Disciplinary Corpus Research for Situated Literacy Instruction

Sarah Blazer
Fashion Institute of Technology (SUNY)

Sarah E. DeCapua
University of Connecticut

Abstract: In this article, the authors demonstrate one example of how corpus research can prepare disciplinary outsiders to support faculty and students engaged in graduate-level reading and writing of disciplinary genres. The corpus study answers the question: What are the most common high-frequency phrases that appear in a corpus of public health (PH) research articles, and what do they mean? Because students often struggle as much or more interpreting the phrases that make up the connective tissue of a text—its subtechnical language—as they do with content or specialized vocabulary and phrases, the former are of particular interest in this study. Students pursuing graduate-level disciplinary study need precise understanding of the language their field most frequently uses to express relationships among key terms and concepts. The authors discuss concrete pedagogical applications for their corpus research findings and connect sociocultural theory to corpus linguistics (CL) research and materials development to discuss how the latter can assist in students' mediation and internalization of discipline-specific linguistic and conceptual knowledge.

Keywords: Public Health (PH), Literacy Development, Subtechnical Language, Disciplinary Discourse, Writing Center, WAC, WID, Corpus Linguistics (CL), Concordancing Program, AntConc, Sociocultural Theory (SCT), Mediation, Internalization

Writing Across the Curriculum (WAC) and writing center scholars engaged in supporting student literacy development are often recruited across disciplinary contexts to guide faculty in their development of course materials and curriculum as well as to develop and teach workshops or courses for students. Outsiders to a field may be challenged to determine what disciplinary genres and discourse conventions could be taught to students and how to provide appropriately situated instruction (Curry, 2016), as advocated for by Chris Thaiss and Terry Zawacki (2006), David Russell (1995, 2002), and others. When possible, one might prepare to guide or teach outside of familiar disciplinary territory by following models like Stoller, Jones, Costanza-Robinson, and Robinson (2005); this team of applied linguistics and

DOI: https://doi.org/10.37514/ATD-B.2020.0407.2.13

chemistry faculty at Northern Arizona University employed corpus research in a lengthy project to systematically explore writing in chemistry and redesign curriculum and instructional materials (see, as well, Caplan, this collection; Tribble & Wingate, 2013). However, when time or budget constraints result in the absence of opportunities for such rich, extended collaboration, engaging in smaller-scale research can support investigation of unfamiliar genres and discourses. Digital corpus research[1] is an accessible, flexible way for writing center and WAC/WID faculty to generate knowledge and teaching materials in support of discipline-specific literacy instruction for the purposes of course or workshop design.

In this chapter, we provide an overview of corpus research in literacy teaching and situate corpus-informed teaching as compatible with sociocultural theories (SCT) of language learning. We find SCT provides a useful framework for understanding how corpus research helps us facilitate graduate students' development of insider discourse knowledge in the discipline of public health. We then demonstrate how one small corpus study helped Sarah Blazer develop discourse and genre knowledge, as well as inquiry-based exercises relevant to the needs of graduate students enrolled in a Public Health graduate reading and writing course at Lehman College, The City University of New York.[2] For the Masters in Public Health (MPH) program at Lehman, we focused our corpus study and materials development on research articles representative of those that students in this program were expected to gain facility reading and producing. Within a corpus of research articles from the *American Journal of Public Health*, we focused on subtechnical language, a feature of disciplinary discourse characterized by abstract, low-imagery vocabulary and phraseology used to create logical connections among concepts (Baker, 1988; Heltai, 1996).

Corpus Research for Teachers of Reading and Writing

Discourse knowledge is developed in part through participation in a community, but individuals learning in an academic setting also benefit from explicit literacy instruction situated within disciplinary and genre contexts (Aull, 2015; Curry, 2016; LaFrance & Corbett, this collection; Samraj, 2002; Swales & Feak, 2012; Tribble & Wingate, 2013). Corpus research can and does inform situated literacy instruction; for example, researchers may use an existing corpus like the Corpus of Contemporary American English (Davies, 2012) to investigate language patterns, or they may

[1] A corpus is a principled collection or database of texts compiled from naturally occurring examples of written language or language transcribed from recorded speech. Electronic corpora can be studied quantitatively and qualitatively using corpus software (Hunston, 2002).

[2] Lehman College's MPH program has since merged with other CUNY Public Health programs and is now housed at another site.

create a corpus to target a particular register. The latter may be more appropriate for classroom teaching (Krieger, 2003), as was the case in our study designed to develop instruction and materials for a graduate course in public health (PH).

Digital corpus research tools allow us to study corpora of any size with precision and efficiency. Through a concordance program like the one we used (AntConc; Anthony, 2011), researchers can produce and analyze data in various ways. For example, as we will later illustrate, one might first search a corpus for its most frequent words or phrases with various parameters for length. From the list produced, a word or phrase can be selected and viewed in concordance lines; the number of surrounding words is set by the search to show each instance throughout the corpus with the degree of context needed. It is also possible to view the word or phrase within the entire original text to provide maximum context. Analysis of concordance lines is a basic and pragmatic approach to processing corpus data when the goal is to inform day-to-day teaching (Hunston, 2002).

Daniel Krieger (2003) summarized corpus-derived language investigation for teaching purposes: researchers can look at many language patterns from morphological to lexical to discourse, and they do so with varying agendas. For example, Mona Baker (1988) discussed teaching applications for corpus-derived collocations with a focus on collocations that function rhetorically in a particular genre. From her corpus of medical journal articles, she found that in Discussion sections, "findings" is frequently preceded by "our" and followed by language like "extend" and "raise a question" to convey authors' evaluative commentary. Baker suggested that learners be made aware of frequent collocations and the genre and sections of text in which such phrases frequently appear to help learners gain facility with "whole stretches of language" (p. 104), as opposed to individual words.

From their study of a large corpus of chemistry research articles, Stoller et al. (2005) created a guide to passive voice, past participle verbs frequently seen in Methods sections. Such a list provided advanced chemistry students with access to discipline- and genre-conventional options for varying their use of verbs. This explicit list—including, among others, "was assigned," "was performed," "was filtered," "was washed" (Appendix B)—may expose students to vocabulary they have not yet learned, as well as invite meaning-making questions from curious novices about why particular phrases are so frequent in a particular section of chemistry research articles. Approaching materials development from a different direction, Christopher Tribble and Ursula Wingate (2013) argue that corpora of student writing may be optimal for the development of pedagogical materials that allow students to gain facility understanding and controlling target genres.

In *Academic Writing for Graduate Students*, John Swales and Christine Feak (2012) suggested that students explore academic phraseology of interest by performing basic internet searches and employing digital corpus tools, including AntConc (p. 28-29). Krieger (2003) also acknowledged the potential for student corpus research but argued that corpus research may be most useful for materials development, as teachers can "harness a corpus by filtering the data for students" ("Exploiting a Corpus," para. 1) to focus students' attention exclusively on understanding patterns of language use, as opposed to the process of actually locating relevant patterns.

