Text-linguistic approaches to register variation

Douglas Biber
Northern Arizona University

Douglas Biber, Regents' Professor of Applied Linguistics at Northern Arizona University, authors this article exploring the connections between register and a text-linguistic approach to language variation. He has spent the last 30 years pursuing a research program that explores the inherent link between register and language use, including at the phraseological, grammatical, and lexico-grammatical levels. His seminal book *Variation across Speech and Writing* (1988, Cambridge University Press) launched multi-dimensional (MD) analysis, a comprehensive framework and methodology for the large-scale study of register variation. This approach was innovative in taking a text-linguistic approach to characterize language use across situations of use through the quantitative and functional analysis of linguistic co-occurrence patterns and underlying dimensions of language use. MD analysis is now used widely to study register variation over time, in general and specialized registers, in learner language, and across a range of languages. In 1999, the *Longman Grammar of Spoken and Written English* (Biber et al.) became the first comprehensive descriptive reference book to systematically consider register variation in describing the grammatical and lexico-grammatical patterns of use in English. Douglas Biber's quantitative linguistic research has consistently demonstrated the importance of register as a predictor of language variation. In his own words, "register always matters" (Gray 2013: 360, Interview with Douglas Biber, *English Language & Linguistics*).

Keywords: textual variation, functional variation, lexico-grammar, multi-dimensional analysis, quantitative analysis

https://doi.org/10.1075/rs.18007.bib
Register Studies 1:1 (2019), pp. 42–75. issn 2542-9477 | e-issn 2542-9485

1. How is register conceptualized in text linguistics and in the text-linguistic approach to register variation?

Text linguistics is a research approach that was developed in the 1970s and 1980s, as a counterpoint to the dominant linguistic paradigms of the 1960s and 1970s that focused almost exclusively on the linguistic structure of sentences. Researchers like Van Dijk (1972), Halliday and Hasan (1976), De Beaugrande and Dressler (1981), and Brown and Yule (1983) all focused on the text as an important linguistic construct on a higher level than the sentence. Thus, these researchers argued that it is possible to describe the grammar of texts in a similar way to the more traditional research goal of describing the grammar of sentences. Research in the text-linguistic tradition described the structural and logical organization of texts, with consideration given to the analyses of cohesion (the referential connections among the words in a text), coherence (the underlying logical structure of a text), and information structure / discourse organization (the ways in which the components of a text are organized, reflecting informational concerns like prominence and topicality).

For the most part, studies in this research tradition paid little attention to the linguistic description of text varieties. However, some researchers note the existence of such textual varieties, referred to as 'genres' (Brown & Yule 1983), 'text types' (De Beaugrande & Dressler 1981), or 'registers' (Halliday & Hasan 1976). Hymes – although associated with the 'ethnography of communication' rather than 'text linguistics' – also emphasized the importance of culturally-recognized spoken textual varieties, referred to as 'speech events'. In a series of publications, Hymes described the ways in which speech events could be described for a range of situational characteristics that had functional underpinnings and linguistic correlates (see, e.g., Hymes 1972, 1974: Chapters 1–3).

The notion that language use can be studied at the textual level complements much other research in sociolinguistics and pragmatics, which focuses instead on lower levels of linguistic structure as the primary object of study (e.g., the choice between phonetic pronunciations, morphemes, syntactic variants, or the realization of speech acts). Reflecting this insight, the label 'text-linguistic' was appropriated in Biber (2012) to refer to a theoretical and methodological approach to the study of register variation. In summary, the text-linguistic approach to register variation uses quantitative methods to describe the linguistic characteristics of each text, as the basis for comparing the patterns of register variation across texts.

The text-linguistic approach can be contrasted with the 'variationist approach', which describes the linguistic characteristics of each token of a linguistic feature, to predict the choices among linguistic variants; in this approach, register can function as one of the contextual variables used to predict linguistic choice. It

turns out that these two approaches differ in their underlying research designs and analytical techniques, in addition to their ultimate research goals (see the direct comparison of the two approaches in Biber, Egbert, Gray, Oppliger, & Szmrecsanyi 2016; cf. Biber 2012).

Register in the text-linguistic approach is studied from a quantitative, comparative perspective. Similar to the notion of 'speech event' in the ethnographic framework, 'registers' in the text-linguistic framework are named, culturally-recognized categories of texts. In many cases, there are overt external indicators in the context that signal the register category. But there are three major defining characteristics of the text-linguistic register framework that distinguish it from other related approaches to textual variation:

1. the research goal of describing text categories for both situational characteristics and lexico-grammatical characteristics;
2. the claim that situational characteristics have a systematic functional relationship to lexico-grammatical characteristics; and
3. the claim that those lexico-grammatical characteristics (and possibly also situational characteristics – see Section 5 below) can be described in a continuous quantitative space of variation

To some extent, these defining characteristics were anticipated by earlier researchers in text linguistic and ethnographic research frameworks. Thus, Hymes, Halliday and Hasan, and De Beaugrande and Dressler all note the importance of describing texts and textual categories with respect to both situational and linguistic characteristics; for example:

> The linguistic features which are typically associated with a configuration of situational features […] constitute a REGISTER. (Halliday & Hasan 1976: 22)

> [Text types are] "classes of texts expected to have certain traits for certain purposes". (De Beaugrande & Dressler 1981: 182)

These researchers also recognized the importance of communicative function as the underlying explanation of situational-linguistic correlations. In fact, it could be argued that Hymes was more interested in the study of communicative function than linguistic form; for example:

> "analysis of use [is] prior to analysis of code", taking into account the "gamut of stylistic or social functions" (Hymes 1974: 79)

> for sociolinguistic research, then, what is essential is […] to take functional questions, questions of social meaning and role, as starting point (Hymes 1972: 6)

None of these researchers employed quantitative methods to study patterns of textual variation. However, De Beaugrande and Dressler mention the possibility, in the context of emphasizing the need for functional interpretation:

> We might count the proportions of nouns, verbs, etc. or measure the length and complexity of sentences […] without really defining the [text] type – we need to know how and why these traits evolve. Statistical linguistic analysis of this kind ignores the functions of texts in communication and the pursuit of human goals. Presumably, those factors must be correlated with the linguistic proportions […].
> (De Beaugrande & Dressler 1981: 183)

In a sense, the text-linguistic approach to register variation could be regarded as a framework designed to meet this challenge. Biber and Conrad (2009: especially Chapters 1–3) provides the fullest description of the text-linguistic register framework (simply referred to as 'register analysis' in the book). The central theoretical foundation of this approach concerns the relationship among the three components of situation, function, and linguistic forms, illustrated in Figure 1 (see Biber & Conrad 2009: 6–10; cf. Egbert & Biber 2016).

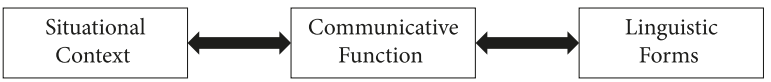

Figure 1. Visual representation of the three-way relationship among situation, function, and linguistic form in the text-linguistic register framework

Similar to speech event analysis, a text-linguistic register analysis involves a full analysis of the situational context, including consideration of the participant identities, relations among participants, channel, production circumstances, setting, and communicative purposes (see Biber & Conrad 2009: 39–46). However, register analysis differs from speech event analysis in its linguistic focus, with the primary goal of describing the lexico-grammatical features that are frequent and pervasive in texts from the target register.

Text-linguistic register analysis further differs from many other sociolinguistic approaches in its foundational claim that linguistic variation is functional rather than indexical or purely conventional. This point is potentially confusing because the term 'function' is ambiguous. In sociolinguistic descriptions, a linguistic variant can 'function' to index a social group, meaning that the group by convention tends to use that variant. But such descriptions do not entail any claims about the underlying communicative functions of the linguistic forms. That is, sociolinguists have traditionally analyzed linguistic variation for the ways in which it conventionally indexes particular social groups or styles, with no attempt to associate linguistic variation with different communicative functions.

This practice holds for most sociolinguists who focus on the study of social dialect variation (see, e.g., Hudson 1980: 191–193 for discussion of the linguistic equality of social dialects) as well as some earlier sociolinguists who focused on the linguistic description of registers and genres (e.g., Ferguson 1994).

In contrast, the text-linguistic approach emphasizes the communicative-functional basis of linguistic variation, claiming that linguistic features are frequent and pervasive in a register because they perform communicative functions required by the situational context. To take a simple example, 1st and 2nd person pronouns function to refer directly to the speaker and addressee, and so they are frequent in highly interactive registers. (See further discussion in Sections 3 and 4 below.)

