1 Introduction

1.1 General Introduction

In this Element, we introduce lexical multidimensional analysis (LMDA), an extension of the multidimensional (MD) analysis framework developed by Biber in the 1980s ("multi-feature multidimensional analysis") to study register variation. Through the identification of (lexical) dimensions or sets of correlated lexical features, LMDA enables the analysis of lexical patterning from a multidimensional perspective. These lexical dimensions represent a variety of latent, macro-level discursive constructs. Although LMDA can be utilized for a range of lexis-based analyses, in this Element the focus is on its application to discourse analysis for the exploration of discourses and ideologies.

The wide extension of corpora makes the analysis macro-level

The authors have independently developed LMDA since the 2010s, initially through Fitzsimmons-Doolan's analysis of language ideologies in a body of educational policy texts (Fitzsimmons-Doolan, 2014, 2019) and Berber Sardinha's analysis of representations of American and Brazilian cultures on Google Books (Berber Sardinha, 2014, 2019, 2020). Since then, the approach has been extended to the analysis of other topics and domains, including US migrant education (Fitzsimmons-Doolan, 2023), the historical development of applied linguistics (Berber Sardinha, 2021, 2022a), popular music (Delfino et al., 2023), the infodemic (Berber Sardinha et al., 2023), and literary style (Kauffmann & Berber Sardinha, 2021), among other domains.

In this Element, we introduce readers to LMDA by focusing on theoretical and operational issues inherent in this approach. On a theoretical level, we explore the relationship of lexis to discourse and ideologies by discussing how lexis serves as markers of discourse formations and ideological alignment. On an operational level, we provide initial guidance on technical issues, from handling frequency counts to the utilization of statistical procedures. Since LMDA includes qualitative analysis of texts, we offer insights into interpreting sets of correlated lexical features from a discourse analytical standpoint.

Two case studies are included to demonstrate the practical application of LMDA in analyzing discourses in different contexts. The first case study illustrates how LMDA can reveal the discourses surrounding climate change on the conservative GETTR social media platform, providing insights into how these discourses manifest in a contested space. And the second case study examines migrant education ideological discourses, focusing on their distribution over time and by register.

https://doi.org/10.1017/9781009335683 Published online by Cambridge University Press

1.2 LMDA's Foundation in Traditional Multidimensional Analysis

Procedurally and theoretically, LMDA is grounded in traditional MD analysis (TMDA). Douglas Biber developed MD analysis in the mid 1980s (Biber, 1988) for the functional description of variation across multiple registers, which, according to Biber and Conrad (2004, p. 42), are "different varieties of language that are associated with different situations and purposes." Since then, MD analysis has evolved to address single-register analysis, including examining variation by authors, social groups, or time periods. As such, the primary goal of TMDA is twofold: first, to identify the intrinsic linguistic parameters, or dimensions, that underlie variation (e.g., by register or style); and second, to delineate the linguistic similarities and differences among texts in relation to these dimensions along a continuous space of variation.

Typically, the basis for the interpretation of linguistic co-occurrence in TMDA is functional. According to Biber (1995, p. 30), "linguistic features co-occur in texts because they reflect shared functions." As a consequence, the dimensions resulting from the co-occurrence of the linguistic features will reflect the communicative functions performed by the texts in particular situational contexts.

Linguistic co-occurrence is captured statistically in TMDA through the computation of correlation coefficients for each pair of linguistic features across the texts in the corpus. Because each observation unit is an individual text, the correlation quantifies how pairs of linguistic features co-occur (positive correlation) or are mutually exclusive (negative correlation) across different texts. However, since in TMDA the association between linguistic co-occurrence and functional realization is predicated on groups of features performing communicative functions, rather than individual pairs of features, it is necessary to rely on multivariate statistical procedures to detect such patterns of association.

Factor analysis leads to the identification of the dimensions, which are the underlying parameters of variation across the texts. The factors are interpreted based on the communicative functions of the co-occurring features and given an interpretive label to capture their essence. Once interpreted communicatively, the factors are considered dimensions.

