *that*-omissions (THATD). Instead, the MFTE adopts, with some minor corrections, the seven semantic verb categories (based on Biber et al., 1999; pp. 361-371) used in Biber's more recent MDA work on university English (Biber et al. 2004; Biber 2006). Though it goes without saying that many of the more frequent verbs have multiple meanings potentially belonging to different semantic domains, e.g., *follo*w can express either a physical activity or a mental process, this semantic classification is based on verbs' core meanings (Biber 2006: 246), which was deemed to be sufficient to compare overall trends between different registers. However, some highly frequent and highly polysemous verbs with the potential to skew the results were removed from Biber's (2006) lists (e.g., *GO* and *GET* which were listed as activity verbs, *LOOK* which was originally counted as an existential/relationship verb and *FIND* which was included in the mental verb category, see Appendix I for details).

In his practical guide to conducting MDAs, Brezina (2018: 192) suggests removing paralinguistic sounds (e.g., *um, er, mhm, mm*) from spoken data because POS-taggers usually misidentify these as nouns, thus potentially vastly overestimating the noun count in natural conversation corpora in which, like the Spoken BNC2014, these vocalisations have been transcribed. Indeed, in the additive MDA conducted by Le Foll (2021b) using the MAT, most of these particles were erroneously identified as nouns. Rather than remove these fillers and hesitations as part of the pre-processing of the corpus texts, however, it was decided to

[9] Biber (1988: 241-242) differentiates between possibility modals (*can, may, might* and *could*), necessity modals (*ought, should* and *must*) and predictive modals (*will, would* and *shall*).

retain them as they can be considered to constitute defining characteristics of natural spoken language. Appendix I provides a full list of the tokens that are identified by the MFTE as interjections or filled pauses (FPUH). These are then counted as a self-contained linguistic variable, which, like all the linguistic features counted by the MFTE, can either be entered in MDA or excluded, depending on the corpora to be analysed and the research questions pursued. Furthermore, all tokens identified as FPUH are excluded when computing the lexical diversity feature (LDE, see below) and total word count (an aspect will be further discussed in 2.4 on the normalisation of feature counts).

Two features were added to the feature portfolio of the MFTE to account for language change since the 1980s: emoji and emoticons (EMO) and web and e-mail addresses (URL). This first version of the MFTE recognises the full list of Unicode emoji characters as of Dec 2018 (Full Emoji List, v.11.0; Unicode.org). Accurately identifying emoticons is more error-prone. First, a long list of emoticons was drawn (the original list was taken from Schnoebelen 2012 and additional emoticons were crowdsourced on Twitter). However, preliminary evaluation rounds revealed that emoticons were frequently being identified when they were none. This issue was particularly prevalent in journal articles of hard sciences which tend to feature formula with many brackets, symbols, and numbers, frequently without spaces, which were then sometimes erroneously identified as emoticons, e.g., (6).

(6) The mercuryonly adsorption isotherms were measured for both FS1 and Smopex®112 at an initial mercury concentration of 100 ppm (Figure **8);** with mercury in the form of the K2Hg(CN)4 salt. The data were fitted to a Langmuir isotherm, shown in Equation (iv), where: qA = mg adsorbate per g adsorbent (mg g1), **b =** adsorption parameter (l mg1), Ce = equilibrium concentration (mg l1) and Q**0 =** maximum capacity (mg g1), which is deemed to be more suitable due to the expected chelation mechanism. <Extract from BNCBAcjT30.txt>

At the same time, not identifying emoticons in e-language registers that do feature them can also lead to severe issues. Indeed, like paralinguistic sounds in transcriptions of spoken language, if they are not recognised, emoticons will be mistagged and this may trigger additional errors further down the line. This is particularly true of emotions featuring letter characters, which the Stanford Tagger will usually identify as nouns. In line with the tagger specifications (see 2.1), a compromise was struck to account for as many emoticons as possible in the feature portfolio of the MFTE (recall) but with as few erroneous tags as possible in texts that do not contain any (precision). Thus, for the purposes of multi-feature register variation in general English, it was decided that precision should prevail over recall because, in most cases, it is expected that texts will either include emoticons or not. In effect, the exact number of emoticons is far less relevant than the binary encoding of the presence or absence of emoticons in any one text to be entered in an MDA. That said, if the MFTE is to be used to differentiate between different internet subregisters, more emoticons may be identified by uncommenting the corresponding lines in the source code of the MFTE. These lines mostly match two-character emoticons such as *:D ;(* and *>3* which were found to be particularly problematic for other registers. Note, too, that this version of the MFTE does not attempt to identify kaomoji such as (ᵔ◡ᵔ) and *¯_(ツ)_/¯*, even though these are increasingly being used in the West (Giannoulis & Wilde 2019). This could be implemented in future versions if texts likely to contain them are to be analysed (see, e.g., Bedrick et al. 2012 for a possible approach to detecting kaomoji in Twitter discourse).

Additional linguistic features, which were not included in Biber's (1988) set of features, include existential *there* (EX), possessive endings (POS), coordinating conjunctions (CC), demonstratives (DEMO), other determiners (DT), politeness markers (POLITE), noun compounds (NCOMP), frequency adverbials (FREQ), particles (RP), lexical density (LDE), as well as *used to* (USEDTO), *HAVE got* (HGOT) and *BE able to* constructions (ABLE).

A final major departure from the feature extraction principle applied by the Biber Tagger and its replications worth mentioning concerns the treatment of multiword items. For example, the

MAT (Nini 2014) tags the first token of the multiword *on the other hand* as a conjunct and assigns a NULL tag onto the remaining tokens. This means that in the final feature count output of the MAT *on the other hand* counts as just one conjunct, i.e., (7). In contrast, the MFTE assigns the conjunct tag (CONJ) in addition to the usual preposition, determiner, adjective and noun tags, i.e., (8). The concept of a NULL tag was 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).

(7) on_CONJ the_NULL other_NULL hand_NULL

(8) on_IN CONJ the_DT other_JJ hand_NN

2.4 Feature normalisation

In general, the texts of a corpus can vary in length. Since text length would skew any results based on raw frequencies (counts per text), feature raw counts must therefore be normalised before they can be further analysed. This process usually involves transforming the raw counts of features within each text to relative frequencies in relation to a normalisation factor. Whilst the previous section outlined how the numerators of these relative frequencies were calculated, this section focuses on the choice and operationalisation of the denominator(s).

The *de facto* normalisation standard in computational linguistics is the division of raw counts by the total number of tokens or words in each text, in other words, text or document percentage. Typically, this figure is then multiplied by 100, 1,000, 10,000 or a million and referred to as a normalised, normed, or relative frequency. Crucially, however, such an approach implies that words are used independently of each other: it does not account for or attempt to model the actual choices that language users make when producing language. In another words, word-based baselines conflate frequency of use and opportunity of use (Wallis 2020: 47–52) simply because, once a language user has chosen one word, their choice of the next word is limited (e.g., in English, a determiner is far more likely to be followed by a noun or adjective than by a verb). In reporting normalised frequencies with a word-based baseline, the frequency of opportunities of use is ignored thus adding a considerable amount of uncontrolled variation to the relative frequencies reported. This, in turn, may mean that the conclusions of studies based on such measures may not be reproducible (Wallis 2020: 74).

Corpus linguistics, in particular, has a long history of using word-based normalisation rates indiscriminatingly. Whilst per-word normalised frequencies can be argued to represent language users' rates of exposure (which may well be what researchers, e.g., lexicographers, are attempting to model), a number of recent publications have pointed to some of the inherent issues with the indiscriminate application of this method. Wallis (2020: 56), for instance, explains how per-word frequencies undermine the assumptions of many of the statistical models used in corpus linguistics which assume that linguistic features follow binomial distributions, i.e., that it is, in principle, possible to observe proportions of 100%. In practice, however, it is highly improbable that a language user would simply repeat a single word or even word class for the entire duration of a text! In sum, not only do word-based baselines pose theoretical and statistical issues, they also increase the risk that the results of such studies may not replicate on different corpus data.

In the Biberian MDA framework, too, the fact that most linguistic feature counts (except average word length and type/token ratio) are normalised on the basis of the total number of words in each text (sample) has remained largely unchallenged. This choice of normalisation unit, however, means that some of the high correlations observed between certain clusters of linguistic features in MDA are, in fact, caused by additional linguistic variables directly mediating these features. In other words, they can be said to be linguistically "obvious correlations" (Evert 2018: 24; cf. Tabachnick & Fidell 2014: 95–96 on "honest correlations").

Altenberg (1989: 173) referred to this phenomenon as a form of "predetermined distribution" and put forward the example of the strong correlation between the normalised frequencies of split auxiliaries and modal verbs in Biber's 1988 model – two features which make positive contributions to Biber's fourth 'persuasive discourse' dimension:

> As Biber himself points out, there does not seem to be anything particularly 'persuasive' about split auxiliaries. Rather, since split auxiliaries often involve modals, the former are bound to cooccur with the latter (Altenberg 1989: 173).

The high positive correlations between the per-1000-word normalised frequencies of contractions, negation and present tense reported in Biber (1988) is another such example: all these features are obviously mediated by the overall frequency of verbs. Similarly, the high positive correlations between the frequencies of nominalisations, determiners and, though to a less extend, prepositions are mediated by the frequency of nouns. It therefore seems evident that not all linguistic features can be meaningfully normalised using the number of words as the denominator. Thus, operationalising the present tense variable as the number of occurrences of the present tense for every 100 finite verbs is likely to be more meaningful than per 1,000 words because the latter option runs the risk of foremost reflecting the number of verbs per 1,000 words. Schegloff (1993: 103) speaks of "environments of possible *relevant* occurrence" (emphasis original) and argues that "quantitative analysis requires an analytically defensible notion of the denominator".

In the context of MDA, the choice of the normalisation units for the linguistic features to be entered in a model are anything but trivial. For a start, according to the statisticians Tabachnick & Fidell (1993: 103), factor analysis (FA) and principal component analysis (PCA), the two statistical methods most widely used to conduct MDAs following Biber's framework, are "exquisitely sensitive to the sizes of correlations". Consequently, the authors warn that "it is critical that honest correlations be employed" (Tabachnick & Fidell 2014: 665). Second, the chosen normalisation units have a direct impact on the (linguistic) interpretability of the results. As Baroni & Evert (2009: 795) note:

> The same observed frequency can have different interpretations (with respect to the corresponding population proportion) depending on the units of measurement chosen as tokens, and the related target population.

Indeed, as stressed by Wallis (2020: 73–75), and decades earlier by Schegloff (1993), choosing an appropriate denominator to report and analyse relative frequencies is inherent to the formulation of a research hypothesis. In other words, changing the denominator will also change the research question(s) that can be answered. That being true, the choice of normalisation unit is by no means a simple one: in practice, it will depend on both the (linguistic) conceptualisation of how one linguistic feature (the numerator) was chosen by the language user over another (the denominator), as well as the feasibility of reliably retrieving that denominator (cf. Wallis 2020: 69–70). For instance, for the normalised frequencies of WH-questions, the total number of sentences or clauses in a given text may seem like the most meaningful option; however, whilst identifying sentences is relatively trivial in written registers, not only can this unit be argued to not make much linguistic sense in spoken registers[10], it is impossible to reliably implement with spoken corpora whose transcription scheme does not include any full stops (e.g., the Spoken BNC2014). As for automatically identifying clauses, this would require dependency parsing, which is currently a near impossible task for transcriptions of spontaneous spoken language without punctuation since current approaches rely on sentence boundaries as a starting point, and which would, in any case, certainly result in units that would be equally difficult to compare across very different text registers.