As noted above, text-linguistic register analyses describe linguistic variation in a continuous, quantitative space, with the goal of identifying the linguistic characteristics that are especially frequent and pervasive in texts from the target register. A methodological variant with the same research goals – the 'corpus-linguistic approach' to register variation – is based on analysis of an entire sub-corpus representing a register. Both of these approaches require comparative analysis, contrasting linguistic rates of occurrence in the target register to other registers. Section 3 below describes the quantitative, corpus-based analytical methods for both approaches.

2. What are the research goals of the text-linguistic approach to register variation?

In the analytical framework developed in Biber and Conrad (2009), the analysis of texts can be approached from a 'register' perspective and from a 'genre' perspective (see especially 15–23). The genre perspective is compatible with earlier work in text linguistics, focusing on the ways in which complete texts are structured and organized. If there are lexico-grammatical features that signal genre characteristics, those features usually occur only once in a text, and they can have a conventional rather than functional relationship to the genre. For example, Introduction-Methods-Results-Discussion sections reflect the genre organization of academic research articles. Those sections are marked linguistically by section headings, and research articles are organized in this way by convention.

In contrast, the analysis of register variation is an extension of the research goals associated with early work in the text-linguistic tradition. Rather than focusing on the structure or organization of a text, text-linguistic register analyses focus on the lexico-grammatical features that are frequent and pervasive in a text. Such analyses are then generalized to a sample of texts that all share

the same situational characteristics, and thus, all represent the same 'register'. The typical lexico-grammatical features that occur in these texts are interpreted as register features because they are functionally associated with the situational contexts of texts. Thus, the register perspective differs from the genre perspective by its focus on linguistic features that are frequent and pervasive, and its focus on features that have a functional (rather than conventional) relationship to the situational context.

Because text-linguistic register studies typically focus on patterns of register variation, analyses are comparative and apply quantitative corpus-based methods (see Section 3, below). As a result, the quantitative research findings relate to the extent to which a linguistic feature is used in one register, in comparison to other registers. Those quantitative patterns are interpreted functionally relative to the situational contexts of the registers.

Research findings from text-linguistic register studies have proven to be important – and often highly surprising – for two reasons: (1) these findings often directly contradict strongly-held beliefs about linguistic patterns of use, and (2) in some cases, these findings have uncovered systematic patterns of use that were not even anticipated by previous theorizing. These two types of contributions are discussed and illustrated below.

First, in addition to their intuitions about grammaticality, native speakers of a language usually have intuitions about language use, reflected in strong beliefs about the linguistic patterns that are normal or typical in a register. However, empirical text-linguistic register studies have repeatedly shown that those intuitions are often wrong – in some cases, dramatically wrong (see Biber & Reppen 2002).

Many patterns of this type are illustrated in the *Longman Grammar of Spoken and Written English* (*LGSWE;* Biber, Johansson, Leech, Conrad, & Finegan 1999), a comprehensive reference grammar with an empirical basis. The LGSWE reports on the patterns of use for the full range of lexico-grammatical features in English from a text-linguistic register perspective (comparing the patterns of use in conversation, fiction, newspaper prose, and academic prose). Methodologically, the analyses for the *LGSWE* actually employed a 'corpus-linguistic' approach to register variation rather than a 'text-linguistic' approach (see discussion in Section 3, below). Because the book is a reference grammar, it provides detailed descriptions of the patterns of variation for each grammatical feature – more in line with the goals of the variationist approach to register. At the same time, though, the book provides comprehensive descriptions of the grammatical characteristics of spoken and written registers, in line with the major research goal of text-linguistic studies.

The *LGSWE* is full of research findings that fly in the face of previously-held intuitions. When a researcher simply reads about these linguistic patterns, it is

easy to think that they are obvious and not surprising. But, when a reader is forced to first commit to their own intuitions, and subsequently see the actual patterns of use, the contrast between the two becomes obvious.

For example, one of the most widely held intuitions about language use among English-language professionals is the belief that progressive aspect is the unmarked choice for verbs in conversation (e.g., *What are you doing?, I'm going to the store.*). One reflection of this belief is the prominent coverage given to progressive aspect verbs (the 'present continuous') in many ESL grammar textbooks (see Biber & Reppen 2002: 203). It turns out that progressive aspect verbs are, in fact, more common in conversation than in other registers (see Biber et al. 1999: 462, Figure 6.4). The contrast with academic prose is especially noteworthy: progressive aspect verbs are quite rare in academic prose but common in conversation. However, when we compare the use of progressive aspect with simple aspect, we see that progressive aspect is certainly not the normal form of the verb phrase used in conversation. Rather, simple aspect is clearly the unmarked choice, occurring more than 20 times more often than progressives in conversation (see Biber et al. 1999: 461, Figure 6.2).

A second case study of this type concerns a set of beliefs held by linguists relating to grammatical complexity and historical change, based on a priori theoretical notions of complexity and intuitions about language use. Those beliefs include the following:

- Conversation is structurally simple, while academic written prose is structurally complex.
- The expression of meaning in conversation is context-dependent, while the expression of meaning in academic writing is maximally explicit.
- Historical change occurs primarily in speech.
- To the extent that historical change does occur in writing, it involves the adoption of colloquial innovations from speech.
- Written registers like academic prose are especially conservative and resistant to historical change.

In a series of studies, Biber and Gray (see especially Biber & Gray 2016; cf. Biber & Gray 2011, 2013; Biber, Gray, & Poonpon 2011) show how text-linguistic register analysis contradicts all of these widely believed generalizations about language use and change. For example, a survey of the LGSWE shows that many grammatical features traditionally associated with complexity are actually more common in conversation than in academic writing. This is especially the case for the use of finite dependent clauses. In contrast, academic writing tends to employ a completely different type of structural complexity, realized as the compression of information in phrasal structures instead of clausal

elaboration. When the evolution of these phrasal features is investigated from a historical perspective, we discover that academic prose has been at least as receptive to historical change as face-to-face conversation. However, these have been changes of an unanticipated nature: towards a much greater use of phrasal modifiers, rather than towards an increased use of colloquial features. Finally, as a result of these historical changes, modern academic prose has become strikingly inexplicit in the expression of meaning, exactly the opposite of commonly expressed stereotypes (see discussion in Biber & Gray 2016: Chapter 6).

Second, as noted above, text-linguistic register studies have also been important because they have uncovered patterns of use that were not even anticipated in previous research. Of course, the studies described above, which directly contradict previously held beliefs/claims, are cases of this type – no researcher 'anticipates' that their beliefs and theoretical claims about language use and change are wrong! But, text-linguistic register studies have also uncovered patterns of use relating to research questions that had simply never been asked.

Research findings of this type are possible because text-linguistic register analyses are often inductive, applying a bottom-up approach to describe the patterns of register variation in a discourse domain. As a result, these studies have discovered linguistic patterns that had not been anticipated in previous theoretical frameworks.

Studies carried out in the research framework known as multi-dimensional (MD) analysis are the most dramatic examples of this type. These studies were originally undertaken to explore theoretical claims about the linguistic differences between speech and writing. However, the scope of analysis was later reconceptualized to encompass analysis of the full range of registers in a language with respect to a comprehensive set of lexico-grammatical characteristics (see, e.g., Biber 1986, 1988, 1995). Sections 3 and 4 below present methodological details and a case study illustrating an MD study.

The major innovation of the MD approach is that it provides a methodology to empirically analyze the ways in which linguistic features co-occur in texts and the ways in which registers vary with respect to those co-occurrence patterns. These goals are accomplished through computational analyses of a large corpus of texts, representing multiple registers, followed by a statistical analysis employing the technique of factor analysis.

The importance of linguistic co-occurrence had been noted in the 1970s by linguists such as Firth, Halliday, Ervin-Tripp, and Hymes. Brown and Fraser (1979: 38–39) observe that it can be "misleading to concentrate on specific, isolated [linguistic] markers without taking into account systematic variations which involve the co-occurrence of sets of markers". Ervin-Tripp (1972) and Hymes (1974) identify 'speech styles' as varieties that are defined by a shared set of co-

occurring linguistic features. Halliday (1988:162) defines a register as "a cluster of associated features having a greater-than-random…tendency to co-occur".

However, despite these theoretical discussions, descriptions of registers and styles during that time were based on consideration of only a few linguistic features, with no empirical analysis of the co-occurrence relations among linguistic characteristics. Several earlier sociolinguistic investigations claimed that registers vary along an underlying linguistic/functional parameter and proposed a set of linguistic features associated with that parameter, thus giving at least implicit recognition to the importance of linguistic co-occurrence. Studies of this type include Ferguson (1959) on 'high' versus 'low' diglossic varieties; Bernstein (1970) on restricted versus elaborated codes; Irvine (1979) on formal versus informal registers; and Ochs (1979) on planned versus unplanned discourse. A few other early researchers went further in proposing specific linguistic co-occurrence patterns associated with two parameters of variation. These include Chafe (1982; Chafe & Danielewicz 1986) and Longacre (1976). Chafe identifies two parameters – integration/fragmentation and detachment/involvement – and posits a number of linguistic features associated with each parameter. Longacre also identifies two underlying parameters – projected time and temporal succession – and posits a group of features associated with each. These studies are important in that they recognize the need for analyses based on linguistic co-occurrence patterns in texts.

