Horizon Europe

D3.4 Annotation Schema

Funded by the European Union

The iRead4Skills – *Intelligent Reading Improvement System for Fundamental and Transversal Skills Development* is a Research & Innovation Action funded by the European Commission, Grant number: 1010094837, Topic HORIZON-CL2-2022-TRANSFORMATIONS-01-07 – Conditions for the successful development of skills matched to needs.

Document Control

Information:

Settings	Value
Deliverable No.	**D3.4**
Document Title:	D3.4 Annotation Schema
Author(s):	Raquel Amaro
Reviewer(s):	Thomas François
Sensitivity:	**Public**
Date:	30/10/2023

Document Location: The latest version of this controlled document is stored in OneDrive-fcsh.unl.pt/iRead4Skills/Project/Work Packages/WP3/Annotation.

TABLE OF CONTENTS

1. Goals

The goal of this annotation schema is to provide information on the decisions and methods followed for the classification and annotation task pursed in Work Package 3 (WP3): *Complexity classification and data*.

The accomplishment of the iRead4skills project encompasses a specific classification and annotation task to validate and correct/improve the final data set of WP3. The goal of this task is to ensure that the final data set is rich and useful for the subsequent train of machine learning algorithms for complexity classification. Besides validating the classification proposed (see Table 1) using texts from the project's first data set (D3.2), the annotation task will allow us to gather additional information on what specific parts of the texts the end-users consider relevant with regard to the text complexity and on the more relevant dimensions (lexicon, structure, topic).

Level	General description
Very easy	Texts that are fully or almost fully understood by everyone, including people with very low schooling (i.e., that did not finish the primary school (ca. 6th year)) and almost no reading experience. It roughly corresponds to CEFR A1 level.
Easy	Texts that are fully or almost fully understood by people with low schooling (i.e., that completed the primary school but do not have more than the 9th year) and have poor reading experience. It roughly corresponds to CEFR A2 level.
Clear	Texts that are understood the first time they are read by people that completed the 9th year and have a functional-to-average reading experience. It roughly corresponds to CEFR B1 level.

Table 1: iRead4Skills complexity levels

This requires a general annotation schema, defined following current and/or well-established proposals (Pustejovsky & Ide, 2017; Hovid & Lavid, 2010) and further specified to account for the needs of the project and to account for the specificities of the annotators (adults with low literacy skills).

We will consider the general annotation pipeline depicted in Figure 1, focusing on the annotation task definition.

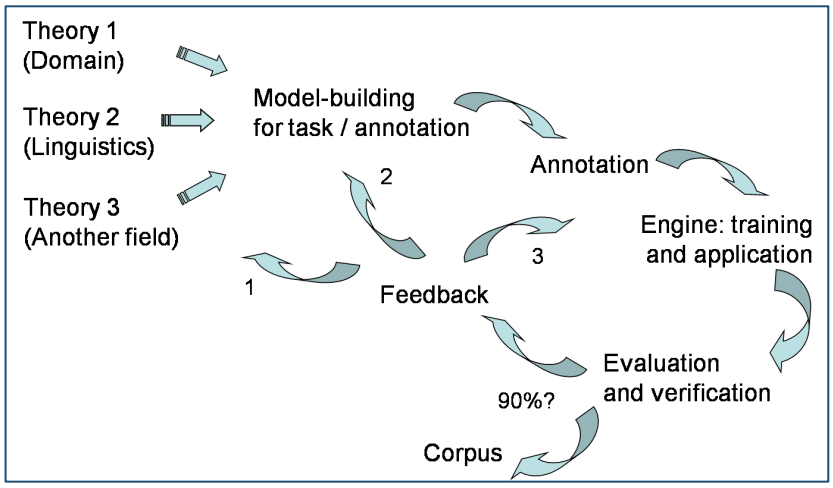

Figure 1: Generic annotation pipeline from Hovid & Lavid (2010: 5)

The model in use for this annotation task stems from the linguistics and education theories that frame adult

skills development and evaluation (e.g., PIACC 2016, ANQEP 2021) and the specific linguistic and communicative descriptors of language proficiency levels (PIACC 2016, ANQEP 2021 and Common European Framework of Reference for languages (CEFR) 2020) explicitly describing reading skills. The levels of complexity defined for the project were object of a specific task (see D3.1 Complexity levels) and provide us with the descriptions that form the basis for the annotation. However, given the target-audience that compose our set of annotators, the phenomena and features considered in the task are be annotated indirectly, i.e., the annotator will not be asked to use specific tags to mark the texts, rather to highlight the parts they consider complex.