Wallis (2020) proposes a "methodology progression of baselines" (see Fig. 3) from the most arbitrary (option 1: the number of words in a text) to the ideal, perfectly accurate, validated

[10] "A sentence is a constituent of writing, while a clause complex is a constituent of grammar" (Halliday 1993: 216).

representation of a language user's choice between possible alternations in a specific context. Wallis (2020: 67) refers to this construct as a methodological continuum that "represents a trade-off between ease of obtaining results and the reliability of conclusions" – thus implying, that in an ideal world, free of practical constraints, the fourth option on Fig. 3 is most desirable. Given that the present tagger is to produce reliable feature counts without any manual intervention, this fourth option is clearly not feasible in the context of this project. However, even if it were, it is not necessarily the case that it would be the most suitable baseline for a broad range of applied linguistic research aims. The case of lexicographers who may genuinely be interested in rates of exposure as opposed to opportunities of use has already been mentioned in passing. But, even when researchers are foremost interested in tapping into language users' opportunities of use, it is arguable whether the full range of options available to a language user at any one point in time can realistically be listed.

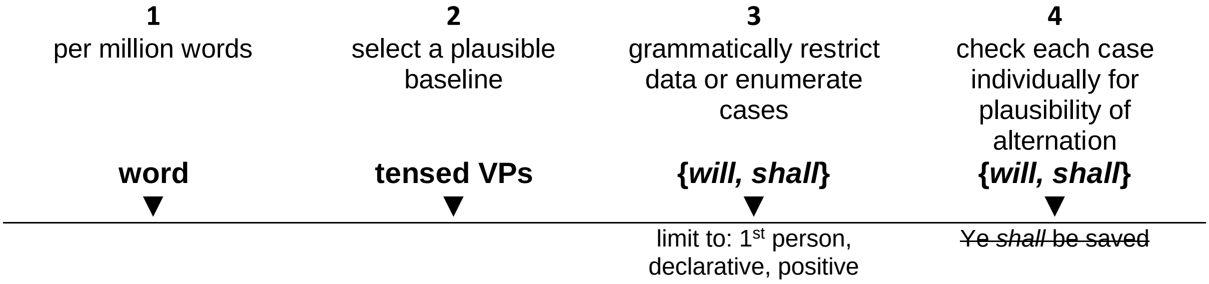

1
per million words
word
2
select a plausible baseline
tensed VPs
3
grammatically restrict data or enumerate cases
{*will, shall*}
limit to: 1st person, declarative, positive
4
check each case individually for plausibility of alternation
{*will, shall*}
~~Ye *shall* be saved~~

Fig. 3: A methodological progression: baselines for the modal auxiliary *shall*, from normalised word frequencies to verified alternation (from Wallis 2020: 62)

For instance, we may ask what might be an appropriate normalisation basis to examine the frequency of verbs in the progressive. The number of words (option 1) can unambiguously be ruled out since it is obvious that a text with very few verbs is also likely to have very few verbs in the progressive and vice versa. Following Wallis' (2020) model, the next most plausible baseline would be finite verb phrases (labelled as "tensed verb phrases" in Fig. 3). This makes intuitive sense because only tensed verbs can be in the progressive. However, it can be argued that many verb types do not or hardly ever occur in the progressive. To overcome this issue, Smitterberg (2005: 46–48) identified certain finite verb phrases, e.g., those with imperative verbs and BE *going to* + infinitive constructions with future reference, that cannot occur in both the progressive and non-progressive and consequently excluded them from the baseline used to measure the frequency of progressives in his variationist study. Such an approach (cf. option 3 on Fig. 3) raises a number of additional questions, however. For instance: On what basis should these exceptions, such as "fixed constructions", be chosen? Or, in the case of progressives, how should finite verb phrases with stative verbs be accounted for, given that, for a long time, they were considered incapable of rendering in the progressive, but are now known to be on the rise (e.g., *I'm thinking*, see, e.g., Deshors & Rautionaho 2018; Rautionaho & Fuchs 2020)? Should they, or should they not be excluded from the baseline count of finite verb phrases to report the frequency of progressives? For this particular problem, Wallis (2020: 51) suggests partitioning the data and calculating separate relative frequencies of progressive occurrences with stative and dynamic verbs.

Whilst creating two progressive vs. non-progressive variables solves this particular problem, another, more fundamental question remains: are we genuinely modelling language users' choices by attempting to have a baseline that accounts for all plausible alternations (i.e., option 4 in Fig. 3)? Variationist linguistics, the tradition in which this fourth option is anchored, sets out to analyse register "in terms of how it influences linguistic choices between functionally equivalent variants" (Szmrecsanyi 2019: 78; cf. Labov 1972: 188). Imagine someone announcing to a colleague that they have decided to call it a day and are going home. They might use a verb in the progressive and say "*I'm leaving now. See you tomorrow!".* Is it likely that at this point in time the speaker chose to use the verb *leave* in the progressive rather than in the non-progressive? Or might they have chosen this utterance over any number of similar routine phrases, e.g., "*I'm off now. Have a good evening!",* to express their intention? If we consider the phrase *I'm off now* to be a plausible alternation to *I'm leaving now*, the most sensible normalisation unit for progressive verbs is more likely to be the clause or the tensed verb phrase. As explained above, identifying clauses requires syntactic parsing for which automatic solutions are not currently reliable enough for (pseudo-)spoken corpora. As a result, it was decided that feature counts such as verbs in the progressive and WH-questions can most plausibly be normalised on the basis of the number of finite verb phrases. The number of finite verb phrases can be approximated to a relatively high degree of accuracy using the MFTE by adding the counts for present tense (VPRT), past tense (VBD), imperatives (VIMP) and all the modal verbs (MDCAN, MDCOU, MDMM, MDNE, MDWO, MDWS) (see Appendix I for operationalisations).

2.5 Tagger outputs

As outlined in 2.4, different linguistic features will likely require different normalisation baselines, and these will also be dependent on the research questions at hand. Consequently, the MFTE outputs three different tables of feature counts:

1. [prefix]_normed_complex_counts.tsv
2. [prefix]_normed_100words_counts.tsv
3. [prefix]_raw_counts.tvs

The prefix to the output filenames is determined by the researcher in the command used to run the MFTE script (see instructions at the top of the MFTE.pl script). All three output files are tab-separated text files. They each consist of a data matrix in which each text of the corpus analysed is represented by a 78-element numeric vector which corresponds to the linguistic features listed in Appendix I. In other words, each row corresponds to a text file from the corpus tagged and each column corresponds to a linguistic feature. Note that this is in contrast to most off-the-shelf corpus tools that generally do not output per-text frequencies but rather only per-(sub)corpus (though this is possible with WordSmith Tools [Scott 2011]; and LancsBox [Brezina, Timperley & McEnery 2021]). This is important because it allows for the description of the extent to which texts vary linguistically within a register (sub)category.

In all three tables, the first five columns of the count tables are identical: the first column lists the filenames of the texts, the second the total number of words in each text (excluding fillers, see Appendix I; this column is used as the basis for the word-based normalisation), the third the average word length, the fourth the lexical density (measured as the type/token ratio, by default for the first 400 words of each text, see instructions in Appendix I on how to change this in the command used to run the script), and the fifth column reports lexical density as expressed by the ratio of content word to total words (see Appendix I). The remaining columns correspond to the rest of the linguistic features listed in Appendix I (except those highlighted in grey at the bottom of the table). Unlike the Biber tagger which, for example, outputs rates of occurrence for three individual categories of modals, as well as for all modals together (Gray 2019: 62), this tagger's feature portfolio was compiled to ensure that none of the features directly overlap. Consequently, each modal verb category is listed separately with no additional total modal count. From an accessibility point of view, it makes more sense for a tagger to output tables of counts that can be, should the researcher be happy with all of the features included, immediately used, rather than one that, in all cases, requires some editing, e.g., to remove overlapping variables. Of course, should a researcher be interested in aggregated counts of all the modals, it is simple enough to add up the frequencies of each modal category to obtain those.

The “complex” normalised feature frequencies are calculated on the basis of the different normalisation baselines listed in the fifth column of the table in Appendix I. This means that, in this table, the present tense variable (VPRT) represents the proportion 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. Though not recommended (see 2.4), for ease of comparability with other taggers widely used for MDA (e.g., the Biber Tagger and the MAT), a table of counts normalised by 100 words is also included.

Furthermore, the tagger outputs a table of raw counts. Unless the texts 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), as well as to easily check the sanity of the data. In addition, for full transparency of the process, it is important that the tagged files themselves also be saved for further exploration, e.g., as part of the interpretation of quantitative results derived from the feature counts. The tagged files are presented in vertical format and the tag codes correspond to those listed in Appendix I.

“All code is guilty until proven innocent.”
- Anonymous

3 Evaluation

There is no doubt that multi-feature analyses of language are highly dependent on the choice and quality of the features entered in such analyses. We know that, even with perfectly well-defined linguistic features, manual identification frequently leads to identification errors caused by inattention or diverging interpretations of the definitions. Although an automated tagger like the MFTE neither suffers from fatigue nor makes any subjective judgements, such automatic approaches are not without their pitfalls, either. As they are far from fail-proof, they can, at best, only provide estimated counts of features. As a result, thorough evaluations of the performance of automatic taggers are necessary to estimate how tagging inaccuracy is likely to impact the interpretation of results based on the output of a tagger. Unfortunately, such evaluations are rarely undertaken.

It would appear that, for the most part, the consequences of tagger performance on the results of MDA studies have not yet been thoroughly considered (cf. Goulart & Wood 2021). However, the accuracy of the tagger used to identify the features to be entered in an MDA is key to the quality of the results. Noteworthy differences of more than 10% in the tagging accuracy of two automatic parsers (the Malt Parser and the Stanford Dependency Parser) and one tagger (the Biber Tagger) were recently reported by Picoral et al. (2021) in a study comparing the precision and recall rates of two phrasal and two clausal features in L1 and L2 novice academic English. Such differences may be larger than can reasonably be expected between different varieties or registers and may therefore call into question the results of MDA studies based on erroneously tagged features. To investigate any potential systematic mismatches between the linguistic constructs investigated and the features as they are actually counted by the MFTE, this chapter evaluates the performance of the tagger by manually checking the accuracy of the MFTE.

When evaluating the performance of a tagger, a number of accuracy measures may be used. On the one hand, the number of correct tags, i.e., true positives, can be counted and, on the other, the number of incorrect tags, i.e., false positives. The simplest accuracy measure is a ratio of true positives to all tags assigned by the tagger. It is called precision because it indicates the proportion of tag labels that are genuinely correct. Recall, by contrast, exemplifies the proportion of relevant features to which the correct tags were assigned. In other words, it takes into account how many features were not tagged as they should have been. In an ideal world, a tagger would have 100% precision (i.e., all assigned tags are correct) and 100% recall (i.e., all features are labelled with all the relevant tags). In practical terms, however, attempts to increase precision on any one feature will usually result in lower recall rate for that particular feature and vice versa. An important aspect of the tagger development phase was therefore to find an 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. However, in the context of an automatic tagger to be used without any manual intervention (see 2.1), both precision and recall are important. This is why a third accuracy measure was calculated: the F1 score, which combines precision and recall (see Table 3). This particular combined measure was chosen because it has already been used in previous studies (e.g., Picoral, Staples & Reppen 2021), thus allowing for better comparisons of the performance of different taggers on different test data (though caveats of comparing the results of different evaluation procedures will be discussed below).

