CHAPTER 2

Talking about (Non-)Canonicity
A Study of Linguistic Terminology

Sven Leuckert and Sofia Rüdiger

2.1 Introduction

What does it mean for a syntactic structure to be 'non-canonical'? How useful is it to even have what appears at first glance to be a binary distinction delineating 'canonical' from 'non-canonical'? And how do these terms differ from related terms and concepts, such as 'markedness' and 'syntactic variation'? Both 'canonical' and 'non-canonical' have been in use in syntactic studies (and other linguistic disciplines) for decades, but pinning down a core definition is not an easy task; they appear to be elusive terms without a clear denotation other than 'non-canonical' referring to something that is less frequent or less 'standard-like' (see also the discussion in Pham et al. 2024). In order to take stock of the terms that are commonly used in academic writing to talk about syntactic (non-)canonicity, we report in this chapter on a corpus study investigating usages of 'canonical' and 'non-canonical' and related terms in a total of 783 contributions to six journals published between 2012 and 2021. These journals, *Corpora*, the *Journal of English Linguistics*, the *Journal of Germanic Linguistics*, the *Journal of Historical Linguistics*, the *Journal of Linguistics*, and *Syntax*, all represent relevant publications to the edited collection at hand and to the study of syntax in English linguistics and beyond, while, at the same time, offering diversity in perspectives and, consequently, assumed diversity in terminology (to some extent). Our aim is (1) to catalogue the frequency of terminology related to '(non-)canonicity' and (2) to provide a terminological point of reference for the following volume contributions.

What we present here is a study of linguistic terms related to 'syntactic canonicity', which, in combination with the Introduction, sets the stage for this edited volume. Other recent meta-studies, such as Kortmann (2021), Larsson et al. (2022), and Buschfeld et al. (2024), have already investigated methodological and terminological preferences related to the 'quantitative turn' in linguistics. These and similar studies (like ours) are

important for multiple reasons. As descriptive linguists, we are interested in the reality of language use and how it is researched, and English for Academic Purposes represents a specialised but important usage context. In addition, taking stock of a scientific field's terminology, if only a small subsection, may raise awareness of its breadth and complexity. In our case, the existence of different definitions of (non-)canonicity is not necessarily an issue by itself, but we know very little about the usage of the terms in the first place and to what extent definitions may vary. This situation can be remedied, at least to a certain degree, by a study of these terms in use (which, we note, might not always be accompanied by a definition).

Following this introduction, we discuss issues in linguistic terminology in relation to (non-)canonicity in Section 2.2. Section 2.3 serves to introduce and present the empirical case study of terminology referring to phenomena of syntactic (non-)canonicity in linguistic journals. In Section 2.4, we discuss and summarise our findings and give an outlook on potential future work.

2.2 Setting the Stage: Terminology and (Non-)Canonicity

In order to set the theoretical stage for the remainder of the chapter and underline the complexity of investigating linguistic terminology, we first discuss general issues in terminology in this section before giving a brief overview of how 'canonical' and 'non-canonical' have been defined in selected dictionaries and major grammars of English.

2.2.1 Some Core Issues in Linguistic Terminology

The use of linguistic terminology across subdisciplines is influenced by a range of factors, as noted by Bugarski (1983), and two of the most crucial ones in this regard are the linguistic paradigm as well as the individual definition applied in a study. A core conflict that results from the inconsistency of how terms are used is the 'distinction between standardisation and unification, allowing for the coexistence of two or more internally unified and partly overlapping standards' (Bugarski 1983: 69).[1] While an outcome of terms emerging in specific paradigms may be a 'typically short

[1] An example in the context of canonicity is the co-existence of multiple terms referring to related information-structural phenomena with sometimes subtle differences in their definitions, including 'topic'/'mental subject', 'focus'/'mental predicate', 'topicalisation'/'fronting'/'preposing', etc.; see Leuckert (2019) for an overview.

lifespan of the terms coined by the proponents of particular theoretical frameworks' (Trask 1993: viii), the situation is probably more complex when semantic ambiguity (as opposed to neologism) is involved.[2]

Another important issue is the applicability of terms across disciplines, languages, time periods, etc. As Vermeer (1971: 14–15) notes, definitions are necessarily products of a given setup of contextual conditions, including space and time. He suggests TABLE as an example: how can we define TABLE without missing some members of this category? His publication on the topic slightly predates prototype theory (Rosch 1973) but is strongly reminiscent of the idea that fuzziness is often inevitable.

While terms may be coined to describe a particular phenomenon, already existing terms may also be redefined (either intentionally or as the result of semantic shift), a practice that 'may or may not have an effect on how the term is applied in practice so that the outcome is either polysemy or inconsistency' (Mugdan 1990: 57). A recent example of intentional change is the reframing of certain concepts in Rüdiger (2019), who suggests using the qualifiers 'plus-' and 'minus-' instead of terms such as 'overuse' or 'underuse' for specific linguistic features (see Rüdiger et al. 2022 for an example of this terminology in use). While achieving 'true' terminological neutrality is probably impossible due to the individual and often purpose-driven definition of linguistic terms, reducing their evaluative or even prescriptive connotations is in line with a move towards descriptivism. This is especially relevant in the context of linguistic disciplines – such as World Englishes and sociolinguistics – that deal with varieties that historically have been or could be subject to discrimination. As this brief summary has shown, linguistic terminology and the encoded knowledge systems are rather complex and the factors influencing them are manifold. In order to gain insight into how the terms central to this edited volume are used, we next provide a brief overview of how the terms are framed in selected dictionaries and grammars before introducing our study.

2.2.2 *'Canonical' and 'Non-Canonical' in Selected Dictionaries and Grammars*

Some of the issues introduced above also apply to the 'canonical' and 'non-canonical' pair of terms. An important concept strongly linked to language

[2] An example of how this issue may be handled can be seen in Trask (1993: viii), who, in his dictionary of grammatical terminology, decided to include only terms that 'have been and are widely used by grammarians of varying theoretical persuasions, and which seem likely to remain current for some time'.

ideology is that of the 'native speaker', as noted by Hackert (2012: 30): 'at least according to a number of authors, what we are looking at if we look at the English native speaker is an imaginary or political construct, something which is discursively constituted and created, and which is about attitudes, affiliation, and social identity rather than about linguistic competence'. 'Canonical' and 'non-canonical' also fall into the category of terms that may carry intended or unintended ideological weight, since they may be consciously or subconsciously linked to ideas of correctness, producing 'native-like' language, following specific rules, and so forth; Hackert (2012: 30) links the native speaker ideology both to the history of Western linguistics and to other ideologies, including standard ideology. A look into the entry for 'canonical' in the online version of the *Oxford English Dictionary* (*OED*) reveals the following definition as the most (and only) relevant one for the syntactic context:[3]

> 4. *gen.* Of the nature of a canon or rule; of admitted authority, excellence, or supremacy; authoritative; orthodox, accepted; standard. (*OED*)

The question to what extent 'canonical' and 'non-canonical' suggest or even imply superiority of one syntactic variant over another obviously depends on how the terms are defined and applied in any given study, but their morphological composition (which includes reference to the notion of the 'syntactic canon') inevitably points to interpretations in the sense of 'according to the canon' or 'not according to the canon'. It is interesting to note, then, that some definitions of related terms include 'canonical' but not 'non-canonical', a case in point being Hickey's entry for 'word order' in *A dictionary of varieties of English*:

> **word order** The arrangement of words in a linear sequence in a sentence. There is normally an unmarked, a so-called 'canonical', word order in a language . . . but usually alternative word orders exist, particularly to allow for emphasis in a sentence such as the fronting of sentence elements for the purpose of topicalization. *See* VERB SECOND. (Hickey 2014: 348; all formatting in the original)

While it would be fascinating to consider how further dictionaries and grammars treat the semantic field of canonicity, a comprehensive overview goes far beyond the scope of this chapter. However, we consider it crucial

[3] The term 'non-canonical' is listed as part of the prefix 'non-', but the provided quotations are exclusively from religious and literary contexts. The prefix 'non-' is used 'to express a neutral negative sense, forming adjectives and occasionally nouns' (*OED*).

to address at least briefly how the three major grammars of English (Quirk et al. 1985; Biber et al. 1999;[4] Huddleston & Pullum 2002) deal with 'canonical', 'non-canonical', and related terms in the context of word order variation. As references with citations in the tens of thousands,[5] all three titles have had a significant impact not only on the dissemination but also on the stabilisation of linguistic terminology across subdisciplines. The first of the three mentions 'canonical' early in the book:

> It is a widely accepted principle . . . that **the simple declarative sentence is in a sense the canonical form of sentence**, in terms of which other types of sentence, including both those which are more complex ('complex' and 'compound' sentences) and those which are more simple ('reduced' sentences), may be explained by reference to such operations as conjunction, insertion, inversion, substitution, and transposition. (Quirk et al. 1985: 78; emphasis added)

The reasoning behind equating the 'simple declarative sentence' with the 'canonical form' remains unclear, however. The term 'non-canonical' is not used in Quirk et al. (1985) at all. Similarly, in their introduction to word order, Biber et al. (1999) use the terms 'unusual' and 'marked' but do not mention 'canonical', 'non-canonical', or 'canonicity'; the 'typical' form of a sentence is addressed with reference to the fixed nature of English word order:

> English word order has often been described as **fixed**. It is certainly true that the placement of the core elements of the clause is strictly regulated. Yet there is variation, even in the core of the clause. Consider this passage from a fiction text:
>
> *It was a beautiful grey stone mellowed by the years. There was an archway in the centre and* ***at the end of the west wing was a tower with battlements and long narrow slits of windows which looked rather definitely out of place with the rest of the house which was clearly of a later period****.* (FICT)
>
> This is a description of a house, and the house is the topical starting-point in both sentences. The portion in bold illustrates an unusual or **marked** choice of word order . . . (Biber et al. 1999: 898–9; emphasis in the original)

Biber et al. (1999: 896) note that 'marked' word order may achieve effects that what we call 'canonical' syntactic patterns cannot; such effects include achieving emphasis, contrast, and structuring the information flow.

[4] We work with this and not the newest (2021) edition, as the articles which we consider in our study were published by 2021 at the latest.

[5] We refrain from giving numbers here, since relevant resources such as Google Scholar can show tendencies but do not provide accurate citation counts.

Finally, in the relevant chapter of the *Cambridge grammar* (Huddleston & Pullum 2002), Ward et al. (2002) contrast 'canonical' and 'information-packaging' constructions:

> Our concern in this chapter is with a number of clause constructions which we refer to collectively as information-packaging constructions, **and which differ syntactically from the most basic, or canonical, constructions in the language**. These information-packaging constructions characteristically have a syntactically more basic counterpart differing not in truth conditions or illocutionary meaning but in the way the informational content is presented. (Ward et al. 2002: 1365; emphasis added)

It is interesting to consider how Ward et al. (2002) contextualise syntactic (non-)canonicity. By pointing out that 'the syntax makes available **different ways of "saying the same thing"**, with the various versions differing in the way the content is organised informationally' (Ward et al. 2002: 1365; emphasis added), they deliberately link their understanding of (non-)canonicity to syntactic variation and in turn to both Labov (1972) and, accordingly, sociolinguistic variation as well as register variation. This framing overtly suggests understanding different syntactic alternatives as variants that may be influenced by intra- and extra-linguistic factors.

As mentioned earlier, it is not possible to trace the entire history of 'canonical', 'non-canonical', 'canonicity', etc. in this section, nor was that the goal. However, our spot check of relevant publications has revealed certain tendencies: (i) 'canonical' appears to be favoured in grammars as a term over 'non-canonical', sometimes to the extent of 'non-canonical' not being named as the counterpart; (ii) 'canonical' is often framed as 'basic', 'essential', or 'close to the standard'; (iii) canonicity is frequently associated with unmarkedness, the standard, and syntactic variation. These observations represent important reference points in our empirical analysis below.

2.3 Case Study: (Non-)Canonicity in Six Linguistic Journals

Following the theoretical deliberations from the previous section, we introduce and present our case study in this section. The aim of our case study is to catalogue the usage of terminology across a range of linguistic journals and to provide a first point of reference for the terminological choices in the subsequent chapters of the edited volume. After a description of the data and method, we first show the results of a quantitative analysis before moving on to a qualitative analysis.

2.3.1 Data and Method

For the case study, we collected all articles published from 2012 to 2021 in the journals *Corpora*, the *Journal of English Linguistics*, the *Journal of Germanic Linguistics*, the *Journal of Historical Linguistics*, the *Journal of Linguistics*, and *Syntax*. An overview of the journals, including the publisher as well as the number of articles contained in our corpus, is provided in Table 2.1.

The selection of the journals to be included was driven by four main factors: (a) the relevance of the journals to the edited collection, (b) the relevance of the journals to the study of syntax more widely, (c) the accessibility of the journals to the authors, and (d) the aim to include research on historical varieties of English as well as languages other than English. *Corpora* is, as the name suggests, devoted to studies in corpus linguistics and, accordingly, has a strong empirical and methodological focus. The *Journal of Germanic Linguistics*, the *Journal of Linguistics*, and *Syntax* share an interest in typological research and a tendency to include contributions from generative grammar and related approaches, although none of the journals is exclusive in this regard. While the former two journals certainly feature a substantial number of contributions dealing with syntactic phenomena, they also include, among others, articles in morphology and phonetics. The *Journal of English Linguistics* is thematically the broadest journal in our corpus and covers topics as diverse as historical linguistics, phonetics, World Englishes, and, of course, syntax. In contrast to the three journals just mentioned, the focus of contributions to the *Journal of English Linguistics* is generally rather usage-based and

Table 2.1 *Corpus for the terminological case study*

Journal/Subcorpus	Abbreviation	Publisher	Number of articles
Corpora	*Cor*	Edinburgh University Press	131
Journal of English Linguistics	*JEngL*	SAGE	121
Journal of Germanic Linguistics	*JGL*	Cambridge University Press	97
Journal of Historical Linguistics	*JHL*	John Benjamins	111
Journal of Linguistics	*JoL*	Cambridge University Press	195
Syntax	*Syn*	Wiley	128

empirical. Finally, the *Journal of Historical Linguistics* invites papers on all facets of historical linguistics across languages, although the website description highlights that 'contributions in areas such as diachronic corpus linguistics or diachronic typology are . . . particularly welcome'.[6]

