This study contributes to ongoing discussions on how measures of lexical diversity (LD) can help discriminate between essays from second language learners of English whose work has been assessed as belonging to levels B1 to C2 of the Common European Framework of Reference (CEFR). The focus is in particular on how different operationalisations of what constitutes a “different word” (type) impact on the LD measures themselves and on their ability to discriminate between CEFR levels. The results show that basic measures of LD, such as the number of different words, the type-token-ratio and the Index of Guiraud (Guiraud 1954) explain more variance in the CEFR levels than sophisticated measures, such as D (Malvern et al. 2004), HD-D (McCarthy and Jarvis 2007) and MTLD (McCarthy 2005) provided text length is kept constant across texts. A simple count of different words was the best predictor of CEFR levels and explained 22 percent of the variance in overall scores on the Pearson Test of English Academic in essays written by 176 adult test takers (mean age 26.82; 39 different first languages). The data also show that the lemma is a more appropriate unit of counting in studies of lexical diversity than the word family. A lemma-based lemmatization principle was found to be more useful to disambiguate student levels on the CEFR than the word family-based principle or using texts that were not lemmatized. Therefore the conclusion was drawn that lemmatization should not erase derivational affixes, as they provide important information about test takers’ knowledge. Disambiguating between homographs did not add value to the measures’ ability to distinguish between CEFR levels because measures which did not take homographs into account obtained higher Eta Squared values.

Keywords: CEFR, lemmatization, lexical diversity, language testing.