Table 3: Summary of the terminology used in tagger performance evaluation

Term	Definition
True positive	Feature correctly tagged by the MFTE as X
False positive	Feature incorrectly tagged by the MFTE as X
False Negative	Feature incorrectly not tagged by the MFTE as X
Accuracy	True positive count / total number of tags in the confusion matrix
Precision	True positive count / (true positive count + false positive count)
Recall	True positive count / (true positive count + false negative count)
F1 score	2 * (precision * recall) / (precision + recall)

Of the 95 MDA studies Goulart & Wood (2021) examined in their meta-analysis, 11 cite the accuracy of the tagger as reported in earlier studies, whilst only 13 report having checked the accuracy of the tagger on their own corpus data: three specifically mention having checked precision and four having checked both precision and recall. Not only are these numbers very low, in practice, such tagging evaluation procedures can be very different and therefore provide more or less informative results. In fact, of the seven MDA studies that mention some attempt to check precision or recall, several merely fleetingly mention the need to verify the accuracy of the automatic tagging procedure (e.g., Asencion-Delaney & Collentine 2011: 7). Thorough tagger evaluations are very labour-intensive and time-consuming, which is, understandably, probably why some scholars opt to only manually check the accuracy of one or a handful of features known to be particularly problematic: for istance, Staples et al. (2017: 5) manually checked all occurrences of the token *that* to fix any tagging inaccuracies associated with this particular token.

Another common strategy is to only 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 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. By contrast, per-sentence accuracy rates of POS-taggers tend to be considerably more modest – hovering around 50–57% – with even lower rates for non-standard varieties and registers (Manning 2011). Nini (2017) is evidently aware of this problem: in an evaluation of his MAT tagger on a random 20% sample of a very small (39,200-word) corpus of forensic texts, he reports an average precision of 96% but specifies that the feature with the lowest precision was 87%, whilst other features, presumably punctuation and determiners, reached 100% precision. Whilst Nini (2017) also reports a relatively small standard deviation of 3.4%, these figures can nevertheless only serve as a broad indication of the high accuracy of the MAT tagger (at least on this particular corpus) since the full list of per-feature accuracy rates is not reported. Scholars interested in using the MAT for their own research therefore have no indication as to which features score some of the lower accuracy rates. In addition, Nini (2017) only reports an average *precision* rate. *Recall* rates are undoubtably more time and labour-consuming to obtain; however, in the context of MDA studies they are just as relevant.

In addition to what accuracy measure is reported and the level of granularity (per token, per sentence, per feature) at which it is measured and reported, tagger evaluation procedures can also differ in terms of the number of tokens, texts, and corpora for which tagging accuracy is (manually) checked. One of the most comprehensive attempts to check the accuracy of a tagger for MDA that the present author is aware of required the recruitment and training of two independent coders as well as the involvement of a project research assistant to complete a line-by-line evaluation of the automatically assigned tags on a 5% sample of a 543,000-word corpus (Biber & Gray 2013: 17–18). It goes without saying that such resources are not always available. When such resources are available and multiple human tagger evaluators are involved, however, it is important that inter-rater agreement scores are calculated and reported.

Another important consideration concerns the actual evaluation procedure. In a technical report on the development of the Penn Treebank Project, Marcus et al. (1993: 7–8) describe an interesting experiment comparing two approaches to tagger evaluation. They report that when two human annotators independently tagged unannotated extracts of the Brown Corpus for POS, inter-annotator disagreement reached 7.2%. By contrast, when two human annotators were given the task of checking the output of an automatic tagger, the disagreement rate dropped to 4.1%. Given these results, we may hypothesise that the results of tagger evaluations based on a gold standard developed independently of the tagger output to be evaluated are likely to be lower than those based on human annotators merely checking and, whenever deemed necessary, correcting the tagger output. Such differences in evaluation procedures may explain the differences reported in Biber & Grey (2013) and Picoral et al. (2021). For instance, Biber & Grey (2013: 98) report 98% precision and 97% recall for attributive adjectives, whilst Picoral et al.'s (2021: 37) performance evaluation with the same tagger (the Biber tagger) only reaches 89% precision and 89% recall. Both evaluations are based on corpora of learner English, but the first paper relies on manual tagger output checking, whilst the latter is based on a gold standard developed entirely independently of tagger output. It is also worth noting that for human annotators to tag sample texts rather than merely check the tagger output much more precise definitions of each linguistic variables must be agreed upon. Indeed, it is quite plausible that different interpretations of what comprises an attributive adjective may well have also contributed to this large difference in reported recall and precision rates for this feature. Such differences often go unnoticed because the annotation schemes of gold standards are rarely fully documented. In Picoral et al.'s (2021: 27–28) evaluation, however, the annotation criteria are made explicit and, for instance, words related to languages or nations in phrases such as *Korean education* are not considered to be attributive adjectives but rather nouns because *Korean* cannot be inflected for comparison nor can it be modified by an adverb. This is in contrast to the Penn Treebank annotation scheme, which considers *English* in an *English sentence* to be an attribute adjective (Santorini 1990: 13–14).

When comparing the results of tagger evaluations, one additional important factor should be considered: some of the most comprehensive tagger evaluations to date (e.g., Biber & Gray (2013b) and Gray (2015)) are based on the same test sample that was originally examined to correct systematic tagging errors to improve tagger performance in fix-tagging procedures. This conflation of tagger *development* or tagger performance *improvement* with tagger *evaluation* risks inflating the final results since many tagging issues are known to be due to problematic topic-specific words and phrases that are very unevenly distributed. If exceptions are added for these either directly in the tagger script or via additional fix-tagging scripts, that test sample will inevitably contain fewer tagging errors afterwards. Indeed, this is the reason why Gray (2015) helpfully reports both "initial reliability rates" and "final reliability rates". However, the uneven distribution of problematic (clusters of) tokens means that texts for which no pre-evaluation has been conducted will very likely feature more tagging errors than those reported in the "final reliability rates". In fact, there is even the risk that such tagger improvements or fix-tagging procedures may generate additional errors in other texts/registers which will not be accounted for if the same test sample is used for both tagger improvement and tagger evaluation.

3.1 Data

The new British National Corpus (hereafter BNC2014; Brezina, Hawtin & McEnery 2021) seemed like an apt choice of evaluation corpus because it matches the specification criteria of the MFTE (see 2.1): it includes a broad range of registers, including transcripts of spoken interactions and different Internet registers (including blogs, forum posts, online reviews, e-mails, Twitter and Facebook data and text messages). It is the most recent balanced corpus of general British English and, like its predecessor the BNC1994 (Burnard 2007), is likely to become a standard reference English corpus for years to come. At the time of writing, the full BNC2014 was not yet available but, at the author's request, a balanced sample[11], the BNC2014 Baby+, was kindly made available for the purposes of this study by Lancaster University. The BNC2014 Baby+ totals some five million tokens, divided across five broad registers: academic writing, newspaper writing, e-language (or Internet English), novels, and spoken conversation. As empirically demonstrated by Biber (1990), the modest size of the BNC2014 Baby+ corpus is more than adequate for the lexico-grammatical representation of such broad registers.

Since one of the five broad registers of the BNC2014 Baby+ is spoken, conversational English, the BNC2014 Baby+ that was made available for this project included a random sample of files from the spoken component of the BNC2014 (Love et al. 2017). However, these files had been stripped of much of relevant metadata and annotations so that a random sample of the full Spoken BNC2014 which has been publicly available since 2019 was used instead. The Spoken BNC2014 consists of orthographic transcriptions of spontaneous conversations. The transcription scheme does not include any punctuation, except question marks. This makes POS tagging particularly difficult. However, the metalinguistic annotation in the untagged XML version of the Spoken BNC2014 includes speaker turns. This version of the Spoken BNC0214 was therefore used for the present tagger evaluation: speaker turns were used as a very approximative marker of utterance boundaries to help improve the tagger accuracy and full stops were therefore inserted at each speaker turn. In addition, the transcriptions of the conversations of the Spoken BNC2014 have been fully anonymised and all personal information has therefore been replaced by anonymising tags. In order for these tags not to be recognised as tokens of the actual conversations by the MFTE, all of the anonymising tags were automatically replaced by a corresponding placeholder of the same POS (e.g., all anonymised place names were replaced by *IVYBRIDGE* and all anonymised female names by *Jill*). All remaining mark-ups (e.g., headers containing metadata) were eliminated and the resulting texts were saved as .txt files in UTF-8 encoding. The R code which was used to pre-process the untagged XML version of the Spoken BNC2014 data, both for utterance boundaries and the pseudo de-anonymisation procedure, can be found in Appendix IV (Pre-processing_BNC2014Baby.Rmd). It relies on the {tm} package (Feinerer, Hornik & Meyer 2008) and a series of regular expressions. The texts of the remaining sub-corpora of the BNC 2014 Baby+ were delivered as .txt files.

A stratified random sample of 24 texts from the BNC2014 Baby+ was tagged with the MFTE for the final tagger performance evaluation. Version 2.9 of the MFTE was ran on perl 5, version 22, subversion 1 (v5.22.1) built for x86_64-linux-gnu-thread-multi. Crucially, these texts had not previously been used to test the performance of the tagger during the tagger development phase. Rather than assess the tagger performance on a smaller number of full texts from the test corpus, whenever the texts were longer than 1,500 tokens, it was decided to extract random samples to capture a broader range of text/topic-specific errors. Indeed, previous rounds of testing have shown that tagging errors are very unevenly distributed across texts and subcorpora because most errors are text/topic-specific and therefore cluster in individual texts. These samples were randomly extracted from the beginning, middle or end of each of the longer text files. The composition of the subsample of the BNC2014 Baby+ used for the tagger performance evaluation is summarised in Table 4. It shows that, in total, 31,311 tags were manually checked as part of this tagger performance evaluation with roughly equal numbers of tags checked across the four main registers: academic, internet, fiction and spoken English.

[11] Note, however, that the version of the BNC2014 Baby+ supplied for this project did not include any magazine samples, television or drama scripts, parliamentary debates, or miscellaneous texts, although these (sub)registers will also form part of the new BNC2014 (see Brezina et al. 2021: Table 5).

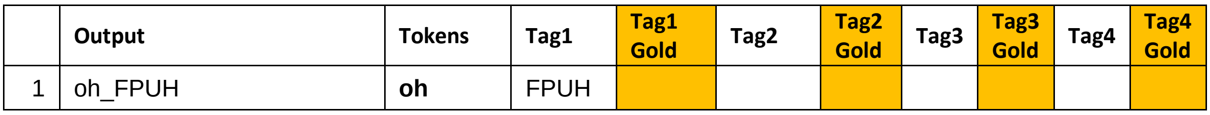
Table 4: List of texts/text samples from the BNC2014 Baby+ tagged and manually evaluated

File name	Register	Subregister	Number of tags
BNCBAcbH_m1	Academic	Academic books: humanities	2,003
BNCBAcjS6	Academic	Academic journals: science	1,073
BNCBAcjM105	Academic	Academic journals: medicine	1,600
BNCBAcjM102	Academic	Academic journals: medicine	1,596
BNCBEBl8	Internet	Blogs	554
BNCBEEm10	Internet	E-Mails	258
BNCBFict_b2	Fiction	–	2,102
BNCBFict_m54	Fiction	–	1,775
BNCBFict_e27	Fiction	–	2,104
BNCBEFor32	Internet	Forum posts	1,305
BNCBERe39	Internet	Product reviews	1,481
BNCBESm3	Internet	Text messages (SMSs)	1,423
BNCBMass16	News	Popular press	1,619
BNCBMass23	News	Popular press	268
BNCBReg111	News	Regional press	1,230
BNCBReg750	News	Regional press	1,275
BNCBSer486	News	Serious press	1,182
BNCBSer562	News	Serious press	738
BNCBEsocFb	Internet	Facebook posts	1,159
S2DD	Spoken	–	1,180
S3AV	Spoken	–	1,126
SEL5	Spoken	–	1,463
SVLK	Spoken	–	1,222
SZXQ	Spoken	–	1,056

3.2 Method

For each file in Table 4, the tagger output was copied into an Excel spreadsheet to facilitate the manual evaluation phase. The 'Data > Text to Columns…' function was used to create tables that allow for the correction of each tag, bearing in mind that any one token may have been assigned up to four tags (see Table 5, extract from the evaluation file S2DD.xlsx). The tagger output is printed without any changes in the far-left column. The 'Tokens' column allows for noise-free vertical reading of the original text. Each tag assigned by the tagger is printed in a separate column to be able to record errors at the tag- rather than the token-level. For instance, in line 18, the token *write* was assigned two tags by the tagger, but only the first, VIMP (for imperative) is incorrect (and was corrected to VB in the 'Tag1Gold' column), whereas the second tag, COMM, referring to the semantics of the verb, can stay as is. This approach also means that it is possible for human evaluators to add missing tags in the corresponding "gold" column (see line 6, Table 6).