However, there are three major theoretical differences between these earlier investigations of register variation and the MD approach. First, apart from the Chafe and Longacre frameworks, most previous studies analyzed register variation in terms of a single underlying parameter, suggesting that there was a single basic situational distinction among registers. Second, most previous studies assumed that register variation could be analyzed in terms of simple, dichotomous distinctions, so that varieties are either formal or informal, planned or unplanned, etc. And finally, none of these early approaches applied empirical methods to identify sets of co-occurring linguistic features. Rather, researchers proposed sets of features that seemed to work together, based on their perceptions and intuitions.

MD analysis differs in all three respects. MD studies have demonstrated that no single parameter or dimension is adequate in itself to capture the full range of variation among registers in a language. Rather, different dimensions are realized by different sets of co-occurring linguistic features, reflecting different functional underpinnings (e.g., interactiveness, planning, informational focus). Second, the dimensions in MD studies are quantitative, continuous parameters of variation, which distinguish among a continuous range of texts or registers. For this reason, dimensions can be used to analyze the extent to which registers are similar (or dif-

ferent). And finally, sets of co-occurring linguistic features (which comprise the dimensions) are identified empirically using quantitative statistical techniques in the MD approach. In contrast, there is no guarantee that groupings of features proposed on intuitive grounds actually co-occur in texts, and in fact, subsequent MD analyses show that neither Longacre's parameters nor Chafe's parameters are completely accurate in identifying sets of linguistic features that actually co-occur regularly in English texts. In contrast, the statistical techniques used in MD studies provide a precise quantitative specification of the co-occurrence patterns among linguistic features in a corpus of texts.

There was, however, one major precursor to MD analysis that employed statistical analysis of linguistic co-occurrence patterns, and amazingly, that study was carried out more than 20 years before these other sociolinguistic investigations: John Carroll's (1960) study on 'vectors of prose style'. Carroll analyzed 39 'objective' linguistic measures and 29 'subjective' perceptual ratings. His corpus seems small by present-day standards: 150 prose text samples, consisting of c. 300 words each. But when we consider the fact that Carroll apparently did all linguistic analyses by hand, the corpus is impressively large! Each linguistic feature was counted in each text sample, and then each text sample was perceptually rated for 29 stylistic evaluations by eight judges. Carroll then applied a statistical factor analysis to reduce those variables to six underlying parameters of linguistic style. This study is remarkable in that it was carried out before the days of large computational corpora, automated tagging software, and computer-based statistical analysis packages. Although it is not framed as a study of register variation, Carroll's 1960 study can in many respects be regarded as the first multi-dimensional investigation of linguistic variation.

The findings from MD studies in the 1980s were interpreted relative to previous research claims about speech and writing. MD studies found no absolute linguistic difference between speech and writing, but rather found systematic patterns of linguistic variation within each mode, and some overlap for the range of variation across spoken and written registers. At the same time, though, these studies found a more general pattern of difference between the two modes: while there is an extensive range of linguistic variation among written registers, there is a much more restricted range of variation among spoken registers. As a result, there is a notable difference between the two modes: writers can produce texts that range from highly colloquial discourse styles to informational dense styles that are completely unlike anything found in speech. In contrast, the production circumstances are much more constrained in speech, and for that reason, the linguistic characteristics of spoken sub-registers are all relatively similar, regardless of differences in interactivity or communicative purpose.

Over the last three decades, there have been dozens of MD studies carried out to investigate the patterns of register variation in different languages (e.g., English, Spanish, Czech, Korean, Chinese, Somali) as well as particular discourse domains in English (e.g., university spoken and written registers, written academic research articles). Biber (2014) surveys MD studies carried out before 2014, and Barbieri and Wizner (in press) identify several MD studies carried out over the last 5 years. The cumulative evidence from those MD studies has gone far beyond tests of previous claims relating to speech and writing, instead providing strong evidence for the existence of universal parameters of register variation. From both theoretical and methodological perspectives, it is to be expected that each MD analysis would uncover specialized dimensions that are peculiar to a given language and/or discourse domain. After all, each of these studies differs with respect to the set of linguistic features included in the analysis, and the set of registers represented in the corpus for analysis. Given those differences, it is reasonable to expect that the parameters of variation that emerge from each analysis will be fundamentally different. And to some extent, this expectation is met, with specialized dimensions emerging in nearly all MD analyses.

However, given this background, the much more surprising and more important finding is the existence of dimensions of variation that emerge in nearly all MD studies. Two such dimensions are especially noteworthy: a dimension associated with 'oral' versus 'literate' discourse, and a dimension associated with narrative discourse. These dimensions of variation have emerged in MD studies regardless of the language or discourse domain of focus.

The robustness of narrative dimensions across languages and discourse domains indicates that this rhetorical mode is basic to human communication, whether in speech or in writing. Rhetoricians and discourse analysts have long argued for the central role of narration in communication. MD studies confirm that claim, showing the importance of this rhetorical mode in virtually all discourse domains (spoken and written; interpersonal and informational; etc.).

But, the most surprising pattern discovered through MD analysis is the oral/literate opposition, which emerges as the very first dimension in nearly all MD studies (see especially the discussion in Biber 2014). In studies based on general corpora of spoken and written registers, this dimension clearly distinguishes between speech and writing. However, other studies show that this is not a simple opposition between the spoken and written modes. In fact, this dimension has emerged consistently in studies restricted to only spoken registers, as well as studies restricted to registers in the written mode.

In terms of communicative purpose, the 'oral' registers characterized by this dimension focus on personal concerns, interpersonal interactions, and the expression of stance. In contrast, 'literate' registers focus on the presentation of

propositional information, with little overt acknowledgement of the audience or the personal feelings of the speaker/writer. These registers are usually produced in situations that allow for extensive planning and even revising and editing of the discourse.

Linguistically, this oral/literate dimension opposes two discourse styles: an 'oral' style that relies on pronouns, verbs, and adverbs, versus a 'literate' style that relies on nouns and nominal modifiers. The oral style relies on clauses to construct discourse – including a dense use of dependent clauses. In contrast, the complexity of the literate style is phrasal. This finding, replicated across languages and across discourse domains, is especially surprising, because it runs counter to assumptions about syntactic complexity held by many linguists. But, it is perhaps the most important and robust finding to emerge cross-linguistically from MD studies: spoken registers (and 'oral' written registers) rely on clausal discourse styles, including a dense use of dependent clauses; written registers (and 'literate' spoken registers) rely on phrasal discourse styles, especially the dense use of phrasal modifiers embedded in noun phrases (see also the Biber and Gray studies discussed above).

In sum, the patterns of variation observed across MD studies support the likelihood of universal parameters of register variation as well as the existence of unique dimensions of variation in each language and/or discourse domain. Section 4 below presents a short case study illustrating this research approach.

3. What are the major methodological approaches that are used to analyze or account for register in the text-linguistic approach?

As noted in Section 1 above, a text-linguistic register analysis has three major components: the situational analysis, the quantitative-linguistic analysis, and the functional interpretation. Biber and Conrad (2009: Chapters 1–3) provides a description of the methodological decisions associated with all three components.

The situational analysis requires consideration of all aspects of the situational context, including participants, interactivity, channel, production circumstances, setting, and communicative purposes (see Biber & Conrad 2009: Table 2.1 and following discussion). The analysis of these characteristics can be based on the researchers' observations and previous experiences, interviews with expert informants, previous research studies, and direct consideration of texts from the register (Biber & Conrad 2009: 37–39).

The quantitative-linguistic analysis requires a comparative approach, and usually involves a corpus and the analytical tools associated with corpus linguistics. The corpus is a sample of texts deliberately designed and collected to represent a

register. For studies of register variation, the corpus is designed to represent the range of registers in a discourse domain (e.g., academic research articles across disciplines; see Gray 2015).

In quantitative text-linguistic studies of register variation, each text is treated as an observation. Rates of occurrence for each linguistic feature are computed for each text. Subsequently, the overall mean rates of occurrence are computed for all texts from a register, coupled with a computation of dispersion (usually a standard deviation), reflecting the extent to which there is linguistic variation among the texts within a register.