Ideally, the annotation task will involve the participation of trainers and students per Adult Learning (AL)/Vocational Education Training (VET) centre. It may require preparatory meetings targeting the trainers and teachers involved, preferentially on site, to assure conditions of equipment. The trainers and teachers will also assist the adult learners annotators in the task.

2. Annotation schema

Data to train machine learning systems requires training data, manually annotated and/or validated.

This process consists of the following general steps (Hovy & Lavid, 2010; Finlayson & Erjavec 2017):

1. Determining the model (tag set, annotation units and formats).

2. Identifying and preparing a selection of material to be annotated.

3. Finding the appropriate annotation tools, platforms (i.e., system on which the tool can be run such as desktops, laptops, mobile phones) and access conditions (i.e., online or offline).

3. Preparing the instructions for the annotators (annotation guidelines).

4. Annotating experiences on fragments of the training corpus, to calibrate the instantiation and the guidelines.

5. Determining the relevant/satisfactory/possible level of agreement (low agreement means higher inconsistency in the annotation).

6. Completing the main annotation task.

7. Distributing the annotated data set, including export formats, results transformation (cleaning annotation, conversions, etc.), licensing, repositories, and data storage.

Other relevant issues also need to be addressed (Finlayson & Erjavec 2017):

i. How will the documents be prepared, including: formatting layouts (lines, paragraphs, page breaks, headings, highlighted text), file formats, normalization, etc.).

ii. How will annotators access the documents, including: document order, annotation order, number of annotators per document, etc.)

iii. When will inter-annotator agreement be measured and how.

iv. Which annotation tool/interface will be used and to which degree is it intuitive, easy to use, bug-free.

Table 2 presents the parameters and decisions taken for each of the general steps described above, including whenever possible information on the other relevant issues listed. More specific information is discussed in the following sections.

Process stage	Parameters and decisions
1. Model	Tag set: 1.1 document level: *Very easy; Easy; Clear; Complex* 1.2 text level: highlighting difficult words / expressions / lines 1.3 Additional information on complexity: word level; text structure level; topic level
2. Data to be annotated	(At least) 20% of the compiled data set 1 (ca. 400 texts) per language: 2.1 100 texts per complexity level (cf. 1.1) 2.2 5-10 texts from each communication domain (whenever possible) covered by the data set (e.g., personal communication, social media, fiction, didactic/scientific materials, institutional communication, etc.) Texts will be displayed in plain txt format, maintaining lines, paragraphs, and titles, with no highlights. Texts will be normalized to correct typos, spelling errors by the compilation team.
3. Tool and platform	Qualtrics (https://www.qualtrics.com/) Online survey with three types of questions 3.1 Attribute a level to the text (single choice) (trainers and students) 3.2 Mark in the text the difficult words/parts (select and highlight) (students) 3.3 Inform on the features related to the text's complexity concerning the words, the sentences and/or the topic (multiple choice) (trainers).
4. Annotator instructions	*D3.5 Annotation Manual* Simple and short guidelines to help annotators (focusing on low literacy adults).
5. Pre-task testing	The survey will be tested by trainers and students in AL and VET centres cooperating with MEC in Portugal (at least 5 people) and corrected whenever possible before application.
6. Inter-annotator agreement (IAA)	For the trainers annotations, IAA will be measured for the 3.1 subtask, where a given classification will be considered if it is attributed by 2 annotators at least (66,55%-100% IAA). IAA is not relevant for the annotations of the students (adults with low literacy skills). In this case, psycholinguistic measures apply and a higher level of variation is expected (Pirali et al. 2022: 48). Texts that are classified in 3 different levels will be further evaluated, considering the qualitative information collected in subtasks 3.2 and 3.3. Information collected in subtasks 3.2 and 3.3 will be analysed statistically, considering each set of results and considering the results per level of complexity.
7. Annotation task	At least 3 annotators per text; 10 to 50 texts per annotator, in a total of 24 to 120 annotators per language, depending on the availability. Texts of different levels of complexity will be displayed randomly. Classification options (1.1) will be presented in pre-defined order, from easier to more complex. Answer options concerning additional information (1.3) will be presented in

	pre-defined order: words, structure, topic. The task will be performed concurrently in AL and VET centres in Belgium, France, Portugal, and Spain, from November to January. It will also be disseminated in the project website and social media channels.
8. Distribution	Text/XML/CSV formats. Open access in Zenodo, by CC licensing, according to FAIR Data Principles (as per iRead4Skills Data Management Plan).

