What Can the History of the English Language Research Offer? A Diachronic Corpus-Based Approach to Research in Writing Studies[1]

Wen Xin, University of Kansas

Abstract: Recent scholarship has called for deeper communications and collaborations between writing studies and language studies because such interdisciplinary connections can facilitate the growth of both fields. While research has explored the potential exchanges between writing studies and language-related fields (such as applied linguistics, second language writing, and sociolinguistics), few studies have focused on the intersections between English historical linguistics and writing studies. This article partially fills the gap by demonstrating how a diachronic corpus-based approach from history of the English language research can be an effective methodological tool for historical research in writing studies, which has been of great interest to writing studies scholars. This article first offers three reasons why a diachronic corpus-based approach can methodologically contribute to historical research in writing studies, and then it illustrates the practice of such an approach by studying the changing disciplinary trends in writing studies from 1997 to 2022 in a self-built, diachronic corpus. To identify the disciplinary trends, word frequency lists and keyness analysis (TF-IDF as procedure) were used. The results indicate both perennial interests and five-year, periodical activities in writing studies over the past 26 years.

Introduction

The relationship writing studies has with language studies has been complex and sometimes incongruous.[2] According to MacDonald (2007), the focus on language had been strong in writing studies until the mid-1970s, during which it was not uncommon to see scholarship from language studies appearing in writing studies research; it was also not uncommon to encounter educators as well as researchers in writing studies with a background in language studies. Such attention to language, as MacDonald charts, slumped over the next thirty years (p. 589), and "the erasure of language" unsurprisingly also resulted in dismissals of language scholarship and training for scholars of writing studies. Interestingly, the decline of attention has rebounded in writing studies, and the attention perhaps has even reached its peak in the past decade or so, which can be seen in the enormous number of studies in the past ten years or so where language-related topics are being discussed. Indeed, "the entire field as a whole seems to be moving toward a better understanding of how language and writing intersect…" (Matsuda, 2013, p. 131). While the engagement with language is evident, Brewer and di Gennaro (2022) also find that many graduate instructors from a writing studies program are provided with insufficient language instruction, and the field in general does not

Across the Disciplines
A Journal of Language, Learning and Academic Writing
10.37514/ATD-J.2024.21.2-3.09

wac.colostate.edu/atd
ISSN 554-8244

seem to have consistent connections to language research outside the field, such as applied linguistics and TESOL.

Recent scholarship in writing studies therefore has called for more training in language and systematic attention to language scholarship from other language-related fields because doing so, among other benefits, sets the foundation to cultivate critical language awareness in students (Gere et al., 2021; Shapiro, 2022), enables students to understand how language works from both social and linguistic perspectives (Aull, 2023), and helps teachers more effectively respond to students' language-related questions in first-year writing classrooms (Ferris & Eckstein, 2020). In fact, not only can more connections to language research better inform language-related teaching practices and scholarly discussions in writing studies, but deep exchanges between both fields, as Donahue (2018) suggests, can "lead to new disciplinary partnerships or at least to mutually respectful growth" (p.132). Although the collaborations between writing studies and language studies seem to be still developing, existing research has unveiled the values of learning from each other.

For example, integrating frameworks from English for academic purposes (EAP) genre studies into rhetorical genre studies (RGS), Aull (2015) provides a more nuanced understanding of how first-year students write. Donahue (2018) discusses how second language research has the potential to offer insights into writing transfer. Devitt (2015) argues how rhetorical conceptions of context allow her "to offer complex and still testable explanations of how and why language varies…" and how RGS has the potential to contribute to different areas of language studies "through enriching the notion of which genres to include and how to define them" (p. 337).[3] Methodologically, a number of studies have employed corpus analysis, a methodology or arguably a subfield from language studies, to investigate writing or issues in writing classrooms (e.g., Gere et al., 2013; Lancaster, 2016; Aull, 2020; Davila, 2022); on the other hand, corpus linguistics research has also discovered how established corpus methodologies can be complemented by rhetorically informed approaches (e.g., Brown & Wetzel, 2023).

Clearly, the values of learning from each other have been demonstrated. Nevertheless, further investigations of what both fields can learn from each other are also necessary to advance more interdisciplinary communications and collaborations. For example, as I have mentioned above, corpora have been an effective methodological tool for writing studies scholars to explore student writing or issues in writing classrooms; however, few studies in writing studies have used historical or diachronic corpora, which are usually compiled to study "how language has changed [over time]" (McEnery & Hardie, 2012, p. 95), nor has much writing studies research looked at how corpora, especially diachronic corpora, can be used to explore topics other than student writing or writing-related issues. In this article, I join the call for more exchanges between writing studies and language studies with the goal of filling these two gaps to partially further demonstrate how both fields can be highly complementary to each other. I focus specifically on how methodologies from English historical linguistics, particularly a diachronic corpus-based approach, can be effective in pursuing certain research questions in writing studies. Although English historical linguistics has been brought up much less often in writing studies than other language-related fields, such as applied linguistics, second language writing, and perhaps sociolinguistics, as I discuss below, some of its methodological frameworks can be helpful in answering historical questions in writing studies, which has been of great interest to scholars in writing studies (e.g., Phillips et al., 1993; Goggin, 2000; Fulkerson, 2005; Mueller, 2012).

In what follows, I first offer an overview of English historical linguistics and a diachronic corpus-based approach. Then I discuss the usefulness of a diachronic corpus-based approach to historical research in writing studies.[4] Finally, I provide an example where I explore the changing disciplinary trends in writing studies in a self-built, diachronic corpus that consists of research articles published in *College Composition and Communication* from 1997 to 2022. Doing so temporally complements

previous studies that have focused on the development of writing studies from the mid-twentieth century to the early twenty-first century (Phillips et al., 1993; Goggin, 2000; Mueller, 2012).

English Historical Linguistics and a Diachronic Corpus-Based Approach

"English [h]istorical [l]inguistics is a subfield of linguistics which has developed theories and methods for exploring the history of the English language" (Kytö & Pahta, 2016, p. iii). History of the English language (HEL) research focuses on every aspect of the English language, including but not limited to orthography, phonology, morphology, lexicon, syntax, and pragmatics, and the correlation with extralinguistic factors (see Bergs & Brinton 2012 for a detailed overview of what HEL research involves). Despite a broad range of focus, at the heart of all HEL research is the diachronic change of the English language. In other words, HEL research not only concerns a particular stage of English in the past, such as Old English and Middle English, but it also explores changes that have happened to the language over time. Another characteristic of HEL research is its heavy reliance on written data. Although language change is assumed to often start with variations in spoken language, HEL scholars most of the time have to use written material to explore language variation and change because spoken evidence simply does not exist until the recent past (Milroy, 1992, p. 5). For example, Walker (2007) studies the use of the second person pronoun *thou* and *you* in speech-related written texts, including trials, depositions, and drama comedy, across the period 1560-1760 in order to understand how those forms may have been used in the spoken version of early modern English.

Because of both characteristics of HEL research, studies in this field are often restrained by the availability, accessibility, and representativeness of written data as well as data-related extralinguistic factors. For example, Smitterberg (2016) notes that research questions are often subject to what material is available to use. The fact that written texts were produced by speakers with at least some level of literacy and that historical texts sometimes are not published for linguistic purposes or have editorial interventions also bring challenges to HEL research (pp. 190-191). Nevertheless, he also indicates that these methodological challenges can be at least partially solved by the use of diachronic corpora.

Diachronic corpora are principled collections of machine-readable natural texts "ordered systematically by temporal dimensions" (Stratton, 2021, p. 202). A principled diachronic corpus attempts to represent a particular language variety or language use in a particular genre over certain time periods. If a diachronic corpus represents a particular language variety, it is called a general or multi-purpose corpus. For example, as a multi-purpose corpus, the *Corpus of Historical American English* (*COHA*) covers the diachrony of American English from the 1820s to 2010s (Davies, 2010). On the other hand, if a diachronic corpus is designed to represent only the language use in a particular genre over time, it is called a specialized or genre-specific corpus. The *Corpus of Early English Medical Writing* (*CEEM*) is considered as a specialized diachronic corpus because it is composed of only medical texts from 1375 to 1800 (Taavitsainen & Pahta, 2010).

For a multi-purpose diachronic corpus to achieve its representativeness, according to Biber et al. (1998), compilers need to cover a) a broad range of genres, both written and speech-based, in a given historical period so that some characteristics of both written and spoken language in that period can be captured, b) enough samples of each genre, and c) enough words in each sample (pp. 252-253). The first criterion does not need to be taken into consideration if a genre-specific diachronic corpus is compiled.[5] It must be noted that the process of compiling a diachronic corpus and making it representative is often not as straightforward as what the principles above describe. For example, Kytö and Smitterberg (2015) mention that the absence of certain genres in a certain historical period and the evolution of one genre into another over time are among a plethora of the common issues compilers have to deal with.

The systematicity of diachronic corpora has made the diachronic corpus-based approach "the standard method for analysing and explaining the diachronic development…from beginnings of English to our own time" (Rissanen, 2012, p. 197). Specifically, using a diachronic corpus as a methodological tool to study HEL has at least three benefits. First, because a diachronic corpus is designed to be representative, the results yielded from an analysis of such a corpus are often generalizable and can be replicated. In other words, the results often allow researchers to make a conclusion of language patterns that a group of language users of a given time period may have followed, and the same results will be found if the same elements are searched for in the corpus. Second, Kohnen (2001) argues that genres "are catalysts for language change: they accelerate the spreading of a construction which already exists" (p. 115). Because genre is a key parameter for the design of a diachronic corpus, researchers can easily investigate how genre, as an extralinguistic factor, has an impact on language variation and change, which in some way can respond to the overarching question HEL research pursues— "why that particular change was initiated and diffused at some particular time and place" (Milroy, 1992, p. 20). Third, electronic diachronic corpora enable researchers to run searches for a variety of linguistic patterns in a large quantity of text collections, identify linguistic patterns that are likely to be unnoticed in a single text or unstructured materials, and answer different research questions.

Suitability of a Diachronic Corpus-Based Approach to Research in Writing Studies

As I mentioned earlier, despite the call for more reciprocity between writing studies and language studies, HEL has been brought up less often in writing studies than other language-related fields. However, writing studies and HEL have more convergence than divergence, and I argue that a diachronic corpus-based approach from HEL can be suitable for particular research questions in writing studies. I discuss three reasons why such an approach can methodologically contribute to writing studies.

First, it goes without saying that HEL focuses primarily on the historical aspect of the English language. A considerable amount of research in writing studies, in fact, also has a historical focus (e.g., D'Angelo, 1984; Stewart, 1985; Phillips et al., 1993; Scott, 1998; Goggin, 2000; Haswell, 2005; Clary-Lemon, 2009; Peary, 2009; Mueller, 2012; Longaker et al., 2022). For example, Crowley (1985) traces the (d)evolution of invention procedure in writing instruction from the last half of the nineteenth century to the late twentieth century. Yancey (1999) reviews the history of writing assessment across the second half of the twentieth century (1950-1999). Ritter (2008) looks at basic writers at Yale between 1920 and 1960. Fredlund (2021) examines a feminist and antiracist history of composition and rhetoric at Oberlin College from 1846 to 1851. Smith et al. (2021) carry out a diachronic study where they examine student writing development over five years at Northeastern University.

