RESEARCH ISSUES

TESOL Quarterly publishes brief commentaries on aspects of qualitative and quantitative research. For this issue, we asked two researchers to discuss corpus-based research in TESOL.

Edited by PATRICIA A. DUFF
University of British Columbia

Corpus-Based Research in TESOL

Quantitative Corpus-Based Research: Much More Than Bean Counting

DOUGLAS BIBER
Northern Arizona University
Flagstaff, Arizona, United States

SUSAN CONRAD
Iowa State University
Ames, Iowa, United States

■ The first language corpora were compiled decades ago (e.g., the Brown Corpus was begun in 1962; see Francis & Kucera, 1979), and many corpus-based linguistic studies have been conducted since that time (see Altenberg, 1991, for a bibliography). Some of the earliest uses of corpus linguistics were for applied purposes, especially the compiling of dictionaries (see, e.g., Sinclair, 1987). More recently, an increasing number of corpus-based studies have made important connections with TESOL. In fact, in the past 4 years three contributions to *TESOL Quarterly* have used corpus-based techniques (Conrad, 2000; Coxhead, 2000; Hughes & McCarthy, 1998).

The unifying characteristics of corpus-based research include the use of a large, representative electronic database of spoken or written texts, or both (the corpus), and the use of computer-assisted analysis techniques. (For an introduction to corpus linguistics, including the importance of corpus design, see Biber, Conrad, & Reppen, 1998; Kennedy, 1998.)

Although corpora are valuable for providing natural examples of words or grammatical features in context, corpus linguistics offers a unique perspective because of its use of quantitative analyses, which allow researchers to investigate patterns of language use that are otherwise impossible to ascertain. Contrary to the appearance that quantitative corpus analyses consist of elaborate bean counting, our investigations point to two major generalizations that are crucial for ESL/EFL teaching:

1. *the centrality of register for studies of language use.* Strong patterns of use in one register often represent only weak patterns in other registers. If linguists are to undertake a complete analysis of grammatical patterns, they must consider the patterns of use across registers, and learners can often benefit from this information.
2. *the unreliability of intuitions about use.* Teachers, authors, and testing professionals constantly rely on their intuitions to choose the most important words and structures to focus on. However, corpus studies show that such intuitions about use are often incorrect.

We illustrate the usefulness of quantitative corpus-based research with two analyses adapted from the *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad, & Finegan, 1999). The analyses are based on approximately 20 million words from four registers: conversation, fiction, newspaper language, and academic prose (see Biber et al., 1999, chapter 2, for a complete description of the corpus). The first example—identifying the most common verbs—illustrates how some simple counts can have important implications for language teaching. The second example looks at a grammatical distribution—simple, progressive, and perfect aspect—with equally important implications for language pedagogy.

COMMON LEXICAL VERBS ACROSS REGISTERS

There are literally dozens of common lexical verbs in English. For example, nearly 400 different verbs occur over 20 times per million words (see Biber et al., 1999, pp. 370–371). These include many everyday verbs, such as *pull, throw, choose,* and *fall.*

Given this large inventory of common verbs, one might assume that no individual verbs stand out as being especially frequent—and it would be impossible to determine whether this assumption is true without a corpus-based study. However, calculating the frequency of verbs is a simple task for corpus research. Surprisingly, only 63 lexical verbs occur more than 500 times per million words in a register, and only 12 verbs occur more than 1,000 times per million words (Biber et al., 1999, pp. 367–378). These 12 most common verbs are *say, get, go, know, think, see,*

make, come, take, want, give, and *mean*. (The primary verbs *be* and *have* are also extremely common.)

With texts within the corpus coded by register, we can also easily compare the frequency of particular verbs across registers. The analysis shows that the 12 most common verbs are especially important in conversation, where they account for almost 45% of the occurrences of all lexical verbs. In contrast, these verbs account for only 11% of lexical verbs in academic prose.