Our study exhibits how a concordance program can prepare faculty—disciplinary outsiders or insiders—to develop instruction and materials by analyzing carefully chosen corpora to identify highly frequent disciplinary genre and discourse conventions, vocabulary, and phraseology (Hunston, 2002; Hyland & Tse, 2007; Stoller et al., 2005). Indeed, corpus research allows one to see patterns even members of the disciplinary community may not intuit (Liu, 2003; Stoller et al., 2005). Thus, with preparation through corpus research, those recruited to teach from outside a discipline may be in a uniquely useful position to guide graduate students working to develop more insider perspectives on their discipline.

Sociocultural Theory and Disciplinary Discourse Teaching and Learning

Literacy and writing studies faculty in disciplinary outsider positions can use corpus research to prepare situated literacy instruction that facilitates students' social acculturation toward more insider status. As disciplinary discourse is a complex, evolving social construction, sociocultural theory (SCT) helps us understand how students, regardless of language background (Curry, 2016), learn and internalize the discourse of an academic discipline and subsequently affect it, too. In SCT, meaning is located in the dialogue between human beings engaged in goal-directed behavior, not only in the signs or language itself (Lantolf & Thorne, 2006). For this reason, teaching highly frequent subtechnical phrases within a specific discipline should be based on a corpus of texts that reflects current and situated goal-directed behaviors so that findings are relevant to students studying a particular body of research, and relevant to their entrance and acclimation to the discipline or field.

Following Pál Heltai (1996) and Baker (1988), we characterize subtechnical language as abstract, low-imagery vocabulary and phraseology that is frequently used across academic discourses. Subtechnical language—*take effect* and *with respect to*, for example—may prove particularly difficult to understand and employ with precision because it is difficult to visualize and even define in some cases (Heltai, 1996). And though subtechnical language can be found across disciplines, which increases the chances students have encountered it, it may also function in unique ways depending on the disciplinary context (Casanave, 2008; Hyland & Tse, 2007).

The focus of discourse instruction, then, should be on patterns of meaning and "meaning potential" of phrases (Lantolf & Thorne, 2006, p. 9) within a corpus of relevant texts, not on a single definition divorced from the meaning-making context. Ken Hyland and Polly Tse's (2007) recent corpus research findings support teaching subtechnical language. From various widely used genres across disciplines, their analysis revealed that highly frequent items "are not used in the same way and do not mean exactly the same thing in different disciplinary contexts" (p. 249), thus challenging the notion that a general academic vocabulary exists across disciplinary environments and can or should be taught as such. For example, the word *expression* often characterizes emotional and/or verbal behavior, but in the phrase *gene expression*, the word characterizes a physical manifestation (Baker, 1988). Further, Casanave (2008) makes the noteworthy point that first language speakers may struggle more with "common words used in specialized ways" than second language speakers, "given the persistent connections [they make] of individual common words with their everyday connotations" (p. 20).

Highly Frequent Subtechnical Phrases as Scientific Concept

Unlike other theories concerned with the social context of learning, SCT is concerned with the psychological changes individuals undergo in the process of learning and *internalizing* what are known as *scientific* or *non-spontaneous concepts* through culturally constructed *mediating tools and artifacts* or "symbolic, communicative, and material resources" (Lantolf & Thorne, 2006, p. 233). Lev Vygotsky (1934/1986) made a key distinction between *spontaneous concepts* and *scientific concepts*. The former concepts are the product of the individual's everyday experiences; their development "know[s] no systematicity and goes from the phenomena upward toward generalizations" (p. 148). In other words, this is conceptual knowledge we develop—often unconsciously—by virtue of our life experiences or participation in a community of people who share certain goals. By contrast, scientific concepts are propositional, codified, documented knowledge that is "publicly accepted as a principled way of understanding phenomena within a particular discourse community" (Johnson, 2009, p. 15). We generally acquire scientific concepts through more explicit or purposeful instruction.

Corpus-derived, highly frequent subtechnical phrases can be characterized as scientific concepts if they have not been acquired unconsciously or through ongoing experiences. Such phrases in a particular discourse community can be understood as a conceptual group, as types of lexical units that students can be aware of as they build knowledge of the research in their field. Highly frequent subtechnical phrases that are introduced as the focus of instruction—and understood by students to be relevant to their own goals—are "conscious (and consciously applied)" (Swain, Kinnear, & Steinman, 2011, p. 52). They are "systematic" and "not bound to context" (p. 52) in that they are used in closely related ways throughout a corpus of disciplinary texts. These uses are accepted by the discourse community as demonstrated through their patterned deployment. They are contextualized within the disciplinary knowledge and thus carry meaning for the discipline. We agree with Hyland and Tse (2007): "we need to identify students' target language needs as well as we can" and address them by "introducing, making salient, and practicing the specialized vocabulary of their fields or disciplines" (p. 249). Thus, highly frequent phrases like those identified in our corpus of PH research articles must be taught if not already known, and they must be understood by participants in this discourse community (Swales, 1990) if meaningful knowledge construction is the shared goal of students and faculty.

Understanding the meaning and sense—including the disciplinary function—of highly frequent disciplinary collocations provides learners with more conceptual understanding so they can "function appropriately in the range of settings in which they may find themselves" (Johnson, 2009, p. 14). As students develop deeper conceptual understanding of the highly frequent subtechnical language that allows researchers and scholars in their field to express relationships between and among complex ideas and factors, they "reframe the way they describe and interpret" (Johnson, 2009, p. 15) their experiences and knowledge: learners can apply a greater understanding of these patterned features of discourse in their field to engaging in more efficient and/or systematic approaches to reading and producing disciplinary work. Instruction can support learners as they begin to apply new concepts to concrete activity, thereby merging their conceptual and everyday knowledge (Johnson, 2009). As they internalize new concepts, they develop tools for building knowledge of research in their field. Concepts can be accessed consciously, metacognitively, until understanding becomes fully internalized.

Internalization Through Mediation

Internalization, defined from an SCT perspective, is "the internal reconstruction of an external operation" (Vygotsky, 1978, p. 56). That is, what is first learned through social interaction and is thus interpsychological next appears intrapsychologically (Vygotsky, 1978). Using internalized concepts, individuals engage in a continual process of reexternalizing internalized concepts such that those individuals have not only been psychologically constructed by culture but also contribute to its construction. As Karen Johnson (2009) describes, internalization occurs through activity that is "initially mediated by other people or cultural artifacts but later comes to be controlled by [the individual] as he or she appropriates and reconstructs resources to regulate his or her own activities" (p. 18).[3]

It is useful, then, to consider the concepts one aims to teach, as well as materials

[3] In SCT, a concept which is initially taught and internalized by way of mediating artifacts ultimately becomes a mediating artifact itself.

and methods, in terms of *mediating* tools required to facilitate internalization. To teach corpus-derived, highly frequent subtechnical phrases, one can use mediating tools like corpus research results, concordance lines, and problem-posing activities to facilitate learning and internalization, where the latter is understood not as simple appropriation of concepts or knowledge, but as a dialogic process whereby individuals engage in activity that leads to "transformation of self *and* activity" (Johnson, 2009, p. 18). The agentive individual influences his or her internalization process and how it contributes to further growth and action (Johnson, 2009). Thus, internalization is understood as a dynamic, bi-directional process which is not about simply appropriating a copy of the external concept learned. Rather, internalization is about "making something one's own" (Lantolf & Thorne, 2006, p. 162), which can then be reexternalized to contribute to further cultural meaning-making. As students develop understanding of new concepts, their existing conceptual knowledge should serve to mediate their development. In learning highly frequent subtechnical phrases from a discipline-specific corpus, students can and should connect new concepts to existing knowledge.