In addition, a special section, namely *Histories*, was dedicated for reviews of some history of writing studies in the September and December volumes of the flagship journal of writing studies, *College Composition and Composition* (*CCC*), in 2009. In my own example of how to carry out a diachronic corpus-based analysis below, we see that the word *history* is among the top 50 most frequently used words in *CCC* between 1997 and 2022, which seems to be another piece of evidence that suggests writing studies has a perennially historical focus.[6] The historical focus and the fact that writing studies most of the time work with written texts instead of spoken ones make a diachronic corpus-based analysis suitable for some research in writing studies. In particular, a diachronic corpus-based approach has great potential of yielding fruitful results for research that concerns the changing shape of the field over time, which seems to have been a long tradition in writing studies (e.g., Phillips et al.,

1993; Goggin, 2000; Fulkerson, 2005; Mueller, 2012). For example, Phillips et al. (1993) and Mueller (2012) explore the development of writing studies by examining the frequency of citations appearing in *CCC* from 1950 to 1993 and from 1987 to 2011, respectively.

Second, replicability and generalizability are something that is scanty in writing studies. Haswell (2005) laments the severe decline of RAD studies, which refer to "scholarly investigation that is **r**eplicable, **a**ggregable, and **d**ata supported" (p. 201, my emphasis), in two flagship professional organizations of writing studies—the *National Council of Teachers of English* (NCTE) and the *Conference on College Composition and Communication* (CCCC). According to Haswell, RAD research transparently systematizes data sampling, data collection, and data analysis, and the lack of RAD studies impairs NCTE/CCCC's ability "to deflect outside criticism with solid and ever-strengthening data" (p. 219). Raucci (2021) echoes that the absence of RAD scholarship in writing studies "considerably narrows the scope and generalizability of research in the field" (p. 441). When it comes to writing program administration (WPA), Anson (2008) argues that the paucity of RAD research makes administrative decisions and practices rely heavily on personal beliefs or studies that "were conducted twenty, thirty, or even forty years ago, under different conditions, with different populations raised and schooled with different values and experiences, and before the advent of technology and digital media" (p. 20). As I discussed above, a diachronic corpus-based study is replicable, and the results yielded from such a study are often generalizable. As a RAD methodology, diachronic corpus-based analysis can facilitate "verification, invention, collaboration, transparency, and revision" in writing studies (Raucci, 2021, p. 441).

Third, one consequence of the absence of RAD research is the heavy reliance on unverified beliefs rather than empirical evidence, as I reviewed above. However, beliefs, more often than not, are untenable and easy to be disparaged, and it is possible that some beliefs actually do not hold anymore because the contexts where those beliefs were developed have changed. For example, Matsuda (1999) discusses how the assumption of teaching writing to ESL students being the sole responsibility of ESL specialists and in turn the removal of second-language components in writing studies in the 1960s had brought issues and challenges to writing instructors in the 1990s because of the increasing number of international students. A diachronic corpus-based analysis can help us identify changing patterns over time that sometimes contradict our intuitions because of its methodologically empirical and systematic nature. In addition, a diachronic corpus-based analysis also enables us to discover elements that are likely to be unnoticed. As Mueller (2012) puts it, distant reading methods, which roughly refer to looking at a large series of texts simultaneously, "help us engage with patterns of disciplinary activity that would otherwise be difficult to discern, particularly for newcomers to the field" (p. 197).[7]

Diachronic Corpus-Based Analysis: Disciplinary Foci in Writing Studies from 1997 to 2022

To illustrate the practice of a diachronic corpus-based analysis, I studied the changing disciplinary foci in writing studies from 1997 to 2022. By disciplinary foci, I mean focal points, activities, and interests in a discipline revealed by language elements in a given time period. For example, as I show below, translingualism seems to be one of the focal points in writing studies between 2017 and 2022 because the word *translingual* is used more frequently in *CCC* between 2017 and 2022 than between 1997 and 2016. By extending the time period to include recent years, this study also aims to temporally complement previous studies that have focused on the development of writing studies from the mid-twentieth century to the early twenty-first century (Phillips et al., 1993; Goggin, 2000; Mueller, 2012).

Corpus and Analytical Techniques

The cornerstone of any diachronic corpus-based studies is a corpus where research questions can be appropriately explored. Despite a large number of existing multi-purpose diachronic corpora, such as *COHA* (Davies, 2010) and the *Corpus of English Dialogues* (*CED*; Culpeper & Kytö, 2006), no existing corpus was appropriate for my research question, so I compiled a new specialized, diachronic corpus suited for my purposes. As mentioned earlier, two key considerations for compiling a diachronic corpus are representativeness and time dimension. To be temporally representative at least to some degree, my corpus consists of research articles published in *CCC* between 1997 and 2022 because previous studies seem to agree that *CCC*, as the flagship journal in writing studies, can index, at least partially, what is happening and what current scholarly conversations are in the field.[8] For example, Phillips et al. (1993) and Mueller (2012) rely exclusively on *CCC* to study the disciplinary activities over time, and *CCC* is also included in Goggin (2000), in addition to some other journals, to investigate changing disciplinary patterns from 1950 to 1990.

To identify the disciplinary foci in the corpus, two techniques were carried out in the programming language R (R Development Core Team, 2021).[9] One technique is a word frequency list, which often presents how many times each individual word appears in a corpus (or token frequency). As one of the most important concepts in corpus-based studies, word frequency can tell us the focal points of the corpus because the more frequently a lexical word appears, the more likely it is a common focus in the corpus. Word frequency lists can also help us understand how two corpora are different when some normalization is carried out (how many times a word occurs to a standard scale, for example, every 10,000 words). While it is perhaps more common to see word frequency lists where the frequency of each individual word (unigram) is shown, it is also possible to present the frequency of word clusters (ngrams), such as a two-word sequence of words (bigram) and a three-word sequence of words (trigram), which sometimes may provide further insights into what topics are discussed in the corpus. For example, the phrase *critical language awareness* contains three unigrams (*critical*, *language*, and *awareness*), two bigrams (*critical language* and *language awareness*), and one trigram (*critical language awareness*).

In this study, I present unigram, bigram, and trigram word lists. When the unigram word list was developing, I took out all the grammatical words, which roughly refer to words that display the relationships between content words in a sentence but often do not have much lexical meaning, such as determiners (e.g., *the*, *a*, *his*), prepositions, (e.g., *of*, *with*, *off*, *at*), conjunctions (e.g., *because*, *for*, *and*), and so on. Doing so helps us better see the topics that the corpus is centered on (Baker, 2006, p. 55).

It is also worth noting that lemmatization was not conducted in the corpus. Lemmatization is the process of grouping words together according to their more generalized forms (lemmas). For example, if lemmatization is performed, the words *English* and *Englishes* will be counted as two occurrences of the same word *English* instead of one occurrence of two different words, and likewise, the words *language* and *languaging* will also be considered as two occurrences of *language*. However, word groups like these two sometimes contain different meanings and connotations in writing studies, and it is for this reason that lemmatization was not carried out in the corpus.

The other analytical technique used to identify the disciplinary foci in the corpus is keyness. A keyness analysis identifies keywords that are most distinctive of one corpus against another corpus or multiple corpora by comparing the frequency of word forms. In other words, a keyword is found if it appears very frequently in one corpus but much less frequently in its reference corpus. When an appropriate metric is adopted, a keyness analysis "would reflect the size of the frequency difference…. the larger the difference, the more 'key' a word would be" (Gabrielatos, 2018, p. 229).[10] To carry out a keyness analysis, I further divided the corpus into five sub-corpora by time periods

(1997-2001, 2002-2006, 2007-2011, 2012-2016, 2017-2022), and the keywords, including both unigrams and bigrams, were investigated in each sub-corpus with the other sub-corpora being references.[11]

As a keyness analysis procedure, term frequency-inverse document frequency (TF-IDF) was carried out in this study (Kilgarriff, 2001).[12] Borrowed from the field of information retrieval and text mining, the TF-IDF measure indexes the degree of importance a word or ngram has in a corpus in comparison to other corpora. A word or ngram will receive a high score (highest being 1) if it appears very frequently in one corpus but does not appear at all in its reference corpora; conversely, if a word occurs very frequently not only in one corpus but across its reference corpora, it will receive a very low score (lowest being 0).[13] For example, grammatical words, such as *the*, *of*, *on*, *but*, usually have a TF-IDF score close to 0 because they are too common to differentiate one corpus from other corpora.

Once the TF-IDF score of each word and bigram was computed in each sub-corpus, TF-IDF > 0 was set to be the threshold to select the keywords.[14] The main reason why this threshold was set is that almost no word can have a high or even relatively high TF-IDF value especially when only four or five corpora are in comparison.[15] Setting the threshold of above zero allows the study to keep as many keywords as possible and to more appropriately map the changing disciplinary foci of writing studies over the past two decades or so.

The unigram, bigram, and trigram frequency lists can help us see what the field of writing studies revolves around between 1997 and 2022, and the keyness analysis realized through TF-IDF enables us to unveil some unique periodical disciplinary foci in the field. That is, some disciplinary conversations that appear often in one particular five-year period but less often in the rest of the time periods. To discuss some of the ngrams on the frequency lists and keywords, a concordance analysis was also conducted. Concordance, which is another commonly used corpus technique, refers to "a list of all the occurrences of a particular search term in a corpus, presented within the context that they occur in" (Baker, 2006, p. 71). By presenting an ngram in its surrounding text, concordance lines provide us with a qualitative lens of understanding how the ngram is used in the corpus. For example, as I illustrate below, concordance lines can help us understand that the word *kairos* in the sub-corpus of 2007-2011 is used to either refer to a rhetorical term or the name of a journal.[16]

Frequency Lists

Figure 1 shows the top 50 most frequent content words (unigrams) in the corpus.