Table 5: Extract 1 from the evaluation file S2DD.xlsx



	Output	Tokens	Tag1	Tag1 Gold	Tag2	Tag2 Gold	Tag3	Tag3 Gold	Tag4	Tag4 Gold
1	oh_FPUH	**oh**	FPUH							

2	you_SPP2	**you**	SPP2							
3	have_MDNE	**have**	MDNE							
4	to_TO	**to**	TO							
5	write_VB COMM	**write**	VB		COMM					
6	the_DT	**the**	DT							
7	topic_NN	**topic**	NN							
8	?_.	**?**	.							
9										
10	you_SPP2	**you**	SPP2							
11	have_VPRT	**have**	VPRT	MDNE						
12	to_IN	**to**	IN	TO						
13	._.	**.**	.							
14										
15	yeah_DMA	**yeah**	DMA							
16	._.	**.**	.							
17										
18	write_VIMP COMM	**write**	VIMP	VB	COMM					
19	the_DT	**the**	DT							
20	subjects_NN	**subjects**	NN							
21	down_RP	**down**	RP							
22	so_SO	**so**	SO							
23	like_LIKE	**like**	LIKE							
24	what_WHSC	**what**	WHSC							
25	we_FPP1P	**we**	FPP1P							
26	've_VPRT CONT	**'ve**	VPRT		CONT					
27	been_PEAS	**been**	PEAS							
28	talking_PROG COMM	**talking**	PROG		COMM					
29	about_IN STPR	**about**	IN		STPR					
30	._.	**.**	.							

The 'TagGold' columns were therefore used by the author to correct existing tags or add missing tags. Whenever a tag was incorrectly assigned and no tag should replace it (as shown in line 26 of Table 6), the TagGold label NULL was added.

Table 6: Extract 2 from the evaluation file S2DD.xlsx

	Output	**Tokens**	**Tag1**	**Tag1 Gold**	**Tag2**	**Tag2 Gold**	**Tag3**	**Tag3 Gold**
1	does_VPRT DOAUX	**does**	VPRT		DOAUX			
2	n't_XX0 CONT	**n't**	XX0		CONT			
3	look_VB	**look**	VB					
4	like_LIKE	**like**	LIKE					
5	it_PIT	**it**	PIT					
6	is_VPRT	**is**	VPRT			BEMA		
7	actually_DMA	**actually**	DMA					

8	in_IN	**in**	IN					
9	there_EX	**there**	EX	PLACE				
10	no_DMA	**no**	DMA					
11	I_FPP1S	**I**	FPP1S					
12	do_VPRT DOAUX	**do**	VPRT		DOAUX			
13	n't_XX0 CONT	**n't**	XX0		CONT			
14	believe_VB MENTAL	**believe**	VB		MENTAL			
15	so_SO	**so**	SO					
16	._.	**.**	.					
17								
18	probably_HDG	**probably**	HDG					
19	about_IN HDG	**about**	IN		HDG			
20	half_QUAN	**half**	QUAN					
21	an_DT	**an**	DT					
22	hour_NN	**hour**	NN					
23	till_IN	**till**	IN					
24	the_DT	**the**	DT					
25	sun_NN	**sun**	NN					
26	sets_NN NCOMP	**sets**	NN	VPRT	NCOMP	NULL		
27	is_VPRT	**is**	VPRT					
28	n't_XX0 CONT	**n't**	XX0		CONT			
29	it_PIT	**it**	PIT					
30	?_. QUTAG	**?**	.		QUTAG			

As in Table 6, the NULL tag was mostly used to denote erroneous tags. However, it was also occasionally used to denote erroneous tokenisation. Most words containing hyphens are tokenised as individual tokens by the Stanford Tagger (e.g., /he/ /was/ /panic-stricken/) but occasional tokenisation issues led to the use of NULL tags to denote superfluous tokens, e.g., in Table 7, lines 19–21, *cross-referencing* should have been tokenised as one token, to be assigned the tag VBG.

Table 7: Extract from the evaluation file BNCBAcbH_m1.xlsx

	Output	**Tokens**	**Tag1**	**Tag1 Gold**	**Tag2**	**Tag2 Gold**	**Tag3**	**Tag3 Gold**
1	The_DT	**The**	DT					
2	UN_NN	**UN**	NN					
3	command_NN NCOMP	**command**	NN		NCOMP			
4	produced_VBD ACT	**produced**	VBD		ACT			
5	a_DT	**a**	DT					
6	map_NN	**map**	NN					
7	to_TO	**to**	TO					
8	support_VB	**support**	VB					
9	its_PIT	**its**	PIT					
10	identification_NN	**identification**	NN					
11	of_IN	**of**	IN					

12	the_DT	**the**	DT					
13	line_NN	**line**	NN					
14	but_CC	**but**	CC					
15	western_NN	**western**	NN	JJAT				
16	journalists_NN NCOMP	**journalists**	NN		NCOMP	NULL		
17	found_VBD	**found**	VBD					
18	themselves_TPP3P	**themselves**	TPP3P					
19	cross_VB	**cross**	VB	NULL				
20	-_:	**-**	:	NULL				
21	referencing_VBG	**referencing**	VBG					
22	the_DT	**the**	DT					
23	three_CD	**three**	CD					
24	reporters_NN	**reporters**	NN					
25	over_IN	**over**	IN					
26	the_DT	**the**	DT					
27	precise_JJAT	**precise**	JJAT					
28	demarcation_NN	**demarcation**	NN					
29	line_NN NCOMP	**line**	NN		NCOMP			

In addition, it was necessary to find a way to deal with “missing” tokens. This is an issue that mostly occurs in informal, non-edited texts such as text messages, forum posts, internet reviews and social media posts. Consider the tagger output in (9). The evaluation procedure designed as part of this project allows, on the one hand, for the positive evaluation of the tag PIT (for *it* as a personal pronoun), as well as the manual correction of the analysis of this token by adding the tags VPRT (present tense), CONT (contraction) and BEMA (*BE* as a main verb) to the TagGold2, TagGold3 and TagGold4 columns respectively.

(9) I_FPP1S
think_VPRT MENTAL
its_PIT
a_DT
brilliant_JJAT
Idea_NN

Tricky POS grammar issues were solved in as consistent a manner as possible, e.g., all forms such as *focussed* and *motivated* as in (10) were considered to be erroneously tagged when they were assigned to the VBN or PASS features by the MFTE and annotated. Instead, they were annotated as predicative adjectives (JJPR) in the gold standard.

(10) But_CC
clever_JJAT
packing_NN
can_MDCA
help_VB CAUSE
make_VB ACT
you_SPP2
feel_VB MENTAL
focussed_VBN
and_PHC
motivated_VBN
._.

Systematic decisions also had to be made concerning the names of things (e.g., book and film titles with capitalised letters) and places. Thus, *White* in *the White House* was considered in the gold standard as an attributive adjective (JJAT) and was therefore flagged as erroneously tagged when it was identified as a noun by the MFTE. Similarly, *united* which was tagged as JJAT by the MFTE was considered false in *it felt good to put on a united shirt again* and *the united fans* as, in this context, *united* refers to the name of a football team. Finally, there were rare instances in which tokens could not be meaningfully interpreted. Again, this mostly occurred in internet registers. In such cases, UNCLEAR was entered in the Tag1Gold column.

3.3 Results

3.3.1 Accuracy

Of the 31,311 manually checked tags, 1,335 (4.26% [4.04–4.50%]) were found to be erroneous and 45 (0.14% [0.11–0.19%]) were deemed to be ‘unclear’. To begin, recall, precision, and the combined accuracy measure of F1 (see Table 3) were calculated for each feature across all 24 evaluation files. Next, problematic features with particularly low precision and/or recall rates were identified. These are listed in Table 8, ordered by precision rate. Note, however, that not all of the features included in this tagger performance evaluation are relevant for MDA: this includes the foreign word (FW) and symbol (SYM) tags, both of which are listed in Table 8, but neither of which are included in the tables of counts generated by the MFTE. Their precision and recall rates are nevertheless examined as part of this tagger performance evaluation because they were sometimes erroneously assigned this tag (e.g., an emoticon was not recognised as such, leading to a lower SYM precision), or because another tag, which is included in the tables of counts for MDA, was erroneously assigned instead of FW or SYM (e.g., an emoticon was detected when, in fact, it was simply a series of mathematical symbols, leading to a lower SYM recall).

Table 8: Features with either precision, recall and/or F1 below 0.8

Feature	Precision	Recall	F1
FW	0.34	0.50	0.40
USEDTO	0.33	1.00	0.50
VIMP	0.72	0.53	0.61
SYM	0.93	0.48	0.63
VBN	0.54	0.87	0.67
THRC	0.78	0.68	0.72
QLIKE	0.62	0.89	0.73
WHQU	0.76	0.76	0.76
PGET	0.89	0.73	0.80
PHC	0.92	0.72	0.81
THSC	0.75	0.99	0.85
RP	0.96	0.80	0.87

Three consequences were drawn from the results displayed in Table 8. First, the feature USEDTO was removed from the MFTE’s feature portfolio due to both poor precision and very low overall frequency, the latter meaning that, even with a much better precision, the feature would most likely not be usable in MDAs. Second, Biber’s PHC feature (phrasal coordination) was merged with the more general coordination (CC) feature due to both poor recall and frequent difficulties in disambiguating between the two as a human annotator. In total, the PHC feature was erroneously tagged 44 times as CC, and 17 occurrences of *and* or *or* as CC were erroneously tagged as PHC. This represents over a fifth (21.94%) of the total number of tagging errors! Since all true occurrences of PHC can be subsumed under the CC variable, it was possible to remove the offending feature simply by adding the PHC counts to the CC counts. Third, as a result of the disappointedly low precision rate, the QLIKE variable was removed from the MFTE's feature portfolio and all occurrences of *like* in the quotative construction BE + LIKE were instead subsumed in the all-encompassing LIKE category (see 2.3). Needless to say, these actions also considerably increased the precision and recall of other features, in particular: CC and LIKE.

Having merged PHC with CC and QLIKE with LIKE, and ignoring UNCLEAR tokens, the overall accuracy of the MFTE, as calculated by the caret::confusionMatrix function (Kuhn 2020), is 96.17% [95.95–96.38%, Kappa = 0.96]. As expected, the MFTE performs less well on internet and spoken registers than on professionally written and edited texts such as those typically found in fiction and academic writing (see Table 9). Some newspaper articles posed considerably more problems than others, especially those that included headlines and other phrases in all-caps (see 3.3.3). The code provided in Appendix V (TaggerTestResults.Rmd) also includes a function to obtain accuracy values for each individual file. It shows that the file with the highest accuracy rate is an extract of an academic journal (BNCBAcjM105.txt) with an accuracy of 98.25% [97.48–98.83%] and that the file that the MFTE struggled with most was a collection of Facebook posts (an extract of BNCBEsocFb.txt) with an accuracy rate as low as 89.61% [87.69–91.33%]. It should, however, be noted that this file was tagged as delivered and the evaluation procedure revealed that it includes passages in languages other than English, which, in addition to elliptical sentences and typos contributed to this low tagging accuracy.