SCT guides both the rationale for using corpus research to inform our teaching of disciplinary discourse and the subsequent development of corpus-based pedagogical applications. Next, we describe the methods and results of the corpus study we engaged in to inform instruction in a graduate reading and writing course for MPH students at Lehman College.

Context for a Public Health Corpus Study

For several years, including when this study was conducted,[4] Lehman College's Masters in Public Health (MPH) program offered a two-credit elective course in which students focused exclusively on developing their reading proficiency with disciplinary texts and their writing of discipline-situated summary, paraphrase, and analysis. Students in this program presented a range of literacy strengths and needs in terms of their experience with the language of their discipline, language characteristic of American academic English, and English for general communicative purposes. The goals and tight focus of the course meant that students benefited from our reading and writing discussions and exercises, regardless of their language backgrounds, a factor that might have been more significant in another context.

Many students in Lehman's program held field positions and had returned to school to improve their chances for career advancement; their strengths often lay in the practical knowledge they brought to their studies rather than in their command

[4] Lehman College's MPH program has since merged with other CUNY public health programs and is now housed at another site.

of more formal academic literacies. They were professional insiders, knowledgeable from work experiences about key public health (PH) issues like diabetes, and generally able to use ubiquitous terminology like "*socioeconomic risk factors*" with ease. But back in school, they often struggled to interpret and command phrases comprising the connective tissue of their discipline's texts, what we address in our study as subtechnical language. For example, students recognized phrases including words like "*incidence*" and "*probability*" but often struggled to fully comprehend these phrases in context and struggled to use them confidently and precisely in discussion and writing.

While the majority of students in this MPH program did not go on to pursue careers in academic research, as graduate students, they were expected to engage with the field's research and scholarship on an advanced level. Outside of school, they would benefit from being able to more effectively read and draw on the research that could more fully inform their daily work. Regardless of professional goals, students pursuing advanced disciplinary study need precise understanding of the fundamental language their field most frequently uses to express relationships among key terms and concepts. For these reasons, the highly frequent subtechnical language became a particularly important aspect of PH texts to investigate, and it became clear that corpus research could help answer a question relevant to teaching students to own more of the fundamental language of their discipline: What are the most high-frequency subtechnical phrases that appear in a corpus of PH research articles, and what do they mean?

From hundreds of three- to five-word phrases ranked by frequency, several of the most frequently occurring phrases became the focus of inquiry for classroom application: *more likely to*, *was associated with*, and *the effect of.* These are three of a number of phrases we might have chosen to focus on; they are not the only three from our corpus worthy of consideration for classroom teaching. At first glance, phrases like *more likely to* may seem trivial; however, from usage patterns, it becomes evident that such phrases are subtle but important carriers of the discipline's ways of thinking and knowing. According to Hyland and Tse (2007), "all academic representations shape and manipulate language for disciplinary purposes, often refashioning everyday terms so that words take on more specific meanings" (p. 247). It is difficult to know whether highly frequent language in a given discipline is or is not used in unique ways without investigating it and comparing it to others. So, while *more likely to* may express the same type of relationship in PH as it does in sociology, its highly frequent use in PH is significant in and of itself because it exemplifies the complexity of phenomena being studied in PH. Since there are always numerous factors influencing any given phenomena of interest in PH, causation is virtually impossible to establish among the range of issues PH researchers address.

Results of corpus studies like the one described here can inform what key language and rhetorical moves we guide students to pay attention to in the texts they engage with as students and emerging professionals in their fields, both as critical readers and producers of text. This study contributed to the development of more disciplinarily situated instruction for students in the MPH program at Lehman College, including exercises that engaged students in interactive inquiry into disciplinary practices.

Methods and Results

Here, we describe the methodology and results of our small PH corpus study, followed by a discussion of socioculturally situated pedagogical applications of our corpus findings for students in the program that sparked the inquiry. The methodology provides a set of guidelines for outsiders or insiders engaged in corpus research for the development of disciplinary discourse knowledge and materials. The results of this study may also be relevant to others teaching PH students since the subtechnical language we focus on is derived from *The American Journal of Public Health*, a prominent publication in the field of public health.

Using AntConc (Anthony, 2011), free digital concordance software, our first research question can be answered with ease: What are the most high-frequency subtechnical phrases that appear in a corpus of public health texts? By carefully analyzing concordance lines in which the phrases appear, we can answer our second question: What do these phrases mean? Below, we outline our methods for developing the corpus, setting search parameters, choosing subtechnical phrases to focus on, and determining the meaning of those phrases. We combine the methods and results here because our methods for choosing and determining meaning of the subtechnical phrases are best understood alongside the results of our searches.

Corpus

Our selection of corpus texts was entirely pragmatic and specific to the goal of supporting students in a particular MPH reading and writing course at Lehman College. For several reasons, *The American Journal of Public Health* was an appropriate source from which to draw corpus texts: faculty in this graduate program drew on it frequently; the journal includes articles on a wide range of PH topics, as opposed to focusing on a particular area (e.g., *Journal of Health Politics, Policy, and Law*); and the journal includes traditional empirical research articles which faculty in this MPH program identified as very challenging for many students.[5]

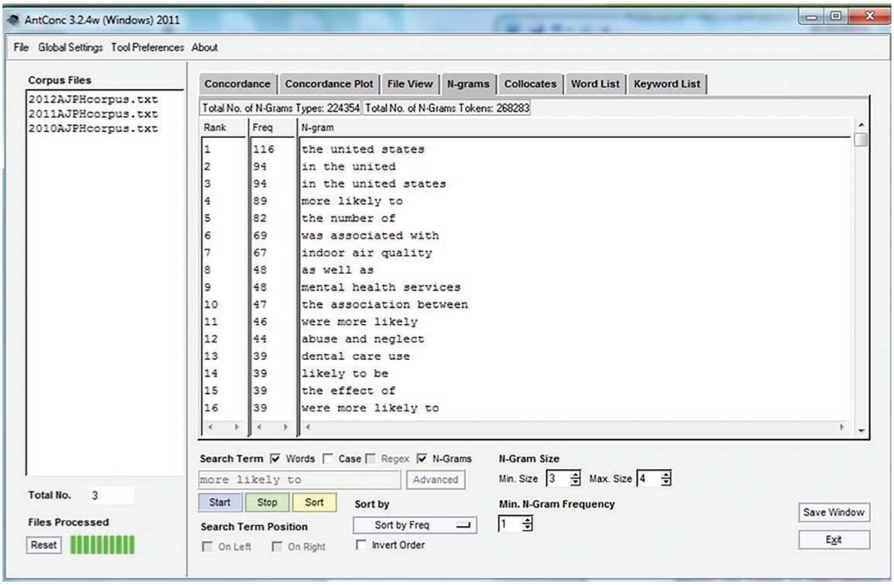
From the "Research and Practice" section of *The American Journal of Public*

[5] Other sections of *The American Journal of Public Health* include PH scholarship; a study of these articles may yield different results in terms of most frequent phraseology.

