1-1-2002

Developing Linguistic Literacy: Perspectives from Corpus Linguistics and Multi-Dimensional Analysis

Douglas Biber

Randi Reppen

Susan M. Conrad
Portland State University

Citation Details

Biber, Douglas, Randi Reppen, and Susan Conrad. "Developing linguistic literacy: Perspectives from corpus linguistics and multi-dimensional analysis." Journal of Child Language 29.02 (2002): 449-488.

DOI: 10.1017/S0305000902235345

Developing linguistic literacy: perspectives from corpus linguistics and multi-dimensional analysis

DOUGLAS BIBER, RANDI REPPEN AND SUSAN CONRAD
Northern Arizona University and Portland State University

INTRODUCTION

In their conceptual framework for linguistic literacy development, Ravid & Tolchinsky synthesize research studies from several perspectives. One of these is corpus-based research, which has been used for several large-scale research studies of spoken and written registers over the past 20 years. In this approach, a large, principled collection of natural texts (a 'corpus') is analysed using computational and interactive techniques, to identify the salient linguistic characteristics of each register or text variety. Three characteristics of corpus-based analysis are particularly important (see Biber, Conrad & Reppen 1998):

- a special concern for the representativeness of the text sample being analysed, and for the generalizability of findings;
- overt recognition of the interactions among linguistic features: the ways in which features co-occur and alternate;
- a focus on register as the most important parameter of linguistic variation: strong patterns of use in one register often represent only weak patterns in other registers.

Corpus studies have documented the linguistic differences among spoken and written registers in English and other languages. Further, by analyzing systematic corpora produced by students at different stages, these same techniques have been used to track the patterns of extended language development associated with literacy.

Two major patterns emerge from studies in this research tradition: (1) adult written language is dramatically different from natural conversation; and (2) written language is by no means homogeneous: rather, there are major linguistic differences among written registers. Thus, the developmental acquisition of linguistic literacy requires control over the patterns of register variation, in addition to a mastery of the mechanics of the written mode.

Corpus studies of individual linguistic features in speech and writing

Over the past 20 years, there have been numerous research papers and books using corpus-based techniques to document the linguistic characteristics of spoken and written registers. More recently, the *Longman Grammar of*

Spoken and Written English (LGSWE; Biber, Johansson, Leech, Conrad & Finegan, 1999) systematically describes the grammar of English giving equal attention to aspects of structure and use. The descriptions of language use in the LGSWE are based on empirical analysis of a 20-million-word corpus representing four spoken and four written registers: conversation, fiction, newspaper language, and academic prose (see Biber *et al*., 1999, chapter 2, for a description of the corpus).

Interestingly, many linguistic features show a fundamental spoken/written difference but also reflect particular patterns of variation among written registers. For example, lexical verbs and phrasal verbs are common in conversation and relatively rare in written academic prose, but they are actually most common in written fiction (LGSWE Fig. 5.2, pp. 367–69; Table 5.13, p. 409). Appositive noun phrases and relative clauses are typical characteristics of formal writing (and rare in conversation), but they turn out to be most common in newspaper writing rather than academic prose (Fig. 8.13, p. 606).

These corpus-based findings highlight the fundamental importance of mode (spoken vs. written) for descriptions of language use. However, they also show that register is a second fundamentally important factor, accounting for much of the variation within each mode.

Multi-dimensional studies of spoken and written registers

While some researchers have focused on the use of individual linguistic features, the multi-dimensional (MD) analytical approach was developed to describe the overall linguistic characteristics of a register, and to compare two or more registers. This corpus-based analytical approach is based on computational analysis of texts from spoken and written registers, to identify the most important patterns of linguistic co-occurrence: the 'dimensions' (identified statistically using factor analysis). Each dimension comprises a distinct set of co-occurring linguistic features, and each has distinct functional underpinnings. Registers can be compared in this multi-dimensional space, enabling empirical analysis of both the extent and the ways in which any two registers are different. Early MD studies investigated the synchronic relations among spoken and written registers in English (e.g. Biber, 1988), while later studies focused on the diachronic development of written registers and register variation in other languages (e.g. Korean and Somali, see Biber, 1995).

MD analyses have resulted in many unanticipated findings about the linguistic nature of spoken and written discourse. Although these studies have documented major linguistic differences between 'oral' and 'literate' registers (e.g. conversation vs. academic prose), they have not identified any absolute differences between speech and writing generally (Biber, 1988, 1995). The absence of absolute differences is due mostly to the extreme

versatility of the written mode. That is, there is comparatively little linguistic variation among spoken registers, apparently because they are all constrained by real-time production circumstances. In contrast, written registers range from the extremely dense informational styles of scientific exposition to the colloquial styles of personal letters and dialogue in fiction (Biber, 2001).