Table 2: iRead4Skills complexity levels

2.1. Tools, interface and results

The tool providing the annotation user interface should be intuitive, easy to use, and bug-free. The choice of the tool should be careful since it can constrain many subsequent decisions. Besides, according to Finlayson & Erjavec (2017:174), a tool for annotation should also provide information or functions for:

a. Capturing metadata, for instance which annotator annotated which document, which documents are already annotated, how long the annotator took, etc.

b. Data visualization: visualization of the tag sets, the data or the results is quite useful for the annotation task itself (that can be tedious), but also for managing tasks, such as knowing the current state of the annotation, perceiving errors, or assessing results.

c. Exporting to publication-quality formats: also related to data visualization and analysis is the ability to easily transforming annotated data into quality figures.

Although not originally devised for linguistic annotation/classification tasks (such as TANGO, MAE or LX-Sense Annotator[1], for instance), Qualtrics demonstrated to be a very user-friendly tool, intuitive for the developer and quite versatile in terms of data visualization for the annotators. It fulfils all the above requirements:

i) It allows for IAA measuring, by presenting absolute and relative results for each answer (Figure 2).

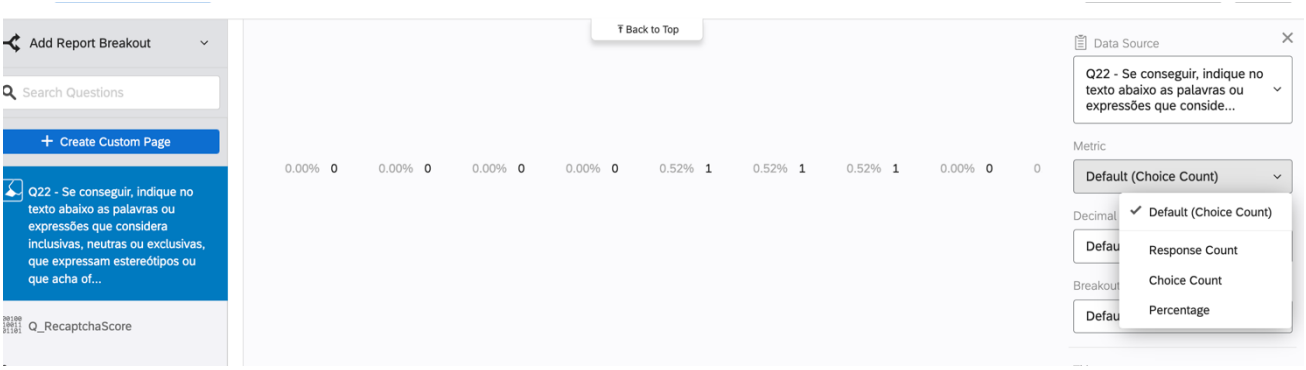

Figure 2: Qualtrics report on results with response count and percentage (from previous UNL project)

ii) It captures and reports on annotators and annotation process metadata (which documents were annotated

[1] Verhagen et al. (2006), Stubbs (2011), Neale et al. (2015), respectively.

by a specific annotator, when where the data annotated; how long did the annotator take, etc.).

iii) It allows for presenting the data for the annotator in several formats (different colours, text boxes, text sizes and fonts, etc.), and it also allows for several types of visualizations of the results (Figure 3).