An issue in compiling the corpus has been the question of what to include and what to exclude, since the journals also publish text types other than research articles that may or may not be relevant for our study. We decided to include all research articles, introductions to special issues, and remarks/short notes and similar smaller publications, but to exclude book reviews, editorial notes, and annotated bibliographies (published separately, for instance as the final part of special issues). This resulted in a corpus of 783 publications, with an average of 130.5 publications per journal (SD = 30.98). The articles were originally available in PDF format and were then processed via OCR to create txt-files suitable for corpus analysis. Random inspection of the corpus files created in this way revealed some errors in the text files which were due to the automatic processing of the data (e.g., OCR artefacts). Overall, however, these problems seemed to be only marginal in nature, and we decided to forgo manual correction in light of feasibility. In addition, the automatised conversion process included page numbers and page headers and we therefore do not present word counts for the subcorpora and the corpus overall but instead use articles as basis for normalisation. Even though article length might vary, we believe this provides a reasonable basis for comparison for the purposes of the current study.

For the analysis, we extracted all instances of the words listed in Table 2.2 using *AntConc*. The first three groupings consist of oppositional pairs: 'canonical' and 'non-canonical', 'standard' and 'non-standard', and 'marked' and 'unmarked'.[7] The searches for 'canonical' and 'non-canonical' also included the adverb forms 'canonically' and 'non-canonically', but most of the uses recorded were either as attributive or as predicative adjectives. While 'non-standard' is typically used as an adjective as well, 'standard' may also be employed as a noun, which also explains the higher general frequency observed for this item in the analysis. As the data were not part-of-speech tagged, we present no further details on word class usage, but see Section 2.3.3 for a collocational analysis of these lexical

[6] See https://benjamins.com/catalog/jhl.

[7] The searches for 'non-canonical' and 'non-standard' included both the hyphenated and joint word forms (i.e., 'noncanonical', 'nonstandard'). In this text, we use the hyphenated spelling to refer to both orthographic realisations.

Table 2.2 *Terms considered in the analysis*

Grouping	Terms
Pair 1	*canonical*
	non-canonical[a]
Pair 2	*marked*
	unmarked[b]
Pair 3	*standard*
	non-standard
Additional terms	*canonicity*
	markedness
	syntactic variation

[a] The term 'uncanonical' occurred only once in the dataset and was thus not considered for further analysis.
[b] The term 'non(-)marked' occurred only three times in the dataset and was thus not considered for further analysis.

items. The nouns 'canonicity' and 'markedness' reflect the semantic relation between 'canonical' and 'marked' as well as their negated counterparts; however, in most general terms, 'canonicity' refers to the 'typical', whereas 'markedness' refers to the 'atypical'.

Of course, further terms and constructions, such as 'alternative / basic / different / infrequent / (a)typical / uncommon / (un)usual word order', are potentially used (near-)synonymously with the lexical items investigated in our analysis. For reasons of space, we concentrate here on the target items as specified above and as listed in Table 2.2, but further research on terminological choices should account for these alternatives as well as the actual terminological definitions given in the papers.

It needs to be pointed out that our analysis is based on the articles in their entirety, that is, including the bibliography (but also abstract, bionote, etc.). At times, our lexical target items do indeed occur in the reference section of an article, for example, when an article cites Jenny Cheshire's 1987 *Linguistics* article with the title '**Syntactic variation**, the linguistic variable, and sociolinguistic theory' (emphasis added). As the bibliography is an essential part of academic manuscripts and referencing specific linguistic works evokes the words used in their titles and thus plays a role in the reinforcement of linguistic terminology, we decided to not exclude the bibliography sections from the corpus.

Last but not least, as will also become clear in the analysis section, most of the terms, particularly those in pairs 1, 2, and 3, are also in usage outside

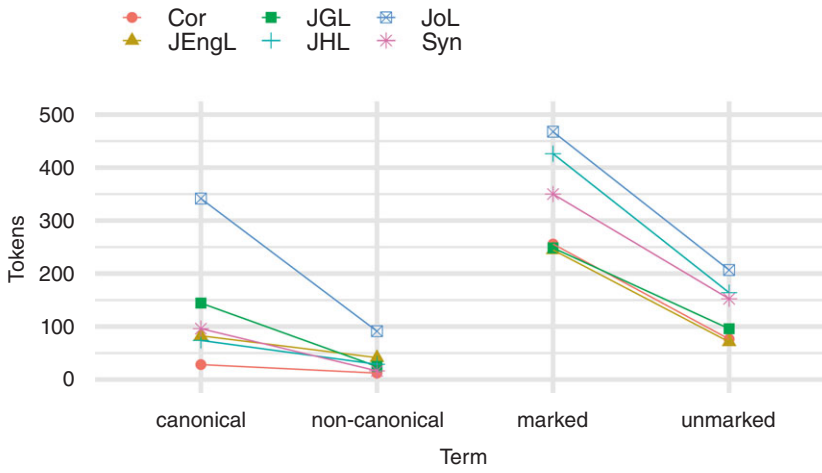

Figure 2.1 Term frequency per 100 articles per journal for 'canonical'/'non-canonical' and 'marked'/'unmarked'

the realm of syntax – and can, for example, also be found in the description of semantic phenomena. In addition, 'marked' and 'standard' (and to a certain degree '(non-)canonical') have further non-specialist meanings (e.g., *marked* used as a past tense verb form in the sentence 'The presentation of the Dynamic Model in 2007 *marked* a major milestone in the rapidly emerging field of World Englishes', *JEngL2015*)[8] or are used in fixed specialised constructions which are unrelated to specific fields (e.g., 'standard deviation'). While such uses affect the results to a certain extent, the majority of them pertain to the present study. Thus, manually weeding out irrelevant hits (an extremely time-consuming task) was not considered beneficial. We will expand on these aspects further in Section 2.3.3 on the qualitative analysis of our data.

2.3.2 *Quantitative Analysis*

First, we consider the frequencies of the terms listed in Table 2.2. Figure 2.1 shows the normalised frequencies of the first two pairs across all six journals.[9] We normalised the figures per 100 articles instead of per a

[8] Please note that we reference datapoints in our corpus with identifiers giving the journal and year of publication. This also means that we do not cite the respective articles.

[9] All figures were created in *R*, using either the *ggplot2* or the *quanteda* (see Benoit et al. 2018) packages.

certain word count (see Section 2.3.1); the values for all terms as well as their range (i.e., occurrence across articles in each journal) are listed in Appendix A (Table 2.5).

Figure 2.1 shows that, for both term pairs, the non-negated term occurs more frequently than the negated term. Comparing 'canonical' and 'non-canonical' across all six journals using a chi-squared test reveals statistically highly significant differences ($\chi^2 = 24.074$, df $= 4$, $p < 0.005$, Cramer's $V = 0.130389$). However, with the exception of the *Journal of Linguistics*, the differences in frequency between 'canonical' and 'non-canonical' appear less extreme than the differences between 'marked' and 'unmarked'. For this pair, the statistical difference across all journals is also significant, albeit to a lesser degree when compared to 'canonical' and 'non-canonical'; the effect size is noticeably small for the frequency difference ($\chi^2 = 10.755$, df $= 4$, $p < 0.05$, Cramer's $V = 0.05684054$).