An alternative research design is commonly employed for linguistic descriptions of a register, referred to here as the 'corpus-linguistic approach' to register variation. In this research design, there is only one observation for each register: i.e., an entire sub-corpus. Texts are usually not recognized as relevant constructs in the corpus-linguistic approach (from either linguistic or statistical perspectives). It is possible to compute overall rates of occurrence for linguistic features in the corpus, but studies from the corpus-linguistic approach typically do not include any measures of dispersion.

Text-linguistic studies of register variation are referred to as Type B research designs in Biber (2012), while corpus-linguistic studies of register variation are referred to as Type C research designs (cf. Biber, Conrad, & Reppen 1998: 269–274; Biber & Jones 2009; Type A designs are variationist, as described in Section 1, above). The independent variables in both designs are register categories (and other situational or social parameters), while the dependent variables are rates of occurrence for lexico-grammatical features.

The linguistic analysis in both designs begins with an annotation of the texts in the corpus, identifying occurrences of lexico-grammatical features. The features are marked with codes, referred to as 'tags'. This process is initially accomplished through automated software (a 'tagger'), and the tags are subsequently edited by hand to ensure high accuracy.

The quantitative analysis then begins by counting the number of occurrences for each linguistic feature. This step is also accomplished automatically (by a 'tag-count' program). Those counts are then converted to normed rates of occurrence (e.g., the number of nouns per 1,000 words) to adjust for the fact that texts and sub-corpora can differ in their lengths. In the text-linguistic design (the statistically and theoretically preferred methodology), this step is based on analysis of each text. Then, mean scores and standard deviations can be computed for all of the texts representing a register, allowing description of the central tendency as well as the degree of variation among texts within a register category. However, it has often been more convenient for researchers to employ a corpus-linguistic design, basing quantitative analyses on a single sub-corpus for each register. (The

Longman Grammar of Spoken and Written English [Biber et al. 1999] is a good example of a large-scale comparison of spoken and written registers that employs the corpus-linguistic approach.) In that case, the results simply provide a single rate of occurrence for each linguistic feature in each sub-corpus, with no indication of the extent of variation among texts within the register.

In both designs, a comparative approach is necessary to evaluate the importance of quantitative results. That is, the absolute frequencies of linguistic features are not in themselves meaningful, simply because they serve different grammatical functions. At the same time, though, features vary in frequency across registers, reflecting the situational contexts of those registers. For example, verbs occur frequently in all registers, simply because English sentences require a verb. In academic writing, there are c. 80,000 lexical verbs per million words (see the Biber et al. 1999: 65, Figure 2.2; cf. Biber & Conrad 2009: 92, Figure 4.1). In isolation, this sounds like a really high frequency, and so we might be tempted to conclude that lexical verbs are an important characteristic of academic writing. However, a comparison to other registers shows that this would be an incorrect conclusion. For example, in conversation, lexical verbs are 50% more frequent than they are in academic writing (a rate of c. 120,000 verbs per million words; *ibid.*). Thus, it turns out that the rate of 80,000 per million words actually means that lexical verbs are relatively rare in academic writing, because this linguistic feature is used much less commonly than in other registers. The rates themselves are essentially meaningless for the purposes of register analysis. Rather, it is the comparative rates that tell us the importance of a feature for characterizing the register.

As noted above, the text-linguistic research design (based on analysis of each individual text) is preferable to the corpus-linguistic design (based on analysis of entire sub-corpora). This is because the former permits description of the extent to which texts vary linguistically within a register category. It turns out that registers are more or less well-defined linguistically, so descriptions of both the central tendency and variation within the category are important. The text-linguistic research design, which treats each text as an observation, enables descriptions that capture both statistical patterns.

MD analysis is a special type of text-linguistic register analysis, which entails additional methodological considerations beyond those described above. First of all, MD analyses are designed to provide comprehensive linguistic characterizations of registers, and thus they are based on an extensive set of linguistic features (as many as 150 lexico-grammatical features in recent analyses). It is not feasible to separately analyze the distribution of each linguistic feature, and not possible to uncover general patterns of use if such analyses were done. Thus, MD analysis is based on the concept of linguistic co-occurrence – represented by underlying

linguistic 'dimensions' – with the goal of describing how registers can be more or less similar along different dimensions (see discussion in the preceding section).

In the MD approach, co-occurrence patterns are identified statistically: First, computer programs are used to analyze the distribution of linguistic features in a large corpus of texts, and then a statistical technique – factor analysis – is used to identify the sets of linguistic features that frequently co-occur in these texts. This is a bottom-up analysis. The researcher does not decide ahead of time which linguistic features co-occur, or which functions are going to be the most important ones. Rather, empirical corpus-based analysis is used to determine the actual patterns of linguistic co-occurrence and variation among registers, and subsequently, the researcher interprets those patterns in functional terms.

In a factor analysis, a large number of original variables (i.e., the rates of occurrence for linguistic features) are reduced to a small set of derived, underlying variables – the factors or 'dimensions' of variation. Each dimension represents a group of linguistic features that tend to co-occur in texts. Once the dimensions have been identified, it is possible to compute a quantitative measure for each dimension in each text: the dimension score. These dimension scores then allow comparisons of the similarities and differences among registers in a multi-dimensional space.

The MD approach is much easier to understand when illustrated through an actual case study. Thus, the following section presents the results from a large-scale MD analysis of spoken and written registers that occur in American universities, while at the same time explaining the analytical procedures in greater detail.

4. What does a typical study of register variation look like in the text-linguistic approach?

Biber (2006) applied text-linguistic register analysis to describe the patterns of linguistic variation among university spoken and written registers, including both analysis of individual linguistic features as well as an MD analysis of the overall patterns of register variation (cf. the summary in Biber & Conrad 2009: Chapter 8). The study was based on the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL Corpus; see Biber, Conrad, Reppen, Byrd, & Helt 2002). The T2K-SWAL Corpus is relatively large (2.7 million words) and representative of the range of university registers that students encounter during an American university education. Table 1 shows the overall composition of the corpus by register category.

Table 1. Composition of the T2K-SWAL Corpus

Register	Number of texts	Number of words
Spoken:		
Class sessions	176	1,248,811
Classroom management	40	39,255
Labs/In-class groups	17	88,234
Office hours	11	50,412
Study groups	25	141,140
Service encounters	22	97,664
Total speech:	**291**	**1,665,516**
Written:		
Textbooks	87	760,619
Course packs	27	107,173
Course management	21	52,410
Other campus writing	37	151,450
Total writing:	**172**	**1,071,652**
TOTAL CORPUS	**423**	**2,737,168**

The first quantitative-linguistic step in a text-linguistic register analysis is to analyze the distribution of all linguistic features that might be associated with register differences. As noted above, MD studies incorporate a much larger set of linguistic characteristics than in other traditional studies. For the present study, over 90 linguistic features were analyzed, including:

1. vocabulary distributions (e.g., common vs. rare [technical] nouns);
2. part-of-speech classes (e.g., nouns, verbs, first and second person pronouns, prepositions);
3. semantic categories for the major word classes (e.g., activity verbs, mental verbs, existence verbs);
4. grammatical characteristics (e.g., nominalizations, past tense verbs, passive voice verbs);
5. syntactic structures (e.g., *that*-relative clauses, *to* complement clauses);
6. lexico-grammatical combinations (e.g., *that*-complement clauses controlled by communication verbs vs. mental verbs);

A computer 'tagging' program identified and counted each of these features in each of the 423 texts of the T2K-SWAL Corpus.

In a traditional text-linguistic register study, registers are compared for their use of individual lexico-grammatical features. For example, Figure 2 shows that

there are striking linguistic contrasts among university registers in their reliance on the four content word classes (nouns, verbs, adjectives, adverbs). Written registers use nouns to a much greater extent than any other content word class. In contrast, spoken registers use nouns and verbs to about the same extent. As a result, verbs are much more common in the spoken registers than in the written registers. Adjectives and adverbs are distributed in a similar way: adjectives are used more commonly in the written registers, while adverbs are favored in the spoken registers.