Since the dimensions represent a continuum of variation, registers can be systematically compared along the dimensions. The similarity between registers is determined by how similarly they use the features that co-occur within these dimensions. Since no single dimension can fully capture the range of similarities and differences among registers, a multidimensional conceptualization of register variation is needed.

The multidimensional nature of the approach is premised on the assumption that multiple parameters of variation act simultaneously on the texts, shaping them to perform a particular job in a particular communicative situation. This means that each single text reflects each dimension to a particular degree, and that no text is free from the incidence of any dimension. The extent to which a text is shaped by the incidence of a dimension is referred to as the extent of its markedness on a dimension. Consequently, different texts will be marked by different dimensions at varying degrees, resulting in a distinctive multidimensional profile of each text. Because functional variation among texts is largely predicted by register (Biber, 2012), texts from the same register will tend to have similar multidimensional profiles in TMDA.

The linguistic features used in TMDA are lexico-grammatical, predominantly comprising structural elements such as tense, aspect, subordination, phrasal structures, modalization, and coordination. Additionally, lexical features are categorized into grammatical classes (such as downtoners, hedges, amplifiers) or semantic categories that differentiate within word classes, including nouns (e.g., abstract, animate, technical), adjectives (e.g., color, evaluative, time), and verbs (e.g., communication, mental, existence). This feature set is selected for its ability to describe the underlying communicative parameters of language from a functional perspective. Though the procedures and underlying assumptions about variation are shared with TMDA, LMDA uses only lexical features and, thus, the resulting dimensions are theorized as macro-level discursive constructs such as discourses, ideologies, or themes.

1.3 Discourses and Ideologies

In contrast to TMDA which identifies functional variation in corpora, LMDA is a method for identifying a different type of variation in a corpus – namely that of latent, macro-level discursive structures. Among such structures, this Element focuses on discourses and ideologies, which we elaborate on in this subsection. We use *ideological discourses* as an umbrella term which includes a variety of constructs that exist in the "socio cognitive" space bounded by and between ideologies and "text or talk" (van Dijk, 2018, p. 242). Discourses (Baker, 2010), language ideologies (Kroskrity, 2004; Schieffelin et al., 1998), ideological discourses (Fitzsimmons-Doolan, 2023), and representations (Berber Sardinha, 2019, 2020) have all been identified in this space. Examples of entities identified under the umbrella of ideological discourses include assumed ideological positions such as *immigrants are threats*, *a people group shares a common language* (e.g., Germans speak German), and *growth is always desirable*. Other entities are less transparently ideological, such as *the discourse of*

https://doi.org/10.1017/9781009335683 Published online by Cambridge University Press

educational practice. When we use the term *ideological discourses* in this Element, we are referring to a range of macro-level discursive constructs that share many common features but can be distinguished on some parameters.

Ideological discourses can be expressed about a range of topics and, though usually highly recognizable as concepts in their explicit form, are rarely expressed explicitly. For our purposes, by ideological discourses, we mean socially shared, socially situated representations of real-world phenomena conveyed implicitly through language use. Because they are shared, ideological discourses also constrain or limit how real-world phenomena are represented. By socially situated, we mean that ideological discourses are developed through social practice and social experience. Because they represent real-world phenomena, ideological discourses make meaning.

Ideological discourses allocate social power (Kroskrity, 2004). They may also be thought of in terms of dominance. That is, when actions consistent with an ideological discourse are taken, some individuals benefit while others do not (or lose) in terms of resource allocation. Dominant discourses are widely accepted and naturalized (Kroskrity, 2004). They tend to be expressed and perceived as "facts." Nondominant discourses can be referred to as *resistant* or *alternative*.

As mentioned earlier, the entities of ideological discourses tend not to be expressed explicitly, but are identified with repeated patterns of wording (Stubbs, 1996, p. 158, 2001) or evaluative stances (Hunston, 2011). However, register differences also mean that these entities may be expressed differently in different texts (Berber Sardinha, 2021). Corpus linguistics studies are typically used to identify these patterns through measures of relative frequency, repetition, and association.

1.4 Corpus Linguistics Approaches to Ideological Discourses

In this subsection, we focus on two influential approaches to discourse analysis that have been integrated with corpus tools and methods to study ideological discourses: critical discourse analysis (CDA) and corpus-assisted discourse studies (CADS).