Health, we selected a total of 36 issues and 108 articles.[6] Charts, tables, and bibliographic entries were not included in our corpus. Three separate files were created in order to run separate analyses on each year of the corpus, but we were able to run the analysis on all three files at once to see the most frequent phrases across the whole corpus and still see clearly the frequency and location of phrases by year of publication.

Concordance Lines

To determine which phrases appeared most frequently in our corpus, we used the AntConc Clusters tool and set the "n-gram" parameters for frequently occurring phrases at a minimum of three words and a maximum of four words (see Figure 13.1); it is simple and useful to play with these parameters in the context of materials development.



Corpus Files
2012AJPHcorpus.txt
2011AJPHcorpus.txt
2010AJPHcorpus.txt

Total No. 3

Files Processed

Total No. of N-Grams Types: 224354 Total No. of N-Grams Tokens: 268283

Rank	Freq	N-gram
1	116	the united states
2	94	in the united
3	94	in the united states
4	89	more likely to
5	82	the number of
6	69	was associated with
7	67	indoor air quality
8	48	as well as
9	48	mental health services
10	47	the association between
11	46	were more likely
12	44	abuse and neglect
13	39	dental care use
14	39	likely to be
15	39	the effect of
16	39	were more likely to

Search Term ☑ Words ☐ Case ☐ Regex ☑ N-Grams
more likely to
Search Term Position ☐ On Left ☐ On Right
Sort by: Sort by Freq ☐ Invert Order
N-Gram Size: Min. Size 3 Max. Size 4
Min. N-Gram Frequency 1

Figure 13.1. N-gram data.

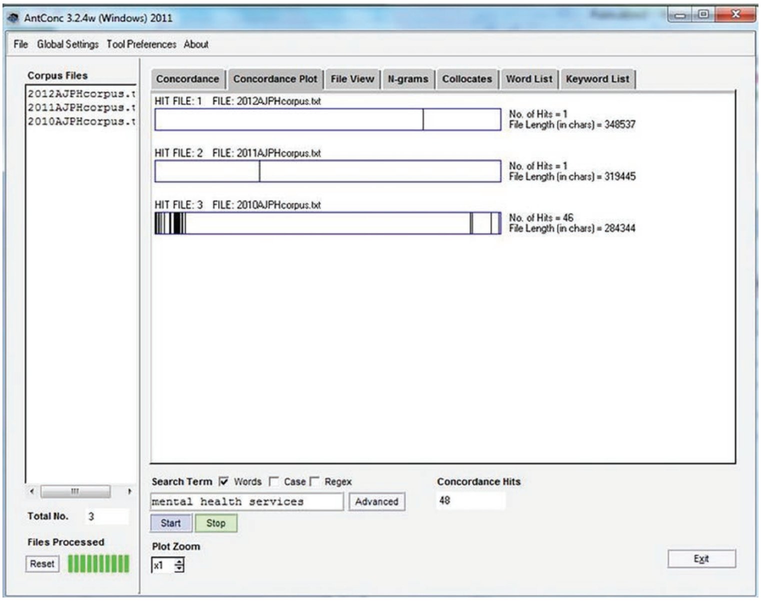
After reviewing results of the query for most frequently occurring three- to four-word phrases, we discussed whether each expressed content or functioned as subtechnical language (see Table 13.1 for a list of the first 15 phrases) and reviewed concordance plots to verify that subtechnical phrases appeared throughout and content phrases appeared in concentrated parts of the corpus (see Figure 13.2). For example, the phrase *more likely to* functions as subtechnical language. It provides a means to express probability, which is of primary concern in issues of PH, and it appeared with

[6] This study was conducted in 2013.

great frequency throughout the corpus, regardless of the focus of a given article. The concrete phrase *mental health services* expresses content, and AntConc's concordance plot tool confirmed that the phrase indeed appeared with great frequency in only one part of the corpus, the focus of one of the 108 articles (see Figure 13.2).

Table 13.1. Fifteen most frequent three- and four-word phrases;

Three- to Four-Word Phrases	Frequency	Content or Subtechnical Language Classification
the united states	116	Content
in the united	91	Content
in the united states	91	Content
more likely to	89	Subtechnical
the number of	74	Subtechnical
was associated with	69	Subtechnical
indoor air quality	67	Content
as well as	48	Subtechnical
mental health services	48	Content
the association between	47	Subtechnical
were more likely	46	Subtechnical
abuse and neglect	44	Content
dental care use	39	Content
likely to be	39	Subtechnical
the effect of	39	Subtechnical



Corpus Files
2012AJPHcorpus.t
2011AJPHcorpus.t
2010AJPHcorpus.t

HIT FILE: 1 FILE: 2012AJPHcorpus.txt
No. of Hits = 1
File Length (in chars) = 348537

HIT FILE: 2 FILE: 2011AJPHcorpus.txt
No. of Hits = 1
File Length (in chars) = 319445

HIT FILE: 3 FILE: 2010AJPHcorpus.txt
No. of Hits = 46
File Length (in chars) = 284344