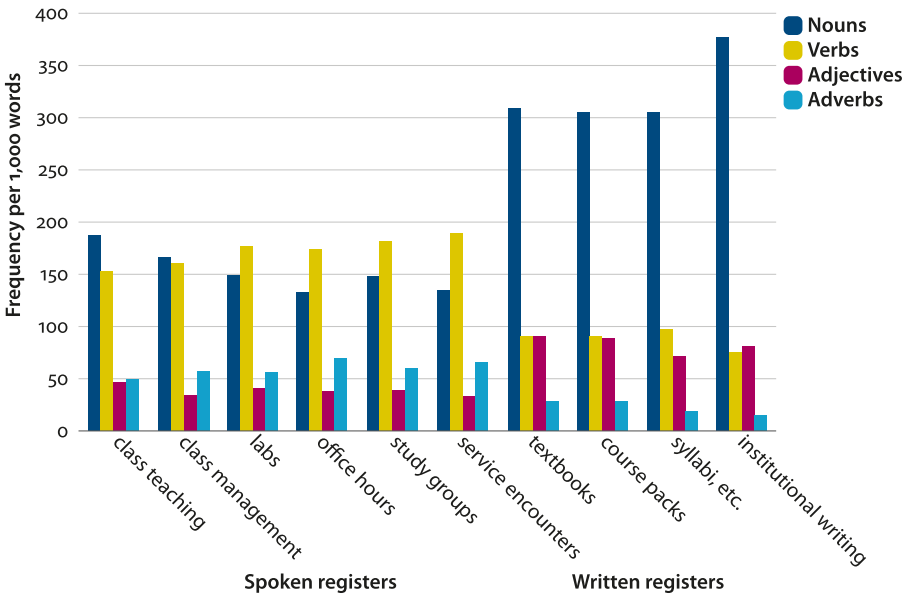

Figure 2. Content word classes across university registers

Text Sample 1, from a service encounter, illustrates the dense use of verbs in spoken registers. (Nouns are underlined, and ***verbs*** are given in bold italics.)

Text Sample 1. Service Encounter (copy shop)

clerk:	Hey there.
customer:	Hi.
clerk:	How's it ***going***?
customer:	OK. I ***want*** these, uh, ***copied***, just as they ***are***.
clerk:	Mhm.
customer:	[2 sylls] and the holes ***punched*** and the whole bit.
clerk:	OK. How many copies?

customer: Tabs, you don't have to ***worry*** about the tabs I'll ***worry*** about the tabs. ***Wait*** you ***need*** to ***mark*** where the tabs ***go*** though. I'd ***put*** a pink sheet or something where every tab ***is***.
clerk: OK.
customer: Or something. So I ***know*** where the tabs ***go***.

This short, spoken interaction includes 8 main clauses and 7 dependent clauses, each with a main verb. These verbs communicate much of the essential information: the required actions (*copied, punched, mark, put*) and the speakers' attitudes and desires (*want, worry, need, know*). In contrast, nouns are comparatively rare and add relatively little new information to the exchange. Note, for example, how the single noun *tab(s)* is used repeatedly in the sample.

At the other extreme, institutional writing represents the densest use of nouns (underlined) of any of these university registers. Text Sample 2, from a brochure for a graduate program, illustrates these patterns:

Text Sample 2. Institutional Writing (brochure for Forestry graduate programs)

Graduate education and research opportunities in the School of Forestry ***provide*** motivated individuals with the knowledge and expertise necessary to successfully ***pursue*** their career objectives in forest land management or research. The School of Forestry and the Department of Geography and Public Planning are ***located*** in the College of Ecosystem Science and Management.

The style of discourse in this register is at the opposite end of the spectrum from the spoken university registers: there are few verbs and clauses, while nearly all important information is packaged in noun phrases. In fact, this register is even more extreme than textbooks in this regard.

Generalizations like the above can be made much more convincingly when they are based on analysis of the co-occurrence patterns among the full set of linguistic features, and this is the primary research goal of MD analysis. The primary statistical technique used for an MD description is factor analysis, which identifies the underlying factors – or 'dimensions'. Each of these dimensions is a group of linguistic features that tend to co-occur in the texts of the corpus. Concretely, this means that the features as a group will all be common in some texts, and they will all be rare in other texts.

Four dimensions were identified in the study of university registers. In a statistical factor analysis, each linguistic feature has a 'factor loading' on each factor. If that loading is sufficiently large, the feature is interpreted as comprising part of the linguistic composition of the underlying dimension. Readers are referred to

previous MD studies for detailed discussions of the statistical analysis (e.g., Biber 1988, 1995; Conrad & Biber 2001).

Table 2 summarizes the important linguistic features that are grouped onto each dimension in the present study (see Biber 2006: Chapter 8). One important point to keep in mind is that the researcher does not decide which features to group together; rather, the statistical analysis identifies the groupings that actually co-occur in texts.

Table 2. Summary of the linguistic features grouped onto each dimension in the MD analysis of university registers

Dimension 1: Oral vs. literature discourse	
Positive features:	
Pronouns:	Demonstratives, *it,* 1st person, 2nd person, 3rd person, indefinite
Verb tense/aspect:	Present tense, past tense, progressive aspect
Verbs:	Mental, activity, communication
Adverbials:	Time, place, certainty, likelihood, hedges, discourse particles
Adverbial clauses:	Causative, conditional, 'other' adverbial clauses
Finite complement clauses:	*Wh*-clauses, *that*-clauses controlled by certainty verbs, likelihood verbs, and communication verbs, *that*-omission
Other:	Contractions, *wh*-questions, clause coordination, stranded prepositions
Negative features:	
Noun types:	Non-derived nouns, nominalizations
Semantic categories of nouns:	Abstract, group, human, mental
Word choice:	Word length, type/token ratio
Adjectives:	Attributive, relational
Passives:	Agentless, *by*-phrase, postnominal modifiers
Relative clauses:	*Wh*-relatives with prepositional fronting; *wh*-relatives with subject gaps
To-clauses:	Controlled by stance nouns, controlled by adjectives
Other:	Prepositional phrases, phrasal coordination
Dimension 2: Procedural vs. content-focused discourse	
Positive features:	
Modals:	Necessity, future
Verbs:	Causative, activity
Pronouns:	2nd person
Nouns:	Group
To-clauses:	Controlled by verbs of desire, controlled by 'other' verbs
Adverbial clauses:	Conditional

Table 2. *(continued)*

Negative features	
Rare, technical words:	Adjectives, nouns, adverbs, verbs
Verbs:	Simple occurrence
Adjectives:	Size
To-clauses	Controlled by probability verbs
Passives:	*By*-phrase
Dimension 3: Reconstructed account of events	
Positive features:	
Pronouns:	3rd person
Verbs:	Past tense, communication, mental
Nouns:	Human, mental
That-clauses:	Controlled by communication verbs, controlled by likelihood verbs, controlled by stance nouns, *that*-omission
Negative features	
Nouns:	Concrete, technical+concrete, quantity
Dimension 4: Teacher-centered stance	
Positive features:	
That-relative clauses	
Stance adverbials:	Certainty, likelihood, attitudinal
Adverbial clauses:	Conditional, other
That-clauses:	Controlled by stance nouns
Negative features	
Wh-questions	
Stranded prepositions	

All four of the dimensions summarized in Table 2 have both 'positive' and 'negative' features. These are actually two groupings of features: the positive features occur together frequently in texts, and the negative features occur together frequently in texts. The two groupings constitute a single dimension because they occur in complementary distribution: when the positive features occur with a high frequency in a text, that same text will have a low frequency of negative features, and vice versa.

In a subsequent analytical step, the dimensions are used to analyze the linguistic characteristics of texts and registers by computing a 'dimension score' for each text. Conceptually, a dimension score represents a simple sum of all linguistic features grouped on a dimension. For example, the Dimension 1 score is computed by adding together the frequencies of contractions, demonstrative

pronouns, pronoun *it*, first person pronouns, present tense verbs, etc. – the features with positive loadings on Factor 1 (from Table 2) – and then subtracting the frequencies of nominalizations, word length, moderately common nouns, prepositions, etc. – the features with negative loadings. Once a dimension score is computed for each text, it is possible to compare the average dimension score for each register. In the present case, all four dimensions are statistically significant, and further, they are all strong or important predictors of register differences (see Biber & Conrad 2009: 229, Table 8.3). The final major step in an MD analysis is to interpret each dimension in functional terms. This analytical step reflects the basic premise of all text-linguistic register studies that linguistic variation has a functional relationship to situational factors (see Section 1 above). Thus, the linguistic co-occurrence patterns uncovered in an MD analysis are considered to be functional: linguistic features occur together in texts because they serve related communicative functions. The interpretation of a dimension is based on (1) analysis of the communicative function(s) most widely shared by the set of co-occurring features, and (2) analysis of the similarities and differences among registers with respect to the dimension. Functional labels are thus assigned to each dimension to summarize this interpretation:

Dimension 1. Oral vs. literate discourse
Dimension 2: Procedural vs. content-focused discourse
Dimension 3: Reconstructed account of events
Dimension 4: Teacher-centered stance

There is not space in the present treatment to fully describe the results of this MD analysis. Rather, for the sake of illustration, I focus on two patterns coming out of this study: the relations among general university registers with respect to Dimension 1, and the relations among academic disciplines with respect to Dimension 2.

