Available online at www.sciencedirect.com

ScienceDirect

Journal of Second Language Writing 26 (2014) 28–41

JOURNAL OF SECOND LANGUAGE WRITING

Quantifying the development of phraseological competence in L2 English writing: An automated approach

Yves Bestgen, Sylviane Granger*

Centre for English Corpus Linguistics, Université catholique de Louvain, Belgium

Abstract

Based on the large body of research that shows phraseology to be pervasive in language, this study aims to assess the role played by phraseological competence in the development of L2 writing proficiency and text quality assessment. We propose to use CollGram, a technique that assigns to each pair of contiguous words (bigrams) in a learner text two association scores (mutual information and *t*-score) computed on the basis of a large reference corpus, the Corpus of Contemporary American English. Applied to the Michigan State University Corpus of second language writing, CollGram shows a longitudinal decrease in the use of collocations made up of high-frequency words that are less typical of native writers. It also shows that the mean MI scores of the bigrams used by L2 writers are positively correlated with the quality of the essays, while there is a negative correlation between the quality of the texts and the proportion of bigrams that were absent in the reference corpus, most of which were shown to be erroneous. The conclusion discusses the marked differences in the effects revealed by the longitudinal and pseudolongitudinal analyses, the limitations of the study, and some potential implications for the teaching and assessment of second language writing.

Keywords: Phraseology; *n*-Gram; Collocation; Association measure; L2 learner corpus; Writing assessment

Introduction

Second language acquisition (SLA) has traditionally focused more on how L2 learners acquire morphology and grammar than lexis:

> the focus has been on how learners acquire grammatical sub-systems, such as negatives or interrogatives, or grammatical morphemes such as plural {s} or the definite or indefinite articles. Research has tended to ignore other levels of language. A little is known about L2 phonology, but almost nothing about the acquisition of lexis. (Ellis, 1985, p. 5)

Although the situation has started to change in recent years, lexical indices of language development are still less frequently used than syntactic measures such as T-unit length or percentage of error-free T-units. In other fields, however, lexis has come to occupy a central position. Corpus linguistics, for example, is largely lexical, probably

* Corresponding author at: Centre for English Corpus Linguistics, Université catholique de Louvain, 1 Place Blaise Pascal, B-1348 Louvain-la-Neuve, Belgium. Tel.: +32 10474947.

E-mail address: sylviane.granger@uclouvain.be (S. Granger).

http://dx.doi.org/10.1016/j.jslw.2014.09.004

because of the ease with which lexical items and lexico-grammatical patterns can be extracted, sorted, and analyzed. In the field of foreign language teaching, Lewis's (1993) "Lexical Approach" which is based on the idea that "language consists of grammaticalized lexis, not lexicalized grammar," has led to a growing lexicalization of the teaching syllabus. The notion of lexis that underlies these approaches is phraseological; in other words, it goes beyond the study of single words to include a wide range of multi-word units. The field of phraseology, that is "the study of the structure, meaning and use of word combinations" (Cowie, 1994, p. 3168), has undergone a profound transformation in recent years. Long confined to the fringes of language study, it is now moving centre stage. There is growing recognition that besides being governed by grammatical and semantic rules, language production also largely relies on pre-patterned segments, a tendency that Sinclair (1991) has termed the "idiom principle," in opposition to the "open choice principle," and defined as follows: "the principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (p. 110). Corpus linguistic tools and methods have helped uncover a much wider range of word combinations than has previously been analysed: Besides traditional units such as idioms (*to spill the beans*), compounds (*red tape*) or phrasal verbs (*give up*), which are characterized by a high degree of syntactic fixedness and semantic non-compositionality, corpus techniques have brought to light several types of sequences that stand out by their high degree of co-occurrence and recurrence rather than their fixedness or opacity. These include collocations, that is, words that co-occur frequently within a short distance of each other in a text (Sinclair, 1991, p. 170), like *grow* + *old, turn* + *blue, dramatic* + *increase*, and lexical bundles, that is, the most frequent recurring sequences of words in a register (Biber, Johansson, Leech, Conrad, & Finegan, 1999, Chap. 13), for example *you see what I mean* in conversation or *it should be noted* in academic writing.

If, as demonstrated by corpus linguistic studies, phraseology is pervasive in language, it is essential to study its role in L2 writing development. As pointed out by Li and Schmitt (2009), "learning to write well also entails learning to use formulaic sequences appropriately," and "L2 learners' failure to use native-like formulaic sequences is one factor in making their writing feel nonnative" (p. 86). More precisely, it has been shown that L2 writers use less diverse formulaic sequences than native writers (De Cock, Granger, Leech, & McEnery, 1998) and overuse the ones they master best (Granger, 1998; Li & Schmitt, 2009).

Coxhead and Byrd (2007, pp. 134–135) advance three reasons that justify a stronger focus on formulaic sequences in L2 academic writing classes based on the analysis of corpus data: (1) using ready-made sequences is easier for students than composing sentences word by word; (2) formulaic sequences are defining markers of fluent academic writing; (3) being at the boundary between lexis and grammar, formulaic sequences are much easier to detect on the basis of corpus data than through the analysis of individual texts.

The kinds of questions we need to address with respect to the role played by phraseology in L2 writing include the following: Do L2 writers use phraseological units? What types of units do they use? How does phraseological competence develop over time? To what types of difficulties do multiword units give rise? Are phraseological errors due to transfer from the learners' mother tongue? A wide range of studies have attempted to answer these questions in recent years (for an overview, see Ebeling & Hasselgård, in press; Ellis, Simpson-Vlach, Römer, Brook O'Donnell, & Wulff, in press; Paquot & Granger, 2012). A large number of these rely on computer learner corpora (i.e., large electronic collections of texts produced by foreign or second language learners), and make use of automatic techniques to extract multiword units. The *n*-gram method, which consists in extracting contiguous sequences of *n* words – two words for bigrams, three words for trigrams, etc. – is growing increasingly popular and has resulted in a large body of research on the use of lexical bundles by L2 writers. The data used is usually a combination of native and learner corpus data. Using a widely used method referred to as Contrastive Interlanguage Analysis (Granger, 1996), the learner corpus data is set against comparable native data with a view to uncovering the specificities of learner use, or against other samples of learner data in order to assess their degree of generalizability. A range of L2 English learner populations have been investigated in this way: French (De Cock et al., 1998), Lithuanian (Juknevičienė, 2009), Swedish (Groom, 2009), Japanese (Ishikawa, 2009) and Chinese (Chen & Baker, 2010), to cite just a few. Some studies compare written and spoken production (De Cock, 2000, 2007). Although the results of these studies are not directly comparable as they make use of different criteria to identify the relevant units, some general tendencies emerge: L2 writers rely on a more limited repertoire of lexical bundles than native writers; they overuse the bundles they are familiar with, often calqued on similar sequences in their L1, and underuse many of the native-like bundles; they also prove to have difficulty with register, introducing speech-like bundles in their formal writing.