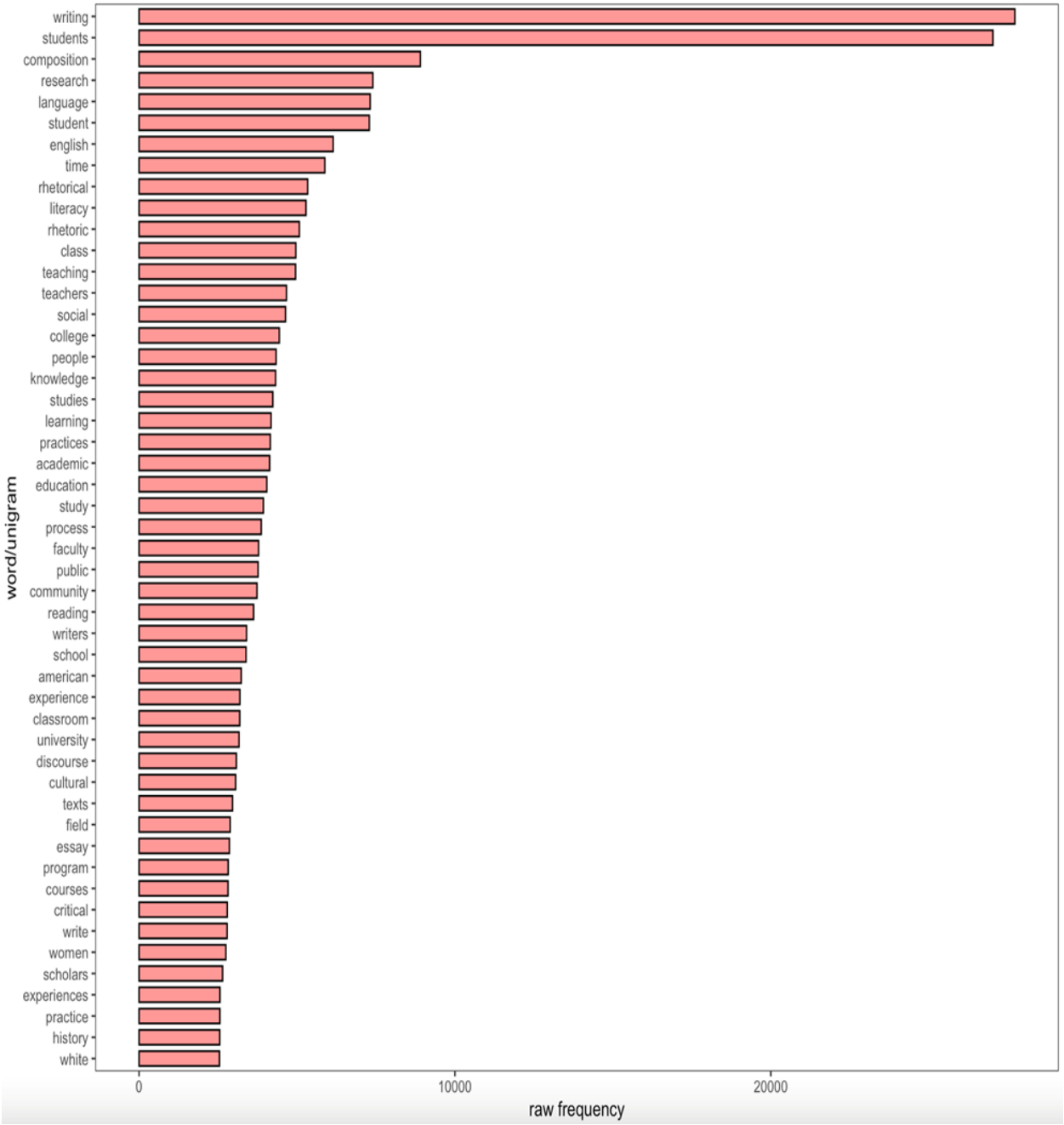

Figure 1: Top 50 words in CCC from 1997 to 2022

As we can see from Figure 1, the two words *writing* and *students* dominate the corpus, which seems to suggest that the field of writing studies has been highly focused on writing instruction over the past 26 years. This is not surprising given the aims and scope of *CCC*: the journal publishes research "that supports college teachers in reflecting on and improving their practices in teaching writing and that reflects the most current scholarship and theory in the field" (NCTE, 2024). A further comparison across the first, second, and fourth most frequent words over years, as shown in Figure 2, reinforces such a focus: while the trends of the frequencies of all the three words fluctuate over time, *writing* and *students* are always used more or even much more frequently than *research*.

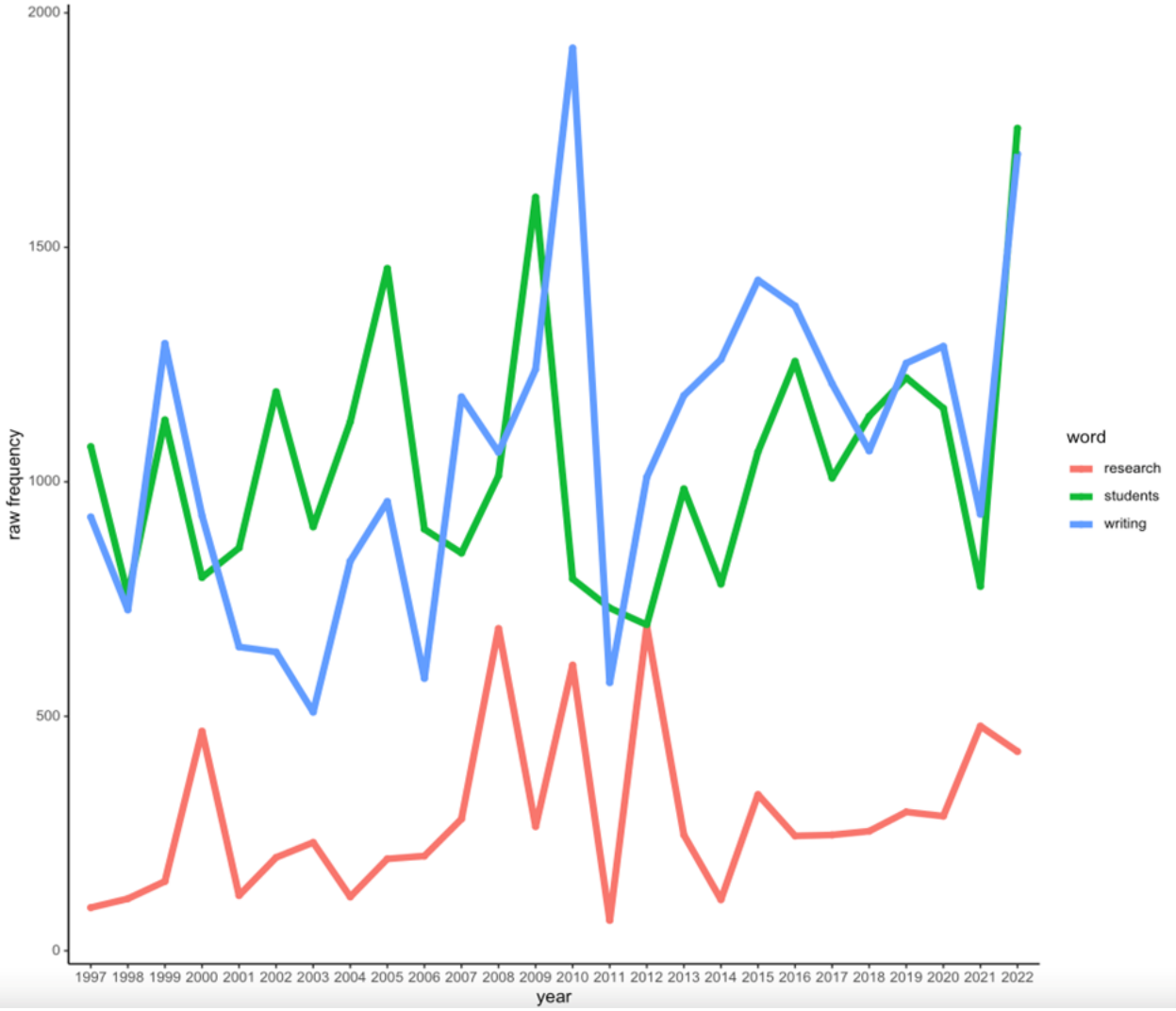

Figure 2: Frequency of writing, students, and research in CCC from 1997 to 2022

The strong focus on writing instruction can also be extrapolated by many other top 50 most frequent words that are related to writing instruction, such as *composition*, *student*, *literacy*, *teaching*, *teachers*, *college*, *knowledge*, *learning*, *academic*, *education*, *reading*, *school*, *classroom*, *university*, and *essay*. All of these words seem to suggest that the educational nature of the field has remained pretty constant since 1997. On the other hand, some other top 50 words seem to imply the diversity of pedagogical approaches to teaching or discussing writing, which, according to Fulkerson's (2005) term, is "axiological consensus and pedagogical diversity" (p. 655).[17] For example, words such as *critical*, *White*, *women*, *public*, and *cultural* seem to imply a critical, cultural studies approach, as can be seen in corpus examples 1, 2, 3 below, whereas *rhetorical* and *rhetoric* seem to implicate a rhetoric-oriented approach, as in corpus examples 4 and 5 below . Both approaches have been found in writing classrooms since the early 2000s (Fulkerson, 2005), and these words seem to tell us that both approaches have still been active after two decades or so.

1. …white composition and rhetoric scholars such as myself have much decolonial work to do.

2. … women writers at both sites negotiate domestic and public identities pertaining to gender and composing…
3. …disregarded the anticipation of emotion as a cultural construct that builds affect, dispositions, and identities.
4. …rhetorical decisions to engage with this audience by listing specific information…
5. When we bring an understanding of digital rhetoric to our electronic classrooms…

Of course, a quantitative mapping like Figure 1 is only suggestive because all the words were generated based on forms instead of meanings and cannot tell in what contexts those top frequent words are used, although many of them are often employed in a disciplinarily conventional way. For example, words like *class*, which is the twelfth most frequent word in the corpus, can reference social classes or classes students attend or sometimes can even be used as a verb, and we are not certain which sense this word has nor what part of speech this word is on the list. Some concordance analysis or close examination is always recommended, as in (1), (2), (3), (4), (5) above, if our goal is to understand how *class* is used and why it appears as one of the most frequent words on the list.

Importantly, Figure 1 also illuminates what words are not on the list and what words surprisingly appear on the list. What words, for instance, would we expect over the past two decades or so but didn't make their way? Given that the timespan of the corpus in this study (1997-2022) roughly corresponds to the timeframe of the cultural turn in writing studies (Silva & Leki, 2004, p. 4), it might be interesting for scholars in the field to look into the words or terms that do not appear on the list. For instance, it might be interesting to compare the trend of *English*, which is the seventh most frequent word on the list, to the trend of *Englishes*, which doesn't appear on the list. On the other hand, what are some of the words that surprise us by their appearances on the list? For example, as I mentioned earlier, research indicates that consistent attention to language has not come back in writing studies perhaps until the past decade or so after its decline in the 1970s (e.g., Matsuda, 1999, 2013; MacDonald, 2007; Aull, 2023). Yet, the word *language* is the fifth most frequently used word between 1997 and 2022. If we take a closer look at how often *language* is used by years, as shown in Figure 3, we see that the trend moves up and down with the word being used most frequently in 2022 and second-most frequently in 1999. While we can see some decrease in the use in the early 2000s, a similar pattern can also be observed between 2010 and 2022, which, to some extent, does not seem to align with the tendency suggested by the scholarship. It would be interesting to further examine to what degree the trend in Figure 3 reflects disciplinary attention on language-related topics over time.

