Exploring multiple profiles of L2 writing using multi-dimensional analysis

Eric Friginal [*], Sara Weigle [1]

Department of Applied Linguistics and ESL, Georgia State University, PO Box 4099, Atlanta, GA 30302, United States

Abstract

In this paper, we explore the application of corpus-based multi-dimensional analysis (MDA) pioneered by Biber (1988, 1995, 2006) in understanding microscopic linguistic variation in an L2 writing corpus. The primary goals of our study are: (1) to identify the functional dimensions of L2 academic essays from the corpus collected for this special issue of the *Journal of Second Language Writing*, and (2) to analyze linguistic variation in the corpus across parameters of time and average assessment scores (in language, vocabulary, and total average score). The corpus was tagged for part-of-speech and additional semantic categories (e.g., semantic categories of verbs and nouns) using the Biber tagger (Biber, 2006). Rates of occurrence were computed for over 80 linguistic features across 209 essays based on tag counts for each text. The values of these variables were then subjected to an exploratory factor analysis. A total of four functional dimensions were identified and interpreted. These are: (1) Involved vs. Informational Focus, (2) Addressee-Focused Description vs. Personal Narrative, (3) Simplified vs. Elaborated Description, and (4) Personal Opinion vs. Impersonal Evaluation/Assessment. Overall, our study shows a successful application of MDA on a micro-level producing a range of functional profiles along different parameters in L2 writing.
Published by Elsevier Inc.

Keywords: Multi-dimensional analysis; L2 academic essays; Writing profiles

Introduction

Many corpus-based studies in the past 20 years have explored the distributions of linguistic features (e.g., modal verbs, prepositions, linking adverbials, structures of subordination and complementation, etc.) that may predict quality scores given by instructors/raters or distinguish differences between the second language (L2) proficiency levels of writers. The papers in this issue examine syntactic features across developmental and assessment variables (cf. Crossley & McNamara) and phraseological competence in learner corpora (cf. Bestgen & Granger). In Corpus Linguistics (CL), written assessment scores and linguistic frequency distributions obtained from automated taggers and parsers have provided the datasets needed to easily employ statistical tests such as correlations and multiple regressions. For example, a study conducted by Ferris (1994) identified the lexico-syntactic features of English as a second language (ESL) writing that could predict assessment scores of academic texts written by ESL students from

* Corresponding author at: 50 Peachtree St., NW #802, Atlanta, GA 30303, United States.
  E-mail addresses: efriginal@gsu.edu (E. Friginal), sweigle@gsu.edu (S. Weigle).
[1] Permanent address: 115 Scott Blvd., Decatur, GA 30030, United States.

four first language (L1) backgrounds (Arabic, Chinese, Japanese, and Spanish) across two proficiency levels (low and advanced). Ferris's seminal work reported that 28 linguistic variables accurately grouped L2 texts into two proficiency levels with five variables established as good predictors of assessment scores: (1) number of words, (2) synonymy/antonymy, (3) word length, (4) passives, and (5) 3rd person/impersonal pronouns. Research along these lines is popular (see, e.g., Beers & Nagy, 2009; Biber & Gray, 2013; Bloch, 2010; Hinkel, 2002; Hyland, 2002; Hyland & Tse, 2009; McNamara, Crossley, & McCarthy, 2010; Norris & Ortega, 2009) as findings have implications not only for L2 writing assessment, especially in the domain of automated scoring, but also for the production of teaching materials for L2 writing classes (Friginal, Li, & Weigle, 2014).

Another area of CL research goes beyond correlational and predictive data from an a priori set of features. As argued by Jarvis, Grant, Bikowski, and Ferris (2003), evidence of a consistent linear relationship between linguistic distributions and quality ratings in academic essays has been limited or somewhat inconclusive thus far. Genre and contextual differences in writing production have been found to contribute to varying distributions of linguistic features across groups of writers (e.g., native vs. non-native speakers of English), academic disciplines, and topics or prompts (Friginal, Li, & Weigle, 2014). Thus, multivariate tests such as cluster and factor analyses have also been used to investigate multiple linguistic patterns of L2 writing that highlight underlying linguistic co-occurrence features within corpora. In these applications, statistical tests identify the structural components of texts from tagged/parsed counts in a dataset. These approaches to analyzing linguistic variation provide unique information relevant to a range of research goals.

The focus of this paper

In this paper, we explore the application of corpus-based multi-dimensional analysis (MDA) pioneered by Biber (1988, 1995, 2006) in understanding microscopic linguistic variation in an L2 writing corpus. MDA is based on the theoretical assumption that functional dimensions of texts involve underlying patterns of linguistic co-occurrence (Grieve, Biber, Friginal, & Nekrasova, 2010). Data for MDA come from Exploratory Factor Analysis (EFA), which considers the sequential, partial, and observed correlations of a wide range of variables. From these multiple correlations, groups of statistically co-occurring factors are extracted (Tabachnick & Fidell, 2001). This internal patterning of tagged linguistic features in a corpus provides functional categories which correspond to salient practical distinctions within texts. The primary research questions of our study are: (1) What are the functional dimensions of academic essays from the specialized L2 corpus explored in this special issue of the *Journal of Second Language Writing*? (2) How does time of writing (beginning, middle, and end of semester of instruction) relate to the distribution of these functional features in the corpus? (3) How do assessment scores (in language use, vocabulary, and total average scores) relate to learners' use of linguistic features across these dimensions?

Multi-dimensional analysis

MDA was developed by Biber (1988) to identify underlying *dimensions of linguistic variation* from a wide range of spoken and written registers of English. This approach to linguistic variation stresses that systematic differences occur in a corpus because speakers and writers make lexical and grammatical choices appropriate to the register (audience, context, purpose, etc.) in which they are speaking or writing (Biber & Conrad, 2001; Friginal & Hardy, 2012; Hymes, 1984). More importantly, these statistical co-occurrences are not one-dimensional or linear. For example, the functional description of texts as ''informal'' or ''formal'' does not entirely capture the broad range of variation existing within these texts. By identifying these co-occurring patterns in the corpus through MDA, one can then analyze individual texts and groups of texts to better understand the core structural and functional characteristics of a given register or genre of discourse as well as changes over time, as in the English writing in this corpus.

Biber's MDA employs the reduction of a large number of tagged variables (i.e., linguistic features) into a small set of derived variables (i.e., factors) with high shared variance that can be interpreted to show various functional dimensions of corpora. His 1988 book, *Variation across Speech and Writing*, presented a total of six functional dimensions of spoken and written English from 23 sub-registers of the London-Oslo-Bergen Corpus (written texts) and London-Lund Corpus (spoken texts). These six dimensions are: (1) Involved vs. Informational Production, (2) Narrative vs. Non-Narrative Concerns, (3) Explicit vs. Situation-Dependent Reference, (4) Overt Expression of Persuasion, (5) Abstract vs. Non-Abstract Information, and (6) On-Line Informational Elaboration. These dimensions