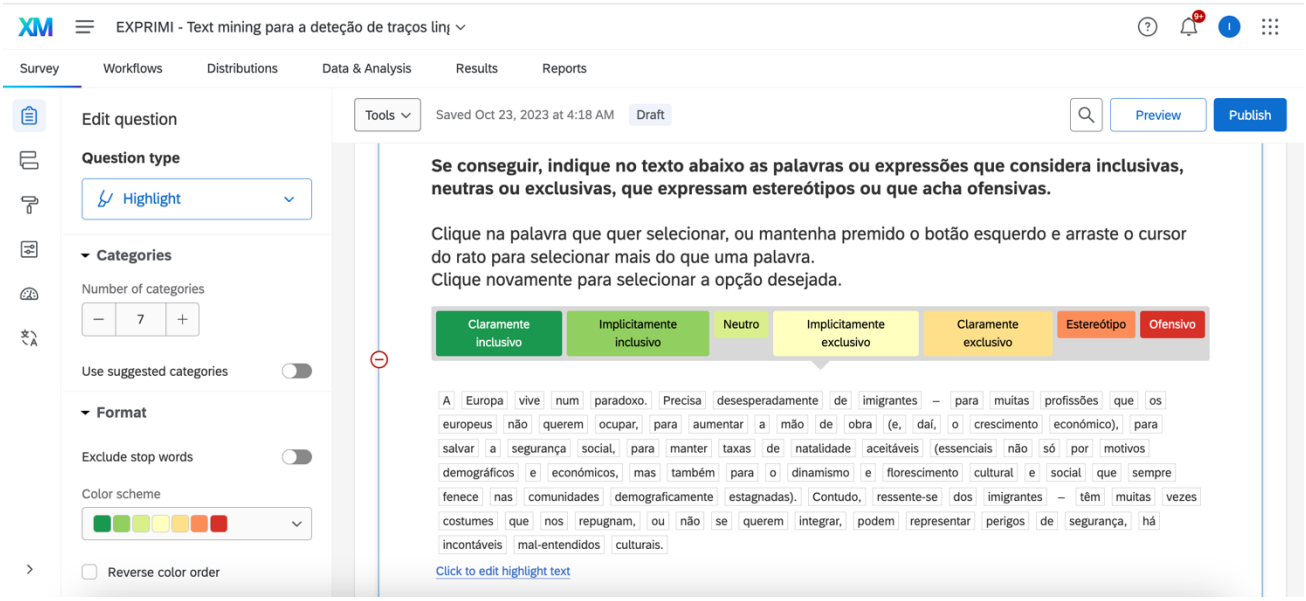

Figure 3: Qualtrics question editor (from previous UNL project)

iv) It exports results in several formats (table XML, CSV, excel, pdf) and provides different types of data visualization of the results (Figures 4 and 5).

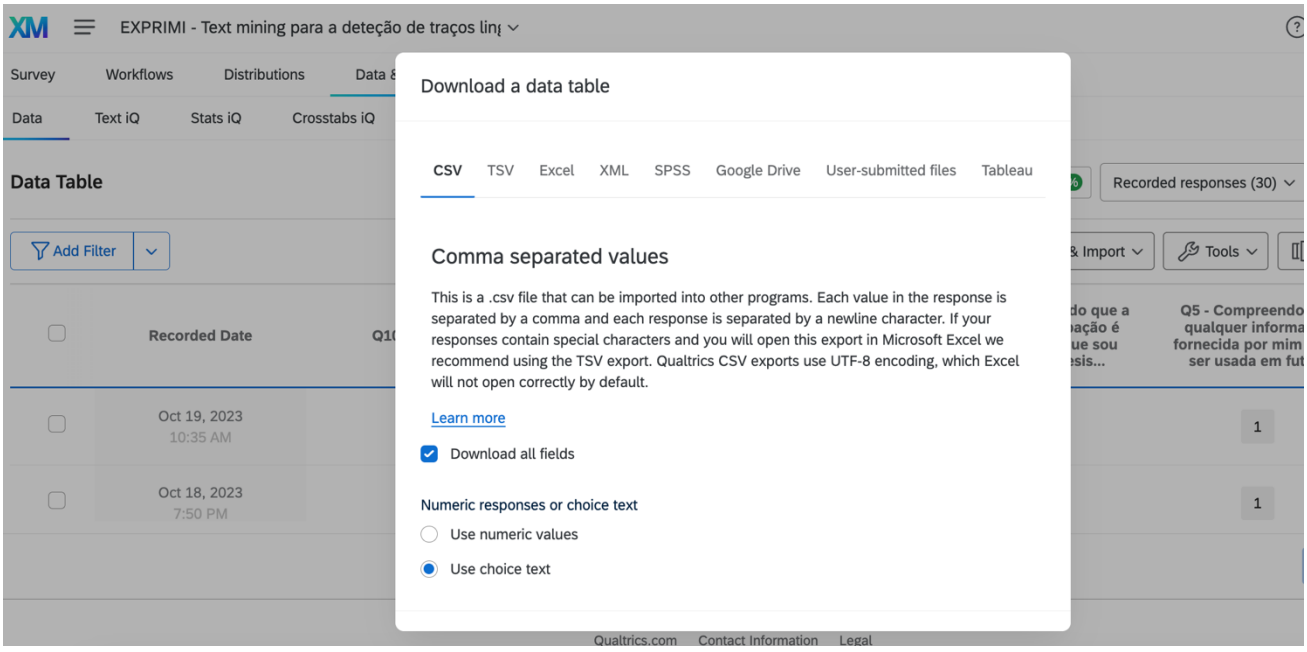

Figure 4: Qualtrics data export formats (from previous UNL project)