Dimension 1 is associated with a fundamental oral/literate opposition. The positive features on Dimension 1 (see Table 2) are associated with several specific functions, but they all relate generally to 'oral' discourse. These include: interactiveness and personal involvement (e.g., 1st and 2nd person pronouns, WH questions), personal stance (e.g., mental verbs, *that*-clauses with likelihood verbs and factual verbs, factual adverbials, hedges), and structural reduction and formulaic language (e.g., contractions, *that*-omission, common vocabulary, lexical bundles). In contrast, the negative features are associated mostly with informational density and complex noun phrase structures (frequent nouns and nominalizations, prepositional phrases, adjectives, and relative clauses) together with passive constructions.

Figure 3 shows that all spoken registers in the university corpus have large positive scores on this dimension, reflecting a frequent use of the positive 'oral'

features. In contrast, all written registers have large negative scores on this dimension, reflecting a frequent use of the negative 'literate' features. This distribution is surprising given that there are major differences in purpose and planning among the registers within each mode. For example, we might expect that classroom teaching – an informational spoken register – would exploit the same styles of informational presentation as textbooks. However, with respect to Dimension 1 features, this is clearly not the case. Instead, there is a fundamental opposition between the spoken and written modes here, regardless of purpose, interactiveness, or other pre-planning considerations.

Service encounters, office hours, and study groups – the registers with the largest positive Dimension 1 scores – are all directly interactive and 'conversational'. Text Sample 1 above illustrated the reliance on verbs (as opposed to nouns) in service encounters; Text Sample 3 illustrates the reliance on the broader constellation of positive Dimension 1 features in a service encounter. Notice the dense use of 1st and 2nd person pronouns (*I, we, you*), contractions (e.g., *we're, don't, I'm, there's*), present tense verbs (e.g., *are, have, get*), time and place adverbials (e.g., *back, there, here, again*), indefinite pronouns (*something*), mental verbs (*think, want*), and causative clauses:

Text Sample 3. Service Encounter (bookstore)

customer: Can I ask you something?
clerk: Yeah.
customer: We're at the previews and of course my book is back there with my husband. Do you have coupons?
clerk: No we don't have any of them here. You guys only get them. Yeah.
customer: OK.
clerk: Did you want to come back? Cos I can hold onto your stuff.
customer: Could you hold all this stuff? Cos I know if I'm getting a big sweatshirt there's one for a sweatshirt and one for a T. shirt.
clerk: Yeah. I'll just hold onto them.
customer: OK.
clerk: I'll go ahead and just put them in a bag.

At the other extreme, institutional writing (e.g., university catalogs) has the largest negative score on Dimension 1, making it even more 'literate' than textbooks or course packs. The following program description for anthropology begins with a friendly, inviting sentence having an extremely simple syntactic clause structure. However, this short sentence is immediately followed by complex sentences with multiple levels of clausal and phrasal embedding. Note

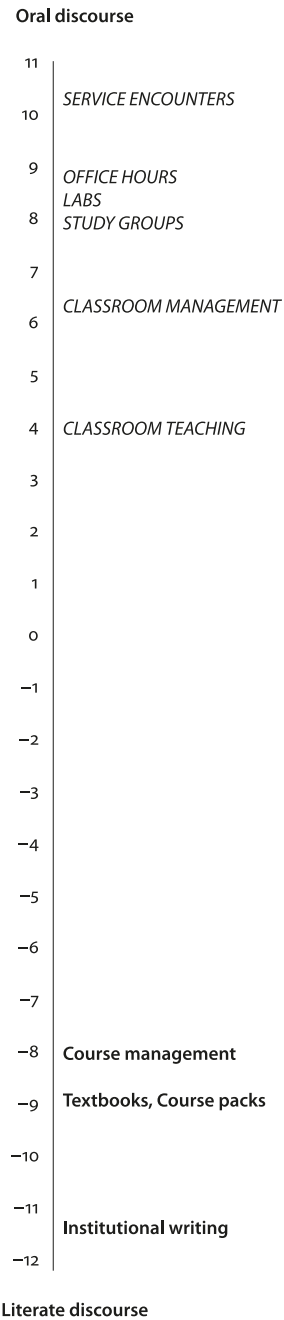

Figure 3. Mean scores of university registers along Dimension 1 – 'Oral vs literate discourse'

especially the dense use of noun phrase structures, often with adjectives and prepositional phrases as modifiers.

Text Sample 4. Institutional Writing (web catalog academic program description, Anthropology)

PROGRAM DESCRIPTION.
Anthropology is the study of people. Its perspective is biological, social and comparative, encompassing all aspects of human existence, from the most ancient societies to those of the present day. Anthropology seeks to order and explain similarities and differences between peoples of the world from the combined vantage points of culture and biology.

Cultural and Social Anthropology deal with the many aspects of the social lives of people around the world, including our own society: their economic systems, legal practices, kinship, religions, medical practices, folklore, arts and political systems, as well as the interrelationship of these systems in environmental adaptation and social change. Physical Anthropology describes and compares world human biology. Its focus is on humans and the primate order to which they belong as part of nature, and it seeks to document and understand the interplay of culture and biology in the course of human evolution and adaptation.

Many of the negative features on Dimension 1 reflect the dense use of nouns and noun modifiers in written informational texts. These features often occur together to build very complex noun phrase structures. For example, the second paragraph in Text Sample 4 begins with a very long sentence, which has only one main verb: *deal with*. Most of this sentence comprises a single noun phrase, functioning as the direct object of *deal with*. In Text Sample 4a, that sentence is marked up below to illustrate this extremely complex syntactic structure with multiple levels of embedding; head nouns of noun phrases are underlined; the main verb is in **bold**; and brackets are used to delimit postnominal modifiers:

Text Sample 4a. Institutional Writing (web catalog academic program description, Anthropology)

Cultural and Social Anthropology **deal with** the many aspects [of the social lives [of people [around the world]]], [including our own society: [their economic systems, legal practices, kinship, religions, medical practices, folklore, arts and political systems], as well as [the interrelationship [of these systems [in environmental adaptation and social change]]]].

Textbooks are similar to institutional writing in their reliance on these 'literate' Dimension 1 features, although they are usually not as densely informational as the above excerpt from a course catalog.

It might be expected that classroom teaching would have an intermediate score on Dimension 1, half way in between written informational registers like textbooks and spoken registers like office hours or study groups. Classroom teaching is similar to conversation in that it is spoken and interactive to some extent, but at the same time it is similar to textbook writing in its primary communicative purpose of conveying information. But, it turns out that classroom teaching is not at all 'literate' in its Dimension 1 score. Rather, it is much more similar to other spoken registers, including study groups and service encounters, than it is to written academic registers like textbooks. This score reflects an extremely dense use of pronouns, verbs and adverbs, questions, finite adverbial clauses, and *that*-complement clauses, as illustrated in Text Sample 5:

Text Sample 5. Classroom Teaching (humanities, Rhetoric, graduate)

> Instructor: I think some of us feel sort of really caught in a bind between agency and acculturation. Sort of um, because you know I think lot of us do want to use writing, use literacy to um, say what we want to say and to help other people say what they want to say but at the same time I think um, we're caught because we, I think we're questioning well, well you know, if, if we, if we teach X-genre are we promoting it? If we don't at the same time question it and dismantle it and kind of take it apart and look at it, and are there, are there other ways?

Findings like this illustrate how the results of MD analysis can run directly counter to our prior expectations. In this case, the pattern along Dimension 1 shows that the real-time production circumstances of classroom teaching are apparently a much more important situational factor than the informational communicative purpose, resulting in a highly 'oral' linguistic characterization.

In contrast to the spoken-written dichotomy identified by Dimension 1, Figure 4 shows that Dimension 2 cuts directly across the spoken/written continuum. Registers with large positive scores on this dimension all have communicative purposes related to the rules and procedures expected in university settings. These include both spoken registers (classroom management, service encounters, and office hours) and written registers (course management and institutional writing). In contrast, only written academic registers with an almost exclusive focus on informational content – course packs and textbooks – have the linguistic characteristics associated with the negative extreme of this dimension. Classroom teaching and study groups have intermediate scores on this dimension.