In the 1990s, CDA emerged as a distinct academic field, marking a development in the study of language and society. It is inherently interdisciplinary, drawing on a diverse array of disciplines including pragmatics, sociolinguistics, philosophy, social psychology, and theoretical linguistics. One of the primary objectives of CDA is to facilitate an intersectional dialog among these disciplines.

As a politically committed field (Caldas-Coulthard & Coulthard, 1996, p. xi), CDA assumes a proactive role in seeking social justice, aligning its analytical

https://doi.org/10.1017/9781009335683 Published online by Cambridge University Press

focus with the pursuit of equitable societal structures. As Forchtner (2013, p. 1439) puts it, CDA does not regard "discourse [as] merely talk," but rather as a constitutive phenomenon that "actually structures conduct" (Webster, 2003, p. 89). Across approaches, CDA scholars are committed to "de-mystifying ideologies and power through the systematic and retroductable investigation of semiotic data" (Wodak & Meyer, 2009, p. 3).

Although CDA is not a corpus-based approach, researchers have experimented with corpus methods, partly in response to methodological criticism concerning rigor and objectivity. A notable critique comes from Widdowson (1995), who contends that CDA analysts often conduct analyses with the primary aim of confirming their pre-existing hypotheses (e.g.,by cherry-picking examples) rather than seeking to gather comprehensive evidence that could potentially contest their views. Similarly, Fowler (1996, p. 8) raises concerns about the scope of CDA, specifically its tendency to engage with a limited range of texts, resulting in evidence that is "fragmentary and exemplificatory." These criticisms stem from the qualitative nature of CDA, which demands deep and interpretative engagement with data, often at the expense of a broader sample size. Addressing the issue of limited text samples, Stubbs (1997) suggests incorporating large text samples into CDA, which can be achieved through various approaches, one of which is to utilize existing precompiled corpora as sources for extracting a more narrowly focused collection of texts that are relevant to the research objectives.

In applying corpus linguistics to CDA, researchers typically utilize tools like concordances, word frequency lists, and keywords. An example is Orpin (2005), who employed concordancing and word frequency counts in the analysis of the semantic domain of *corruption*. The study analyzed the frequencies of collocates of these words, using a corpus of 800 texts, sourced from four newspapers within the Bank of English.

Beyond frequency-based analysis for CDA studies using corpus linguistics, Stubbs (1997) proposes the adoption of methodological principles advocated by MD analysis. First, this would involve the recognition that "registers are very rarely defined by individual features, but consist of clusters of associated features which have a greater than chance tendency to co-occur" (Stubbs, 1997, p. 5). Second, this integration would involve adopting analyses "of co-occurring linguistic features" (Stubbs, 1997, p. 9), a key principle of MD analysis, rather than focusing solely on individual features. Although these suggestions may not have been embraced in the practice of CDA, they highlight the potential for applying MD principles to identify and critique ideological discourses from a corpus linguistic perspective. Essentially, these points lead to

https://doi.org/10.1017/9781009335683 Published online by Cambridge University Press

a multi-way characterization of texts and registers, away from binary distinctions. We argue that both suggestions can be incorporated in corpus-based analyses of discourse through LMDA, as we demonstrate in this Element.

In turn, CADS represents a development within corpus linguistics that integrates corpus-based methods and discourse analysis. Unlike CDA, where the integration of corpus methods was a subsequent development, CADS has incorporated corpus linguistics techniques as a fundamental part of its approach from its inception in the late 1990s and early 2000s. It emerged primarily in the UK and Italy through the pioneering work of researchers such as Paul Baker, Michael Stubbs, Tony McEnery, and Alan Partington. This development was facilitated by the increasing availability of both large corpora and personal corpus analysis software, such as WordSmith Tools (Scott, 1996).

As in corpus-assisted CDA, CADS researchers also rely on mainstream corpus tools such as concordances, keywords, and collocate and word lists, which enable them to both mine the corpus for the most salient linguistic features associated with a discursive issue and identify the patterns surrounding these linguistic features. As Gillings et al. (2023) put it, “corpus assistance helps us to link large-scale social phenomena with linguistic choices at the micro level.” Analysts in CADS concentrate on uncovering recurrent patterns within the corpus, which is in line with the key concept of discursive repetition, “the idea that an attitude or ideology can be transmitted over a long period of time through people’s repeated encounters with words or phrases, eventually resulting in a discourse being uncritically perceived as natural or normal” (Baker, in press).