Moving from quantitative to more qualitative interpretation of use, the corpus analysis can clarify the functions of these verbs. Many, such as *get*, have diverse functions whereas others, such as *say*, have a single primary function related to an activity. In both cases, it makes sense to introduce these verbs to students early on, as these are the ones students will most often hear in their day-to-day interactions with native speakers. However, low-level ESL grammar books tend not to cover these verbs, instead introducing activity verbs like *eat, play, work, run, travel*, and *study*. Although these verbs have more concrete meanings relating to activity, they are much less common. Thus, even simple quantitative analyses can provide important information that teachers and material writers can use to revise current teaching materials.[1]

ASPECT ACROSS REGISTERS

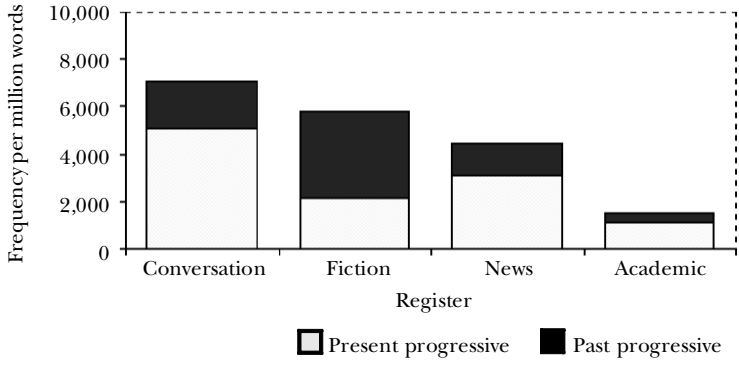
One of the most widely held intuitions about language use among English language professionals is that progressive aspect is the unmarked choice in conversation. As a result, ESL grammars usually introduce the present progressive in chapter 1, often before discussion of the simple present tense. Quantitative corpus-based research is a perfect tool for investigating such a belief.

As Figure 1 shows, progressive aspect is indeed more common in conversation than in other registers (Biber et al.,1999). The contrast with academic prose is especially noteworthy: Progressive aspect is rare in academic prose but common in conversation. However, as Figure 2 shows, it is not at all correct to conclude that progressive aspect is the unmarked choice in conversation. Rather, simple aspect is clearly the unmarked choice. In fact, in conversation simple-aspect verb phrases are more than 20 times as common as progressives are.

[1] One counterargument goes as follows: The most common forms might be easy to acquire, so it makes sense to continue to focus on less common forms. However, current pedagogical practice does not support this argument. Textbooks often present forms that have concrete meanings and are easy to learn. In contrast, many extremely common forms are likely to be less noticeable because they express more subtle meanings, so they are not at all easy to acquire. Thus, verbs *like eat, play, work*, and *run* tend to have their literal activity meanings in conversation whereas the much more common verbs, like *get, go, see*, and *make*, have an extremely wide range of meanings and functions.

FIGURE 1

Frequency of Present Progressive and Past Progressive in Four Registers

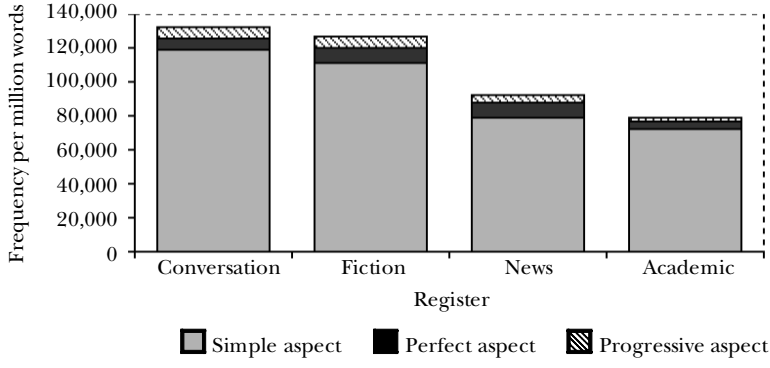


Note. From *Longman Grammar of Spoken and Written English* (p. 461), by D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, 1999, Harlow, England: Longman. Copyright © 1999 by Pearson Education. Adapted with permission.

With a corpus-based study, researchers can analyze associations between grammatical features and lexical items. Thus, it is a relatively simple matter to determine whether there are associations between progressive aspect and particular verbs. In fact, in conversation, a few lexical verbs—including *bleeding, chasing, shopping, starving, joking, kid-*

FIGURE 2

Frequency of Simple, Perfect, and Progressive Aspect in Four Registers



Note. From *Longman Grammar of Spoken and Written English* (p. 462), by D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, 1999, Harlow, England: Longman. Copyright © 1999 by Pearson Education. Adapted with permission.

ding, and *moaning*—occur most of the time with the progressive aspect. However, the norm—even in conversation—is to express verbs with the simple aspect. In marked contrast to the expectations created by many popular grammars, verb phrases such as *I've been having* and *is always telling* are exceptions rather than the rule.