In earlier historical periods, there was considerably less variation among written registers in English (Biber, 1995; Biber & Finegan, 2001). That is, in the 17th and 18th centuries, scientific written registers were relatively similar linguistically to popular registers like fiction. But in the last century, scientific registers have moved away from popular registers, developing linguistic styles with densely packaged information, especially through complex modification of noun phrases. This gradual evolution suggests that the production possibilities of the written mode are not obvious. Rather, it took centuries to recognize that extensive revision and editing in writing can result in the extremely dense informational styles found in academic prose. These diachronic developments parallel the development of literacy skills by school-aged children explored by Ravid & Tolchinsky.

Multi-dimensional studies of literacy development

One important aspect of the framework proposed by Ravid & Tolchinsky is that linguistic development associated with literacy continues well into early adulthood. The MD approach has also been used to track these developmental changes.

For example, Reppen (1995, 2001a, b) uses the MD approach to investigate the patterns of linguistic development in a corpus of elementary student writing (ages 8;0–12;0). As early as 3rd grade (age 8;0), students begin to reflect register differences in their own writing, using linguistic features to distinguish between narrative tasks and expository tasks (e.g. use of past tense verbs vs. longer words and increased use of nominalization). This register awareness continues to be refined over the following years. For example, 6th grade (age 12;0) students begin to develop a distinct linguistic style for argumentative/persuasive writing, although it is still far removed from the decontextualized language used in adult argumentative/persuasive texts. These findings support the descriptions of increasing register or genre awareness in Ravid & Tolchinsky.

The MD analysis of elementary student registers can be compared to the adult MD model to show some of the developmental changes that take place between upper elementary school and adulthood (Reppen, 2001b). Two areas of comparison are noted here:

(1) First, the models can be compared with respect to their dimensions, and the functions represented by those dimensions. Both student and adult models have dimensions that serve the following functions: informa-

tional focus; narration; involvement/stance; argumentation (see Reppen, 2001b, p. 195). At the same time, there are striking differences. For example, student argumentative texts are contextualized and have a high number of second person pronouns, resulting in an 'other-directed' style; and the student 'projected scenario' dimension has no counterpart in the adult model.

(2) The order of the dimensions can also be compared, reflecting their relative strengths. In both the student and adult models, a fundamental oral/literate dimension is the first one to emerge. The second dimension in both models reflects narrative purposes. In contrast, in both models the last two dimensions reflect task-specific concerns, rather than the general production circumstances that are reflected in the first dimension.

The development of early adult literacy skills has also been investigated with the MD approach. Conrad (1996b, 2001) investigates variation across research and summary writing in two academic disciplines, biology and history, and compares the writing of professionals in these disciplines to university students' writing (Conrad, 1996a). Numerous differences exist across the disciplines and registers, but there are certain consistent patterns of writing development as students advance from the introductory undergraduate level through the graduate level. The most notable trend concerns the density of information packaging. In both disciplines and both types of writing, student writing at the introductory level is far less informationally dense than professional writing; but at each level, student writing becomes more informationally dense. Not only do students increase their use of technical terms, as is expected at higher levels, but they also come to control much more complex noun phrase structures generally, so that referents become highly specified. From this perspective, the development of student writing is similar to the development of scientific registers historically – that is, moving towards the extremely dense packaging of information.

Like Ravid & Tolchinsky, this MD study provides insight into the processes that influence students as they learn to write advanced, specialized registers. Specifically, in many cases students seem able to imitate the surface structure of professional writing without yet being able to express their ideas clearly through the use of those structures. For example, academic professionals in history frequently use *wh*-relative clauses for elaborated reference in summary writing. Graduate student writing shows an increasingly dense use of these same features, but in many cases, the students' elaboration makes the referents more confusing. For example, one student writes about Theodore Roosevelt: 'In 1897 his personal faith was that war with Spain would erase the social abyss which was not assured.' In examples like this, we see students imitating (or even exaggerating) the linguistic structures used

by professionals before they fully control the structure's function and can manipulate the structure competently.

As can be seen from this brief survey, corpus linguistics and Multi-Dimensional Analysis provide powerful tools to explore the linguistic developmental changes associated with older learners acquiring a range of spoken and written registers. These studies strongly support the developmental framework proposed by Ravid & Tolchinsky, documenting the important interaction between linguistic patterns of language development and register variation. There remain many areas that need to be explored further as we work to complete the picture of advanced language development.