The higher frequency of the non-negated term appears to be even more extreme in the 'standard'/'non-standard' pair, which is depicted in Figure 2.2 and for which the differences across the journals are, again, statistically highly significant ($\chi^2 = 179.75$, df $= 4$, $p < 0.005$, Cramer's $V = 0.2176041$).

In *Corpora*, 28.24 tokens of 'canonical' and 12.21 tokens of 'non-canonical' occur per 100 articles, which means that the non-negated form occurs slightly more than twice as often. In contrast, 'standard' occurs at 204.58 and 'non-standard' at 50.38 tokens per 100 articles, meaning that the

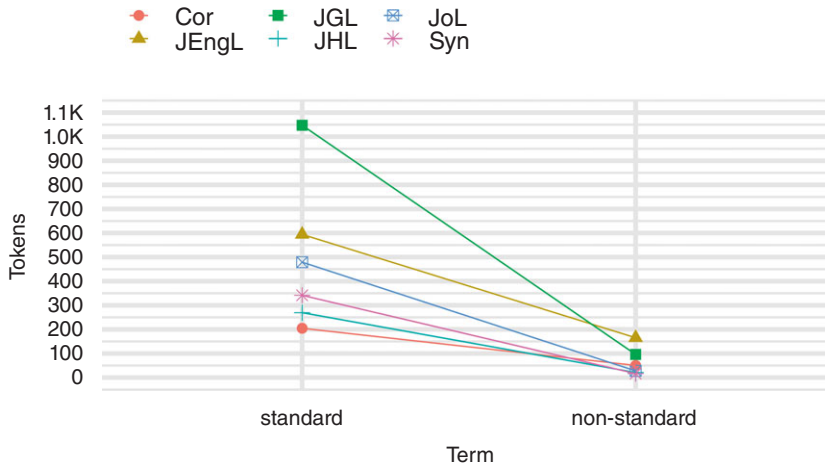

Figure 2.2 Term frequency per 100 articles per journal for 'standard'/'non-standard'

non-negated form occurs four times as often. However, as mentioned above, this is likely due to the multiple word class membership of 'standard'.

While the tendency of the non-negated form being more frequent than the negated form appears consistent throughout, it is important to consider dispersion as an additional measure. When authors decide on a set of terms in their study, it seems likely that they stick to these terms, which means that investigating dispersion may reveal potential imbalances in term usage across journals. To illustrate this phenomenon, Figure 2.3 shows an X-ray plot of the dispersion of 'canonical', 'non-canonical', 'marked', and 'unmarked' in the articles published in *Syntax* in 2021.

Overall, 14 publications are part of this corpus segment. However, as three publications do not contain any of the four items, only 11 files are included in the X-ray plot. Six of the 11 publications (54.55%) use a term from both pairs at least once; in some cases, a clear preference for one pair (e.g., 'marked'/'unmarked' in file 13) is obvious. In addition to the X-ray plot, we calculated dispersion measures based on Gries (2008, 2020) and the corresponding *R* script. More precisely, we calculated the *DP* value (deviation of proportions) for the three pairs in a comparison of all six journals.[10] The advantages of *DP* over other dispersion measures are manifold and explained in detail in Gries (2008) and Gries (2020); a key feature is that *DP* is able to deal with uneven corpus sizes, which makes it appropriate for our purposes.[11] The values are shown below:

- 'canonical': $DP = 0.2943403$
- 'non-canonical' (incl. 'noncanonical'): $DP = 0.4083444$
- 'marked': $DP = 0.08008977$
- 'unmarked': $DP = 0.1218987$
- 'standard': $DP = 0.2161513$
- 'non-standard' (incl. 'nonstandard'): $DP = 0.48796315$

In general, lower *DP* values indicate a more even spread, and higher *DP* values indicate a more uneven spread across the dataset. While the low *DP* value for 'marked' is not surprising given its multi-word-class status, 'unmarked' has the second-lowest frequency, meaning that it is also comparatively evenly dispersed. 'Non-canonical' and 'non-standard'

[10] We would like to thank Stefan Th. Gries and Tobias Bernaisch for their invaluable help in calculating the *DP* values in R.

[11] Note that Gries (2022) outlines new developments in how the relation between dispersion and frequency needs to be factored into calculations.

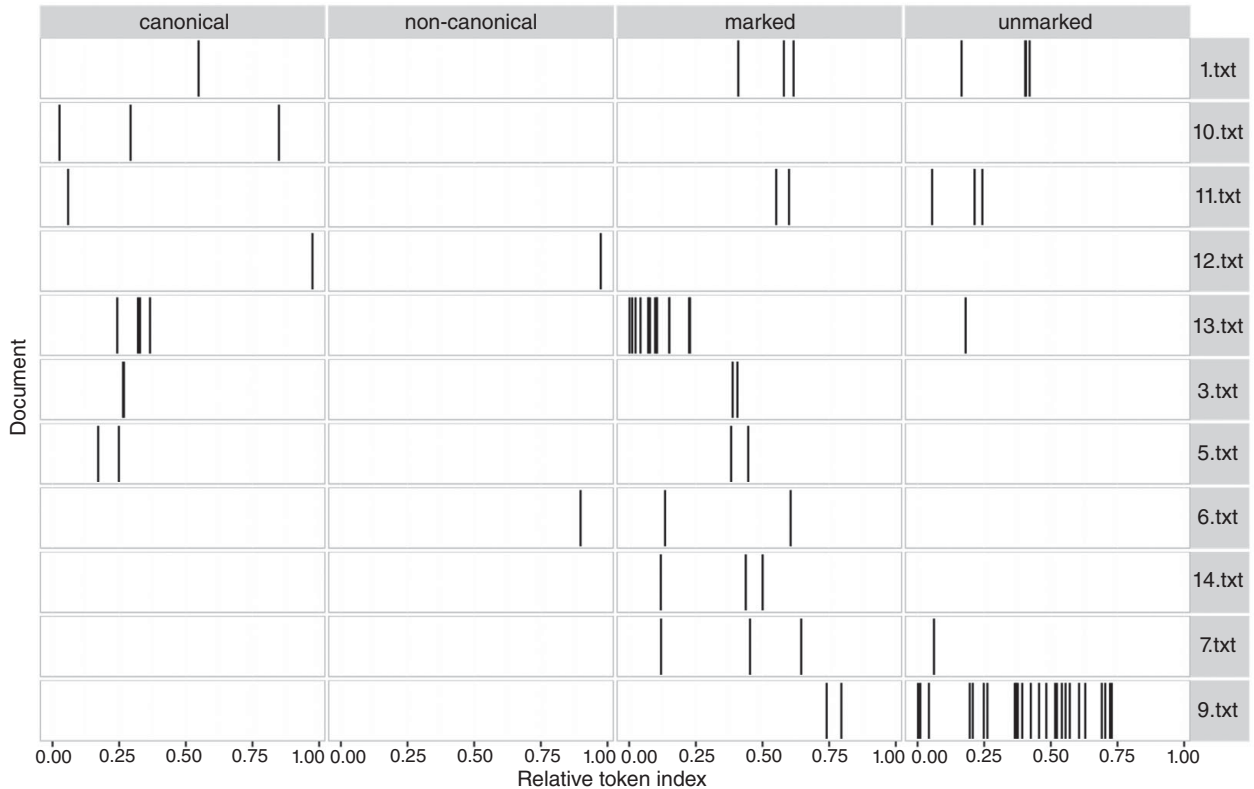

Figure 2.3 X-ray plot showing the dispersion of 'canonical', 'non-canonical', 'marked', and 'unmarked' in articles published in *Syntax* in 2021