Search Term ☑ Words ☐ Case ☐ Regex
mental health services

Concordance Hits 48

Total No. 3

Files Processed

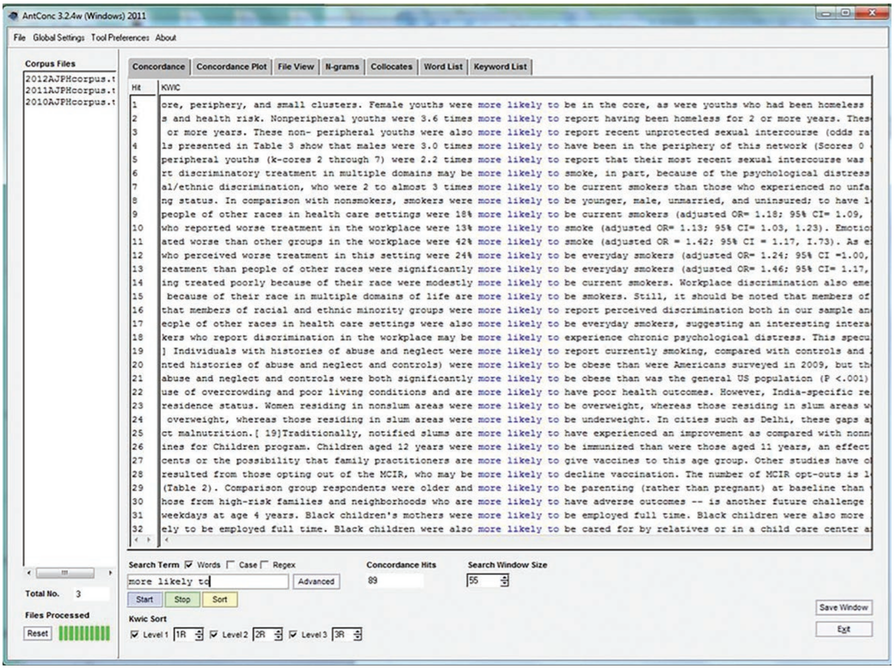
Plot Zoom x1

Figure 13.2. Concordance plot for mental health services.

After determining through our discussions whether the top 75 most frequently occurring three- to four-word phrases expressed content or functioned as subtechnical language, we isolated the subtechnical phrases for further consideration (see Table 13.2), since only phrases classified as subtechnical were relevant to our study. Initially, we determined the meaning of these phrases intuitively; then we studied the concordance lines for each phrase (see Figure 13.3 for a view of *more likely to* concordance lines). For example, while our classification of *more likely to* as an expression of probability did not change after studying the concordance lines, our classification of *on the basis of* did change; we initially guessed that *on the basis of* signals the use of empirical evidence to make a claim, but learned from the concordance lines that the phrase is used more generally to indicate any condition(s) underlying a claim, theory, decision, or action.

Table 13.2. Twenty-five most frequent subtechnical phrases

Three- to Four-Word Phrases	**Frequency**	**Meaning**
more likely to	89	Probability
the number of	82	Quantification
was associated with	69	Relationship between and among factors and outcomes
as well as	48	Expresses concurrence, modification
the association between	47	Relationship between a possible contributing factor and outcome
were more likely	46	Probability
likely to be	39	Probability
the effect of	39	Causality
were more likely to	39	Probability
included in the	38	Containment of something as part of a whole
in the past	35	Temporal
less likely to	35	Probability
we found that	34	Research process
the basis of/on the basis/on the basis of	31	Indicates an underlying condition
more likely to be	27	Probability
significantly associated with	27	Relationship between a possible contributing factor and outcome(s)
the odds of	27	Probability
we did not	27	Research process
with respect to	27	Referential
because of the	25	Causality
we controlled for	25	Research process
data from the	24	Referential
the prevalence of	24	Quantification



Corpus Files
2012AJPHcorpus.t
2011AJPHcorpus.t
2010AJPHcorpus.t

```
Hit KWIC
1   ore, periphery, and small clusters. Female youths were more likely to be in the core, as were youths who had been homeless
2   s and health risk. Nonperipheral youths were 3.6 times more likely to report having been homeless for 2 or more years. Thes
3    or more years. These non- peripheral youths were also more likely to report recent unprotected sexual intercourse (odds ra
4   ls presented in Table 3 show that males were 3.0 times more likely to have been in the periphery of this network (Scores 0
5   peripheral youths (k-cores 2 through 7) were 2.2 times more likely to report that their most recent sexual intercourse was
6   rt discriminatory treatment in multiple domains may be more likely to smoke, in part, because of the psychological distress
7   al/ethnic discrimination, who were 2 to almost 3 times more likely to be current smokers than those who experienced no unfa
8   ng status. In comparison with nonsmokers, smokers were more likely to be younger, male, unmarried, and uninsured; to have l
9   people of other races in health care settings were 18% more likely to be current smokers (adjusted OR= 1.18; 95% CI= 1.09,
10  who reported worse treatment in the workplace were 13% more likely to smoke (adjusted OR= 1.13; 95% CI= 1.03, 1.23). Emotio
11  ated worse than other groups in the workplace were 42% more likely to smoke (adjusted OR = 1.42; 95% CI = 1.17, I.73). As e
12  who perceived worse treatment in this setting were 24% more likely to be everyday smokers (adjusted OR= 1.24; 95% CI =1.00,
13  reatment than people of other races were significantly more likely to be everyday smokers (adjusted OR= 1.46; 95% CI= 1.17,
14  ing treated poorly because of their race were modestly more likely to be current smokers. Workplace discrimination also eme
15   because of their race in multiple domains of life are more likely to be smokers. Still, it should be noted that members of
16  that members of racial and ethnic minority groups were more likely to report perceived discrimination both in our sample an
17  eople of other races in health care settings were also more likely to be everyday smokers, suggesting an interesting intera
18  kers who report discrimination in the workplace may be more likely to experience chronic psychological distress. This specu
19  ] Individuals with histories of abuse and neglect were more likely to report currently smoking, compared with controls and
20  nted histories of abuse and neglect and controls) were more likely to be obese than were Americans surveyed in 2009, but th
21  abuse and neglect and controls were both significantly more likely to be obese than was the general US population (P <.001)
22  use of overcrowding and poor living conditions and are more likely to have poor health outcomes. However, India-specific re
23  residence status. Women residing in nonslum areas were more likely to be overweight, whereas those residing in slum areas w
24   overweight, whereas those residing in slum areas were more likely to be underweight. In cities such as Delhi, these gaps a
25  ct malnutrition.[ 19]Traditionally, notified slums are more likely to have experienced an improvement as compared with nonn
26  ines for Children program. Children aged 12 years were more likely to be immunized than were those aged 11 years, an effect
27  cents or the possibility that family practitioners are more likely to give vaccines to this age group. Other studies have o
28  resulted from those opting out of the MCIR, who may be more likely to decline vaccination. The number of MCIR opt-outs is l
29  (Table 2). Comparison group respondents were older and more likely to be parenting (rather than pregnant) at baseline than
30  hose from high-risk families and neighborhoods who are more likely to have adverse outcomes -- is another future challenge
31  weekdays at age 4 years. Black children's mothers were more likely to be employed full time. Black children were also more
32  ely to be employed full time. Black children were also more likely to be cared for by relatives or in a child care center a
```

Search Term ☑ Words □ Case □ Regex
more likely to
Concordance Hits 89
Search Window Size 55
Total No. 3
Files Processed
Kwic Sort ☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

Figure 13.3. Concordances for more likely to.

Finally, we determined which subtechnical phrases were most interesting to us for the purposes of our development of course materials and instruction. Of the phrases qualifying as subtechnical, we chose to investigate three phrases appearing more than 30 times in the corpus: *more likely to*, *was associated with*, and *the effect of*.