Table 9: Accuracy of the MFTE

Register	Accuracy	Kappa	Upper 95% CI	Lower 95% CI
Academic	96.84%	0.96	96.38%	97.26%
Fiction	97.37%	0.97	96.95%	97.74%
Internet	94.03%	0.94	93.41%	94.61%
News	96.31%	0.96	95.81%	96.76%
Spoken	96.19%	0.96	95.68%	96.66%
Overall	**96.17%**	**0.96**	**95.95%**	**96.38%**

Though it struggled most with e-language, these results show that the MFTE copes relatively well with very different registers. However, overall accuracy rates can be very misleading since the most frequent tokens are usually the easiest to tag accurately and these therefore naturally tend to inflate all overall accuracy measures. This is why Table 10 presents recall, precision and F1 measures for each tag type (note, however, that not all tags are later counted by the MFTE and therefore included in the feature portfolio). The code provided in Appendix V also allows for the calculation of these measures for each of the five broad register categories of the BNC2014 Baby+ corpus.

Table 10: Precision, recall and F1 score for each MFTE tag

	Precision	Recall	F1
-LRB-	99.28%	100.00%	99.64%
-RRB-	99.28%	100.00%	99.64%
,	100.00%	100.00%	100.00%
:	99.61%	99.61%	99.61%
.	100.00%	100.00%	100.00%
''	100.00%	100.00%	100.00%
``	100.00%	96.55%	98.25%
$	100.00%	100.00%	100.00%

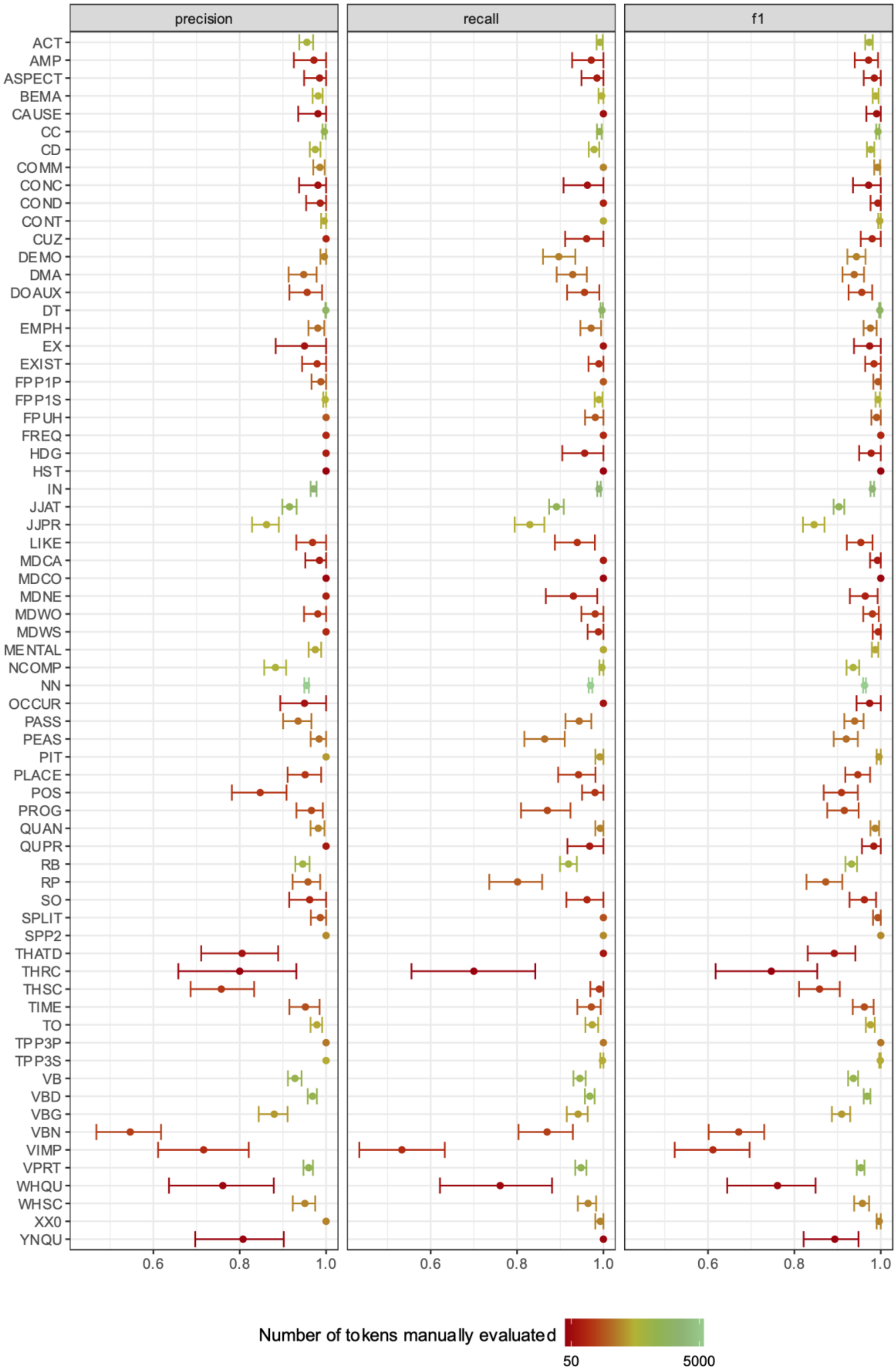
ABLE	100.00%	87.50%	93.33%
ACT	95.56%	99.21%	97.35%
AMP	97.18%	97.18%	97.18%
ASPECT	98.51%	98.51%	98.51%
BEMA	98.16%	99.58%	98.86%
CAUSE	98.11%	100.00%	99.05%
CC	99.60%	99.10%	99.35%
CD	97.50%	97.85%	97.67%
COMM	98.58%	100.00%	99.29%
CONC	98.11%	96.30%	97.20%
COND	98.65%	100.00%	99.32%
CONT	99.55%	100.00%	99.77%
CUZ	100.00%	96.10%	98.01%
DEMO	99.59%	89.67%	94.37%
DMA	94.82%	92.89%	93.85%
DOAUX	95.61%	95.61%	95.61%
DT	99.94%	99.65%	99.79%
DWNT	100.00%	100.00%	100.00%
ELAB	100.00%	93.75%	96.77%
EMO	95.83%	88.46%	92.00%
EMPH	98.08%	97.14%	97.61%
EX	95.00%	100.00%	97.44%
EXIST	97.92%	98.95%	98.43%
FPP1P	98.78%	100.00%	99.39%
FPP1S	99.79%	98.96%	99.37%
FPUH	100.00%	98.11%	99.05%
FREQ	100.00%	100.00%	100.00%
FW	34.00%	50.00%	40.48%
GTO	100.00%	96.43%	98.18%
HDG	100.00%	95.65%	97.78%
HGOT	87.50%	87.50%	87.50%
HST	100.00%	100.00%	100.00%
IN	97.22%	98.99%	98.10%
JJAT	91.69%	89.13%	90.39%
JJPR	86.61%	82.96%	84.75%
LIKE	96.88%	93.94%	95.38%
MDCA	98.48%	100.00%	99.24%
MDCO	100.00%	100.00%	100.00%
MDMM	92.59%	100.00%	96.15%
MDNE	100.00%	93.06%	96.40%
MDWO	98.08%	98.08%	98.08%
MDWS	100.00%	98.84%	99.42%
MENTAL	97.49%	100.00%	98.73%

NCOMP	88.41%	99.64%	93.69%
NN	95.91%	97.03%	96.47%
OCCUR	95.00%	100.00%	97.44%
PASS	93.52%	94.39%	93.95%
PEAS	98.40%	86.38%	92.00%
PGET	88.89%	72.73%	80.00%
PIT	100.00%	99.18%	99.59%
PLACE	95.15%	94.23%	94.69%
POLITE	100.00%	100.00%	100.00%
POS	85.47%	98.04%	91.32%
PROG	96.61%	87.02%	91.57%
QUAN	98.18%	99.26%	98.72%
QUPR	100.00%	96.83%	98.39%
QUTAG	95.24%	100.00%	97.56%
RB	94.74%	91.91%	93.30%
RP	95.80%	80.12%	87.26%
SO	96.20%	96.20%	96.20%
SPLIT	98.66%	100.00%	99.32%
SPP2	100.00%	100.00%	100.00%
STPR	81.82%	100.00%	90.00%
SYM	76.47%	48.15%	59.09%
THATD	80.56%	100.00%	89.23%
THRC	80.00%	70.00%	74.67%
THSC	75.74%	99.04%	85.83%
TIME	95.21%	97.20%	96.19%
TO	97.86%	97.39%	97.62%
TPP3P	100.00%	100.00%	100.00%
TPP3S	100.00%	99.77%	99.89%
URL	100.00%	100.00%	100.00%
USEDTO	33.33%	100.00%	50.00%
VB	92.79%	94.54%	93.66%
VBD	96.86%	96.86%	96.86%
VBG	87.99%	94.13%	90.96%
VBN	54.64%	86.96%	67.11%
VIMP	71.64%	53.33%	61.15%
VPRT	96.22%	94.80%	95.50%
WHQU	76.09%	76.09%	76.09%
WHSC	95.09%	96.44%	95.76%
XX0	100.00%	99.27%	99.63%
YNQU	80.77%	100.00%	89.36%

Although over 30,000 tags were manually checked as part of this tagger performance evaluation, some tags nevertheless occurred relatively rarely across the test texts of the BNC2014 Baby+ corpus (see Table 4). As a result of some of these small sample sizes, the three accuracy metrics reported in Table 10 risk being skewed by one or more outliers. In order to diminish the impact of such outliers, a form of data resampling with replacement called bootstrap simulation was applied to the dataset 1,000 times. 95% confidence intervals (CI) were then calculated on the basis of the results obtained on each of these 1,000 samples, thus reducing the risk of outliers exerting undue influence the results (see Picoral, Staples & Reppen 2021 for a very similar procedure). This procedure was performed in python[12] (see Appendix VI: Bootstrapped_Accuracy.ipynb for the code) and was applied to all tags for which more than 40 tokens were included in the test corpus (thus excluding the following features: USEDTO, HGOT, ABLE, STPR, PGET, ELAB, QUTAG, DWNT, URL, MDMM, EMO, POLITE and GTO). The results of the analysis are plotted in Fig. 10. The colours give an indication of how often each tag occurred in the test files (see Table 4). Note that no confidence intervals can be calculated for metrics with a value of 100%.

[12] Heartfelt thanks go to Luke Tudge who converted my bootstrapping R code (which itself was inspired by the code published in the appendix of Picoral et al. 2021) to python for it to run across all linguistic features within a reasonable timeframe given how computationally expensive this procedure is.

Fig. 4: Precision, recall and F1 score with bootstrapped 95% CI for each linguistic feature



3.3.2 Frequent tagging errors

As well as the recall, precision and F1 score of each feature, the most frequent types of tag corrections were also tallied. Table 11 lists all the combinations of tag corrections that were made ten or more times during the evaluation process. Together, these represent two-thirds of all identified tagging errors.