Table 2.3 *Absolute and normalised frequencies of 'canonicity', 'markedness', and 'syntactic variation' in six linguistic journals as well as range across articles within each journal*

Term		*Cor*	*JEngL*	*JGL*	*JHL*	*JoL*	*Syn*
'canonicity'	AF	8	5	7	0	11	0
	NF	6.11	4.13	7.22	0	5.64	0
	Occurrence	2/131	2/121	3/97	0/111	8/195	0/128
	(%)	(1.53)	(1.65)	(3.09)	(0.00)	(4.10)	(0.00)
'markedness'	AF	53	28	34	123	399	41
	NF	40.46	23.14	35.05	113.89	204.62	32.03
	Occurrence	5/131	12/121	16/97	14/111	55/195	24/128
	(%)	(3.82)	(9.92)	(16.49)	(12.61)	(28.21)	(18.75)
'syntactic variation'	AF	6	43	16	12	18	9
	NF	4.58	35.54	16.49	11.11	9.23	7.03
	Occurrence	4/131	18/121	8/97	8/111	16/195	10/128
	(%)	(3.05)	(14.88)	(8.25)	(7.21)	(8.21)	(7.81)

Note. AF = absolute frequency; NF = normalised frequency; Occurrence = occurrence in number of articles.

have the highest *DP* values, which, in general terms, means that they are the least evenly dispersed of the six terms. This suggests that their dispersion is 'clumpy', that is, they are not used evenly by a range of authors. 'Canonical' and 'standard' fall between the other terms mentioned so far.

Finally, the absolute and normalised frequencies per 100 articles of the terms 'canonicity', 'markedness', and 'syntactic variation' as well as their range across articles within each journal are presented in Table 2.3.

The figures show that, in general, 'canonicity' is not a frequent term in any of the journals, with zero hits in the *Journal of Historical Linguistics* and *Syntax*. 'Markedness', on the other hand, occurs relatively frequently across the corpus, with the *Journal of Linguistics* and the *Journal of Historical Linguistics* boasting the highest frequencies. Finally, 'syntactic variation' is most frequently used in the *Journal of English Linguistics*, which underlines the journal's tendency to favour usage-based, empirical, and often sociolinguistic contributions. However, it is important to be aware that individual preferences of authors, editorial guidelines, text type, etc. may all have an influence on how terms are used and defined; a closer look into the terms in use is provided in the next section.

2.3.3 Qualitative Analysis

In the following, we present the results of an n-gram analysis for the first three target pairs, grouped by the six journals and focusing on the top five bigrams (with the target item occurring in the first slot).[12] The analysis was conducted using *AntConc*'s clusters/n-grams tool, with the minimum frequency set to two occurrences. The table with the full list of results can be found in Appendix B (Table 2.6). We use the results from the top five bigram analysis to inform our further analysis of specific constructions. It should be mentioned that absence from the top five bigram list does not automatically entail that the specific word combination is not used in a journal, as it could merely be less frequent than the top five listed. Where necessary, bigrams were extended to include following items (e.g., to complete noun phrases).

Pair 1: 'Canonical' and 'Non-Canonical'

The top five bigram analysis shows that across the subcorpora 'canonical' is frequently followed by words designating specific word classes (e.g., *verbs*, *transitive* [*verbs*]) or functions (e.g., *subject*, *null-subject*, *complement*). Other nominal right collocates indicate more general constructs which are consequently considered 'canonical' in the articles: *sequence/s*, *use*, *status*, *utterance*, and *position/s*. In *Corpora* and the *Journal of English Linguistics*, we also find right collocates pointing towards usage in the field of semantics: *oppositions*, *oppositional*, and *antonyms* (likewise for 'non-canonical' with *opposition/s*). In two cases (*JEngL* and *JGL*), the conjunction *and* is listed in the top five right collocates. Considering the whole corpus, 'canonical' is followed by the coordinating conjunction *and* 33 times. Most frequently (n = 14, range = 8) this concerns coordination with 'non-canonical' to set up a contrast as demonstrated in (1) and (2).[13]

[12] For reasons of space, we discuss only the right collocates of our target items in this chapter; left collocate searches (i.e., bigrams with the target item occurring in the second slot) turned out to be rather unrevealing, as they mainly resulted in articles and prepositions. An exception was 'marked', for which the left collocate search helped us to determine that it was frequently used as a verb in our corpus (as attested by the frequency of word forms of *be* before 'marked' – *is* [n = 347, range = 188], *are* [n = 238, range = 133], and *be* [n = 185, range = 104] all feature in the top five bigram list for the whole corpus). The use of *more* (n = 47; range = 41) and *most* (n = 27; range = 12) as L1 collocates of 'marked' is also worthy of mention here, as it points to a certain gradability of the concept (for 'unmarked' the frequencies are as follows: *more unmarked* n = 3; range = 3; *most unmarked* n = 4; range = 3).

[13] The emphases in these and the following examples are ours.

(1) The revised typology with **canonical and non-canonical** examples is set out in Table 2. (*JEngL2012*)

(2) On the basis of the observation that mixing **canonical and non-canonical** forms normally proceeds in that order … (*JoL2013*)

Other low-frequency trigrams include 'canonical and clefted', 'canonical and derived', 'canonical and partial', 'canonical and impersonal', 'canonical and inversed', and 'canonical and prototypical'.

As 'non-canonical' is overall much rarer in the corpus than 'canonical' ($n_{non\text{-}canonical} = 321$; $n_{canonical} = 1{,}148$), fewer bigrams were available as well. In some cases, collocations pointed towards specific syntactic phenomena (i.e., *subjects*, *agreement*, *plural*, *passives*, or *case*) or specifically at phenomena related to syntactic structure (i.e., *position/s*, *order*, *word* [*order*]).

Pair 2: 'Marked' and 'Unmarked'

According to our bigram analysis, 'marked' is predominantly used as a verb in our corpus and is usually followed by a preposition. The list of top five right collocates across the subcorpora only contains two non-prepositions: *plural* (n = 21; range = 1) in *Corpora* as well as *and* in the *Journal of Linguistics* (n = 52; range = 15). Across all subcorpora, 'marked and' occurs 125 times. Most frequent is the combination 'marked and unmarked' (n = 31; range = 10), setting up a similar contrast as seen above for 'canonical and non-canonical', see (3).