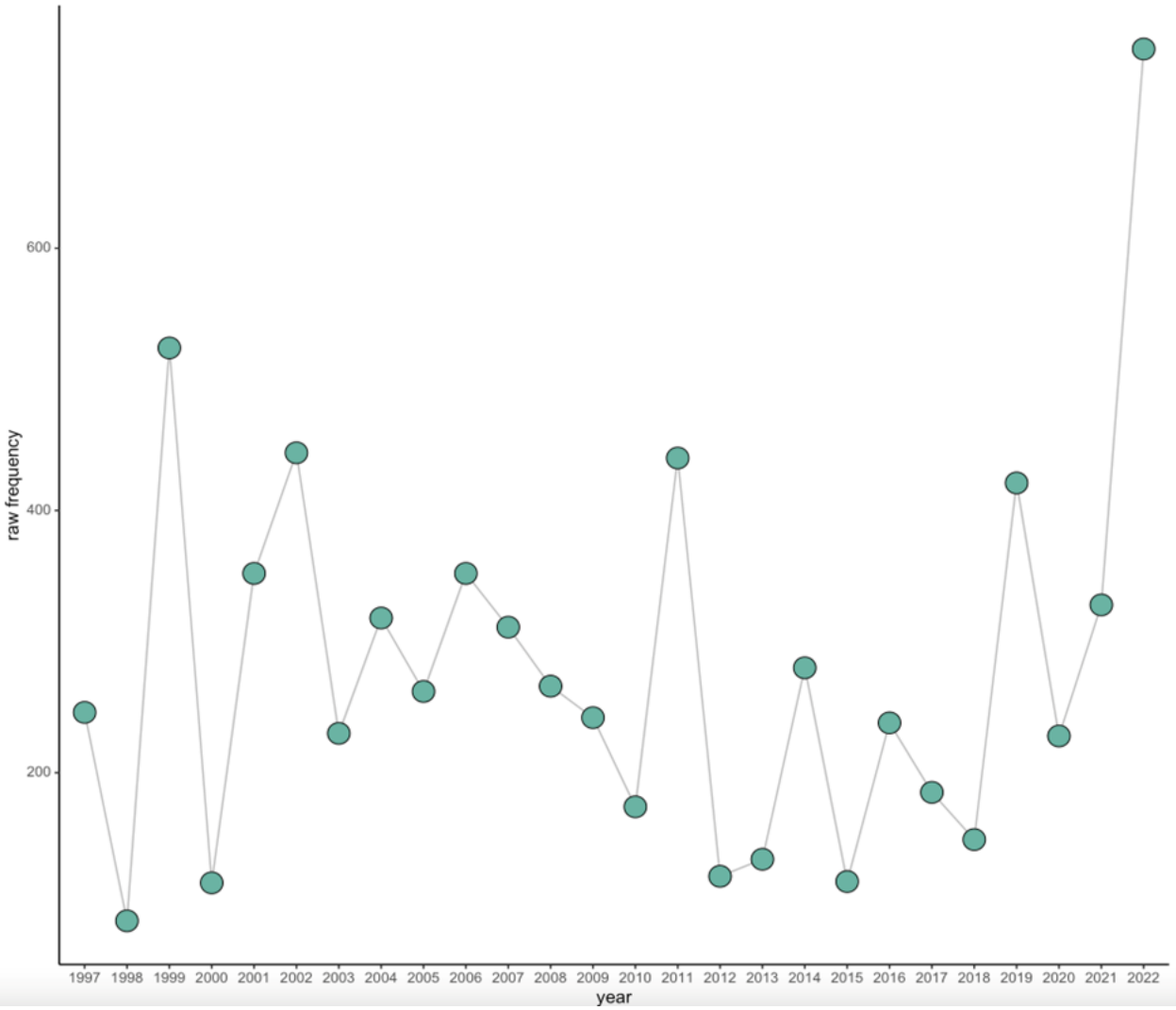

Figure 3: Frequency of language in CCC from 1997 to 2022

Figure 4 presents the top 50 most frequent bigrams (two-word clusters) in the corpus. It is clear that most of the bigrams in the figure consist of only grammatical words, which, as I mentioned earlier, are used to index the grammatical relationships between content words but do not have much lexical meaning. Because bigrams are generated based on word sequences instead of grammaticality or semantics, many bigrams, such as *and to*, *of the*, *in the*, and so on, are not meaningful enough to tell us much about some disciplinary activities in writing studies over time. However, if we look at those bigrams that consist of content words, including *of writing*, *students to*, and *the writing*, they unsurprisingly align with the aforementioned discussions of the unigrams and the aims and scope of *CCC* that writing studies has been highly focused on writing instruction over the past 26 years.