Discussion

We chose to focus our development of pedagogical applications on frequently appearing abstract phrases expressing probability, causality, and relationships between factors and outcomes because these phrases are difficult to explain and understand outside of authentic contexts, and they carry important meaning in the field of public health. Students need to gain facility reading and using these phrases to succeed in their PH coursework, as well as in many PH professional environments. First, phrases including the word *likely* were of interest to us as they express probability and appeared with great frequency and in various collocations in this corpus: *more likely to*, *were more likely*, *likely to be*, *were more likely to*, *less likely to*, and *more likely to be*. The phrases *was associated with* and *the association between* were of interest as these phrases express relationships between and among factors and outcomes that are important to PH researchers. We chose to investigate *the effect of* because we wondered whether concordance lines would reveal that this phrase indicated a greater articulation of certainty than the "associated" and "likely" phrases. The phrases we have selected are by no means the only phrases worthy of close attention. The applications we describe are simply three illustrations of what is possible.

Throughout this corpus of disciplinary texts, the phrases we selected are used in a finite number of closely related ways. Lehman's MPH students needed to understand and gain facility with subtechnical phrases such as these, as well as the more specialized terminology of PH research, in order to contribute to meaningful knowledge construction in this discourse community. Throughout their coursework, they were explicitly introduced to and required to use the language of their field, which they did with varying degrees of success. Oftentimes, those who struggled to produce clear and accurate writing did so because they misused subtechnical language used to express clear and logical relationships between concepts and data. In the reading and writing course we have been discussing, students frequently expressed shock at how differently they and their peers could interpret the same sentence written to express fact. They realized on a new level how difficult, but essential, it is to explain concepts and data precisely. Through this course, students in Lehman's MPH program had a rare opportunity to slow down their thinking about reading and writing, and they were eager to practice carefully interpreting and using more effectively the language that surrounds all of the specialized vocabulary and content they had learned. Corpus research can usefully inform instruction in this environment.

The exercises we discuss next do not represent formulas for the authors, nor are they intended as prescriptions for readers. Rather, each exercise demonstrates how a single phrase or family of phrases could be the basis for explicit teaching in an actual class session. In the primary author's experience, exercises and discussion around phrases of interest were embedded organically throughout course meetings.

Pedagogical Application: *more likely to*

Public health research is focused on identifying and understanding trends in health issues, as well as proposing and studying the effects of preventive programming and interventions, and the phrase *more likely to* (as well as *were more likely*, *likely to be*, *less likely to*, etc.) provides a useful and discourse-familiar expression of probability. It is typically used to discuss trends and outcomes alongside numerical data in Results sections and without numerical data in Discussion sections.

A number of questions about precision and discourse conventions could be posed in a class setting such that students develop greater understanding of the meaning and sense of the highly frequent phrase *more likely to*: What does it mean that someone or something is more likely to be or do something in PH research? Does it mean 51 percent more likely to? Eighty-nine percent? How often is the phrase qualified by numerical values? Do usage patterns differ depending on the section of the article (Results or Discussion)? That is, might the phrase appear more frequently without qualifying numerical data in the Introduction or Discussion sections and more frequently with qualifying numerical data in Results sections?

The questions posed above can provide the basis for a useful interactive exercise. In a class setting, the instructor might present to students various tools to mediate instruction: their finding from the corpus analysis—the highly frequent phrase—as well as concordance lines in which the phrase appears (refer again to Figure 13.3), one or more articles from the corpus to show the phrase in larger context, and guiding questions. The instructor might begin by asking students, "What does this phrase mean to you? Do you notice it often?" And could then establish the relevance of this corpus-based lesson by asking, "What could we figure out by looking more closely at the phrase in the context of research articles from a prominent journal in PH?" With the questions that arise in the corpus analysis process and the full text of one or more articles included in the corpus, the students could then discuss or work collaboratively to observe usage patterns in the texts and report back to the group. Guiding questions could include:

- How often is the phrase qualified by numerical values?
- Are there sections of the article where the phrase is more or less likely to be qualified by numerical values?
- What other patterns are noticeable? What could be significant about them?
- What do your findings suggest about the use of data in your field?
- What does this help us understand about reading and writing research in public health?
- Do your findings cause you to think in new ways about how you read research articles in your field?

This exercise and the subsequent discussion engage students in abstract, critical, and analytical thinking about the nature of their field through the context of a specific lexical unit they will encounter often and must use carefully. Students may be interested to see the utility of a phrase like *more likely to*; from its frequency and usage patterns, it is clear that the phrase is indicative of the discipline's use of large data sets as a way of thinking and knowing about—and describing—important health phenomena. It may be difficult to ask students, novices to a discipline or profession, to consider the nature of their field. This exercise engages them in such an inquiry for pragmatic purposes while also supporting increasing awareness and fluency with a frequently used phrase. Indeed, Lehman MPH students were highly engaged in conversation around patterns of language use in their field, including group translation sessions around statements of research findings and group editing sessions around the students' own attempts to use probabilistic language to paraphrase others' research findings.

Sociocultural theory's notion of language as communicative activity positions "language learning as an emergent process [which] focuses more on doing, knowing, and becoming, rather than on the attainment of a steady state understood as a well-defined set of rules, principles, and parameters, etc." (Lantolf & Thorne, 2006, p. 138). As illustrated above in terms of classroom learning, the process of developing disciplinary discourse knowledge should be seen in terms of doing, knowing, and becoming since discourse is an ever-evolving set of socially constructed conventions and patterns created and used to carry out ever-evolving needs and interests of a given research and professional community. Students must come to understand the dialogic nature of this process, as well, and classroom instruction around disciplinary discourse patterns and conventions should engage them on an active, conceptual level if it is meant to facilitate participation and meaning-making from novices.

Pedagogical Application: *was associated with*

With 69 occurrences in the corpus, *was associated with* can frequently be seen conveying important evidence of relationships among factors and outcomes. The phrase is used to express degrees of association, a key objective of PH research, and is frequently followed by adjective-noun combinations communicating statistical possibility—"*increased odds*" or "*reduced probability*"—as well as actual occurrences—"*increased times*" or "*lower numbers*."