CO-OCCURRENCE

These two simple case studies illustrate how quantitative corpus research can have direct implications for pedagogical practice. Additionally, corpus analyses have been used to describe more complex patterns. The investigation of lexicogrammar is one such area of research, showing how related grammatical constructions tend to be used with different sets of words (e.g., the most common verbs, adjectives, and nouns controlling *that*-clauses vs. *to*-clauses). The discourse factors influencing the choice among grammatical variants can also be studied from a corpus perspective (e.g., the factors favoring the choice of *that* or *which* as a relative pronoun).

In addition, quantitative corpus-based techniques have allowed researchers to entertain new kinds of research questions, investigating issues that had previously been considered intractable. For example, sociolinguists had long recognized that linguistic co-occurrence is central to an understanding of register variation (see, e.g., Ervin-Tripp, 1972), but they lacked research techniques to identify the sets of co-occurring linguistic features. Quantitative corpus research has filled this gap, using multivariate statistical techniques to identify basic dimensions of co-occurring linguistic features and to analyze the similarities and differences among registers with respect to those dimensions (see, e.g., Biber, 1988; Conrad & Biber, 2001).

CONCLUSION

We have identified some of the research made possible by corpus-based techniques together with some of the obvious pedagogical implications of that research (see Conrad, 1999, 2000, for a fuller discussion of these implications). Although we have given only a brief introduction, we have illustrated the surprising findings that emerge from quantitative corpus research: Language professionals often fail to notice the most common forms and, as a result, these forms are often slighted in teaching. In the absence of other compelling factors (e.g., learnability at a given stage or basic knowledge required as a building block for later instruction), we would argue that dramatic differences in frequency should be among the most important factors influencing pedagogical decisions. However, only with the recent availability of corpus-based

findings—including quantitative analyses—are language professionals in a position to begin this synthesis of research and practice.

THE AUTHORS

Douglas Biber is Regents' Professor of English (Applied Linguistics Program) at Northern Arizona University. His research has focused on register variation, English grammar, and corpus linguistics. His books include *Variation Across Speech and Writing* (Cambridge University Press, 1988), *Dimensions of Register Variation: A Cross-Linguistic Comparison* (Cambridge University Press, 1995), and the coauthored or coedited volumes *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge University Press, 1998), *The Longman Grammar of Spoken and Written English* (Longman/Pearson, 1999), and *Multi-Dimensional Studies of Register Variation in English* (Pearson, 2001).

Susan Conrad is an associate professor in the Department of English and Program in Linguistics at Iowa State University. Her publications in corpus linguistics include the coauthored *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge University Press, 1998) and *Longman Grammar of Spoken and Written English* (Pearson, 1999), and the coedited collection *Variation in English: Multi-Dimensional Studies* (Pearson, 2001).

REFERENCES

Altenberg, B. (1991). A bibliography of publications relating to English computer corpora. In S. Johansson & A.-B. Stenström (Eds.), *English computer corpora* (pp. 355–396). Berlin: Mouton de Gruyter.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

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

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

Conrad, S. (1999). The importance of corpus-based research for language teachers. *System, 27,* 1–18.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34,* 548–560.

Conrad, S., & Biber, D. (2001). *Variation in English: Multi-dimensional studies*. London: Pearson.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34,* 213–238.

Ervin-Tripp, S. (1972). On sociolinguistic rules: Alternation and co-occurrence. In J. J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics* (pp. 213–250). New York: Holt.

Francis, W. N., & Kucera, H. (1979). *Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with digital computers*. Providence, RI: Brown University, Department of Linguistics.

Hughes, R., & McCarthy, M. (1998). From sentence to discourse: Discourse grammar and English language teaching. *TESOL Quarterly, 32,* 263–287.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Sinclair, J. (Ed.). (1987). *Looking up: An account of the COBUILD Project in lexical computing*. London: Collins ELT.