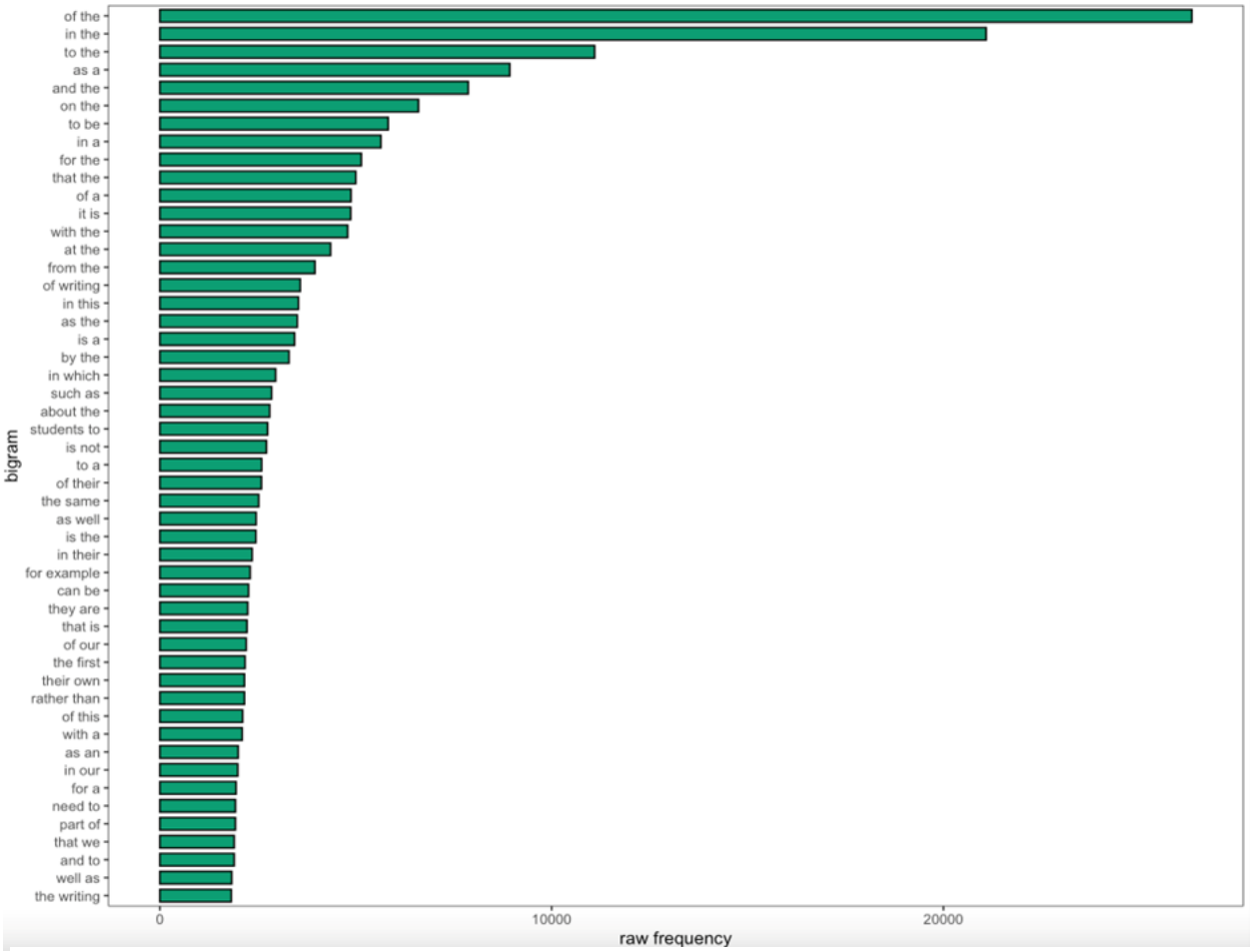

Figure 4: Top 50 bigrams in CCC from 1997 to 2022

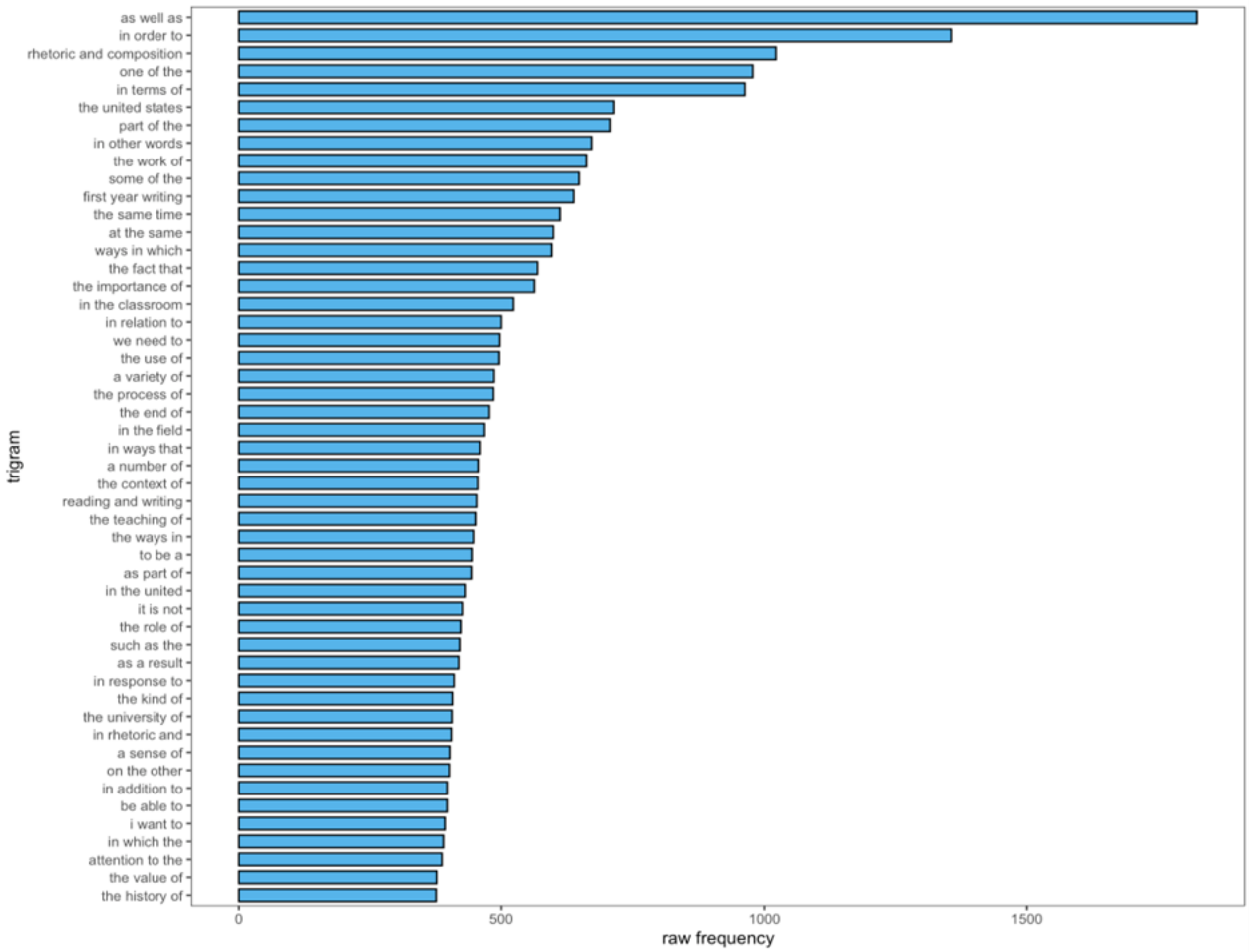

Figure 5: Top 50 trigrams in CCC from 1997 to 2022

Unlike the bigrams above, some of the trigrams (three-word cluster) in this figure are both grammatical and have complete meanings, including *rhetoric and composition*, *the United States*, *first year writing*, *in the classroom*, *reading and writing*. The phrases *in the classroom* and *reading and writing* can further confirm the field's primary commitment to writing instruction between 1997 and 2022 revealed by some unigrams and bigrams above, and the phrase *first year writing* suggests it being the core of writing instruction in writing studies or the site where writing education usually happens over the past 26 years. The phrase *the United States* seems to affirm that "explicit composition instruction (especially in colleges and universities) is mostly a North American phenomenon" (Silva & Leki, 2004, p. 7), or that the focus of *CCC* has been very centered towards college-level writing in U.S. contexts, as can be seen in corpus examples 6, 7, and 8:

6. Contemporary demographic changes in the United States usher in important challenges for writing assessment researchers who are committed to issues of fairness in the assessment of writing.
7. …the next section considers women and rhetorical education in the United States before providing specifics about the students who attended Oberlin…
8. The number of undergraduate students in the United States who grew up with first languages other than English…

The term *rhetoric and composition*, on the one hand, seems to indicate the disciplinary preference for how scholars have addressed their own field in the past two decades or so. While sometimes *composition studies*, *writing studies*, and *rhetoric and composition* can be used interchangeably with some clarifications and delimitations, it seems *rhetoric and composition* has been favored, at least since 1997. This is not very surprising because, for many, the term *writing studies* sounds more inclusive because it covers not only first language but also second language writing (and perhaps beyond, such as business writing and legal writing), and professionals in writing studies are not necessarily from English departments in North America; in contrast, second language elements had not been part of the agenda in rhetoric and composition because of the disciplinary division of labor at least until the early 2000s (Matsuda, 1999), and scholars in rhetoric and composition are "typically housed in English departments [in North America]" (Silva & Leki, 2004, p. 9), and many graduate programs use *rhetoric and composition* to name themselves, as can be seen in corpus examples 9 and 10:

9. …rhetoric and composition graduate programs need to do much more to prepare their students to…
10. …whether we call ourselves composition studies or rhetoric and composition, it is odd that …

On the other hand, the frequent occurrences of *rhetoric and composition* perhaps could be interpreted as authors' attempt to build an insider identity and "seek credit for that position" (Hyland, 2001, p. 222), as in corpus example 11. The occurrences could also be considered as scholars' commitment to promoting the field, distinguishing the field from other disciplines, and establishing credibility for the field, which may be seen in corpus example 12. More broadly, the frequent occurrences of *rhetoric and composition* can be argued as a way that the discipline has used to build or consolidate its identity over the past 26 years.

11. I further argue that, for rhetoric and composition, new media is tied to multimodality and digital composition….
12. [r]hetoric and composition as a discipline also has a more sophisticated and more nuanced understanding of research epistemologies, methodologies, and methods…

It must be noted that a quantitative mapping, such as a word or ngram frequency list, usually is for exploratory purposes and serves as the first step in a diachronic corpus-based analysis even if we have specific words in mind to search for. This is because, as I have mentioned above, we are unsure of in what contexts each word or ngram on the list is used, although it is possible that many words are only employed in a disciplinarily conventional way. For this reason, while the quantitative mapping helps us see the disciplinary foci in the field of writing studies over the past 26 years, a concordance analysis, as we can see, is also necessary to help us verify and explain those disciplinary patterns.

Keywords

Figure 6 presents the key unigrams in every sub-corpus.

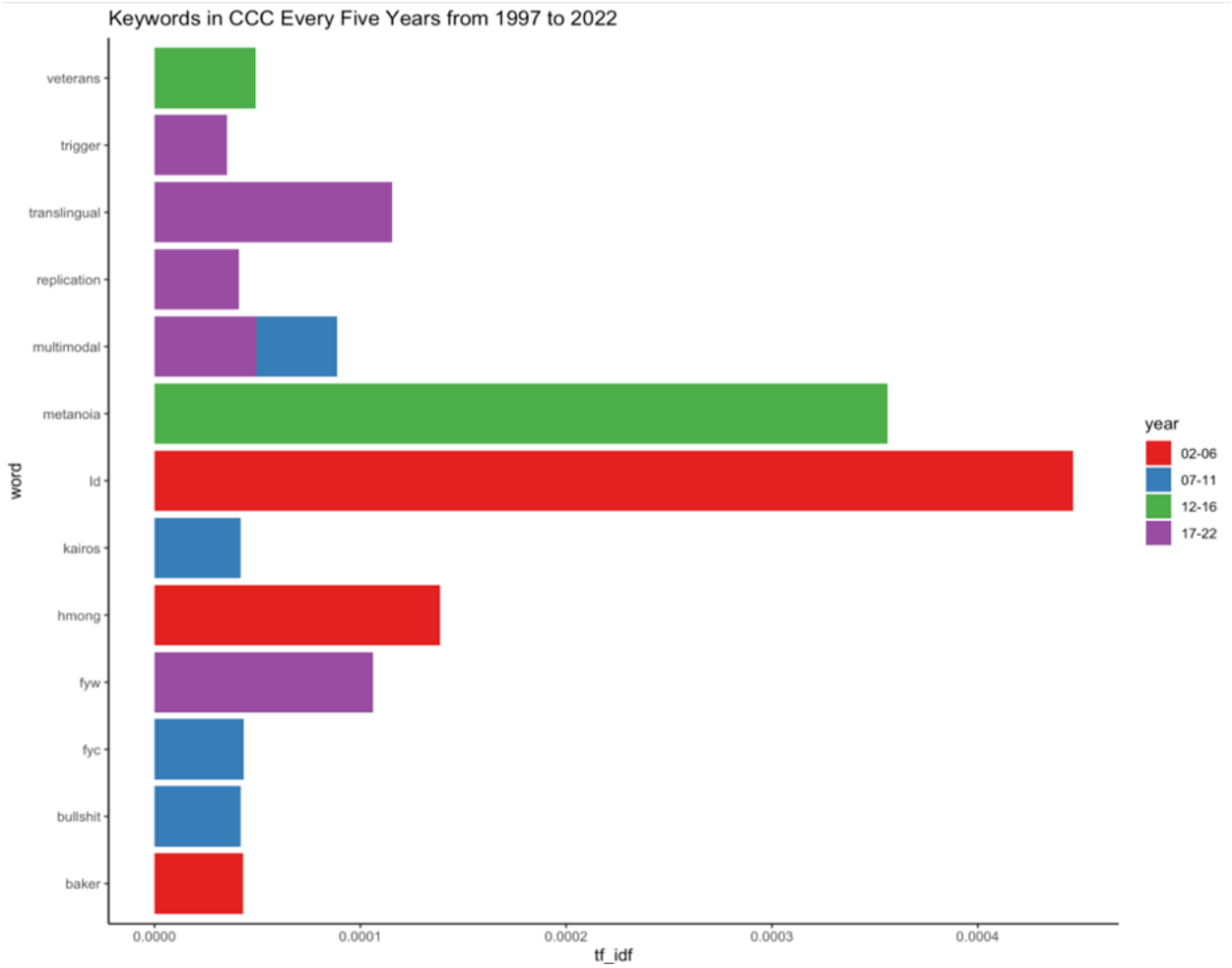

Figure 6: Keywords in CCC every five years from 1997 to 2022

From this figure, we can see that the key unigrams in the sub-corpus of 2002-2006 include *ld*, *hmong*, and *baker*, and the key unigrams of the sub-corpus of 2007-2011 include *multimodal*, *kairos*, *fyc*, and *bullshit*. The key unigrams of the sub-corpus of 2012-2016 are *veterans* and *metanoia*, and the key unigrams of the sub-corpus of 2017-2022 are *trigger*, *translingual*, *replication*, *multimodal*, and *fyw*. The sub-corpus of 1997-2001 does not have any key unigrams, and this means that every single word that appears in that corpus also occurs in the rest of the sub-corpora.[18]

These results have brought up some important observations. First, some of the key unigrams, such *ld*, *baker*, *bullshit*, *hmong*, and perhaps *trigger*, do not tell us much without the contexts where they appear. For example, unless we run concordance checks or go back to the original articles in the sub-corpora, we are unlikely to know that *ld* is the abbreviation of *learning disability* that appears in White (2002), nor would we understand that *baker* refers to George Pierce Baker in Bordelon (2006). When keywords are not meaningful or helpful enough, it would be crucial to look through each concordance where those words show up if we want to have a better understanding of how those words are used and perhaps why they are key in their sub-corpora. Second, keyness computations are likely to be influenced by some idiosyncratic words, such as *baker* and *bullshit*, *hmong*, *metanoia*, in the corpus. Those words are so unique that they appear only in one individual article but are very unlikely to show up in their reference sub-corpora. When these words appear fairly frequently, they are likely

to be considered keywords. There are several possible ways to deal with this issue. The corpus can be further preprocessed to take out all the idiosyncratic words that appear in only one article or in only one year before it is analyzed.[19] It is also possible to set a different TF, IDF, or TF-IDF threshold or try a different keyness technique to avoid those words.

Third, based on the keywords that do not need much further investigation of what they refer to, including *multimodal* (2007-2011 and 2017-2022), *kairos* (2007-2011), *fyc* (2007-2011), *translingual* (2017-2022), *replication* (2017-2022), and *fyw* (2017-2022), we can assume that multimodal composition has received much attention in writing studies since 2007 although such attention might have declined slightly between 2012 and 2016, and a further mapping more or less confirms this assumption, as shown in Figure 7.