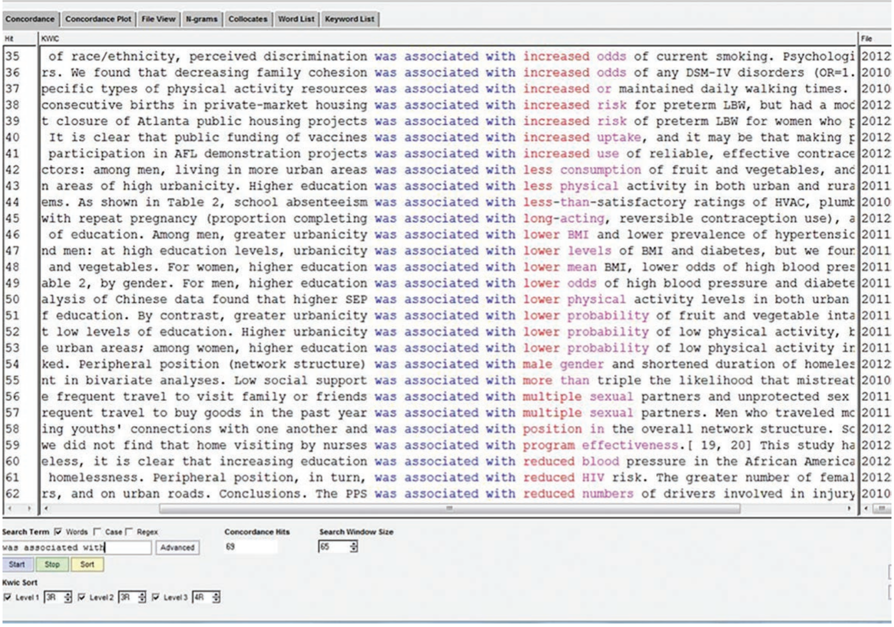
While the phrase carries significant information, students in the MPH course we have been discussing struggled to accurately explain or paraphrase the findings expressed with those phrases, and perhaps more significant, did not realize that they frequently misrepresented findings. One of the course objectives is for students to discuss and write more clearly and accurately about source material. In class discussions, they often attempted to paraphrase information from a research article, and it was not uncommon for others to then disagree and offer counter representations. In these instances, it became clear that many were struggling to understand and/or re-present information from the text.

In some cases, students were unclear of a phrase's meaning; in others, they simply did not yet have access to alternatives that would show they could re-present it in equivalent terms. For students who could benefit from greater understanding of, or exposure to, the phrase *was associated with*, it would be useful to point out its patterns of use and observe them in context. For example, students might be presented with the following list of phrases that follow *was associated with* in the corpus:

was associated with	INCREASED	odds
		times
		probability
		uptake
		use
		risk
	REDUCED	numbers
		risk
		blood pressure
		odds
	BETTER	risk factor profile
	LOWER	odds
		mean BMI
		probability
		activity

One could ask students, then, "What patterns do you see?" And point out, if needed, "There are terms expressing statistical possibility and terms expressing actual occurrences." The differences between the two can be clarified, and then students can look at some of these phrases in context. (See Figure 13.4 for a view of *was associated with* concordance lines.) They can try to explain the statements to one another and determine together—with guidance from the instructor where needed—where they are clear or unclear about the findings expressed. Students in Lehman's MPH program often had work and research experiences they were eager to draw on as they contextualized new concepts; promoting such sharing is particularly useful to encourage as students without these experiences benefit from hearing about those of their peers. If, in this discovery process, students are surprised by their varying interpretations of the same phrase in context, this is in itself a useful realization for them; information of this kind can be communicated back to disciplinary faculty who may consider additional ways to address challenging statistical concepts in their teaching.

Subsequent exercises might engage students in processes of consciously attempting to employ phrases like those discussed here in their own writing, thereby "imitating" (Lantolf & Thorne, 2006, p. 151)—or intentionally modeling—the discipline's use of important phrases in their own written production of meaning. To usefully facilitate students' awareness of and integration into discourse communities, teachers must facilitate carefully mediated inquiry that allows students to draw on their experiences and knowledge as they develop and internalize new linguistic and conceptual knowledge relevant to their professional development.



```
Hit  KWIC                                                                                                                    File
35    of race/ethnicity, perceived discrimination was associated with increased odds of current smoking. Psychologi      2012
36   rs. We found that decreasing family cohesion was associated with increased odds of any DSM-IV disorders (OR=1.       2010
37   pecific types of physical activity resources was associated with increased or maintained daily walking times.       2010
38   consecutive births in private-market housing was associated with increased risk for preterm LBW, but had a mo       2012
39   t closure of Atlanta public housing projects was associated with increased risk of preterm LBW for women who p      2012
40    It is clear that public funding of vaccines was associated with increased uptake, and it may be that making p      2012
41    participation in AFL demonstration projects was associated with increased use of reliable, effective contrace     2012
42   ctors: among men, living in more urban areas was associated with less consumption of fruit and vegetables, and     2011
43   n areas of high urbanicity. Higher education was associated with less physical activity in both urban and rura     2011
44   ems. As shown in Table 2, school absenteeism was associated with less-than-satisfactory ratings of HVAC, plumb     2010
45   with repeat pregnancy (proportion completing was associated with long-acting, reversible contraception use), a     2012
46    of education. Among men, greater urbanicity was associated with lower BMI and lower prevalence of hypertensio     2011
47   nd men: at high education levels, urbanicity was associated with lower levels of BMI and diabetes, but we foun     2011
48    and vegetables. For women, higher education was associated with lower mean BMI, lower odds of high blood pres     2011
49   able 2, by gender. For men, higher education was associated with lower odds of high blood pressure and diabete     2011
50   alysis of Chinese data found that higher SEP was associated with lower physical activity levels in both urban      2011
51   f education. By contrast, greater urbanicity was associated with lower probability of fruit and vegetable inta     2011
52   t low levels of education. Higher urbanicity was associated with lower probability of low physical activity, b     2011
53   e urban areas; among women, higher education was associated with lower probability of low physical activity in     2011
54   ked. Peripheral position (network structure) was associated with male gender and shortened duration of homeles     2012
55   nt in bivariate analyses. Low social support was associated with more than triple the likelihood that mistreat     2010
56   e frequent travel to visit family or friends was associated with multiple sexual partners and unprotected sex      2011
57   requent travel to buy goods in the past year was associated with multiple sexual partners. Men who traveled mo     2011
58   ing youths' connections with one another and was associated with position in the overall network structure. Sc     2012
59   we did not find that home visiting by nurses was associated with program effectiveness.[ 19, 20] This study ha     2012
60   eless, it is clear that increasing education was associated with reduced blood pressure in the African America     2012
61    homelessness. Peripheral position, in turn, was associated with reduced HIV risk. The greater number of femal     2012
62   rs, and on urban roads. Conclusions. The PPS was associated with reduced numbers of drivers involved in injury     2010
```

Search Term ☑ Words □ Case □ Regex: was associated with

Concordance Hits: 69

Search Window Size: 65

Kwic Sort: ☑ Level 1 3R ☑ Level 2 3R ☑ Level 3 4R

Figure 13.4. Concordances for was associated with.

Pedagogical Application: *the effect of*

While *the effect of* is among the most frequently used phrases throughout this corpus of research articles (39 times), it seemed interesting that this phrase expressing causality appeared less frequently than phrases that more generally express relationships—phrases including the word "*association*." A closer look at the concordance lines and full text of the articles revealed that *the effect of* was typically used in the Results and Discussion sections to indicate the impact of one or more factors on a phenomenon of interest. In Results sections, the phrase was used along with indications of statistical significance; in Discussion sections, the phrase was used more generally to refer back to more precise data-based statements found in Results sections.