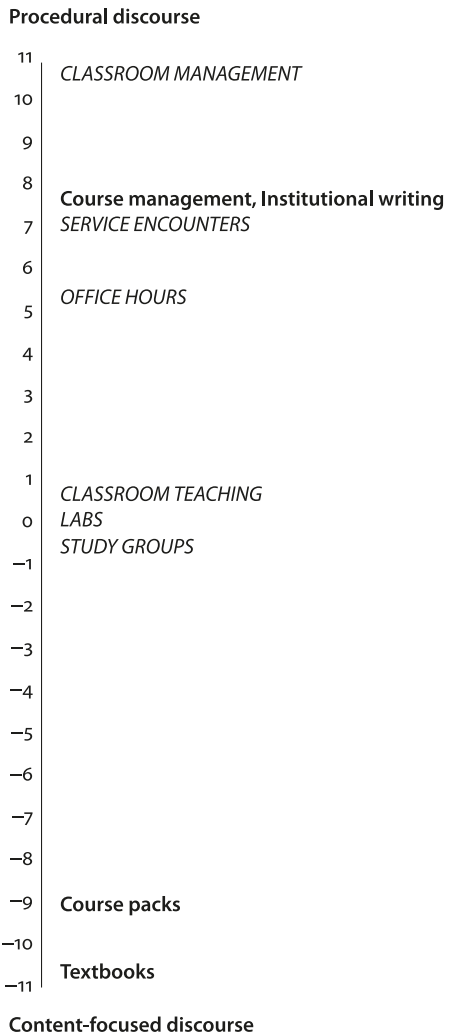

Figure 4. Mean scores of registers along Dimension 2 – 'Procedural vs. content-focused discourse'

Table 2 shows that the linguistic features associated with this dimension include necessity and prediction modal verbs (*must, should, have to, will, would, going to*), 2nd person pronouns, causative verbs, *to*-clauses with verbs of desire (e.g., *want to, would like to*), and *if*-adverbial clauses. Considering these co-occurring linguistic features, together with the distribution of registers, the interpretive label 'procedural vs. content-focused discourse' can be proposed for this dimension.

'Procedural' features are most common in spoken classroom management:

Text Sample 6. Classroom Management (humanities, History, upper division)
(Positive Dimension 2 features are in **bold underlined**)

> um, let's see, **if** a student misses more than one week of classes **you** should talk to me immediately, **if** you know **you**'re **gonna** be gone. Let's say for example **you**'re **gonna** go to Montana for a couple of days this week or something like that **you** might **let** the instructor know **you**'re **gonna** be gone. Uh, **if you**'re, I had a woman who was pregnant one semester and she, said well I'm **gonna** be missing part of the class and I said yeah, I think you probably **will** be. OK, but **let** me know. Um, **you should let** me know **if you** miss more, **if you** miss a test, **you'd have to** bring me some type of written evidence as to why **you** were gone, just so that it's fair for everybody so that they don't **have to** deal with a whole lot of excuses.

The opposite end of Dimension 2 represents the dense use of technical vocabulary, including 'rare' adjectives, nouns, adverbs, and verbs. These are words restricted to a particular discipline, like *adiabatic, arbuscules,* or *autodeliquescence*. Other negative Dimension 2 features include simple occurrence verbs (e.g., *become, happen, change, decrease, occur*), probability verb + *to*-clause constructions (e.g., *seem / appear to…*), and size adjectives (e.g., *high, large*). The dense use of these co-occurring features is restricted to the written academic registers; for example:

Text Sample 7. Textbook (natural science, Chemistry, graduate)

> Up to now we have been concerned with the magnetic resonance of a single nucleus and with explaining the physical basis of an nmr experiment. We will now turn our attention to the nuclear magnetic resonance spectra of organic molecules and in so doing will encounter two new phenomena: the chemical shift of the resonance frequency and the spin-spin coupling. These two phenomena form the foundation for the application of nuclear magnetic resonance spectroscopy in chemistry and related disciplines.

As described above, factor analysis is used in the MD approach to identify groups of co-occurring features, based on the distribution of linguistic features in texts. Register distinctions have no direct influence on the statistical identification of the factors. Rather, the factor analysis identifies the groupings of features that tend to co-occur in texts, regardless of the register of those texts. However, as shown above, these dimensions are usually powerful predictors of register differences, because both linguistic co-occurrence patterns (the basis of factor analysis) and register differences have a functional basis (cf. Egbert & Biber 2016).

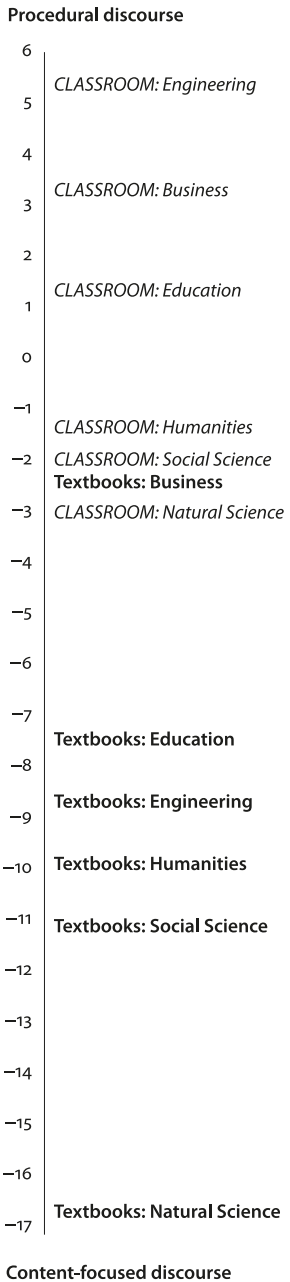

Figure 5. Mean scores of disciplines along Dimension 2 – 'Procedural vs. content-focused discourse'

The fact that dimensions are identified independently of register categories means that they can be used to explore the patterns of variation among any sub-register categories that are represented in the target corpus. Dimension 2 provides a nice example of this type. Figure 5, which plots textbooks and classroom teaching from different academic disciplines along Dimension 2, identifies a surprising pattern. With respect to the other dimensions, engineering and natural science texts are highly similar in their typical linguistic characteristics. However, along Dimension 2, these two technical disciplines are sharply distinguished: engineering is the most "procedural" discipline, within both teaching and textbooks, while natural science is by far the most "content-focused", again within both classroom teaching and textbooks. This distinction reflects the applied focus of engineering, in contrast to the more theoretical and descriptive focus of natural science.

5. What are the most promising areas of future research on register variation from a text- linguistic perspective?

One major topic for future research concerns the triangulation of text-linguistic versus variationist approaches to linguistic research questions. The text-linguistic approach analyzes linguistic rates of occurrence with the goal of characterizing texts and registers; the variationist approach considers register differences with the goal of characterizing linguistic choices. The two approaches also differ in their quantitative research designs and statistical analyses. However, they can both be used to investigate the same linguistic phenomena and same situational phenomena. Thus, it should be possible to triangulate the results to learn more about the patterns of linguistic register variation than would be possible through either approach on its own (see Baker & Egbert 2016 on the importance of triangulated methods). Biber et al. (2016) is one attempt to accomplish this goal, describing grammatical change in the use of English genitive constructions (*'s* versus *of* versus noun-noun sequences). However, that study focuses primarily on the differing research conclusions resulting from text-linguistic versus variationist approaches, with less attention given to the integration of these findings. Future research would benefit from efforts to fully integrate the results of the two approaches, for genitive constructions as well as the range of other linguistic features that can be realized through multiple variants.

A related area of research that would benefit from a triangulated approach is the study of the full range of linguistic variation found in a speech community, including both register variation and social dialect variation. Here again, we find two methodological approaches being applied in previous research, with little to no interaction between the research perspectives. Studies of social dialects have

usually adopted a variationist perspective, focused on the linguistic choices that speakers make rather than characterization of the texts that speakers produce. The primary external variables relate to social characteristics of the speakers, with little attention paid to the situational context. This contrasts with the text-linguistic register approach, which characterizes texts that are categorized by aspects of the situational context, with almost no attention paid to social variables.

However, a worthy future goal would be a comprehensive description of the full range of linguistic variation found in a speech community. This would involve two aspects: (1) application of text-linguistic analysis to describe the dialects found in the speech community, and (2) integration of a comprehensive register analysis and comprehensive dialect analysis into a single study. To our knowledge there has never been a comprehensive linguistic description of a social dialect carried out from a text-linguistic perspective. Such an analysis would survey the full set of lexico-grammatical features, documenting the extent to which each dialect used each feature. National dialects of English have been described from this perspective (e.g., there have been relatively comprehensive grammatical descriptions of American versus British English), but social dialects have been described for only a small set of linguistic characteristics, within the framework of the sociolinguistic variable.

Of course, to make sense, such a description would need to compare the full set of registers across dialects. And so, the description would also need to describe the social distribution of registers across dialect groups: to what extent do different groups employ one or another register? That is, what is the register repertoire of each dialect group? There are obvious demographic differences here. For example, it seems obvious that younger people are more likely than older people to participate in some social media registers. Similarly, people with professional occupations are more likely than manual labor occupations to produce professional reports or memos. But, these are just conjectures based on our own casual observations. To our knowledge, there has never been an empirical study to determine the distribution of registers across dialect groups.