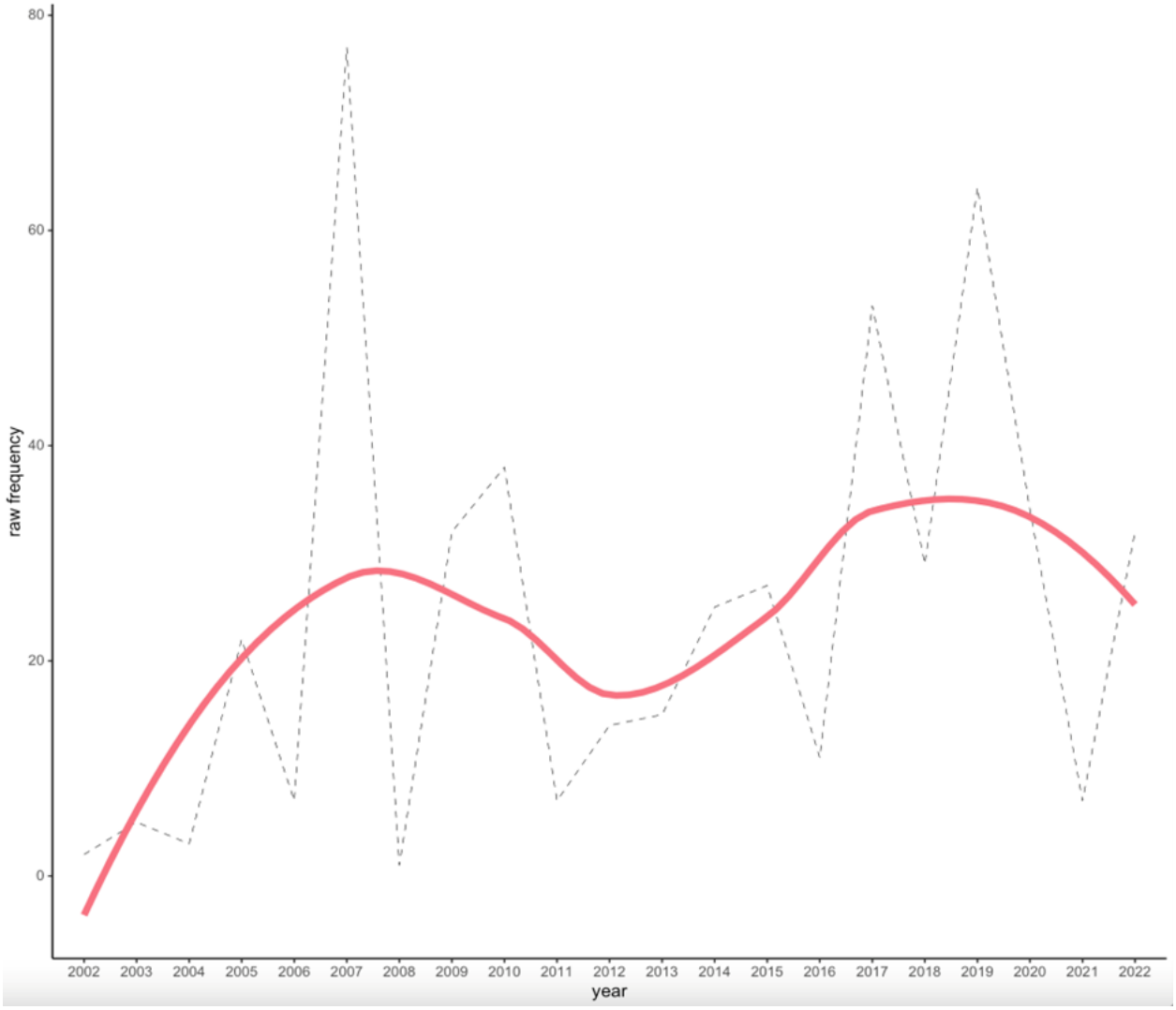

Figure 7: Trend of multimodal in CCC from 1997 to 2022

The words *fyc* and *fyw* seem to tell a terminology shift from the former in early 2000s to the latter in most recent years. A more detailed look, as in Figure 8, suggest *fyc* had remained the preferred term until the emergence of *fyw* in 2011, and since then, *fyc* has given its way to *fyw.*

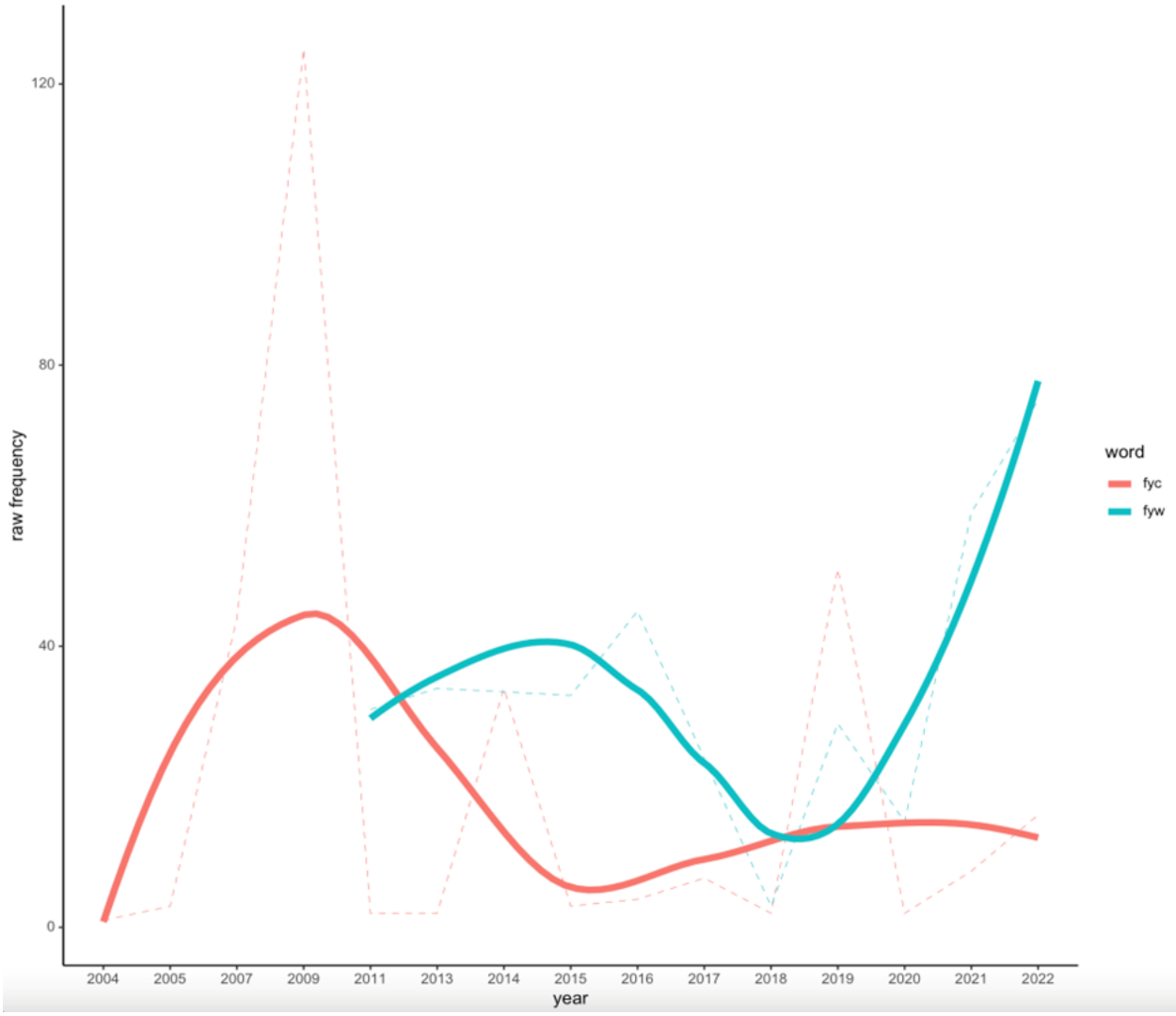

Figure 8: Trends of fyc and fyw in CCC from 1997 to 2022

In addition to translingualism, it looks like writing studies has also been concerned about the replicability of research since 2017. Kairos, seems to be a focus in the field between 2007 and 2011. A closer look reveals that *kairos* was not only used to refer to a rhetorical term, as in (13), (14), and (15), but it is also used to reference the journal *Kairos*, as in (16) and (17).

1. …more nuanced definitions of kairos surface as implicit orientations in the research of award-winning works.
2. The students in this sample seemed aware of the ancient principle of kairos and wrote with a sense of what is appropriate for formal college writing.
3. The kairos of the situation required reading only in the second…
4. …*Computers and Composition*, *Rhetoric Review*, *Journal of Teaching Writing*, *Kairos*…
5. …published online in *Kairos* in which…

Figure 9 presents the key bigrams in every sub-corpus.[20]

