A1 A1+ A2 A2+ B1 B1+ B2 B2+ C1 C1+ C2 C2+

Revolutionary CEFR classification algorithm determines correct level at 90% accuracy

Amsterdam, May 2019

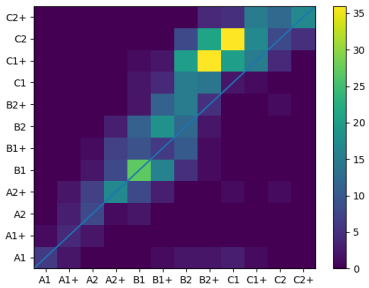
In collaboration with experienced English Language Teaching (ELT) authors, EDIA's artificial intelligence experts have created a revolutionarily precise algorithm that automatically classifies any text on the standard CEFR[1] language difficulty scale.

The training data consists of a variety of English texts ranging from academic papers and news articles to children's books. Each text was rated by at least three ELT experts with significant experience writing textbooks, graded readers, assessments and similar content for renowned ELT publishers.

The result is an algorithm that classifies texts at higher accuracy, consistency, and speed than an individual human expert.

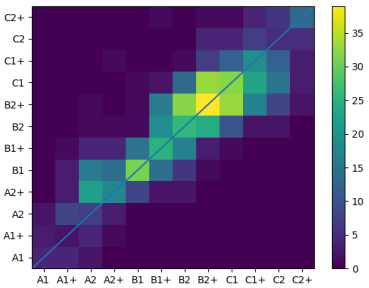
The algorithm is trained to minimise the distance to the median[2] of the domain experts, per text. It takes a variety of measures into account, ranging from role and frequency of individual words to the grammatical structure of sentences.

CLASSIFICATION BY AUTHORS



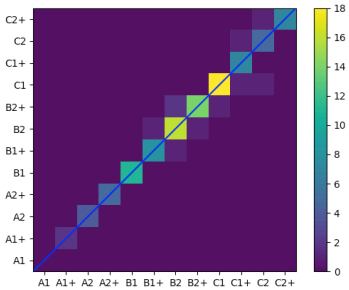
One of the less reliable experts vs the median of 2 other experts per text

CLASSIFICATION BY AUTHORS



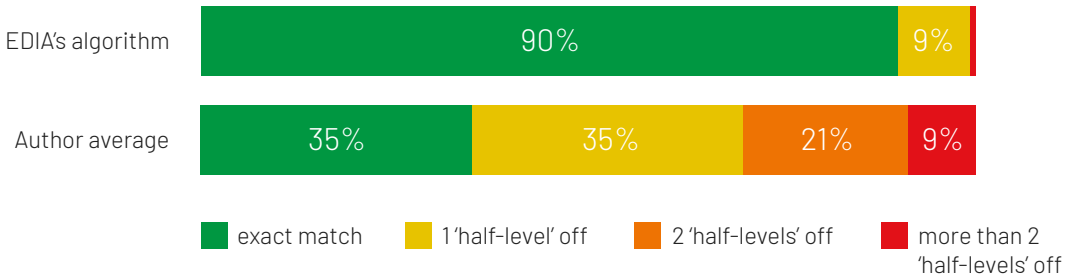
The most reliable expert vs the median of 2 other experts per text

CLASSIFICATION BY ALGORITHM



Our algorithm vs the median of 7 experts (3 experts per document)

To visualise results, we had our algorithm label texts that it has not analysed before and compared it to the median of the labels given by the three ELT experts. We also included a performance overview of the ELT experts compared to each other.



MULTILINGUAL TECHNOLOGY

We have also analysed our algorithm on German and Italian examples, provided by the Merlin[3] dataset. The results were comparable on both, showing that our technology has multilingual potential.

Want to learn more about EDIA's CEFR classifier?

Feel free to contact us via **edia.nl** In the mean time,why not **test it** yourself!

[1] The Common European Framework of Reference for Languages (CEFR) is an international language learning standard set by experts from the Council of Europe. The CEFR scale has levels ranging from A1 (beginner) to C2 (native).
[2] The CEFR median is the middle value of multiple CEFR ratings. For example, if 3 annotators rate a text A1, B1+ and B2 respectively, the median would be B1+ since it is both the second highest and second lowest difficulty rating for the text.
[3] The MERLIN dataset is an EU text corpus aligned to CEFR: https://merlin-platform.eu/C_mcorpus.php