A survey of this type would lay the foundation for a comprehensive linguistic description of the patterns of variation within a speech community: across the full range of dialects and registers, with respect to the full set of lexico-grammatical linguistic features. Analyses of this type would not replace traditional variationist accounts of social dialect variation. Rather, they would be asking fundamentally different kinds of research questions: what are the comprehensive patterns of linguistic variation within a speech community, and specifically, how do the patterns of register variation interact with the patterns of dialect variation? From a theoretical perspective, such a study would also permit empirical evaluation of the relative importance of social versus situational/communicative differences as the

basis of linguistic variation. Linguists are sharply divided on this issue (see, e.g., Finegan & Biber 1994, 2001, and the other papers published in Eckert & Rickford 2001). But surprisingly, there has not been a comprehensive analysis of linguistic variation in a speech community to investigate this issue empirically.

Finally, a third promising area of future research concerns the treatment of registers as continuous (rather than discrete) linguistic and situational constructs. As described in Section 3, linguistic variation across registers has always been studied in a continuous space, based on the rates of occurrence for linguistic features in texts. Thus, registers are described for their central tendencies in the use of linguistic features, as well as for the range of quantitative-linguistic variation among the texts within a register.

In contrast, the registers themselves have traditionally been treated as discrete categories. Most corpora are organized in terms of such discrete, non-overlapping categories (e.g., fiction, academic prose, press reportage, press editorials), with individual texts placed into a single category. However, there is no reason why texts and registers could not be investigated from the outset in a quantitative, continuous situational space (see Sharoff 2018). In a major project to study variation among web registers, Biber and Egbert (2018 – see especially Chapter 9; cf. Biber, Egbert, & Davies 2015) explore this possibility with respect to the indeterminate and hybrid nature of many web documents. That possibility is being further investigated in current research that builds on the earlier study of web registers (see Biber, Egbert, & Keller under review). In that project, each situational parameter is operationalized as a continuous variable. For example, coders evaluate the extent to which a document is interactive or opinionated, rather than simply making categorical 'yes-no' decisions. And then, register categories are determined on the basis of an empirical bottom-up analysis of those quantitative situational parameters. That is, texts are grouped into register categories on the basis of their similarities with respect to situational parameters, and thus the categories themselves are continuous constructs, with individual texts being central or peripheral to the situational characteristics of the register. In our ongoing research, we are exploring the intersections of discrete versus continuous descriptions of register variation on the web, with respect to both situational and linguistic dimensions of variation.

References

Baker, P., & Egbert, J. (Eds.). (2016). *Triangulating methodological approaches in corpus linguistic research*. New York: Routledge. https://doi.org/10.4324/9781315724812

Barbieri, F., & Wizner, S. (in press). Appendix A: Annotations of major register and genre studies. In D. Biber & S. Conrad. *Register, genre, and style*, 2nd ed. Cambridge: Cambridge University Press.

Bernstein, B. (1970). *Class, codes, and control, Vol. I: Theoretical studies towards a sociology of language*. London: Routledge.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, 62, 384–414. https://doi.org/10.2307/414678

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

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519871

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.23

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8, 9–37. https://doi.org/10.1515/cllt-2012-0002

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, 14(1), 7–34. https://doi.org/10.1075/lic.14.1.02bib

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511814358

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511804489

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly* 36, 9–48. https://doi.org/10.2307/3588359

Biber, D., & Egbert, J. (2018). *Register variation online*. Cambridge: Cambridge University Press.

Biber, D., Egbert, J., & Davies, M. (2015). Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora*, 10(1), 11–45. https://doi.org/10.3366/cor.2015.0065

Biber, D., Egbert, J., Gray, B., Oppliger, R., & Szmrecsanyi, B. (2016). Variationist versus text-linguistic approaches to grammatical change in English: Nominal modifiers of head nouns. In M. Kytö & P. Pahta (Eds.), *Cambridge handbook of English historical linguistics*, pp. 351–375. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139600231.022

Biber, D., Egbert, J., & Keller, D. (under review). Reconceptualizing register in a continuous situational space. Ms.

Biber, D., & Gray, B. (2011). Grammar emerging in the noun phrase: The influence of written language use. *English Language and Linguistics*, 15, 223–250. https://doi.org/10.1017/S1360674311000025

Biber, D., & Gray, B. (2013). Being specific about historical change: The influence of sub-register. *Journal of English Linguistics*, 41, 104–134. https://doi.org/10.1177/0075424212472509

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511920776

Biber, D., & Gray, B. & K. Poonpon. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45, 5–35. https://doi.org/10.5054/tq.2011.244483

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.

Biber, D., & Jones, J. K. (2009). Quantitative methods in corpus linguistics. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1286–1304). Berlin: Walter de Gruyter.

Biber, D., & Reppen, R. (2002). What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition*, 24, 199–208. https://doi.org/10.1017/S0272263102002048

Brown, P., & Fraser, C. (1979). Speech as a marker of situation. In K. R. Scherer & H. Giles (Eds.), *Social Markers in Speech* (pp. 33–62). Cambridge: Cambridge University Press.

Brown, G., & Yule, G. (1983). *Discourse Analysis*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511805226

Carroll, J. B. (1960). Vectors of prose style. In T. A. Sebeok (Ed.), *Style in language* (pp. 283–292). Cambridge, MA: The MIT Press.

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.) *Spoken and written language: Exploring orality and literacy* (pp. 35–54). Norwood, NJ: Ablex.

Chafe, W. L., & Danielewicz, J. (1986). Properties of spoken and written language. In R. Horowitz & S. J. Samuels (Eds.), *Comprehending oral and written language* (pp. 82–113). New York: Academic Press.

Conrad, S., & D. Biber (Eds.). (2001). *Variation in English: Multi-dimensional studies*. London: Longman.

De Beaugrande, R. A., & Dressler, W. U. (1981). *Introduction to text linguistics*. London: Longman.

Eckert, P., & Rickford, J. R. (Eds.). (2001). *Style and sociolinguistic variation*. Cambridge: Cambridge University Press.

Egbert, J., & Biber, D. (2016). Do all roads lead to Rome?: Modeling register variation with factor analysis and discriminant analysis. *Corpus Linguistics and Linguistic Theory*. 14(2), 233–273.

Ervin-Tripp, S. (1972). On sociolinguistic rules: Alternation and co-occurrence. In J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication* (pp. 213–250). New York: Holt.

Ferguson, C. (1959). Diglossia. *Word*, 15, 325–340. https://doi.org/10.1080/00437956.1959.11659702

Ferguson, C. A. (1994). Dialect, registers, and genre: Working assumptions about conventionalization. In Biber, D. & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 15–30). New York: Oxford University Press.

Finegan, E., & Biber, D. (1994). Register and social dialect variation: An integrated approach. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 315–347). New York: Oxford University Press.

Finegan, E., & Biber, D. (2001). Register variation and social dialect variation: The register axiom. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 235–267). Cambridge: Cambridge University Press.

Gray, B. (2013). Interview with Douglas Biber. *Journal of English Linguistics*, 41(4), 359–379. https://doi.org/10.1177/0075424213502237

Gray, B. (2015). *Linguistic variation in research articles: When discipline tells only part of the story*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.71

Halliday, M. A. K. (1988). On the language of physical science. In M. Ghadessy, (Ed.). *Registers of written English: Situational factors and linguistic feature* (pp. 162–178). London: Pinter.
Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
Hudson, R. A. (1980). *Sociolinguistics*. Cambridge: Cambridge University Press.
Hymes, D. (1972). Editorial introduction to "Language in Society". *Language in Society*, 1, 1–14. https://doi.org/10.1017/S0047404500006515
Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia, PA: University of Pennsylvania Press.
Irvine, J. (1979). Formality and informality in communicative events. In J. Baugh & J. Sherzer (Eds.) *Language in use: Readings in sociolinguistics* (pp. 211–228). Englewood Cliffs, NJ: Prentice-Hall.
Longacre, R. (1976). *An anatomy of speech notions*. Lisse: Peter de Ridder Press.
Ochs, E. (1979). Planned and unplanned discourse. In T. Givón (Ed.) *Discourse and syntax* (pp. 51–80). New York: Academic Press.
Sharoff, S. (2018). Functional text dimensions for annotation of web corpora. *Corpora*, 31(2), 65–95. https://doi.org/10.3366/cor.2018.0136
Van Dijk, T. A. (1972). *Some aspects of text grammars: A study in theoretical linguistics and poetics*. The Hague: Mouton.

Address for correspondence

Douglas Biber
Department of English
Northern Arizona University
Flagstaff, AZ 86011-6032
USA
douglas.biber@nau.edu