Keyword analysis (Scott, 1996), which identifies words that are used statistically more frequently in one corpus compared to another, is a widely used method in CADS due to its utility in helping researchers sample a subset of words from the entire corpus that merit further investigation. Baker (2014) employed keyword analysis to investigate the gender differences hypothesis in language use, concluding that this hypothesis, as it pertains to lexical choice, was not substantiated by the data. Depending on how a keyword study is designed, the approach can be used to identify discourses or ideologies. For example, Baker and McEnery (2015) identified discourses about government benefits in a corpus of tweets by finding and grouping keywords.

In CADS, as in most keyword studies, the detection of keywords typically relies on frequency counts taken across the entire corpus rather than on a text-by-text basis (but see Egbert and Biber [2019] for a version of keyword analysis that uses text-based counts). This methodological choice can lead to skewed distributions of keyword usage. Such a skew arises because the corpus-wide counts may be influenced by the overuse of certain words by individual speakers or texts, rather than reflecting marked choices across the texts.

Collocational networks are an innovation aimed at flagging groups of collocations through visual displays that represent the connections among different individual collocations in a corpus. Tools like GraphColl, which is part of the LancsBox suite, provide capabilities for constructing collocation networks (i.e., associational relationships among a node's first and additional order collocates). A network is composed of different individual graphs, which can take various forms, including linear graphs, triangles, and quadrilaterals. As Baker (2016) shows, these different graphs can indicate specific linguistic patterns among the words, such as grammatical class membership, lexical bundles, or frames.

Each of these corpus linguistics approaches to discourse and ideologies is based on a theoretical relationship between lexical variables and ideological discursive constructs. The next subsection explores such theories.

1.5 Theories of Lexis

Though Stubbs (2015) indicates that there is no unified theory of lexis, most theoretical models that give prominence to lexis are rooted in collocation. These include theories of semantic prosody and semantic preference – and all variations in nomenclature referring to these ideas, extended lexical units (Stubbs, 2009), lexical priming (Hoey, 2005), and knowledge-free associative patterning (Phillips, 1985). A collocation is a node word and a word that repeatedly and meaningfully co-occurs with that node within a given local span in a text or a corpus. The local span is often four words to the left and four words to the right of the node. "Repeatedly" and "meaningfully" can be operationalized in a variety of ways by the analyst in terms of frequency and association (Brezina et al., 2015). Firth established the theoretical groundwork for collocation and famously claimed that we "shall know a word by the company it keeps" (1957/1968, p. 11). It has been well established that collocations can reveal socially loaded perspectives (Baker, 2010, 2016; Stubbs, 1996), and Baker (2016) shows how analysis of collocational networks can reveal information which may have "ideological significance" (p. 148).

Semantic preference and semantic prosody are two of the primary mechanisms through which collocation creates meaning. Semantic preference is also called semantic association (Hoey, 2005) and generally refers to the lexical set (i.e., thematic set) to which collocates of a node belong (e.g., the domain of medicine or the absence/change of state; Partington, 2004). Semantic prosody has two meanings (Hunston, 2007). The more common meaning is the evaluative (positive or negative) association a node and its collocates convey.

Semantic prosody in this sense is also called discourse prosody (Stubbs, 2001) and evaluative prosody (Partington et al., 2013). Hunston (2007) concludes that meaning derived from semantic preference and semantic prosody can often (but not always) be carried across texts by individual words. Finally, both semantic prosodies and semantic preferences are thought to often demonstrate register association (Partington, 2004).

Hoey's (2005) theory of lexical priming attempts to account for collocation observed in corpora through a psychological process of priming which is also sensitive to register. In this theory, at the local level, based on an individual's language experience, individual words are primed for collocation, semantic association (semantic preference), colligation, and pragmatic functions. Through a nesting operation, multi-word units are created with their own primes. At the level of text, words/multi-word units are primed to co-occur with other words/multi-word units in a text (textual collocation), in particular discourse functions (textual semantic association) and in particular sections of a text (textual colligation). In sum, the theory of lexical priming suggests that a large part of an individual's language can be accounted for through bottom-up processes driven by associational patterns in the lexical system with the text as an important unit of analysis.

