Chapter 48
Constructing a Collocation Learning System From the Wikipedia Corpus

Shaoqun Wu
University of Waikato, New Zealand

Liang Li
University of Waikato, New Zealand

Ian H. Witten
University of Waikato, New Zealand

Alex Yu
Waikato Institute of Technology, New Zealand

ABSTRACT

The importance of collocations for success in language learning is widely recognized. Concordancers, originally designed for linguists, are among the most popular tools for students to obtain, organize, and study collocations derived from corpora. This paper describes the design and development of a collocation learning system that is built from Wikipedia text and provides language learners with an easy-to-use interface for looking up collocations of any word that occurs in Wikipedia. The use of this corpus exposes learners to contemporary, content-related text, and enables them to search for semantically related words for a given topic. The system organizes collocations by syntactic pattern, sorts them by frequency, and links them to their original context. The paper includes a practical user guide to illustrate how to use the system as a language aid to facilitate academic writing.

INTRODUCTION

Collocations are of great importance for second language learners: they play a key role in producing language accurately and fluently. In recent years, corpus-based collocation learning has aroused considerable interest from teachers and researchers (e.g. Boulton, 2010, 2012; Chambers & O'Sullivan, 2004; Chang, 2014; Chen, 2011; Daskalovska, 2015; Yeh, Li, & Liou, 2007; Yoon, 2008). Concordancers,

DOI: 10.4018/978-1-5225-7663-1.ch048

originally designed for linguists, are popular tools for students to explore corpora, particularly with a view to examining collocations. Support for learner use of corpora and concordancing is premised on the fact that exposure to a word and its associated lexical and grammatical patterns in different contexts allows learners to develop a greater sense of its form, meaning and use.

This paper describes the design and development of a collocation learning system, FlaxCLS. FlaxCLS is one of the key elements of the FLAX system (http://flax.nzdl.org), a self-access language learning system documented in Wu (2010), Wu, Franken, and Witten (2009, 2010), and Wu, Witten, and Franken (2010). FlaxCLS has two components: a collocation database built from three million Wikipedia articles comprising three billion words, and a simple interface for looking up collocations. The use of this text base allows learners to inspect typical language use in contemporary, content-related text. Wikipedia articles represent modern English in almost every area of art, life, and science, and includes emerging topics whose vocabulary is not covered by standard corpora such as the British National Corpus.

The term collocation has different definitions in the literature. We take a syntax-oriented approach in this paper that emphasises the grammatical structure of collocation (Firth, 1957; Nation, 2013; Nattinger & DeCarrico, 1992; Nesselhauf, 2004; Sinclair, 1991) and identifies collocations by syntactic structures (e.g. verb + noun, adjective + noun, noun + verb). FlaxCLS first downloads Wikipedia text, parses it, extracts useful syntactic-based word combinations (e.g., verb+noun, noun+noun, adjective+noun), organizes them by syntactic pattern, sorts them by frequency, and links them to their context sentences. Once this comprehensive collocation database is established, an easy-to-use and learner friendly interface is provided through which learners can seek collocations that include any given word and word type (verb, noun, adjective and adverb), or search for combinations of multiple words (e.g., play an extremely important role).

Furthermore, the concept structure of Wikipedia is used to retrieve semantically related words on a given topic, so that learners can seek topic-related key words and their collocations. For example, searching for animal testing yields related words like toxicity, drug, ethical, welfare, treatment, pain, and their collocations, such as toxicity tests, effect of the drug, ethical principles, animal welfare, potential treatment and pain relief.

The paper is organized as follows. First we examine the use of the Web corpus in collocation learning and rationalize the choice of Wikipedia articles as the primary source from which to build a collocation database. Next we consult the literature to see how concordancers are used to facilitate the inspection of collocations. We discuss limitations reported by researchers and teachers, and suggestions that have been made for learner friendly interfaces. We then describe the design principles underlying our system, including how collocations are extracted, organized and presented in a simple manner. Following that we briefly walk through how to use the online interface to explore collocations. Finally, we review a student guide that has been created to demonstrate its use in preparing essays, choosing appropriate words, using hedging and boosting devices, improving formality, and increasing text variation during writing.

USING THE WEB CORPUS

The web, a vast, contemporary, freely available corpus, has the potential to offer language learners authentic, representative language resources (e.g. Boulton, Jul 2012; Hundt, Nesselhauf, & Biewer, 2007; Kilgarriff & Grefenstette, 2003). Various concordance tools have been developed for Web search, including WebCorp (Renouf, Kehoe, & Banerjee, 2007), KwiCFinder (Fletcher, 2007), and WebBootCat