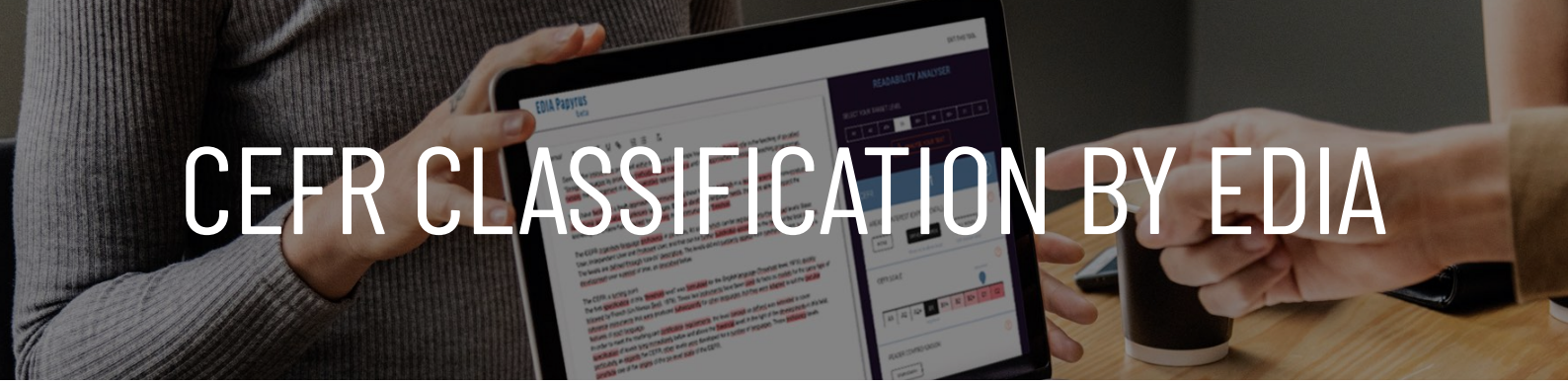
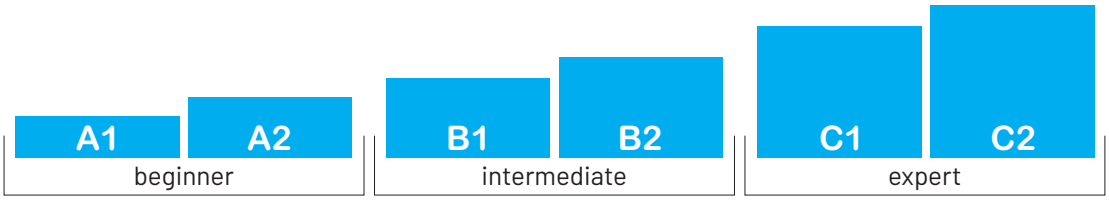


CEFR CLASSIFICATION BY EDIA

What is CEFR classification?

The Common European Framework of Reference (CEFR) is a system of language proficiency and ability. The framework defines readability of a text on a 6 point scale ranging from A1 to C2.



Our CEFR tagger is able to measure the readability of texts on the CEFR scale on a more granular level. That's why we use a 9 point CEFR scale (A1, A2, A2+, B1, B1+, B2, B2+, C1, C2).

Why is CEFR classification important? Who is it for? And what are the benefits?

The readability of a text is important in various ways. It ensures that a text is readable for the people you write for. In education teachers, writers, editors, content managers and publishers can be sure that the material they produce and use is on the right level, not only for ELT (English Language Teaching), but for all written content.

What does make the EDIA CEFR solution unique?

Most CEFR taggers look at the *individual* words in a text to assess the readability. These taggers use a dictionary in which each word is measured from A1 to C2. The readability of the text is concluded by combining the readability of individual words in the text. The words in the dictionary are given a rating without context. This method detects the right readability level in approximately 60% of cases. In essence, this approach is simply counting words. A text is so much more than the sum of the individual words, and individual words do not make the text.

How does the EDIA tagger work?

Our algorithm looks at many aspects of the text, such as parts of words[1], the words themselves, word combinations, sentence structure and the structure of the entire text. It does so by methods of deep learning and pre-trained language models. Instead of a mathematical formula that calculates based on a dictionary using single definitions of words, we believe that advanced neural networks are superior. Such algorithms are capable of taking into account every aspect of the text, not just the words. **Our algorithm is effective in more than 90% of the cases.**

How is the EDIA tagger made?

Our CEFR tagger is trained by taking texts from a variety of sources and reading levels. Each text is evaluated by multiple language experts, and the evaluations combined for our dataset. The CEFR tagger is trained in such a way that 20% of the texts remain unknown to our algorithm. This 20% is used to check the quality of our algorithm. We repeated this with a different 20% of the total texts to make sure the results are not a fluke (i.e. the tagger is good at predicting CEFR for this set of examples, but does not generalize to other texts).

---

[1] Morpheme, prefix, root and suffix

How can your CEFR ‘experts’ be right?

The texts are evaluated by experts, but how do we know these so-called experts are right?
The fact that experts agree on something is the key here, science builds upon the model of scientific consensus. That means that if a large majority of independent experts believe that something is true, it is considered to be true. And we work with the same principle. Furthermore, our machine learning is capable of reproducing such classification of texts to a high level. Validation of results is a crucial component of the scientific method.

How can the EDIA CEFR tagger be better than humans?

Our algorithm is better than humans. How can this be? Let’s consider an example. Radiologists assess your health by analysing MRI scans. In some cases, machine learning has shown to be better on this task. That is because the machine learning is trained by the knowledge of many experts, not just one. Machine learning is trained using a huge dataset, one that is impossible to digest by humans. Machine learning doesn’t need a coffee break and doesn’t have a bad night’s sleep.
As you can see, there are many similarities between this example to our technology, as it is trained by experts on a large, validated dataset and outperforms humans in CEFR classification.

With which software is the EDIA CEFR tagger compatible?

The EDIA CEFR tagger is available in the following content management systems (CMS) and editors: Microsoft Word, Google Docs, Alfresco, PublishOne, FontoXML, EDIA Papyrus etc. In principle, EDIA’s CEFR tagger can be integrated into any system that your organization uses.

How can you activate EDIA CEFR tagger in your workflow?

The EDIA CEFR tagger is available in a form of API that you can easily activate in one of the software solutions mentioned above. All you need to get started is a valid license key which can be requested through the EDIA sales team.

How many costs can I save by using EDIA's CEFR tagger?

On average it takes a human expert 5 minutes to label one piece of content. The hourly labour cost to tag manually 12 content items on average is € 48 (€ 4 per content item).
Our commercial proposition varies per volume of the content that you want to classify with EDIA CEFR tagger. Typically we charge 50 cents per 1 API call (= 1 content item of 200 words or 1,000 characters). The more articles need to be classified, the lower the price per content item.

1,000 content items
costs €40,000
10+ work days

1,000 content items
costs €500
10 minutes

For example, to classify 1,000 content items would cost € 40,000 and more than 10 workdays using manual tagging (excluding hours required to find the grading experts). By using the EDIA automated CEFR tagger it would take approximately 10 minutes for € 500, resulting in a **90% saving** for your organization, as well as significantly increased efficiency and time savings.

Contact information of EDIA sales team

Walter Montenarie
M +31 (0)6 2270 3526
E walter@edia.nl

For more technical explanation what’s behind the EDIA CEFR tagger visit **Papyrus**