Finally, Phillips (1985) hypothesizes that "a distributional analysis of linguistic substance; invoking no knowledge of the semantic content, the syntactic organization, or the lexical meaning of the text; would reveal global patternings in the lexis of the text" (p. 11) that he calls macrostructures. He goes on to test this hypothesis in a textbook, identifying the "aboutness" of chapters and the text as a whole based on frequency and associational measures of collocations, resulting in multiple groups of words he calls "lexical sets." While the macrostructure in question in this study is aboutness, ideological discourses can be similarly categorized as macrostructures (Ellis, 2019).

These theories indicate quite a bit about identifying ideological discourses from lexis. First, examining lexis through corpora reveals socially shared primings, collocations, semantic prosodies, and semantic preferences. As repositories of socially shared language, corpora reveal shared lexical primings, which in turn add to the priming data for authentic users of the language captured in a corpus. Hoey (2005) notes that "priming leads to a speaker unintentionally reproducing some aspect of language, and that aspect, thereby reproduced, in turn primes the hearer" (p. 9). As mentioned earlier, according to Hoey (2005), priming can explain collocation and Partington (2004) describes how socially shared primes account for socially shared semantic preferences and semantic prosodies, which are part of the communicative competence of individual speakers. He also presents a model in which collocations, semantic

https://doi.org/10.1017/9781009335683 Published online by Cambridge University Press

preferences, and semantic prosodies are all derived from text and each other in increasing levels of abstraction (with semantic prosodies being the most abstract). That is, a collocation is identified in a text, a semantic preference is identified from a set of collocations, and a semantic prosody is identified from a set of semantic preferences.

Second, these theories suggest that information about many linguistic levels seems to be accessible from associations among lexical items when text is the unit of analysis. Hoey (2005) shows how it is possible that the lexical system encodes the grammatical system, concluding that "what we think of as grammar is the product of the accumulation of all of the lexical primings in an individual's lifetime" (p. 159). Phillips (1985) is able to identify textual macrostructures from such associations. The underlying associative structure and successful performance of contemporary large language models (e.g., ChatGPT) also empirically validate this claim. Finally, the fact that collocations, semantic preferences, and semantic prosodies are all sensitive to register means that contextual/situational/social information must also be encoded in lexical distribution.

Therefore, taken together, theories of discourse and ideology and lexis, as well as corpus linguistics approaches to ideological discourses discussed in the previous subsection, suggest that examining associational, co-occurrence patterns of lexis through corpora using text as the unit of analysis can reveal ideological discourses. As repositories of socially shared language, depending on the alignment between the design of a specific corpus and the discourses being identified, corpora are ideal data sources. Lexis seems to be the appropriate linguistic level for identifying macrostructures such as ideological discourses conveyed through evaluative language. Partington's (2004) model sets up discourses as being an additional level of abstraction beyond semantic prosodies which can thus be derived from lexical co-occurrence. There is an indication that co-occurring sets of lexical items within and across texts carry ideological information. Finally, register seems to be an important delimiter in terms of both lexical association patterns and ideological expression, or, as Silverstein (1998, p. 126) puts it, "if all cultural and linguistic phenomena are essentially event linked, even where they appear to be manifestations of people's 'intuitions,' they are, as it were, ideological 'all the way down.'"

1.6 Similarities and Differences between Traditional and Lexical MDA

As LMDA is an extension of TMDA and both approaches have roots in the Flagstaff School of Corpus Linguistics (cf. Cortes & Csomay, 2015), their

https://doi.org/10.1017/9781009335683 Published online by Cambridge University Press

procedures are roughly the same and they share foundational assumptions. That is, as with conducting a TMDA, to conduct an LMDA, a researcher (1) constructs a corpus, (2) identifies variables for analysis, (3) counts occurrences of the variables per text, (4) subjects the counts to a multidimensional analysis to identify underlying constructs, and (5) for each dimension in the result, engages in qualitative analysis of texts with high values to interpret the underlying construct based on how the variables are deployed. However, distinct characteristics emerge as each approach is tailored to specific research goals. These differences and similarities will now be outlined, beginning with the common traits.