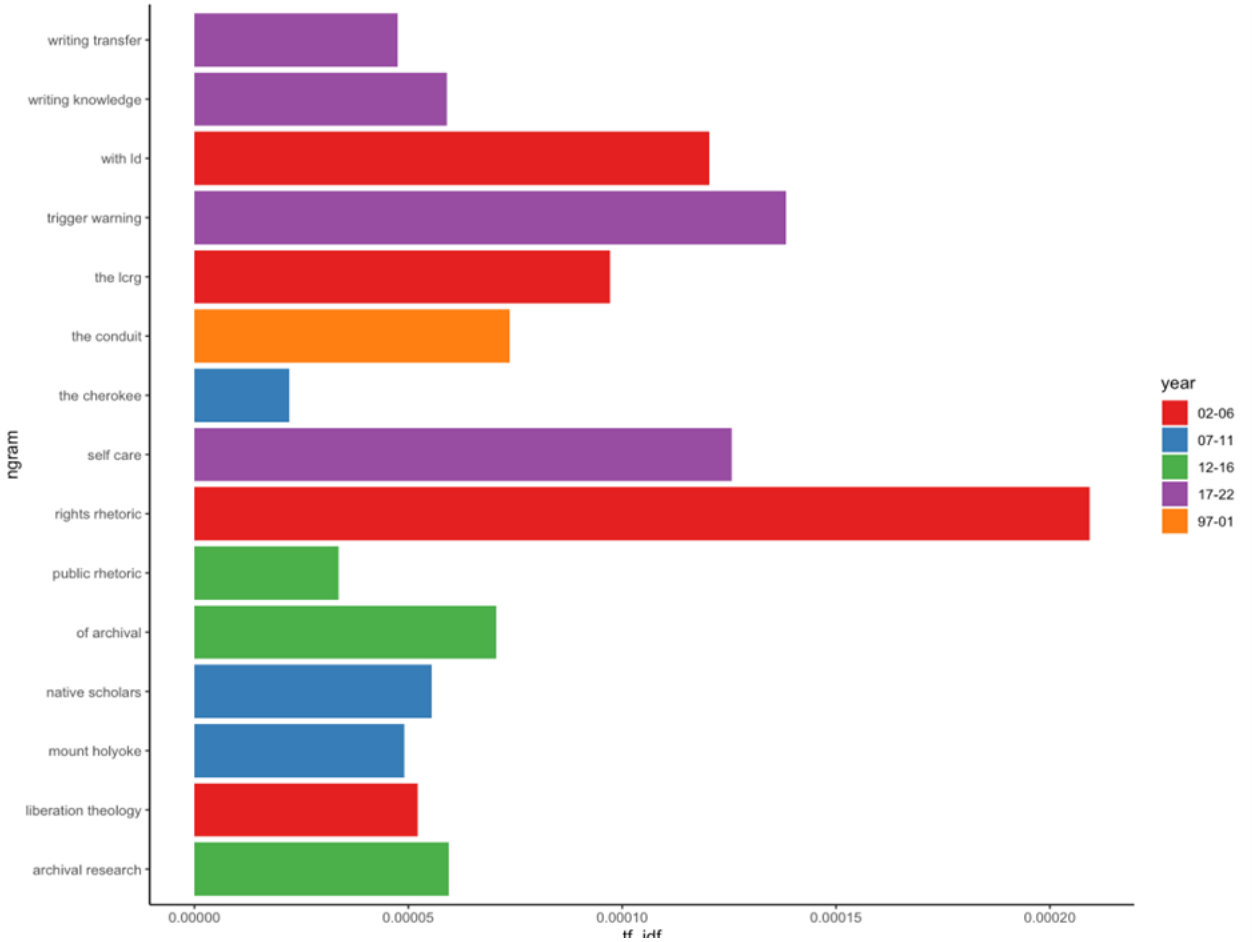

Figure 9: Key bigrams in CCC every five years from 1997 to 2022

From the figure, we see that the key bigrams in the corpus of 1997-2001 include *the conduit*, and the key bigrams in the corpus of 2002-2006 include *with ld*, *the lcrg*, *rights rhetoric*, *liberation theology*. The key bigrams of the corpus 2007-2011 are *native scholars*, *mount holyoke*, *the cherokee*, and the key bigrams of the corpus of 2012-2016 are *public rhetoric*, *of archival*, and *archival research*. *Writing transfer*, *writing knowledge*, *trigger warning*, and *self care* are the key bigrams in the corpus of 2017-2022.

As I said above, unless further concordance investigations are conducted, some bigrams, such as *the lcrg* and *with ld*, alone are not meaningful enough to help us understand what they refer to and why they are identified as key in their sub-corpora. If we focus on those key bigrams that are meaningful, we can assume that *native scholars* and *the cherokee* seem to indicate an interest in connections between writing studies and indigenous studies between 2007 and 2011. The bigrams *public rhetoric* and *archival research* suggest that both were among the focal points in writing studies between 2012 and 2016, and *writing transfer* and *writing knowledge* are among the focal points between 2017 and 2022. It is worth noting that while qualitative analysis, such as concordance, is always recommended to best understand how bigrams are used, as I have mentioned earlier, unlike unigrams, bigrams, if grammatical, often can have only one interpretation, especially when the discipline is specified. For example, while the unigram *kairos* can be understood as either the name of a journal or a rhetorical

element in the field of writing studies with the context where it appears, bigrams, such as *writing transfer* and *archival research*, usually can only be interpreted in one way because the co-occurrence of the two words in the particular sequence constrains the meaning each word can take.

To summarize, it looks like some focal points that the field of writing studies had between 2007 and 2011 include multimodal composing, first-year composition, kairos, and indigenous studies. The field focused on, among others, public rhetoric and archival research from 2012 to 2016. The most recent focal points of the field include translingualism, first-year writing, research replication, multimodal composing, writing transfer, and writing knowledge.

Conclusion

In this article, I have discussed the potential of a diachronic corpus-based approach to research in writing studies. As a RAD methodology, such an approach often can yield fruitful results in historical research in writing studies, especially research that concerns stability and/or change over time. However, this does not mean that a diachronic corpus-based approach is the only methodological framework through which to explore historical research in writing studies. For example, while analysis of a diachronic corpus can reveal terminology shifts over time, as can be seen in my example of *fyc* to *fyw* earlier, it often cannot tell the triggers for such shifts.

Sometimes it is better to conduct interviews in addition to a diachronic corpus-based analysis (multimethodology) to more appropriately answer the research question, and researchers should always choose a methodology according to their research questions. For instance, conducting interviews with journal editors and/or scholars within the field besides the results of my diachronic corpus-based analysis would help us understand the reasons behind the shift from *fyc* to *fyw*. Also, while it is unquestionable that a diachronic corpus-based approach can be suitable for explorations of historical research in writing studies, it may be used for pedagogical purposes as well. For example, graduate faculty can use it to provide a snippet of what the field of writing studies looks like and where the field is moving to newcomers, or a picture of when a particular movement starts, peaks, and ends to students in a topic-based class. It is also possible to rely on the results of a diachronic corpus-based analysis to select textbooks or journals for publication.

I have also shown an example of a diachronic corpus-based approach to the disciplinary trends in writing studies from 1997 to 2022. The analysis has revealed the field's unchanged focus on writing instruction at the U.S. college level over the past 26 years. Dividing the corpus into five sub-corpora according to time periods, I have also uncovered some five-year-period interests in writing studies, such as translingualism and writing transfer between 2017 and 2022, public rhetoric and archival research between 2012 and 2016, and multimodal composing and indigenous studies between 2007 and 2011. It must be reiterated that my example is primarily for illustrative purposes and only exhibits one way of using a diachronic corpus-based approach to one particular research question in writing studies. It is possible to compile a different diachronic corpus in which different research questions can be studied. For instance, a diachronic corpus that consists of journal articles from different journals of writing studies over time would allow us to have a more comprehensive understanding of the trajectory of the field, and the corpus would also enable us to study how the focal points in those journals are similar or different over time. A diachronic corpus that is composed of student writing over a certain time period would help us trace student writing development. It is also possible to carry out different corpus techniques from or in addition to the two I used in my example to analyze the corpus and answer different research questions. For example, as I mentioned earlier, concordance analysis can be useful to help us understand why the word *language* appears frequently in the corpus, although some research claims a considerable decline of attention to language-related topics from the 1970s to the early 2000s. Collocations can also be conducted to help

us see what words are likely to co-occur with *language*, which may provide some further insights into how the word is used. A different keyness metric can also be carried out to study keywords from different perspectives.

In the end, I hope that this article has further demonstrated the connections and potential contributions that language studies, especially HEL research, have to writing studies. I also hope that future research can explore different possibilities of more exchanges between the two fields.

References

Anson, Chris M. (2008). The intelligent design of writing programs: Reliance on belief or a future of evidence. *WPA: Writing Program Administration*, 32(1), 11-36.

Aull, Laura. (2023). Attention to language in composition. *Composition Forum*, 51.

Aull, Laura. (2015). Linguistic attention in rhetorical genre studies and first year writing. *Composition Forum*, 31.

Aull, Laura. (2020). *How students write: A linguistic analysis*. New York, NY: *Modern Language Association*.

Baker, Paul. (2006). *Using corpora in discourse analysis*. London, UK: Continuum.

Bergs, Alexander, & Laurel J. Brinton. (2012). *English historical linguistics: An international handbook* (Vols. 1-2). Berlin, Germany: Mouton de Gruyter.

Biber, Douglas, & Conrad, Susan, & Reppen, Randi. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Bordelon, Suzanne. (2006). A reassessment of George Pierce Baker's "the principles of argumentation": Minimizing the use of formal logic in favor of practical approaches. *College Composition and Communication*, 57(4), 763-788.

Brewer, Meaghan, & di Gennaro, Kristen (2022). Naming what we don't know: Graduate instructors and declarative knowledge about language. *College Composition and Communication*, 73(3), 410-436.

Brown, David West, & Wetzel Danielle Zawodny. (2023). *Corpora and rhetorically informed text analysis: The diverse applications of DocuScope*. Amsterdam, Netherlands: John Benjamins.

Clary-Lemon, Jennifer. (2009). The racialization of composition studies: Scholarly rhetoric of race since 1990. *College Composition and Communication*, 61(2), w1-w17.

Crowley, Sharon. (1985). The evolution of invention in current-traditional rhetoric: 1850-1970. *Rhetoric Review*, 3(2), 146-162.

D'Angelo, Frank. (1984). Nineteenth-century forms/modes of discourse: A critical inquiry. *College Composition and Communication*, 35(1), 31-42.

Davies, Mark. (2010). *The Corpus of Historical American English* (COHA). https://www.english-corpora.org/coha/

Davila, Bethany. (2022). White language supremacy in course descriptions. *College Composition and Communication*, 73(4), 640-664.

Devitt, Amy J. (2015). Motives and habits: Some thoughts on what linguistics can gain from rhetoric and composition. *Journal of English Linguistics*, 43(4), 334-340.

Devitt, Amy J. (2020). The blurred boundaries of genres-in-use: Principles and implications from rhetorical genre studies for English historical linguistics. In Peter J. Grund & Megan E. Hartman (Eds.), *Studies in the history of the English language viii: Boundaries and boundary-crossings in the history of English* (pp. 45-72). Berlin, Germany: Mouton De Gruyter.

Donahue, Christiane. (2018). "We are the 'other'": The future of exchanges between writing and language studies. [Special issue on transdisciplinary and translingual challenges for WAC/WID.] *Across the Disciplines*, 15(3), 130-143.

Ferris, Dana, & Eckstein, Grant. (2020). Language matters: Examining the language-related needs and wants of writers in a first-year university writing course. *Journal of Writing Research*, 12(2), 299-342.

Fredlund, Katherine. (2021). A feminist and antiracist history of composition and rhetoric at Oberlin College (1846-1851). *College Composition and Communication*, 72(3), 413-439.

Fulkerson, Richard. (2005). Composition at the turn of the twenty-first century. *College Composition and Communication*, 56(4), 654-687.

Gabrielatos, Costas. (2018). Keyness analysis: Nature, metrics and techniques. In Charlotte Taylor & Anna Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 225-258). London, UK: Routledge.

Gere, Anne Ruggles, Curzan, Anne, Hammond, J.W., Hughes, Sarah, Li, Ruth, Moos, Andrew, Smith, Kendon, Zanen, Kathryn Van, Wheeler, Kelly L., & Zanders, Crystal J. (2021). Communal justicing: Writing assessment, disciplinary infrastructure, and the case for critical language awareness. *College Composition and Communication*, 72(3), 384-412.

Gere, Anne Ruggles, Aull, Laura, Escudero, Moisés Damián Perales, Lancaster, Zak, & Lei, Elizabeth Vander. (2013). Local assessment: Using genre analysis to validate directed self-placement. *College Composition and Communication*, 64(4), 605-633.

Goggin, Maureen Daly. (2000). *Authoring a discipline: Scholarly journals and the post-World War II emergence of rhetoric and composition*. Mahwah, NJ: Erlbaum.

Grund, Peter J., & Hartman, Megan E. (Eds). (2020). *Studies in the history of the English language VIII: Boundaries and boundary-crossings in the history of English*. Berlin, Germany: Mouton de Gruyter.