With this phrase, it would be useful to have students observe the difference in its uses in the Results and Discussion sections; one article in our corpus provides a particularly clear example of the two different uses, as it employs the phrase numerous times in both sections. Students could be given this article and asked to highlight occurrences of the phrase in the two sections and determine how it is used differently. In fact, this pattern held true for all three phrases we investigated. And because students in the MPH program we have been describing explicitly stated that the differences between Results and Discussion sections are not entirely clear to them, an exercise in which they can see clearly how a single phrase functions differently in these two sections may be a helpful step toward clarifying the distinction in concrete terms.

Conclusion

Graduate students benefit from a range of opportunities to orient to the discourse communities they seek to enter, and they often find such opportunities through field experiences and apprenticeships. Complementing these experiences, corpus-informed systematic and explicit disciplinary discourse instruction can help speed up discourse acquisition already in progress (Russell, 1995).

From a sociocultural perspective, the focus of discourse instruction should arise from the needs and interests of students. Our study investigated highly frequent subtechnical language in a corpus of disciplinary texts relevant to Masters in Public Health students at Lehman College. We focused on subtechnical language because students in the program demonstrated a lack of fluency with it, and this lack of fluency impeded their ability to understand and articulate new concepts and to contribute new knowledge. For Lehman MPH students, studying the language, structure, and rhetorical moves conventional in PH research articles in their field provided opportunities to look closely at a genre in ways they had not experienced before. Students benefited from collaboratively considering the meaning of patterns from the rich perspectives of their varied backgrounds as undergraduates—in chemistry, social work, and nursing, to name a few—and as working professionals—in research laboratories, community health, and non-profit advocacy. Through reading and guided discussion, they gained awareness of some of their field's underlying principles and goals, but also its evolving and flexible dimensions (Thaiss & Zawacki, 2006), especially as they compared and raised questions about language, structure, and rhetorical moves in research articles from various journals. A useful follow-up to our study of discourse features and applications for materials development would be a study of ways in which students' reading, writing, and discussion changed as a result of corpus-informed instruction.

Corpus research can be employed on large and small scales to inform literacy instruction that meets immediate interests and needs of students. It can provide background knowledge for composition, WAC/WID, or writing center faculty recruited to teach or support consultants and fellows outside of their disciplines and inform the development of materials to mediate instruction. For example, developing a corpus and conducting a small-scale corpus study like the one we have described could allow a writing center director, writing fellow, and faculty member from another discipline to prepare for classroom collaboration; from working together to build a relevant corpus, to sharing observations and questions regarding results of analyses, all collaborators would have opportunities to gain knowledge, awareness, and ideas for instruction. Corpus research is accessible and can be used to respond to varying agendas for discourse instruction, allowing outsiders and insiders to develop situated literacy instruction sensitive to the dynamic nature of disciplinary discourse.

References

Anthony, L. (2011). AntConc (Version 3.2.4) [Software]. Waseda University. http://www.antlab.sci.waseda.ac.jp/

Aull, L. (2015). Linguistic attention in rhetorical genre studies and first-year writing. *Composition Forum, 31*. https://compositionforum.com/issue/31/linguistic-attention.php

Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language, 4*(2), 91-105. http://www.nflrc.hawaii.edu/rfl/PastIssues/rfl42baker.pdf

Caplan, N. A. (2020). Genres and conflicts in MBA writing assignments. In M. Brooks-Gillies, E. G. Garcia, S. H. Kim, K. Manthey, & T. G. Smith (Eds.), *Graduate writing across the disciplines: Identifying, teaching, and supporting*. The WAC Clearinghouse; University Press of Colorado. https://wac.colostate.edu/books/atd/graduate

Casanave, C. P. (2008). Learning participatory practices in graduate school: Some perspective-taking by a mainstream educator. In C. P. Casanave & X. Li (Eds.), *Learning the literacy practices of graduate school* (pp. 14-31). University of Michigan Press.

Curry, M. J. (2016). More than language: Graduate student writing as "disciplinary becoming." In S. Simpson, N. A. Caplan, M. Cox, & T. Phillips (Eds.), *Supporting graduate student writers: Research, curriculum, and program design* (pp. 78-96). University of Michigan Press.

Davies, M. (2012). *Corpus of contemporary American English.* http://corpus.byu.edu/coca/

Heltai, P. (1996). Teaching abstract subtechnical vocabulary. *Cuadernos de Filología Inglesa, 5(*2), *7*1-82.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary?" *TESOL Quarterly*, *41*(2), 235-253.

Johnson, K. (2009). *Second language teacher education: A sociocultural perspective*. Routledge.

Krieger, D. (2003). Corpus linguistics: What it is and how it can be applied to teaching. *The Internet TESL Journal, 9*(3). iteslj.org/Articles/Krieger-Corpus.html

LaFrance, M., & Corbett, S. J. (2020). Discourse community fail! Negotiating choices in success/failure and graduate-level writing development. In M. Brooks-Gillies, E.

G. Garcia, S. H. Kim, K. Manthey, & T. G. Smith (Eds.), *Graduate writing across the disciplines: Identifying, teaching, and supporting*. The WAC Clearinghouse; University Press of Colorado. https://wac.colostate.edu/books/atd/graduate

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford University Press.

Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly*, *37*(4), 671-700.

Russell, D. R. (1995). Activity theory and its implications for writing instruction. In J. Petraglia (Ed.), *Reconceiving writing, rethinking writing instruction* (pp. 51-77). Lawrence Erlbaum Associates.

Russell, D. R. (2002). *Writing in the academic disciplines, 1870-1990: A curricular history* (2nd ed.). Southern Illinois University Press.

Samraj, B. (2002). Introductions in research articles: Variations across disciplines. *English for Specific Purposes, 21*(1), 1-17.

Stoller, F. L., Jones, J. K., Costanza-Robinson, M. S., & Robinson, M. S. (2005). Demystifying disciplinary writing: A case study in the writing of chemistry. *Across the Disciplines*, *2.* https://wac.colostate.edu/docs/atd/lds/stoller.pdf

Swain, M., Kinnear, P., & Steinman, L. (2011). *Sociocultural theory in language education: An introduction through narratives.* Multilingual Matters.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J., & Feak, C. (2012). *Academic writing for graduate students: Essential tasks and skills* (3rd ed.). University of Michigan Press.

Thaiss, C., & Zawacki, T. M. (2006). *Engaged writers and dynamic disciplines: Research on the academic writing life.* Boynton/Cook.

Tribble, C., & Wingate, U. (2013). From text to corpus—a genre-based approach to academic literacy instruction. *System, 41,* 307-321.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. M. Cole, V. John-Steiner, S. Scribner & E. Souberman (Eds.). Harvard University Press.

Vygotsky, L. (1986). *Thought and language*. (A Kozulin, Trans.). MIT Press. (Original work published 1934)