(3) One of the most long-standing debates in the generative framework has hinged on the specification of the features and/or principles that motivate **marked and unmarked** syntactic orders. (*Syn2019*)

In terms of prepositional right collocates, the preposition *for* is of interest to us here (occurring in all top five bigrams with 'marked' in the left position), as this potentially shows us *what* is being marked. In Table 2.4, we list all right collocates for 'marked for' and 'unmarked for' which occur at least three times across all subcorpora.

As can be seen in Table 2.4, and in contrast to example (3) above, 'marked for' clearly has the primary meaning of referring to the presence of morphological marking, that is, the authors refer to verbs, for example, being marked for the *progressive*, *tense*, or the *past*, and nouns and other parts of speech for *case*, *gender*, or *definiteness*. The same applies to 'unmarked for', which occurs four times, each followed by *case* or *past*. What is absent here is an underlying notion of typicality/ frequency/ standardness, which contrasts with how 'canonical' and 'non-canonical' are used in the corpus (see above).

Table 2.4 *Right collocates of 'marked for' and 'unmarked for' occurring at least three times in the corpus*

marked for ...		*unmarked for* ...
the (n = 33)	*progressive* (n = 5)	*case*, *past* (n = 4)
	absence, *genitive* (n = 4)	
	dense, *use*, *same* (n = 2)	
	[nonce uses omitted]	
tense (n = 15)		
case (n = 9)		
gender, *number*, *past* (n = 7)		
person (n = 6)		
progressive (n = 5)		
definiteness, *feminine* (n = 4)		
accusative, *aspect*, *deletion*, *force*, *masculine* (n = 3)		

The top five right collocates of 'unmarked' in five of the six journals surveyed contained *form* and/or *forms*. Here, 'unmarked' primarily seems to reference the absence of a particular kind of marking (usually one that would be expected). In (4), for example, an English verb used by a speaker is described as 'unmarked' for the past tense and in (5) the 'unmarked form' refers to a Korean noun which is used without a subject case marker (but which nonetheless is considered 'the better option').

(4) She uses the **unmarked form** *give* to reference this past difficulty. (*JEngL2017*)

(5) Whereas both the case-marked and noncase-marked form are acceptable in (a), the **unmarked form** is the better option in (b). (*JoL2016*)

This seems to contrast with the results for 'un/marked for' given above and indicates that more in-depth qualitative analysis is necessary to further disambiguate the different usages of these terms.

Pair 3: 'Standard' and 'Non-Standard'

The top five right collocates of 'standard' across our subcorpora mainly subsume references to specific standard languages, such as *English* (found in the top five of all journals except *JGL*), *Dutch* and *German* (*JGL*), *Finnish* (*JHL*), and *Polish*, *Arabic*, and *Russian* (*JoL*). In addition, the top five right collocates of three journals contain the generic *standard language* (*JEngL*, *JGL*, and *JHL*), with the top five of the *Journal of Germanic Linguistics* also featuring the combination *standard variety*.

'Standard' also forms part of specific statistical terminology, such as *standard deviation* and *standard error*, and unsurprisingly these items feature prominently in the top five of journals with a largely quantitative focus (*Cor*, *JEngL*, *Syn*). Further combinations related to methodology can be found in *standard reference* [*corpus*/*corpora*] (*Cor*), *standard of comparison* (*Syn*), and *standard analysis* (*Syn*). The only term related to a specific linguistic phenomenon is *negation*, which occurs 73 times in the *Journal of Historical Linguistics* (range = 3). As the coordinative conjunction *and* features in three top five lists (*Cor*, *JEngL*, *JGL*), we also had a closer look at 'standard and' constructions across all subcorpora and beyond the top five bigrams. Altogether, 'standard and' occurs 94 times throughout our corpus (range = 38). Most frequent is the combination 'standard and vernacular' (n = 24), but all of these instances are found in two articles only. More widely dispersed is the combination 'standard and non-standard' (n = 17, range = 10), followed by *varieties* in eight cases (range = 5).[14]

While the top five right collocates of 'non-standard' also feature specific languages, such as *Polish* (*JEngL*) and *English* (*JoL*) or specific groups of languages (*Ibero-Romance*; *JHL*), general references (which might or might not be specific in context) are distinctively frequent: *varieties* can be found in all six top five lists, *language* in two, and *dialects* in one. Terms indicating non-standard usage are also prominent: we thus find 'non-standard' used to specify particular *forms* and *form* (in two top five lists each), *sentences*, *features*, *use*, and *uses* (in one each). In addition, specific phenomena are singled out as being of non-standard nature: *capitalisation* (n = 25; range = 1) and *spellings* (n = 5; range = 4) (both in *Cor*), and *gender agreement*/ *assignment*/ *marking* (n = 6; range = 2; *JGL*). Last but not least, we took the occurrence of *and* in the top five bigram list of the *Journal of English Linguistics* as starting point for a search for this coordinative construction across all corpora. The following items were identified as occurring together with 'non-standard': 'non-standard and uncommon' (n = 15; range = 1), 'non-standard and spontaneous', 'non-standard and stigmatized', 'non-standard and/or non-native', 'non-standard and variable', and 'non-standard and vernacular' (each n = 1). This indicates that non-standard

[14] The other nine combinations only occur once: *accents*, *categories*, *components*, *forms*, *part*, *parts*, *realisations*, *pronominal usage*, and *uses* (ordered alphabetically and extended to full noun phrase if necessary).

features are associated with low prestige (*stigmatised*) and specific modes of production (*non-native*, *spontaneous*, *vernacular*). Analysis of the concordance lines revealed that 'non-standard and uncommon' was not used as a list of adjectives to refer to one phenomenon (i.e., pointing out that non-standard features are also of low frequency), but instead referred to two different constructions – one deemed 'non-standard', the other 'uncommon' (see also the next section, on meta-discussion of terminology).

2.3.4 Meta-Discussion of Terminology

As can be seen in the use of constructions such as 'I/we label' (n = 24), 'I/we term' (n = 25), and 'I/we call' (n = 93),[15] article authors at times motivate their terminological choices or at least draw explicit attention to them. In a few cases, such as (6), this concerns the terms under consideration in this study. In (6), we find the author distinguishing between two agreement patterns, labelling one 'non-standard' and the other 'uncommon'.

(6) Therefore, I label [singular+*don't*] as a **nonstandard** agreement pattern and [plural+*doesn't*] as **uncommon**. (*JEngL2014*)

The author continues to set up a three-way contrast, between 'standard', 'non-standard', and 'uncommon' (see (7)). In this case, the terminological distinction is motivated by the 'variable levels of exposure' by speakers to the specific constructions.

(7) In experiment 1, participants read sentences in the four possible combinations of subject number and verb form: two that can be considered **"standard,"** one that can be considered **"nonstandard,"** and one that I am labeling **"uncommon."** This creates a point of comparison across structures that speakers may have had variable levels of exposure to as part of their sociolinguistic experience. (*JEngL2014*)

Labelling one construction specifically as 'uncommon' might lead to the assumption that the other two are 'common' (albeit one of them is also non-standard) and it would be interesting to know why other potential terminological choices were rejected in this specific case (e.g., 'canonical'/ 'non-canonical').

[15] Please note that these have not been manually disambiguated and thus contain some hits which do not refer to the discussion of terminology (e.g., use in examples such as *Who in heaven's name (could I call)?* [*Syn2021*]).