Table 11: Most frequent tagging errors and their corrections

Tag correction (MFTE tag -> corrected tag)	Number of occurrences
NCOMP -> NULL	71
NN -> JJAT	57
JJAT -> NN	53
NN -> VB	32
NN -> VPRT	31
IN -> RP	30
VBN -> JJAT	29
ACT -> NULL	23
VB -> VIMP	22
JJPR -> JJAT	21
NN -> JJPR	21
VB -> NN	21
THSC -> DEMO	19
VBG -> NN	18
VBN -> JJPR	16
VBN -> VBD	16
NN -> VBG	15
JJPR -> NN	14
THATD -> NULL	14
VBG -> PROG	14
VPRT -> NN	14
FW -> NN	13
FW -> SYM	13
RB -> NN	13
THSC -> THRC	12
CD -> NN	11
IN -> RB	11
IN -> TO	11
JJAT -> JJPR	11
JJAT -> RB	11
JJPR -> RB	11
MENTAL -> NULL	11
NN -> CD	11
PASS -> JJPR	11
VBN -> PEAS	11

WHSC -> WHQU	11
VB -> VPRT	10
VBD -> PEAS	10
VBD -> VBN	10
VPRT -> VIMP	10
WHQU -> WHSC	10
YNQU -> NULL	10

If the most frequent tagging errors listed in Table 11 involve so many nouns (in the form of noun-noun compounds, NCOMP, and individual noun tokens, NN) that is because the Stanford Tagger, which serves as the basis for the MFTE, defaults to noun whenever it encounters unknown alpha tokens. Thus, many of the tokens that required tag corrections of the type NN -> JJAT, NN -> VB and NN -> VPRT also involved the removal of unnecessary NCOMP tags because one or more of the "nouns" in the noun-noun compounds identified by the MFTE were, in fact, not nouns at all. NCOMP were also sometimes added to strings of proper nouns, especially when these were not capitalised, e.g., *phil collins*, even though noun-noun compounds are defined in the MFTE feature portfolio (see Appendix I) as strings of two or more nouns whereby only the first of any compound may be a proper noun (thus allowing for *Monday afternoon* and *Hollywood stars* but not *Phil Collins*).

Some of the frequent tagging errors can be said to be symmetrical: e.g., VPRT tags are frequently assigned instead of VIMP and vice versa. Others, on the other hand, are not: e.g., whilst particles (RP) are frequently mistagged as prepositions (IN), the opposite is much rarer. Most erroneous FW tags are the result of misspellings or OCR errors. In addition, many of the most frequent tagging error types listed in Table 11 may also be seen as different, rather than outright erroneous classifications of certain linguistic phenomena, e.g., many of the tag corrections of the types VBN -> JJAT, VBG -> NN and VBN -> JJPR could be interpreted either way; however, in the context of this tagger performance evaluation, the scheme outlined in 3.2 was consistently applied thus leading to relatively high numbers of errors of these types in Table 11.

3.3.3 Common sources of tagging errors

Very generally, the first principal source of errors arises from the form of the text files fed into the tagger. Thus, many mistags are the result of transcription errors in spoken data, OCR errors in poorly scanned written documents, or of very informally written texts that contain non-standard spellings, typographical errors and/or that largely lack punctuation. Examples of transcription errors in the Spoken BNC2014 include *your* instead of *you're* and missing question marks, e.g., *yeah but what are the subject*, meaning that the *what* is incorrectly assigned the tag WHSC instead of WHQU by the MFTE. A large proportion of source-text induced errors involves capital letters. As already mentioned in 2.3, the heavy use of capital letters for common nouns in a range of registers led to the dropping of the distinction between common and proper nouns because recall and precision for both categories were far too low. The problem, however, is not restricted to the disambiguation of common vs. proper nouns. Sometimes an oddly placed capital letter or a missing punctuation mark may mean that a verb token is recognised as a (proper) noun, e.g., in *Emma Have you ever tried thought of trying your luck here?* the capitalised *Have* is identified as a noun rather than a verb. This, in turn, triggers a further error: *thought* is not assigned the perfect aspect tag (PEAS) but rather the past tense tag (VBD). The use of all-caps, in particular in newspaper writing and online reviews, also causes strings of tagging errors in the POS-tags assigned by the Stanford Tagger. An extreme example is presented in (11). The only tokens to be correctly tagged are those overwritten by the MFTE (*very, although, they, me, lot*) because the corresponding regular expressions in the MFTE script allow for both lower- and upper-case words.

(11) VERY_AMP
HAPPY_NN
ALTHOUGH_CONC
I_NN
WISH_NN
THEY_TPP3P
WOUKD_NN
TELL_NN
ME_FPP1S
I_NN
WAS_NN
A_NN
LOT_QUAN
LIGHTER_NN

Problems associated with upper-case letters where capitalisation would be expected are somewhat rarer but also recurrent: e.g., in (12), the first token is assumed by the Stanford Tagger to be a list marker (LS) rather than a first person pronoun and in (13) *may* is erroneously identified as a modal verb due to the lack of capital letter.

(12) i_LS
also_CC
think_VPRT MENTAL

(13) this_DEMO
wednesday_NN
the_DT
5th_CD
of_IN
may_MDMM

The second main source of errors is due to erroneous tokenisation or ambiguous tokens being falsely disambiguated by the Stanford POS-tagger and not rectified by the MFTE. As explained in 2.3, the MFTE was designed to counter some of the frequent and systematic errors that the Stanford Tagger was found to make in earlier development phases: e.g., identifying the *'s* in *let's* as a possessive marker, failing to disambiguate between past participles as past tenses, and tagging most filled pauses in transcripts of spoken English as nouns. Although many such systematic errors are corrected by the MFTE, some basic tokenisation and POS-tagging errors nonetheless remain. Tokenisation issues frequently arise in hyphenated compound words, e.g., *slow-motion* and *liquid-light* in (14), whilst the most frequent POS errors arise when a single word form can occur with different POS and the least frequent of the two possible POS is present, e.g., *throw* in (15).

(14) like_LIKE
a_DT
slow_JJAT
-_:
motion_NN
wave_NN NCOMP
,_,
liquid_JJPR
-_:
light_JJAT
patterns_NN

(15) DeAndre_NN
Yedlin_NN
was_VBD BEMA
guilty_JJPR
of_IN
a_DT
foul_JJAT
throw_VB ACT

Finally, the third source of tagging errors stems from the way in which the features are operationalised. For many of the more complex grammatical features, in particular, the algorithms described in Appendix I and implemented in the MFTE (see Appendix III) are best approximations: they are designed to reach a good recall-precision compromise (see 3.2), but they cannot be expected to be 100% accurate in all contexts. For example, the loops designed to capture *that*-omissions (THATD) only do so for a specific set of verbs (listed in Appendix I) and *yes-no-* (YNQU) and WH-questions (WHQU) are limited to 15 tokens in length to avoid too low precision rates, but this inevitably means that longer questions are not recognised (thus leading to lower recall rates). On some features, the tagger performs less well in specific registers: for example, in spoken English as a result of the many fragments of speech, e.g., (16), repetitions and interruptions that characterise spontaneous conversation. The MFTE can also be led astray by particularly long and complex sentence structures in some literary works, or by the insertion of footnote indices or in-text references in academic texts.

(16) You_SPP2
going_VBG
to_IN
stroke_NN
the_DT
cat_NN
?_.

In general, it is also fair to say that, whatever their origin, tagging errors tend to cluster. For example, in (16), the omission of the auxiliary BE leads to a total of three erroneous tags (VBG, IN and NN) and one missing tag (YNQU). Similarly, in (17), the Stanford Tagger erroneously tagged *'s* as a genitive marker (POS) instead of a present tense verb (VPRT) thereby leading to three tags failing to be assigned by the MFTE: CONT and BEMA for *'s* and JJPR for *terrible*.

(17) That_DEMO
photo_NN
's_POS
terrible_JJAT

The token *like* had already been identified as a particularly problematic word in the tagger development phase and even with the addition of the special LIKE variable (see 2.3), errors remain and they, too, can trigger clusters of errors. For example, in *There's a place actually in IVYBRIDGE where you can like jump through the trees*, this occurrence of *like* was tagged as a base form verb (VB) by the Stanford Tagger as a result of it following the modal *can*. This triggered an additional error by the Stanford Tagger: *jump* is identified as a noun, and a further erroneous tag added by the MFTE as *like* is additionally assigned the mental verb tag (MENTAL). In some cases, it was possible to "over-engineer" the regular expressions used to identify some features to avoid one error triggering additional ones. An example of this can be seen in (15), where the erroneous VB tag on *throw* would have led to the incorrect tagging of *foul* as a predicative adjective (JJPR) had it not been for the fact that the JJAT tag (for attribute adjectives) is assigned to both adjectives occurring immediately before a noun and adjectives occurring immediately after a determiner. In (15), it is the latter pattern that leads to the correct identification of *foul* as an attributive adjective.

3.4 Discussion

As explained in 2.4, reporting per-token accuracy rates or reporting precision only can be very misleading. In the results of the present tagger performance evaluation, it was therefore decided to report both precision and recall for each linguistic feature across the full evaluation subsample, as well as measures broken down per register and a detailed list of all the most frequent errors detected. The reasons for doing this are threefold. First, such detailed evaluation results will help researchers interested in using the MFTE to identify features that they may not want to include in their analysis if they consider precision or recall too low for their research aims, in particular in certain registers (e.g., PGET whose recall and precision rates were <90% and is relatively infrequent). Similarly, it can help them identify features that they may want to check post-automatic tagging and (semi-)manually correct as part of fix-tagging procedures. Third, the full list of error patterns can be used by researchers to decide when it may be meaningful to merge two or more feature categories into one where low precision and recall are due to the tagger confusing a small set of features (e.g., a contender for such a merge would be the THRC and THSC categories). Thus, publishing such a detailed tagger performance evaluation makes it possible to increase the overall accuracy of the tagger for specific research purposes without necessarily having to temper with the script or perform (semi-)manual fix-tagging. Of course, since the script and the full evaluation is available, it is equally possible to adapt the tagger to the needs of a specific corpus and/or research question(s).

Some of the frequent sources of errors outlined in 3.3.2 and 3.3.3 can be averted as part of a more thorough text pre-processing phase. For instance, on the basis of this evaluation, it is recommended that, whenever possible, additional scripts are run to correct systematic OCR errors in any scanned documents to be tagged. In particular, it is important to understand that the MFTE considers line breaks (\n) as the start of a new utterance regardless of whether there is a corresponding punctuation mark. This feature is mostly very helpful but can be problematic if the texts to be tagged have retained end-of-page line breaks from their original print format. Similarly, end-of-line hyphens are also highly problematic. Of course, both of these issues can, and whenever it is feasible to do so, should be resolved pre-tagging (see 3.1 and Appendix IV). It is highly recommended that files are carefully examined for such systematic formatting issues before proceeding with the tagging.

Since repetition in transcriptions of spoken language also tends to cause tagging errors: e.g., in *yeah had a bit of yeah… had … had lunch with them*, the doubling of the token *had* means that the MFTE recognised the third *had* occurrence as a perfect aspect verb. Thus, researchers may consider removing such repetitions before proceeding with automatic tagging (cf. Brezina 2018: 172 who also recommends removing paralinguistic sounds and incomplete words before proceeding with POS tagging for the same reason). However, as such repetitions, like paralinguistic sounds and hesitations, can also be argued to be a defining feature of spoken language, such procedures should only be carried out if the research questions allow for them.

Finally, a word of warning about the evaluation procedure itself: although the tagger performance evaluation was carried out on 24 short texts or extracts of the longer texts from the BNC2014 Baby+ (see Table 4) rather than fewer longer, full texts, it remains the case that text- or topic-specific errors account for a substantial proportion of tagging errors. For instance, one short review text discussed *alpha and beta builds* at length, leading to eight occurrences of *build* or *builds* being incorrectly identified as a verb, with many of these mistags triggering further tagging errors within this one short text extract (see BNCBERe39.xlsx in Appendix V). In this context, evaluating the tagger on even shorter text extracts may have proven even more pertinent. That said, even within very short text extracts, problematic words and phrases tend to cluster. This problem is well illustrated in an extract of an academic book (see BNCBAcjM102.xlsx in Appendix V) in which the word *rugby* was incorrectly tagged as an adjective ten times within this short extract in compounds such as *rugby club* and *rugby player*, triggering twice as many tagging errors as occurrences of the word *rugby* since the NCOMP tag (for noun compound) was also systematically missing as a result of *rugby* not being recognised as a noun.

