Characteristic Sentence Stems in Academic English Texts: Definition, Identification, and Extraction

Authors : Jingjie Li, Wenjie Hu

Abstract : Phraseological units in academic English texts have been a central focus in recent corpus linguistic research. A wide variety of phraseological units have been explored, including collocations, chunks, lexical bundles, patterns, semantic sequences, etc. This paper describes a special category of clause-level phraseological units, namely, Characteristic Sentence Stems (CSSs), with a view to describing their defining criteria and extraction method. CSSs are contiguous lexico-grammatical sequences which contain a subject-predicate structure and which are frame expressions characteristic of academic writing. The extraction of CSSs consists of six steps: Part-of-speech tagging, n-gram segmentation, structure identification, significance of occurrence calculation, text range calculation, and overlapping sequence reduction. Significance of occurrence calculation is the crux of this study. It includes the computing of both the internal association and the boundary independence of a CSS and tests the occurring significance of the CSS from both inside and outside perspectives. A new normalization algorithm is also introduced into the calculation of LocalMaxs for reducing overlapping sequences. It is argued that many sentence stems are so recurrent in academic texts that the most typical of them have become the habitual ways of making meaning in academic writing. Therefore, studies of CSSs could have potential implications and reference value for academic discourse analysis, English for Academic Purposes (EAP) teaching and writing.