Variation: Both TMDA and LMDA are founded on the principle that language use inherently varies depending on the context. This means that language cannot be treated as a homogeneous entity; rather, its usage is shaped by the specific historical and contextual factors in which it occurs. Consequently, linguistic descriptions within these frameworks must account for systematic variation in language use.

Comprehensiveness: Both TMDA and LMDA assume a comprehensive approach to linguistic description, as opposed to a reductionist one. This means that their descriptions are based on a varied set of linguistic features, rather than starting off with just a few elements. This comprehensive approach allows for a more detailed and inclusive analysis of language use.

Co-occurrence: The need for a large and varied pool of linguistic features arises from the need to model linguistic co-occurrence; in turn, the relevance of co-occurrence arises from the fact that it reflects shared function (a communicative function for TMDA and a discursive function for LMDA). Since linguistic co-occurrence plays such a central role in MD analysis, it has achieved "formal status in the Multi-Dimensional approach to register variation" (Biber, 1995, p. 30).

Dimensionality: Both TMDA and LMDA share the hypothesis that latent constructs underlie language usage, shaped by the conditions in which language is used in natural settings. This hypothesis posits that these underlying constructs manifest as "dimensions" – sets of co-occurring linguistic features across texts.

Multidimensionality: Given that language variation is patterned by dimensions, and that multiple dimensions are needed to account for variation, both approaches are inherently multidimensional. This means they presuppose the simultaneous action of various dimensions on texts, shaping them to perform specific communicative functions in TMDA or to convey particular discourses or ideological formations in LMDA.

https://doi.org/10.1017/9781009335683 Published online by Cambridge University Press

Parsimony: While both TMDA and LMDA utilize a large and varied set of linguistic features, their objective is to identify the smallest number of dimensions that account for language variation. This approach reduces the extensive initial set of individual characteristics into a few cohesive groups of variables, collectively explaining the variation observed across texts.

Comparative stance: Both TMDA and LMDA foster comparisons, as they highlight similarities and differences between various language varieties (TMDA) or social contexts (LMDA). By comparing different categories along the dimensions in extended study design, these categories can be more sharply portrayed.

Statistical foundation: Since a reliance on statistical methods is a defining trait of the Flagstaff School of Corpus Linguistics, both types of MD analysis depend on statistical analysis. These methods are essential for detecting latent phenomena, that is, constructs that while predicted are not directly observable. The primary statistical procedure in MD analysis is correlation, which is used to measure the systematic co-variation of variables. Given the comprehensive approach to the array of linguistic features and the goal of identifying dimensions of variation, MD analysis employs multivariate statistical techniques, including dimensionality reduction methods like factor analysis. The factors identified in such analyses represent sets of correlated variables, corresponding to patterns of cross-text variation.

Qualitative interpretation: Despite their strong quantitative foundation, both approaches necessitate qualitative interpretation of texts to assist in unveiling the underlying communicative functions (TMDA) or discourses (LMDA). Without careful interpretation, based on the consultation of numerous text samples, dimensions cannot emerge from factors.

Despite their similarities, TMDA and LMDA can be distinguished based on differing research goals, feature sets, and interpretive foci.

Research goals: TMDA is particularly relevant for research goals that focus on the functional aspects of language. This approach is grounded in the idea that shared linguistic features indicate a shared function. It is typically employed to describe register variation along functional lines, essentially detailing the differences and similarities across various registers in a language or domain. If a researcher's objective involves analyzing texts from a functional perspective through structural, syntactic, or morphological classes, then TMDA is the appropriate method.

In contrast, as presented in this Element, LMDA is designed to cater for the identification of latent, macro-level constructs encoded in discourse. The range of research goals that can be addressed with a focus on

https://doi.org/10.1017/9781009335683 Published online by Cambridge University Press

discourse is vast, covering such aspects as ideologies, representations, identities, themes, motifs, schemas, and many other conceptual systems. Thus, if the research goal includes describing the lexical materialization of such discourse-based constructs, then LMDA is the necessary method over TMDA.