Finally, we reproduce in (8) an exceptionally long passage explaining the choice of an author to use the terms 'canonical' and 'non-canonical'. Even though the author applies these terms to describe semantic phenomena, we find his reasoning of particular interest for our study as well. The terminological choice of 'canonical' and 'non-canonical' is motivated by the terms being perceived as (1) describing phenomena which are considered 'more or less stable', (2) less judgmental than the terms 'good' or 'bad', and (3) non-mutually exclusive (i.e., gradable). In addition, the terms are considered 'necessarily imprecise' – a quality considered both essential and problematic by the manuscript's author.

(8) Therefore, although many standard studies of lexical semantic relations label these types of oppositional pairs as 'antonyms' (sometimes narrowed down to refer to gradable opposites), in this article I use the term opposition, which encapsulates a broader sense of this type of relation. **I also use the terms canonical and noncanonical** – adapted from Murphy's (2003) pragmatic approach to lexical semantic relations – to refer to **oppositions that have a more or less stable basis in the linguistic system in which they participate**.

This is **to avoid judging oppositional pairs as "good" or "bad" examples**, for, as I argue, if an unusual oppositional pair (e.g., "cream" / "spleen") resides in one of the frames common to conventional pairs, then in that instance it is not a bad opposition, just a context-bound one. **It is also important to note that the terms canonical and noncanonical are not intended to treat oppositions as two mutually exclusive categories.** The canonical status of oppositions ranges in **a gradable cline from canonical to noncanonical**, so **the terms are necessarily imprecise.** At the same time, this demonstrates the difficulties in avoiding representing ideas and concepts in anything other than a binary fashion, even in the realm of academic discourse. (*JEngL2012*)

2.4 Discussion and Conclusion

In this study, we set out to investigate usage patterns of related terms in the context of syntactic canonicity across six high-profile linguistic journals. To this end, we compiled a corpus consisting of contributions to the journals published between 2012 and 2021 and subjected these contributions to quantitative and qualitative analysis. The quantitative analysis revealed that non-negated forms outmatch the negated forms in the case of the three pairs 'canonical' vs. 'non-canonical', 'marked' vs. 'unmarked', and 'standard' vs. 'non-standard', which, as the collocation analysis has shown, is also due to 'marked' being used as a past tense verb form and 'standard' being used as a noun. It needs to be noted that 'marked' conceptually corresponds to 'non-canonical' and

'non-standard', which means that the non-negated term refers to the deviation in the 'marked' vs. 'unmarked' pair. An investigation of the nouns 'canonicity', 'markedness', and 'syntactic variation' revealed journal-based differences. However, despite certain trends becoming evident, the situation is complex: individual authors may prefer certain terms; and terms being in use does not mean that they are used in the same way across publications.

As the analysis of bigrams of the three pairs ('canonical' vs. 'non-canonical', 'marked' vs. 'unmarked', and 'standard' vs. 'non-standard') has shown, they are used with partially overlapping but also distinctive meanings, implying that it might be necessary for authors to reflect explicitly in writing as to why specific terminology has been adopted (as we found in a rare instance in example (8)). The terminological pairs are also often used to set up a contrast between the canonical and the non-canonical, the marked and the unmarked, and the standard and the non-standard, that is, reflecting an ideological underpinning that a specific construction, sequence, etc. is either the one or the other. In some cases, a continuum of options falling between the two poles may be assumed but, if present, is frequently implicit.

It is neither advisable nor reasonable to assume that linguists stick to a fixed set of terms with fixed definitions (at least across article boundaries). However, in light of parallel developments such as globalisation and decolonisation, developing higher awareness of the potential ideological dimensions of terminology is fundamental. While this is more clearly apparent with terms such as 'mother tongue' and 'native speaker', reflection is necessary whenever language variation and change are involved. This is not at all a call against using 'canonical' and 'non-canonical'; instead, it is a suggestion to be aware of both a term's explicitly linguistic scope and the values transmitted more or less subtly by it.

Due to limitations of space, we considered only a selection of journals as well as a clearly defined timeframe. Future work following up on this case study may investigate diachronic trends in the use of terminology related to syntactic canonicity and incorporate other journals with additional foci and, most promisingly, add further qualitative insight into the use of these terminological items (i.e., which meaning is evoked for each usage case). Moreover, considering additional variables such as text type and author (such as individual author usage profiles across one but also several articles) may also provide further insights into terminological choices.

Appendix A: Normalised Frequencies of Target Items

Table 2.5 *Normalised frequencies (per 100 articles) and range (% of articles in which the term occurred) of 'canonical', 'non-canonical', 'marked', 'unmarked', 'standard', and 'non-standard' in six journals*

Journal		'canonical'	'non-canonical'	'marked'	'unmarked'	'standard'	'non-standard'
Cor	NF	28.24	12.21	255.73	76.34	204.58	50.38
	Range	6.11%	3.05%	64.12%	10.69%	61.83%	15.27%
JEngL	NF	82.64	41.32	244.63	71.07	594.21	165.29
	Range	10.74%	4.96%	71.90%	21.49%	72.73%	33.06%
JGL	NF	144.33	24.74	248.45	95.88	1047.42	95.88
	Range	17.53%	6.19%	69.07%	26.80%	83.51%	19.59%
JHL	NF	73.87	28.83	426.13	163.96	269.37	20.72
	Range	20.72%	6.31%	71.17%	40.54%	62.16%	5.41%
JoL	NF	341.54	91.28	467.69	206.67	478.97	27.69
	Range	37.44%	15.89%	74.87%	31.28%	78.46%	10.26%
Syn	NF	96.09	16.41	350.00	152.34	341.41	16.41
	Range	37.50%	8.59%	72.66%	28.91%	76.56%	10.16%

Note. NF = normalised frequency.