"The pursuit of truth is desirable, but often this constitutes trying to develop a model of reality, an explanation of events employing abstract and intangible concepts."
– Thomas R. Black, 1999 (cited in Neumann 2014: 43)

4 Conclusion

Examining the results of the evaluation, this concluding summary attempts to identify the extent to which the specifications elaborated in 2.1 have been met. As a reminder, the specifications are re-printed below.

The new tagger should:

1. Identify a broad, as comprehensive as possible, range of lexico-grammatical features of English
   a. that can each be meaningfully interpreted
   b. to a satisfactorily high degree of accuracy (precision and recall rates of > 90%)
   c. without the need for human intervention
   d. in a broad range of English registers with standard American or British orthography
   e. to examine register variation using multivariable methods in a broad range of general English registers.
2. Output both raw and normalised counts per text in a standard format.
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. in an accessible format with adequate documentation for researchers with basic computer skills to be able to run the programme.

With a total of 75 features covering lexical density and diversity, part-of-speech classes, verb tense and aspect, some of the most frequent lexico-grammatical constructions, and semantic categories of verbs among others, the MFTE's final feature portfolio can be said to meet the first criterion of "a broad […] range of lexico-grammatical features". The number of features is comparable to various versions of the Biber Tagger. The systematic approach taken to selecting the features (see 2.2) aimed to make it "as comprehensive as possible". As documented in 2.3 and Appendix I, all features were operationalised such that they could be "meaningfully" – in other words, functionally and linguistically – interpreted. Care was taken to avoid the overlap of feature categories. Furthermore, the considered use of three different normalisation units for different features served to avoid "obvious correlations" (see 2.4).

The formal accuracy of the MFTE was evaluated in 3.3. Whilst the vast majority of the 75 features that form part of the MFTE feature portfolio reached a "satisfactorily high degree of accuracy", 26 features did not reach "precision and recall rates of > 90%". Consequently, three of these (QLIKE, PHC and USEDTO) were removed from the MFTE's feature portfolio in versions 3.0+ (see Appendix VII: MFTE_3.0.pl). As the counts for QLIKE and PHC could be merged with the two feature categories with which these features were most often confused (LIKE and CC), these operations reduced the total number of problematic features with recall and/or precision rates of < 90% to 21. Of those, ten have F1 scores below 0.9: VIMP (F1 = 0.61), VBN (F1 = 0.67), THRC (F1 = 0.75), WHQU (F1 = 0.76), PGET (F1 = 0.80), JJPR (F1 = 0.85), THSC (F1 = 0.86), HGOT (F1 = 0.88), THATD (F1 = 0.89) and YNQU (F1 = 0.89). Note that the metrics for two of these features, PGET and HGOT, are not particularly reliable: they did not occur often enough in the evaluation data (see 3.1) to calculate bootstrapped 95% confidence intervals (see 3.3.1) and would therefore warrant further manual checks. Others, such as THRC, THSC, THATD and VBN are known to be problematic features to automatically tag and belong to those for which users of the Biber Tagger have frequently reported having to perform (semi-)manual fix-tagging. These low accuracy metrics may be the price to pay in order to meet specification criteria 1c ("without the need for human intervention") and 1a (because if linguistic features are to be "meaningful", in the sense of functionally and linguistically interpretable, it is important to distinguish, for example, verbs in the imperative and those in the infinitive form).

As already mentioned in the discussion of individual features (e.g., discussion on the operationalisation of emoticons in 2.3), it would certainly be possible to improve the tagging accuracy of certain features by making minor adjustments to the code depending on whether the texts to be tagged are fully punctuated (e.g., professionally written and edited written texts) or not (e.g., most Internet registers and the transcriptions of the Spoken BNC2014). However, the present specifications specifically called for one unique tagger to perform lexico-grammatical tagging "in a broad range of English registers". In practice, many of the Internet texts that formed part of the evaluation corpus did not (fully) conform to "standard American or British orthography" which, as shown in Table 9, considerably lowered the overall accuracy metrics reported in 3.3.1.

The MFTE not only produces a table of raw counts (thus allowing researchers to apply their own normalisations) and two tables of normalised counts (see 2.5) it also saves the tagged texts for detailed examinations of the texts themselves. This is important to ensure full transparency of the tagging process and to check the accuracy of the tagger outputs. In order to meet the final bullet point in the specification list, the MFTE source code and all the additional evaluation materials have been uploaded onto GitHub under a GPL-3.0 license for further use by the research community and public scrutiny. Criterion 6b will be met by publishing step-by-step instructions on how to install and run the MFTE together with the present document that details the development and evaluation of the tagger in the same repository.

The results of a conceptual replication of Biber's (1988) widely used multi-dimensional model of register variation in general spoken and written English using the BNC2014 Baby+ as tagged by the MFTE will soon be published.

Acknowledgments

This document is based on revised, selected chapters from an M.Sc. thesis submitted for the degree of *Master of Science in Cognitive Science* at the Institute of Cognitive Science, Osnabrück University (Germany) on 5 November 2021. I would like to thank Peter Uhrig and Michael Franke for supervising my M.Sc. thesis. Many thanks to Andrea Nini for releasing the MAT under an open-source licence. Heartfelt thanks also go to Stefanie Evert, Muhammad Shakir, and Luke Tudge who contributed advice and code in various ways (see comments in code for details) and to Larissa Goulart for her insights into the Biber Tagger. Finally, I would also like to thank Dirk Siepmann for supporting this project.

UNIVERSITÄT
OSNABRÜCK

References

Altenberg, Bengt. 1989. Review of Douglas Biber (1988) Variation across speech and writing. *Studia Linguistica* 43(2). 167–174. https://doi.org/10.1111/j.1467-9582.1989.tb00800.x.

Amasyalı, M. Fatih & Banu Diri. 2006. Automatic Turkish Text Categorization in Terms of Author, Genre and Gender. In Christian Kop, Günther Fliedl, Heinrich C. Mayr & Elisabeth Métais (eds.), *Natural Language Processing and Information Systems* (Lecture Notes in Computer Science), 221–226. Berlin, Heidelberg: Springer. https://doi.org/10.1007/11765448_22.

Argamon, Shlomo. 2019. Register in computational language research. *Register Studies* 1(1). 100–135. https://doi.org/10.1075/rs.18015.arg.

Asencion-Delaney, Y. & J. Collentine. 2011. A Multidimensional Analysis of a Written L2 Spanish Corpus. *Applied Linguistics* 32(3). 299–322. https://doi.org/10.1093/applin/amq053.

Baroni, Marco & Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing* 21(3). 259–274. https://doi.org/10.1093/llc/fqi039.

Baroni, Marco & Stephanie Evert. 2009. Statistical methods for corpus exploitation. In Merja Kytö & Anke Lüdeling (eds.), *Corpus Linguistics. An International Handbook*, vol. 2, 777–803. Berlin. Mouton de Gruyter.

Bartlett, Tom & Gerard O’Grady (eds.). 2017. *The Routledge handbook of systemic functional linguistics* (Routledge Handbooks in Linguistics). London; New York, NY: Routledge.

Bateman, John A. 2017. The place of systemic functional linguistics as a linguistic theory in the twenty-first century. In Tom Bartlett & Gerard O’Grady (eds.), *The Routledge handbook of systemic functional linguistics* (Routledge Handbooks in Linguistics), 11–26. London; New York, NY: Routledge.

Bedrick, Steven, Russell Beckley, Brian Roark & Richard Sproat. 2012. Robust kaomoji detection in Twitter. In, 56–64. https://aclanthology.org/W12-2107.pdf (25 October, 2021).

Berber Sardinha, Tony. 2017. Lexical priming and register variation. In Michael Pace-Sigge & Katie J. Patterson (eds.), *Lexical Priming: Applications and Advances* (Studies in Corpus Linguistics), vol. 79, 190–230. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.79.08ber. https://benjamins.com/catalog/scl.79.08ber (30 January, 2020).

Biber, Douglas. 1984. *A model of textual relations within the written and spoken modes*. University of Southern California PhD.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024.

Biber, Douglas. 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, Douglas. 1995. *Dimensions of Register Variation*. Cambridge, UK: Cambridge University Press.

Biber, Douglas & Susan Conrad. 2001. Quantitative corpus-based research: Much more than bean counting. *TESOL quarterly* 35(2). 331–336.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay & Alfredo Urzua. 2004. *Representing Language Use in the University: Analysis of the TOEFFL 2000 Spoken and Written Academic Language Corpus* (TOEFL Monograph Series). Princeton, NJ: Educational Testing Service.

Biber, Douglas & 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. http://doi.wiley.com/10.1002/j.2333-8504.2013.tb02311.x (30 January, 2020).

Bohmann, Axel. 2017. *Variation in English world-wide: Varieties and genres in a quantitative perspective*. Austin: University of Texas.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. 1st edn. Cambridge University Press. https://doi.org/10.1017/9781316410899.

Brezina, Vaclav, Abi Hawtin & Tony McEnery. 2021. The Written British National Corpus 2014 – design and comparability. *Text & Talk*. De Gruyter Mouton 41(5–6). 595–615. https://doi.org/10.1515/text-2020-0052.

Brezina, Vaclav, Matt Timperley & Tony McEnery. 2021. *LancsBox*. Available from: corpora.lancs.ac.uk/lancsbox.

Burnard, Lou (ed.). 2007. Reference Guide for the British National Corpus (XML Edition). http://www.natcorp.ox.ac.uk/XMLedition/URG/ (17 September, 2021).

Conrad, Susan & Douglas Biber (eds.). 2013. *Variation in English: Multi-Dimensional Studies* (Studies in Language and Linguistics). New York: Routledge.

Crossley, Scott A., Max M. Louwerse, Philip M. McCARTHY & Danielle S. McNAMARA. 2007. A Linguistic Analysis of Simplified and Authentic Texts. *The Modern Language Journal* 91(1). 15–30. https://doi.org/10.1111/j.1540-4781.2007.00507.x.

Crossley, Scott & Max M. Louwerse. 2007. Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics* 12(4). 453–478. https://doi.org/10.1075/ijcl.12.4.02cro.

Deshors, Sandra C. & Paula Rautionaho. 2018. The progressive versus non-progressive alternation: A semantic exploration across World Englishes. *English World-Wide* 39(3). 309–337. https://doi.org/10.1075/eww.00016.des.

Diwersy, Sascha, Stephanie Evert & Stella Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*, 174–204. Berlin, Boston: De Gruyter.

Egbert, Jesse, Tove Larsson & Douglas Biber. 2020. *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. 1st edn. Cambridge University Press. https://doi.org/10.1017/9781108888790. https://www.cambridge.org/core/product/identifier/9781108888790/type/element (5 November, 2020).

Egbert, Jesse & Shelley Staples. 2019. Doing Multi-Dimensional Analysis in SPSS, SAS, and R. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 125–144. Bloomsbury Academic. https://doi.org/10.5040/9781350023857.

Evert, Stephanie. 2018. Statistics for Linguists with R – A SIGIL Course: Unit 7: A multivariate approach to linguistic variation. FAU Erlangen-Nürnberg. http://www.stephanie-evert.de/SIGIL/sigil_R/ (4 November, 2021).

Evert, Stephanie & Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In

Gert De Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical Translation Studies*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110459586-003.

Feinerer, Ingo, Kurt Hornik & David Meyer. 2008. Text mining infrastructure in R. *Journal of Statistical Software* 25(5). 1–54.

Giannoulis, Elena & Lukas R. A. Wilde (eds.). 2019. *Emoticons, Kaomoji, and Emoji: The Transformation of Communication in the Digital Age*. New York: Routledge. https://doi.org/10.4324/9780429491757.

Goulart, Larissa & 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. https://doi.org/10.1558/jrds.18454.

Gray, Bethany. 2019. Tagging and counting linguistic features for multi-dimensional analysis. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 43–66. Bloomsbury Academic. https://doi.org/10.5040/9781350023857.