Linguistic features: The features typically used in a TMDA are lexico-grammatical, comprising structural, syntactic, and morphological classes. The exact features to be used in a TMDA project depends on previous consideration of the features of relevance for the specific research goals. On the other hand, LMDA utilizes the actual words in the texts as its primary units of analysis, contrasting with the broader lexico-grammatical classes employed in TMDA. Thus, in an LMDA, the features used are entirely lexical, including the actual words, their base forms (lemmas), semantic categories, collocations, or n-grams.

Interpretive focus: TMDA primarily focuses on identifying functional parameters of variation in language. By "function," we refer to the communicative roles that linguistic features play, enabling users to perform specific tasks with language. As Biber and Conrad (2019, p. 2) state:

> The underlying assumption of the register perspective is that core linguistic features (e.g., pronouns and verbs) serve communicative functions. As a result, some linguistic features are common in a register because they are functionally adapted to the communicative purposes and situational context of texts from that register.

The dimensions in TMDA, which are correlated sets of linguistic features, correspond to the underlying macro communicative function of the texts. Researchers determine these underlying macro functions through factor interpretation, linking the linguistic patterns to the situational characteristics of the registers. Consequently, a functional interpretation of the patterns within these dimensions is essentially "an account of *why* these patterns exist" (Biber & Conrad, 2019, p. 69; emphasis in the original text).

Conversely, in LMDA, the interpretive focus is on unearthing the latent discourse constructs materialized in the texts. The interpretation taps into the potential of lexical features as signposts or entry points to the analysis of discourse, as acknowledged in corpus-assisted approaches to discourse analysis. For instance, according to Stubbs, lexical keywords are "nodes around which ideological battles are fought" (Stubbs, 2001, p. 188). Similarly, Mautner describes a word such as *entrepreneurship* as a "carrier of key values" (Mautner, 2005, p. 96), providing "focal points around which current discourses ... crystallize" (Mautner, 2005, p. 111), in the context of educational

https://doi.org/10.1017/9781009335683 Published online by Cambridge University Press

entrepreneurialism. In turn, Krieg-Planque (2010, p. 9) considers that particular lexical expressions, which she refers to as "formulas," have a dual role of both constructing and crystallizing political and social issues. Meanwhile, TMDA offers limited entry points to the discursive layers of language because of its goal of describing variation at the functional level of language use.

1.7 Overview of Element

This section has presented the foundation of LMDA for identification of ideological discourses, demonstrating that the approach is grounded in (1) procedures of TMDA and CADS and (2) theories of lexis and discourse studies. Following this introduction, in Section 2, the major studies to date using LMDA will be synthesized. The synthesis will address variation in design and constructs identified, as well as lessons learned both in terms of methodology and theoretical advances. This section will explicitly and robustly attend to the range of meaning systems encoded in lexis that are identifiable by application of LMDA.

In Section 3, step-by-step guidance on how to perform an LMDA will be provided. The major methodological steps will be presented and illustrated, including corpus design, part-of-speech tagging and lemmatization, feature selection and counting, statistical analysis, and the interpretation of the results from both a qualitative and a quantitative perspective.

In Section 4, Case Study 1 will demonstrate how LMDA can be used to detect discourses in social media, more specifically on the conservative platform GETTR. The analysis focuses on the discourses around climate change underlying thousands of messages challenging environmental activism.

In Section 5, Case Study 2 will showcase how this approach allows researchers to explore the distribution of the constructs identified in LMDA over time or over other variables using inferential statistics. This case study uses four ideological discourses about twenty-first-century migrant education in the US.

Finally, Section 6 will briefly summarize the major points presented, consider the potential of the approach, and explore some of its possible future developments.

2 Synthesizing Existing LMDA Scholarship

2.1 Introduction

This section will focus on the LMDA studies conducted thus far. First, early LMDA studies directly grown out of TMDA will be presented, followed by more recent LMDA studies which have identified latent discursive constructs, explored the distribution of such constructs, derived additional measures from the latent discursive construct data, or some combination of these outcomes (Table 1).