Appendix B: Top Five Bigrams with Target Item on the Left

Table 2.6 *Top five bigrams with the target item in the left position*

	Cor	*JEngL*	*JGL*	*JHL*	*JoL*	*Syn*
canonical	*sequence* (6/2) *antonyms* (5/1) *agreement/ clause-initial/ emphatic/ psi/ sequences/ use* (4/1)	*antonyms* (9/2) *oppositions* (8/1) *oppositional* (6/2) *and* (5/2) *status* (5/1)	*verbs* (26/1) *transitive* (17/1) *EO.ACC* (17/1) *EO* (6/1) *and* (5/1)	*subject* (27/4) *grammatical* (5/1) *transitive* (3/2) *complement* (3/1) *utterance* (3/1)	*gender* (60/1) *morphosyntactic* (37/2) *case* (24/7) *agreement* (24/3) *typology* (23/6)	*subject* (11/7) *perfective* (9/1) *position* (8/4) *order* (7/6) *null-subject/ positions* (3/2)
non-canonical	*oppositions* (2/1)	*oppositions* (17/3) *examples* (5/1) *opposition* (3/2) *textual* (3/1) *ones* (2/1)	*order* (6/1) *word* (2/1) *clauses* (2/1) *marking* (2/1) *sentences* (2/1)	*subjects* (14/3) *subject* (8/2) *marking* (2/2)	*case* (21/2) *agreement* (9/6) *behaviour* (7/1) *morphosyntactic* (6/1) *passives* (5/4)	*plural* (3/1) *agreement* (2/2) *positions* (2/2) *position* (2/1)
marked	*as* (30/20) *with* (29/20) *by* (26/15) *for* (22/8) *plural* (21/1)	*by* (34/27) *for* (34/7) *in* (21/16) *as* (19/11) *with* (15/7)	*by* (33/23) *for* (30/13) *as* (27/17) *in* (18/12) *on* (12/8)	*by* (94/37) *with* (78/22) *in* (48/19) *for* (29/16) *as* (23/6)	*by* (114/54) *with* (102/40) *and* (52/15) *for* (51/22) *as* (41/30)	*with* (100/38) *by* (43/22) *as* (30/13) *for* (20/12) *in* (15/9)
unmarked	*forms* (21/2) *and* (13/3) *form* (9/1) *marked* (5/1) *ministro* (4/1)	*in* (6/3) *form* (4/2) *speakers* (4/2) *or* (3/2) *order* (3/2)	*verb* (8/2) *for* (7/4) *form* (6/4) *gender/ in/ order* (4/2)	*transitive* (20/2) *construction* (11/2) *in* (9/6) *for* (5/4) *p-prominent* (5/2)	*subject* (28/1) *nominal* (15/3) *causative* (14/1) *with* (13/2) *form* (12/2)	*case* (32/8) *argument* (10/1) *and* (7/3) *objects* (7/2) *option* (6/4)

standard	*deviation* (26/10) *deviations* (16/11) *and* (12/5) *English* (11/8) *reference* (7/6)	*English* (143/37) *deviation* (37/14) *and* (29/13) *language* (24/12) *error* (18/5)	*Dutch* (246/23) *German* (230/24) *language* (84/17) *and* (32/6) *variety* (27/9)	*negation* (73/3) *Finnish* (26/1) *language* (10/9) *ModGr* (10/1) *English* (6/4)	*Polish* (63/4) *English* (62/11) *Arabic* (51/13) *OT* (39/9) *Russian* (24/6)	*assumptions* (22/17) *English* (21/7) *of* (21/1) *error* (14/8) *analysis* (13/8)
non-standard	*capitalisation* (25/1) *spellings* (5/4) *language* (3/3) *varieties* (3/3) *forms* (3/2)	*varieties* (25/11) *and* (16/7) *forms* (13/7) *sentences* (9/1) *form* (6/4)	*language* (17/4) *varieties* (11/7) *features* (9/3) *gender* (6/2) *form* (5/5)	*varieties* (8/2) *European* (2/1) *Ibero-Romance* (2/1) *Polish* (2/1) *primarily* (2/1)	*English* (4/3) *use* (4/1) *varieties* (3/3) *uses* (3/2) *Lucas/in/ never* (3/1)	*varieties* (4/2) *assumption* (3/2) *dialects* (2/2)

Note. Numbers in brackets give raw frequency and range; items are sorted by frequency; in case two or more items have the same frequency, the one with the higher range is given first (everything being equal, items are listed in alphabetical order).

REFERENCES

AntConc (2020). Computer software by Laurence Anthony. Version 3.5.9. Tokyo: Waseda University. Retrieved from www.laurenceanthony.net/software.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, & Akitaka Matsuo (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774.

Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad, & Edward Finegan (1999). *Longman grammar of spoken and written English*. London: Longman.

Bugarski, Ranko (1983). Sociolinguistic issues in standardizing linguistic terminology. *Language in Society*, 12(1), 65–70.

Buschfeld, Sarah, Sven Leuckert, Claus Weihs, & Andreas Weilinghoff (2024). How *real* is the quantitative turn? Investigating statistics as the *new normal* in linguistics. *ICAME Journal*, 48(1), 1–22.

Cheshire, Jenny (1987). Syntactic variation, the linguistic variable, and sociolinguistic theory. *Linguistics*, 25(2), 257–82.

ggplot2: Elegant graphics for data analysis (2016). *R* package by Hadley Wickham. New York: Springer. Retrieved from https://ggplot2.tidyverse.org.

Gries, Stefan Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–37.

Gries, Stefan Th. (2020). Analyzing dispersion. In Magali Paquot & Stefan Th. Gries, eds, *A practical handbook of corpus linguistics*. Cham: Springer, 99–118.

Gries, Stefan Th. (2022). What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies*, 5(2), 171–205.

Hackert, Stephanie (2012). *The emergence of the English native speaker: A chapter in nineteenth-century linguistic thought*. Berlin: Mouton de Gruyter.

Hickey, Raymond (2014). *A dictionary of varieties of English*. Malden: Wiley-Blackwell.

Huddleston, Rodney & Geoffrey K. Pullum, eds (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Kortmann, Bernd (2021). Reflecting on the quantitative turn in linguistics. *Linguistics*, 59(5), 1207–26.

Labov, William (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Larsson, Tove, Jesse Egbert, & Douglas Biber (2022). On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora*, 17(1), 137–57.

Leuckert, Sven (2019). *Topicalization in Asian Englishes: Forms, functions, and frequencies of a fronting construction*. London: Routledge.

Mugdan, Joachim (1990). On the history of linguistic terminology. In Hans-Josef Niederehe & Ernst F. K. Koerner, eds, *History and historiography of linguistics: Proceedings of the fourth international conference on the history of the*

language sciences (ICHoLS IV), Trier, 24–28 August 1987. Vol. 1: Antiquity–17th Century. Amsterdam: John Benjamins, 49–62.

OED Oxford English Dictionary Online (2023). Oxford: Oxford University Press. Retrieved from www.oed.com/.

Pham, Teresa, Sven Leuckert, Gea Dreschler, Sandra Götz, Christine Günther, Kathrin Kircili, Claudia Lange, Louise Mycock, Theresa Neumaier, & Sofia Rüdiger (2024). Defining non-canonicity: An integrated approach to modelling syntactic variation. *OSF Preprints*. Retrieved from https://osf.io/preprints/osf/92zhg.

quanteda: An R package for the quantitative analysis of textual data (2018). *R* package by Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, & Akitaka Matsuo. Retrieved from https://quanteda.io.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, & Jan Svartvik (1985). *A comprehensive grammar of the English language*. London: Longman.

R: A language and environment for statistical computing (2024). Vienna: R Foundation for Statistical Computing. Retrieved from www.R-project.org/.

Rosch, Eleanor H. (1973). Natural categories. *Cognitive Psychology*, 4(3), 328–50.

Rüdiger, Sofia (2019). *Morpho-syntactic patterns in spoken Korean English*. Amsterdam: John Benjamins.

Rüdiger, Sofia, Leimgruber, Jakob R. E., & Tseng, Ming-i Lydia (2022). English in Taiwan: Expanding the scope of corpus-based research on East Asian Englishes. *English Today*, 39(2), 100–9.

Trask, Robert L. (1993). *A dictionary of grammatical terms in linguistics*. London: Routledge.

Vermeer, Hans J. (1971). *Einführung in die linguistische Terminologie*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Ward, Gregory, Betty Birner, & Rodney Huddleston (2002). Information packaging. In Rodney Huddleston & Geoffrey K. Pullum, eds, *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press, 1363–448.