REFERENCES

Biber, D. (1988). *Variation across spoken and written English*. Cambridge: CUP.

Biber, D. (1995). *Dimensions of register variation : a cross linguistic comparison*. Cambridge: CUP.

Biber, D. (2001). On the complexity of discourse complexity: a multidimensional analysis. In S. Conrad & D. Biber (eds), *Variation in English : multi-dimensional studies*. London: Longman. (Reprinted from Discourse Processes (1992) **15**, 133–63).

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics : exploring language structure and use*. Cambridge: Cambridge University Press.

Biber, D. & Finegan, E. (2001). Diachronic relations among speech-based and written registers in English. In S. Conrad & D. Biber (eds), *Variation in English : multi-dimensional studies*. London: Longman.

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

Conrad, S. (1996a). Academic discourse in two disciplines: professional writing and student development in biology and history. Unpublished doctoral dissertation, Northern Arizona University.

Conrad, S. (1996b). Investigating academic texts with corpus-based techniques: an example from Biology. *Linguistics and Education* **8**, 299–326.

Conrad, S. (2001). Variation among disciplinary texts: a comparison of textbooks and journals in Biology and History. In S. Conrad & D. Biber (eds), *Variation in English : multi-dimensional studies*. London: Longman.

Reppen, R. (1995). A multi-dimensional comparison of spoken and written registers produced by and for students. In B. Warvik, S. Tanskanen & R. Hiltunen (eds), *Organization in Discourse* Vol. 15, 477–86. Turku, Finland: University of Turku Press.

Reppen, R. (2001a). Elementary student writing development: corpus-based perspectives. In Rita Simpson & John Swales (eds), *Corpus linguistics in North America : selections from the 1999 Symposium*, 211–25. Ann Arbor: University of Michigan Press.

Reppen, R. (2001b). Register variation in student and adult speech and writing. In S. Conrad & D. Biber (eds), *Variation in English : multi-dimensional studies*. London: Longman.

cognitive prerequisites of learning to read and write and the sociocultural context in which such learning processes take place.

REFERENCES

Barton, D. (2001). Literacy in everyday contexts. In L. Verhoeven & C. Snow (eds), *Literacy and motivation. Reading engagement in individuals and groups*. Mahwah, NJ: Erlbaum.

Butler, D. L. & Winne, P. H. (1995). Feedback and self-regulated learning: a theoretical analysis. *Review of Educational Research* **65**, 245–81.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.

Durgunoglu, A. Y. & Verhoeven, L. (eds) (1998). *Literacy development in a multilingual context : cross-cultural perspectives*. Mahwah, NJ: Erlbaum.

Elley, W. B. (2001). Literacy in the present world: realities and possibilities. In L. Verhoeven & C. Snow (eds), *Literacy and motivation. Reading engagement in individuals and groups*. Mahwah, NJ: Erlbaum.

Gee, J. (1990). *Social linguistics and literacies : ideology in discourse*. London: Falmer Press.

Geva, E. & Verhoeven, L. (2000). Basic processes in early second language reading. *Scientific Studies of Reading* **4**, 261–353. (Special issue).

Goldman, S. R. & Rakestraw, J. A. (2000). Structural aspects of constructing meaning from text. In M. L. Kamil Mosenthal, P. B. Pearson, P. D. & R. Barr (eds.), *Handbook of reading research*, Vol III. Mahwah, NJ: Erlbaum.

Hymes, D. (1971). On communicative competence. Philadelphia: Philadelphia University Press.

Levy, C. & Ransdell, S. (1996). *The science of writing : methods, individual differences and applications*. Hillsdale, NJ: Erlbaum.

National Reading Panel (2000). *Teaching children to read*. Washington, DC: The National Institute of Child Health and Human Development.

Perfetti, C. (1998). Learning to read. In P. Reitsma & L. Verhoeven (eds), *Literacy problems and interventions*. Dordrecht: Kluwer.

Snow, C. E., Burns, M. S. & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Snowling, M. J. (1998). Reading development and its difficulties. *Educational and Child Psychology* **15**, 44–58.

Street, B. V. (1994). Cross-cultural approaches to literacy. In L. Verhoeven (ed.), *Functional literacy. Theoretical issues and educational implications*. Amsterdam/Philadelphia: Benjamins.

Verhoeven, L. (1994). Modeling and promoting functional literacy. In L. Verhoeven (ed.), *Functional literacy. Theoretical issues and educational implications*. Amsterdam/Philadelphia: Benjamins.

Verhoeven, L. (1996). Vernacular literacy in nonmainstream communities. *International Journal of the Sociology of Language* **119**, 1–152. (Special issue).