Collocation in beginner learner writing: A longitudinal study

Anna Siyanova-Chanturia[*]

School of Linguistics and Applied Language Studies, Victoria University of Wellington, Kelburn Parade, Wellington 6140, New Zealand

ABSTRACT

The fact that second-language (L2) learners have problems with collocation is widely attested. However, few studies have investigated the development of L2 collocational knowledge longitudinally. Fewer still have employed more than a handful of participants. In addition, almost all studies to date have looked at advanced learners of English. Other L2 proficiencies and backgrounds have, by and large, been disregarded. The lacuna left by this paucity motivated the present study. Thirty-six Chinese beginner learners of Italian wrote a composition at the beginning (Level 1), in the middle (Level 2), and at the end (Level 3) of an intensive course. A small corpus of L2 Italian was compiled with the aim of investigating N + Adj combinations. Analyses revealed that the number of combinations at the beginning and at the end of the course was comparable; the number of higher frequency items, however, was greater in Level 3 compositions than Level 1 compositions. Importantly, analyses showed an increase in strongly associated collocations in Level 3 writings compared to Level 1 writings. Thus, Level 3 compositions contained not only more higher frequency combinations, but also more strongly associated collocations than did Level 1 compositions. Taken together, the study provides new insights into the development of L2 collocational competence.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Recent years have seen a surge in work on collocation in learner corpora. This is hardly surprising given that the appropriate use of collocation is now widely considered to be one of the key prerequisites for proficient language use (Cowie, 1998; Sinclair, 1991; Wray, 2002). Collocations — and other types of multi-word expressions — are ubiquitous in language; they have been found to comprise anywhere from 20 to over 50 percent of spoken and written native-speaker discourse (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Erman & Warren, 2000; Foster, 2001; Howarth, 1998; Sorhus, 1977). Biber et al. (1999) reported that multi-word speech constituted 28% of the spoken and 20% of the written discourse analysed. Erman and Warren (2000) and Howarth (1998) estimated that multi-word speech (collocations, idioms, multi-word verbs, etc.) amounted to 52.3% and 40% respectively of the written discourse they looked at. According to Pollio, Barlow, Fine, and Pollio (1977) and Glucksberg (1989), four multi-word expressions — of one kind or another — are produced by a native speaker in every minute of spoken discourse. Collocation is what makes native speakers' speech idiomatic, fluent and natural. It is also what often renders second language (L2) learners' speech awkward, unnatural and even odd. Indeed, it has been established that L2 learners have problems with collocation in their written and spoken language (Granger, 1998; Howarth, 1998; Nesselhauf, 2003, 2005). Some have argued that L2 learners rely on creativity and make "overliberal assumptions about

the collocational equivalence of semantically similar items" (Wray, 2002: 201—202). Others have suggested that collocations present a major problem for L2 learners because collocational competence does not develop in parallel with general vocabulary knowledge (Bahns & Eldaw, 1993). As a result, learners often underuse or misuse native-like expressions and use atypical word combinations instead.

2. Literature review

By far the most common method of investigation of L2 collocational patterns has been to compare word combinations extracted from a written learner corpus with those found in a comparable native speaker (L1) reference corpus, with the aim of evaluating learner collocations qualitatively (correctness and appropriateness) and/or quantitatively (number of L1 combinations vs. L2 combinations). Many of the studies employing this method have focused on learners of L2 English from the same first language (L1) background: Chinese (Chen & Baker, 2010), French (Gilquin, 2007; Granger, 1998; Henderson & Barr, 2010), German (Lorenz, 1999), Hebrew (Laufer & Waldman, 2011; Levitzky-Aviad & Laufer, 2013), Swedish (Bartning, Forsberg, & Hancock, 2009; Forsberg, 2010), and Russian (Siyanova & Schmitt, 2008). Fewer studies have compared the patterns of collocation usage in learners from a number of different L1 backgrounds. For example, Waibel (2008) and Alejo Gonzalez (2010) compared the use of phrasal verbs in argumentative essays written by German versus Italian, and Spanish versus Swedish learners, respectively; while Cross and Papp (2008) looked at Chinese versus German versus Greek learners' use of V + N collocations in writing (also see Paquot, 2010, for a small-scale analysis of collocation use across learner groups representing ten L1 backgrounds). However, it is not only L2 written production that has been of interest to researchers in the context of collocation learning and use; there have also been studies, albeit relatively few, into collocation use in learner speech. For example, Crossley and Salsbury (2011) recorded casual conversations to investigate learners' use of lexical bundles, while De Cock (2004) focussed on two-to-six word recurrent strings in a spoken corpus of L2 English. In another study, Gotz and Schilk (2011) compared three-word lexical bundles in English as a foreign language (EFL) versus English as a second language (ESL) varieties of spoken English. Finally, Bartning et al. (2009) and Forsberg (2010) analysed conventional sequences in advanced L2 spoken French — a rare example of the target L2 being non-English.

