
Researching lexical thresholds and lexical profiles across the Common European Framework of Reference for Languages (CEFR) levels assessed in the Aptis test

This is a summary of a report by Nathaniel Owen, Prithvi Shrestha and Stephen Bax as part of the ARAGs Research Online Series. See the full report at: www.britishcouncil.org/exam/aptis/research/publications/

WHAT WE LOOKED AT:

This project explored language profiles of test-takers' writing in the British Council's Aptis Writing test at different ability levels, benchmarked to the Common European Framework of Reference for Languages (CEFR). Specifically, the project explored vocabulary and metadiscourse (organisational features) use by students across a range of language proficiencies. We did this using quantitative measures of language use to describe the levels of the CEFR and to identify boundaries between levels of the framework. Examples of quantitative measures used in the study include simple descriptive measures, such as number of words and number of sentences, through to more complex measures of vocabulary use derived from corpus data. The research used the free online tool, Text Inspector (www.TextInspector.com), to analyse the learner writing. An additional rationale was to explore how useful quantitative measures might be in an automated scoring engine. Text measures which showed differences across levels of the CEFR were used to develop a statistical model to see how well they could predict which CEFR level each text was scored at.

The study addressed two research questions:

Research question 1: What are the lexical thresholds and lexical profiles of candidates taking the Aptis writing test across the Common European Framework of Reference for Languages (CEFR) levels?

Research question 2: To what extent and in what ways are the measures identified in RQ1 of value, or deficient, for the purposes of automated assessment of learner writing?

HOW WE DID IT:

To answer the research questions, this project used 6,407 samples of Aptis candidates' writing. This data, representing 65 countries, together with candidates' score data, were analysed in terms of their language use depending on the CEFR level awarded to the samples. Measures and datasets used in the analysis included measures such as number of words, number of sentences, number of syllables, average sentence length and average word length. The study also examined vocabulary use by comparing vocabulary used by test-takers to data from the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), the English Vocabulary Profile (EVP), the Academic Word List (AWL), and a bespoke corpus of metadiscourse markers used by *Text Inspector*. For a full list of measures used, please visit www.textinspector.com.

Researching lexical thresholds and lexical profiles across the Common European Framework of Reference for Languages (CEFR) levels assessed in the Aptis test

WHAT WE FOUND:

To address the first research question, a series of statistical tests were carried out and the results revealed differences in learner writing across levels of the CEFR. The findings reveal that the Aptis writing test provides evidence that text complexity changes systematically as the CEFR level of learners increases. The research investigated 110 measures of text complexity used by Text Inspector. Twenty-six measures were significant across all CEFR boundaries, including measures of text length (sentence, token and type count), and measures of sophistication (syllable count and number of words with more than two syllables). Fourteen of the 26 metrics represented vocabulary use. The study also presents evidence that analysis of metadiscourse markers independently of vocabulary is justified. Four measures of metadiscourse use were able to discriminate across all CEFR thresholds.

The second research question was investigated through a statistical test which used measures identified in relation to research question 1 to predict the CEFR level of each text. Twenty measures which were significant across multiple CEFR thresholds were used to build a statistical model which was trained on a sample of the data which held nationality data constant. This model was then used to predict the CEFR band of another sample of the data which also held nationality data constant. The statistical model revealed that lexical use metrics from the Cambridge Learner Corpus (CLC) were the most successful at predicting CEFR level, and the model was most successful in identifying the lowest and highest (A1 and C) level responses. However, the model struggled to accurately differentiate samples at A2, B1 and B2 levels of the CEFR, suggesting that other, organisational variables play a significant role in human judgements, which are not covered in this study.

This research study of the Aptis writing test suggests that Text Inspector and similar other software programs may be useful for test developers in several ways. They may be used to develop automated scoring engines. Alternatively, they may help to ensure better agreement between raters with regard to observable characteristics of learner writing. However, examination boards need to determine *a priori* which measures they believe are suitable. Data about learner writing could be introduced into test item writer guidelines or rating scale descriptors to assist human assessors in awarding scores. Information from text analysis may also prove useful in rater training to help those involved better understand test task characteristics which represent different proficiency levels at different levels of the CEFR.