Halliday, Michael A. K. 1993. *Language as social semiotic: The social interpretation of language and meaning*. 1. publ. in paperback, 8. impr. London: Arnold.

Halliday, Michael Alexander Kirkwood & Ruqaiya Hasan. 1991. *Language, context and text: aspects of language in a social-semiotic perspective* (Social Semiotic). 3. impr. Oxford: Oxford Univ. Press.

Huddleston, Rodney D & Geoffrey K Pullum. 2016. *The Cambridge grammar of the English language*.

Hymes, Dell. 1984. Sociolinguistics: Stability and consolidation. *International Journal of the Sociology of Language*. Mouton Publishers 45. 39–45.

Kanaris, Ioannis & Efstathios Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *19th IEEE international conference on tools with artificial Intelligence (ICTAI 2007)*, vol. 2, 3–10. https://doi.org/10.1109/ICTAI.2007.107.

Kuhn, Max. 2020. *caret: Classification and regression training*. https://CRAN.R-project.org/package=caret (2 September, 2021).

Labov, William. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press.

Le Foll, Elen. 2021a. *A New Tagger for the Multi-Dimensional Analysis of Register Variation in English*. Osnabrück University: Institute of Cognitive Science Unpublished M.Sc. thesis.

Le Foll, Elen. 2021b. Register Variation in School EFL Textbooks. *Register Studies* 3(2). https://doi.org/10.1075/rs.20009.lef.

Le Foll, Elen. in preparation. *Textbook English: A Corpus-Based Analysis of the Language of EFL textbooks used in Secondary Schools in France, Germany and Spain*.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014. *International Journal of Corpus Linguistics* 22(3). 319–344. https://doi.org/10.1075/ijcl.22.3.02lov.

Luyckx, Kim. 2010. *Scalability issues in authorship attribution*. Brussels: University Press Antwerp.

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* (Lecture Notes in Computer Science), vol. 6608, 171–189. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-19400-9_14.

Marcus, Mitch, Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. *Building a Large Annotated Corpus of English: The Penn Treebank*. Fort Belvoir, VA: Defense Technical Information Center. https://doi.org/10.21236/ADA273556. http://www.dtic.mil/docs/citations/ADA273556 (3 October, 2021).

Matthiessen, Christian M.I.M. 1995. *Lexicogrammatical Cartography: English Systems* (Textbook Series in the Language Sciences). Tokyo: International Language Sciences Publication.

Matthiessen, Christian M.I.M. 2019. Register in Systemic Functional Linguistics. *Register Studies* 1(1). 10–41. https://doi.org/10.1075/rs.18010.mat.

Neumann, Stella. 2014. *Contrastive register variation: a quantitative approach to the comparison of English and German* (Trends in Linguistics. Studies and Monographs volume 251). Berlin; Boston: De Gruyter Mouton.

Neumann, Stella & Stephanie Evert. 2021. A register variation perspective on varieties of English. In Elena Seoane & Douglas Biber (eds.), *Corpus-based approaches to register variation* (Studies in Corpus Linguistics volume 103), 144–178. Amsterdam; Philadelphia: Benjamins.

Nini, Andrea. 2014. *Multidimensional Analysis Tagger (MAT)*. http://sites.google.com/site/multidimensionaltagger (18 September, 2019).

Nini, Andrea. 2017. Register variation in malicious forensic texts. *International Journal of Speech Language and the Law* 24(1). 99–126. https://doi.org/10.1558/ijsll.30173.

Nini, Andrea. 2019. The Multi-Dimensional Analysis Tagger. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 67–96. New York: Bloomsbury.

O'Donnell, Mick. 2017. Interactions between natural- language processing and systemic functional linguistics. In Tom Bartlett & Gerard O'Grady (eds.), *The Routledge handbook of systemic functional linguistics* (Routledge Handbooks in Linguistics), 761–574. London; New York, NY: Routledge.

Picoral, Adriana, Shelley Staples & Randi Reppen. 2021. Automated annotation of learner English: An evaluation of software tools. *International Journal of Learner Corpus Research* 7(1). 17–52. https://doi.org/10.1075/ijlcr.20003.pic.

Popescu, Marius. 2011. Studying Translationese at the Character Level. In *Proceedings of Recent Advances in Natural Language Processing*, 634–639. Hissar, Bulgaria.

Rautionaho, Paula & Robert Fuchs. 2020. Recent change in stative progressives: a collostructional investigation of British English in 1994 and 2014. *English Language and Linguistics* 25(1). 35–60. https://doi.org/10.1017/S136067431900042X.

Ruette, Tom, Dirk Geeraerts, Yves Peirsman & Dirk Speelman. 2014. Semantic weighting mechanisms in scalable lexical sociolectometry. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating Dialectology, Typology, and Register Analysis*. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110317558.205.

Santorini, Beatrice. 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project*. University of Pennsylvania, School of Engineering and Applied Science. https://www.nilsreiter.de/assets/2019-09-06-reflected-text-analysis/Penn-Treebank-Tagset.pdf (1 January, 2021).

Schegloff, Emanuel A. 1993. Reflections on Quantification in the Study of Conversation. *Research on Language & Social Interaction* 26(1). 99–128. https://doi.org/10.1207/s15327973rlsi2601_5.

Schnoebelen, Tyler. 2012. Do You Smile with Your Nose? Stylistic Variation in Twitter Emoticons. *University of Pennsylvania Working Papers in Linguistics* 18(2). https://repository.upenn.edu/pwpl/vol18/iss2/14.

Scott, Mike. 2011. *WordSmith Tools*. Stroud: Lexical Analysis Software.

Sinclair, John McH. 1992. The automatic analysis of corpora. In Jan Svartvik (ed.), *Directions in Corpus Linguistics*, 379-400. Berlin, New York: De Gruyter Mouton.

Smitterberg, Erik. 2005. *The progressive in 19th-century English: A process of integration* (Language and Computers 54). Amsterdam: Rodopi.

Spoustová, Drahomíra, Jan Hajič, Jan Raab & Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron POS tagger. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, 763–771. Athens, Greece: Association for Computational Linguistics. http://portal.acm.org/citation.cfm?doid=1609067.1609152 (12 August, 2021).

Staples, Shelley, Geoffrey T. Laflair & Jesse Egbert. 2017. Comparing Language Use in Oral Proficiency Interviews to Target Domains: Conversational, Academic, and Professional Discourse. *The Modern Language Journal* 101(1). 194–213. https://doi.org/10.1111/modl.12385.

Szmrecsanyi, Benedikt. 2013. Analysing aggregated linguistic data. In Manfred Krug & Julia Schluter (eds.), *Research Methods in Language Variation and Change*, 433–455. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt. 2019. Register in variationist linguistics. *Register Studies* 1(1). 76–99. https://doi.org/10.1075/rs.18006.szm.

Tabachnick, Barbara G. & Linda S. Fidell. 2014. *Using multivariate statistics* (Always Learning). Pearson new international edition, sixth edition. Harlow: Pearson.

Taylor, Charlotte. 2010. Science in the news: a diachronic perspective. *Corpora* 5(2). 221–250.

Teich, Elke, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes & Ekaterina Lapshinova-Koltunski. 2016. The linguistic construal of disciplinarity: A data-mining approach using register features. *Journal of the Association for Information Science and Technology* 67(7). 1668–1678. https://doi.org/10.1002/asi.23457.

Toutanova, Kristina, Dan Klein, Christopher D. Manning & Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In, 173–180. Association for Computational Linguistics.

Toutanova, Kristina & Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics* -, vol. 13, 63–70. Hong Kong: Association for Computational Linguistics. https://doi.org/10.3115/1117794.1117802. http://portal.acm.org/citation.cfm?doid=1117794.1117802 (18 August, 2020).

Unicode.org. Full Emoji List, v.11.0. https://unicode.org/emoji/charts-11.0/full-emoji-list.html (25 October, 2021).

Volansky, Vered, Noam Ordan & Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 98–118. https://doi.org/10.1093/llc/fqt031.

Wall, Larry. 1994. *The PERL Programming Language*. http://www.fxjyzy.com:8080/ebook/%E5%B9%BF%E4%BF%A1%E4%B9%A6%E5%BA%93/1211/gjfd/ts005085.pdf (2 November, 2021).

Wallis, Sean. 2020. *Statistics in Corpus Linguistics Research: A New Approach*. 1st edn. Routledge. https://doi.org/10.4324/9780429491696.

Whitelaw, Casey & Shlomo Argamon. 2004. Systemic Functional Features in Stylistic Text Classification. *AAAI Technical Report* (7).

R packages and python libraries used

Feinerer, Ingo & Kurt Hornik (2020). tm: Text Mining Package. R package version 0.7-8. https://CRAN.R-project.org/package=tm

Harris, C.R., Millman, K.J., van der Walt, S.J. *et al.* (2020) Array programming with NumPy. Nature 585, 357–362. DOI: 10.1038/s41586-020-2649-2.

Henry, Lionel & Hadley Wickham (2020). purrr: Functional Programming Tools. R package version 0.3.4. https://CRAN.R-project.org/package=purrr

Hornik, Kurt (2020). NLP: Natural Language Processing Infrastructure. R package version 0.2-1. https://CRAN.R-project.org/package=NLP

Kuhn, Max (2020). caret: Classification and regression training. https://CRAN.R-project.org/package=caret

Müller, Kirill (2020). here: A Simpler Way to Find Your Files. R package version 1.0.1. https://CRAN.R-project.org/package=here

Müller, Kirill & Hadley Wickham (2021). tibble: Simple Data Frames. R package version 3.1.5. https://CRAN.R-project.org/package=tibble

Makowski, D., Ben-Shachar, M.S., Patil, I. & Lüdecke, D. (2020). Automated Results Reporting as a Practical Tool to Improve Reproducibility and Methodological Best Practices Adoption. CRAN. Available from https://github.com/easystats/report

Pandas development team (2020). pandas-dev/pandas: Pandas, https://doi.org/10.5281/zenodo.3509134

Pedregosa, Fabian *et mult. al.* (2011). Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research12, pp. 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Perry, Patrick O. (2021). utf8: Unicode Text Processing. R package version 1.2.2. https://CRAN.R-project.org/package=utf8

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Revelle, William (2020) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA. Version = 2.0.12. https://CRAN.R-project.org/package=psych

Signorell, Andri *et mult. al.* (2021). DescTools: Tools for descriptive statistics. R package version 0.99.40.

Ushey, Kevin, JJ Allaire & Yuan Tang (2021). reticulate: Interface to 'Python'. R package version 1.22. https://CRAN.R-project.org/package=reticulate

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Virtanen, Pauli *et mult. al.* (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.

Wickham, Hadley (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, Hadley (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Wickham, Hadley (2021). tidyr: Tidy Messy Data. R package version 1.1.3. https://CRAN.R-project.org/package=tidyr

Wickham, Hadley & Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

Wickham, Hadley & Jim Hester (2021). readr: Read Rectangular Text Data. R package version 2.0.2. https://CRAN.R-project.org/package=readr

Wickham, Hadley, Romain François, Lionel Henry & Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

Wickham *et mult. al.* (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Appendices

All the appendices as well as additional materials can be found on: https://github.com/elenlefoll/MultiFeatureTaggerEnglish. For the Rmd. files, a *renv.lock* lockfile is also enclosed. Executing renv::init() will install the exact same versions of the R packages used to run the code provided here.

Appendix I

tables/ListFullMDAFeatures_3.0.xlsx (see also pdf version, below)

Appendix II

tables/Tagger_Example_PEAS_PASS.pdf (see also below)

Appendix III

code/MFTE_2.9.pl **(not the latest MFTE version!)**

Appendix IV

code/Pre-processing_BNC2014Baby.Rmd

Appendix V

code/TaggerTestResults.Rmd

Appendix VI

code/Bootstrapped_Accuracy.ipynb

Appendix VII

code/MFTE_3.0.pl **(recommended MFTE version)**