Haswell, Richard H. (2005). NCTE/CCCC's recent war on scholarship. *Written Communication*, 22(2), 198-223.

Hyland, Ken. (2001). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes*, 20(3), 207-226.

Kilgarriff, Adam. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), 97-133.

Kohnen, Thomas. (2001). Text types as catalysts for language change. In Hans-Jürgen Diller & Manfred Görlach (Eds.), *Towards a history of English as a history of genres* (pp. 111-124). Heidelberg, Germany: The Universitätsverlag Winter.

Kytö, Merja, & Culpeper, Jonathan. (2006). *Corpus of English Dialogues 1560-1760* (CED).

Kytö, Merja, & Pahta, Päivi. (Eds). (2016). *The Cambridge handbook of English historical linguistics*. Cambridge, UK: Cambridge University Press.

Kytö, Merja, & Smitterber, Erik. (2015). Diachronic registers. In Douglas Biber & Randi Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 330-345). Cambridge, UK: Cambridge University Press.

Lancaster, Zak. (2016). Do academics really write this way? A corpus investigation of moves and templates in "they say/I say." *College Composition and Communication*, 67(3), 437-464.

Leech, Geoffrey. (2007). New resources, or just better old ones? The holy grail of representativeness. In Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer (Eds.), *Corpus linguistics and the web* (pp. 133-149). Amsterdam, Netherlands: Rodopi.

Longaker, Mark Garrett, Kreuter, Nate, Dadugblor, Stephen Kwame, Foltz, Hannah, Hooker, Tristin Brynn, Karnes, Martha Sue, Radcliff, Bethany Caye, Schaeffner, KJ, & Walker, Kiara. (2022). Archiving our own: The digital archive of rhetoric and composition at the University of Texas at Austin, 1975-1995. *College Composition and Communication*, 73(4), 774-805.

MacDonald, Susan. (2007). The erasure of language. *College Composition and Communication*, 58(4), 585-625.

Marín, María José. (2014). Evaluation of five single-word term recognition methods on a legal English corpus. *Corpora*, 9(1), 83-107.

Matsuda, Paul Kei. (1999). Composition studies and ESL writing: A disciplinary division of labor. *College Composition and Communication*, 50(4), 699-721.

Matsuda, Paul Kei. (2013). It's wild out there: A new linguistic frontier in U.S. college composition. In Suresh Canagarajah (Ed.), *Literacy as translingual practice: Between communities and classrooms* (pp. 128-138). New York, NY: Routledge.

McEnery, Tony, & Hardie, Andrew. (2012). *Corpus linguistics: Method, theory, and practice*. Cambridge, UK: Cambridge University Press.

Milroy, James. (1992). *Linguistic variation and change*. Oxford, UK: Blackwell.

Mueller, Derek. (2012). Grasping rhetoric and composition by its long tail: What graphs can tell us about the field's changing shape. *College Composition and Communication*, 64(1), 195-223.

NCTE. (2024). About College Composition and Communication. https://publicationsncte.org/content/journals/ccc

Peary, Alexandria. (2009). The licensing of the poetic in nineteenth-century composition-rhetoric textbooks. *College Composition and Communication*, 61(2), pp. 149-176.

Phillips, Donna Burns, Greenberg, Ruth, & Gibson, Sharon. (1993). *College Composition and Communication*, 44(4), 443-465.

R Development Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing. http://www.r-project.org

Raucci, John. A replication agenda for composition studies. (2021). *College Composition and Communication*, 72(3), 440-461.

Rissanen, Matti. (2012). Corpora and the study of the history of English. In Merja Kytö (Ed.), *English corpus linguistics: Crossing paths* (pp.197-220). Amsterdam, Netherlands: Rodopi.

Ritter, Kelly. Before Mina Shaughnessy: Basic writing at Yale, 1920-1960. (2008). *College Composition and Communication*, 60(1), 12-45.

Scott, Zaluda. (1998). Lost voices of the Harlem Renaissance: Writing assigned at Howard University, 1919-31. *College Composition and Communication*, 50(2), 232-257.

Shapiro, Shawna. (2022). *Cultivating Critical language awareness in the writing classroom*. New York, NY: Routledge.

Silva, Tony, & Leki, Ilona. (2004). Family matters: The influence of applied linguistics and composition studies on second language writing studies—past, present, and future. *The Modern Language Journal*, 88(i), 1-13.

Smith, Kevin G., Girdharry, Kristi, & Gallagher, Chris W. (2021). Writing transfer, integration, and the need for the long view. *College Composition and Communication*, 73(1), 4-26.

Smitterberg, Erik. (2016). Extracting data from historical materials. In Merja Kytö & Päivi Pahta (Eds.), *The Cambridge handbook of English historical linguistics* (pp. 181-199). Cambridge, UK: Cambridge University Press.

Stewart, Donald C. (1985). The status of composition and rhetoric in American colleges, 1880-1902: An MLA perspective. *College English*, 47(7), 734-746.

Stratton, James M. (2021). Corpora and diachronic analysis of English. In Eric Friginal & Jack A. Hardy (Eds.), *The Routledge handbook of corpus approaches to discourse analysis* (pp. 202-218). New York, NY: Routledge.

Stubbs, Michael. (1996). *Text and corpus analysis: Computer assisted studies of language and culture*. London, UK: Blackwell.

Taavitsainen, Irma, & Pahta, Päivi. (2010). *Corpus of Early English Medical Writing* (CEEM).

Walker, Terry. (2007). *Thou and you in early modern English dialogues*: *Trials, depositions, and drama comedy*. Amsterdam, Netherlands: John Benjamins.

White, Linda Feldmeier. (2002). Learning disability, pedagogies, and public discourse. *College Composition and Communication*, 53(4), 705-738.

Yancey, Kathleen Blake. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, 50(3), 483-503.

Notes

[1] I would like to thank Peter Grund, Chris Palmer, and the anonymous reviewer for their helpful comments on earlier versions of this article.

[2] Scholars may have different understandings of what the term *writing studies* refers to. In this article, *writing studies* references the field that focuses on college-level composition instruction in the United States, and the term is roughly equivalent to *rhetoric and composition* or *composition studies*.

[3] See Devitt (2020) for discussions of how RGS can contribute specifically to HEL research.

[4] In this article, the words *history* (and its adjectival form *historical*) and *diachrony* (and its adjectival form *diachronic*) not only refer to a particular stage in the past but also to changes that happened over certain time periods.

[5] Note that scholars may have different perspectives on how a diachronic corpus achieves its representativeness. For example, Leech (2007) is different from Biber et al. (1998) in terms of whether proportional sampling is necessary.

[6] As explained below, lemmatization was not performed in my study. The frequency only represents how often the word *history* appears rather than how often *history* in addition to other words that share similar forms as *history*, such as *historical*, *histories*, *historicize*, and *historicizing*, appear in the corpus.

[7] A diachronic corpus-based approach belongs to one of the distant reading methods defined by Mueller (2012).

[8] Works cited pages were not included in the corpus because they could have skewed the results of the analysis.

[9] The main reason why R was used instead of an existing corpus analysis tool is that R enabled keyness analysis to be conducted using TF-IDF.

[10] A discussion of different metrics for keyness is beyond the scope of this study. See Gabrielatos (2018) for a detailed review.

[11] It must be noted that my periodization is subjective and mainly for illustrative purposes of the practice of a diachronic corpus-based study. If a different time boundary is set, the results are likely to be different. See Grund & Hartman (2020) for more discussions (p. 9). In addition, a more traditional way to study keyness is to investigate keywords by comparing a corpus to a different corpus, often a multi-purpose corpus, instead of dividing a corpus into sub-corpora and measuring the keyness of each sub-corpus against the other sub-corpora. However, it is only through dividing my corpus into sub-corpora and measuring the keyness of each sub-corpus against the others that the changing disciplinary foci over time can be appropriately identified.

[12] TF refers to how many times a word appears in a document (in this study, in a sub-corpus), and IDF refers to the number of documents the word appears in a collection of documents (in this study, the number of the sub-corpora).

[13] TF-IDF scores are calculated by a multiplication of TF (term frequency) and IDF (inverse document frequency); when a word appears frequently enough in a given corpus, it will have a high TF, but when this word shows up not only in that corpus but also in its reference corpora, its IDF score will be 0 (DF is 1 when the word appears in not only the corpus but all its reference corpora; IDF is a logarithmic score, and log1=0). A detailed discussion of how TF-IDF works mathematically is beyond the scope of this study. See Kilgarriff (2001) and Marín (2014) for more details.

[14] See different types of keyness thresholds in Gabrielatos (2018).

[15] A word must have both high TF as well as high IDF in order to have a high TF-IDF value, but for a large size corpus, no words can have high TFs (TF=word frequency/total words in the corpus, and a value of TF is almost always lower than 0.05 even for those extremely high frequency grammatical words, such as *their*, *on*, *by*, *also*, and *but*, whose TF-IDF values are always 0), and since the highest IDF value in this study can only be around 1.60943 ($\ln 5/1 \approx 1.60943$), a reasonable TF-IDF range is likely to be only between 0 to 0.0005.

[16] Due to the length limit and illustrative purpose of this study, concordance analysis was not performed for all the most frequent ngrams in the corpus nor all the keywords in each sub-corpus. Concordance analysis

was only carried out for ngrams and keywords that have the potential to further reveal something about the changing disciplinary trends between 1997 and 2022. See Chapter 4 in Baker (2006) for a detailed overview of how to carry out a concordance analysis.

[17] Note that Fulkerson (2005) actually argues for an axiological divergence in writing studies at the turn of the twenty-first century, which cannot be clearly observed in my exploratory study.

[18] The keywords model used in this study could arguably disadvantage the sub-corpus of 1997-2001 and advantage the sub-corpus of 2017-2022 because later publications could model their language based upon earlier ones but not vice versa. However, according to Stubbs (1996), "[n]o terms are neutral. Choice of words expresses an ideological position" (p. 107). While it is possible for later publications to model the language used in earlier publications, they also don't have to. This can also explain why there seems to be a terminology shift from *fyc* to *fyw* in the most recent years.

[19] One way to preprocess the corpus is to look at the dispersion of the potential keywords in the corpus. Dispersion indicates how evenly a language element is distributed throughout a corpus. If the distribution of a potential keyword is highly concentrated in a certain year or a certain article, it can be excluded from the corpus. See Chapter 3 in Baker (2006) for details about dispersion. To illustrate some common issues a keyness analysis may encounter, this study chose not to preprocess the corpus.

[20] A keyness analysis of trigrams was also carried out, but the results were not fruitful enough to be included in this study.

Contact

Wen Xin
Assistant Teaching Professor
Department of English
University of Kansas
Email: wenxin@ku.edu

Complete APA Citation

Xin, Wen. (2024, December 31). What can the history of the English language research offer? A diachronic corpus-based approach to research in writing studies. [Special issue on *Confluences of Writing Studies and the History of the English Language*] *Across the Disciplines*, 21(2/3), 213-236. https://doi.org/10.37514/ATD-J.2024.21.2-3.09