Clearly, the studies into L2 collocational knowledge differ in a number of methodological ways. Some have focussed on L2 written production, while others analysed L2 speech. Many compared L2 learners of the same mother tongue against an L1 reference corpus, whereas others compared L2 learners of two or even three different L1 backgrounds. In addition, researchers have focussed on a variety of multi-word expressions: Premodifier + N collocations (Durrant & Schmitt, 2009; Siyanova & Schmitt, 2008), V + N collocations (Altenberg & Granger, 2001; Cross & Papp, 2008), Intensifier + Adj collocations (Granger, 1998), lexical bundles (Chen & Baker, 2010; Crossley & Salsbury, 2011), phrasal verbs (Alejo Gonzalez, 2010; Waibel, 2008), lexical verbs (Granger & Paquot, 2009), and conventional sequences (Bartning et al., 2009; Forsberg, 2010). Learner corpus studies, however, also differ vastly in the way in which the corpus data were extracted and analysed. For example, some have used statistical measures, such as frequency, *t*-score and mutual information to identify and compare collocations in L2 and L1 corpora (Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Groom, 2009; Lorenz, 1999; Siyanova & Schmitt, 2008); others have employed an automated frequency-driven approach (Chen & Baker, 2010), or a frequency index (Crossley & Salsbury, 2011). A different approach altogether has been to use L1 judgements and collocation dictionaries to evaluate the formulaicity, appropriateness, and correctness of L2 collocations (Bartning et al., 2009; Forsberg, 2010; Howarth, 1996, 1998; Foster, 2001; Laufer & Waldman, 2011; Nesselhauf, 2005).

Because learner corpus studies markedly differ in the methods of data extraction and analysis, they correspondingly vary in the formal definition of collocation adopted. Some have relied on native speaker judgements of 'formulaicity' in their definition of collocation (Foster, 2001). Others have defined collocation as word combinations that are necessarily restricted in some way (Howarth, 1998; Nesselhauf, 2005). However, L1 intuitions would inevitably play a role in this approach, too. Other researchers have adopted a purely 'frequency-based' approach in which intuitions and native-speaker judgements play no role (Durrant & Schmitt, 2009). The frequency-based tradition defines collocation as "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991: 7). That is, two words can be considered a collocation if they are found together more often than their individual frequencies would predict (Jones & Sinclair, 1974; Manning & Schutze, 1999; also see Hoey, 2005; Sinclair, 2004; Stubbs, 1995).

As pointed out by Paquot and Granger (2012), the studies into collocational knowledge in learner corpora have yielded a wealth of interesting findings. However, methodological and conceptual differences, such as different types of corpora used and definitions of collocation adopted, various approaches to data extraction and analysis, a wide range of collocations and other kinds of multi-word expressions investigated, and the type of L2 learner make it difficult to directly compare many of the studies in question. Despite this heterogeneity in both data and methods, nevertheless, a number of common findings can be identified. Arguably, the most commonly reported finding is under-, over- or misuse of target L2 items compared to L1 reference data (Altenberg & Granger, 2001; Chen & Baker, 2010; De Cock, 2004; Durrant & Schmitt, 2009; Gilquin, 2007; Granger, 1998; Granger & Paquot, 2009; Henderson & Barr, 2010; Howarth, 1998; Laufer & Waldman, 2011; Nesselhauf, 2005; Paquot, 2010). For example, L2 learners in Durrant and Schmitt (2009) were found to underuse less frequent, strongly associated collocations, identified using mutual information (*densely populated*), but their use of frequent collocations, identified using frequency and *t*-score (*hard work*), was found to be comparable to that of L1 speakers. Patterns of underuse in L2 writing have also been observed within a specific domain. For example, Granger and Paquot (2009) investigated the use of lexical verbs in English for Academic Purposes writing of advanced EFL learners from a range of L1