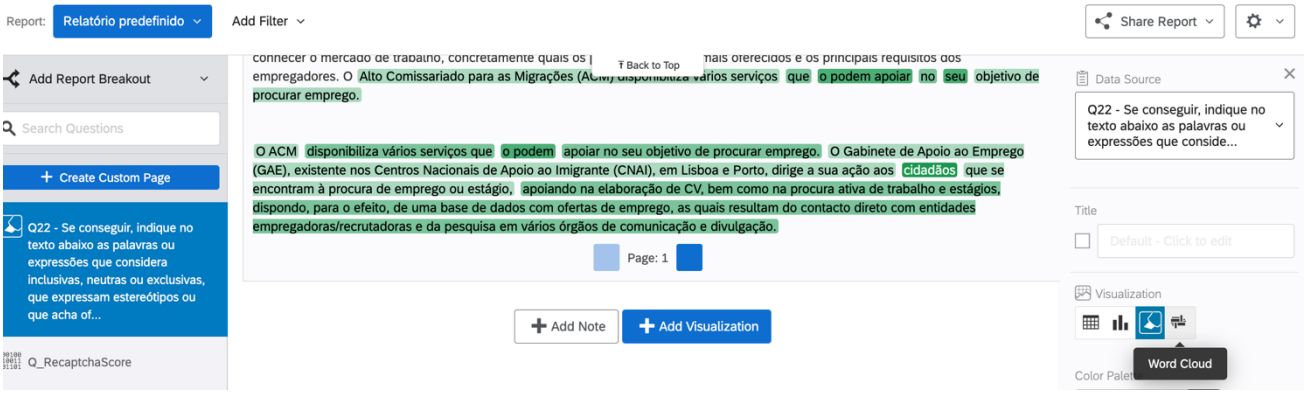

Figure 5: Qualtrics report builder (from previous UNL project)

The choice of the tool considered all these features, as well as the fact that it provides a stable web-based interface and user support (https://www.qualtrics.com/support/).

2.2. Representation formats

A generalized format for annotated data extends beyond the schemes designed for a specific tool or goal (Ide et al. 2017), as it should consider future uses of the resources. Usually, data format must be capable of:

- representing all relevant linguistic data, including text, annotations and relevant metadata,
- provide the means for transduction to and from other formats,
- enable easy and incremental addition, modification, deletion, and merging of annotations from different sources,
- accommodate widely used formats and technologies, such as XML or RDF/OWL.
- provide mechanisms for documentation of the resulting resource (Ide at al. 2017: 74: 97-98).

The choice of representation format imposes limits to the type and complexity of the information in the resource, as well as in its readability (e.g., conflation of information in undecipherable tags).

Considering that a representation format is a mean to associate linguistic information to parts of the data by means of annotation tags (identifiers indicating what the data in a specific region is), we defined an arbitrary flat tag scheme considering the marking of difficult words / expressions / lines in the text (cf. 1.2).

These annotations will be formatted as stand-off annotation, in XML table format, where the tag ('difficult') is attributed to a specific token of the texts or to a sequence of tokens, as allowed by the annotation tool: the tool produces a new document with all tokens numbered and ordered, one per column/row and attributes the tags to specific tokens. In the case of the annotations by the trainers, an attribute with the category of the 'difficulty' can be added. Stand-off annotation can co-exist with different annotations for the same phenomenon, as well as with different annotation labels and features for diverse phenomena (e.g., part-of-speech, multiword expressions, semantic features, etc.).

The resulting annotated data can be converted into in-line annotation if necessary.

3. References

Alves, M. J. and Lameira, S. (coord.) (2021). Referencial de Competências-chave de Educação e Formação de Adultos – Nível Básico, ANQEP, I.P.

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, www.coe.int/lang-cefr.

Finlayson, M. A. and Erjavec, T. (2017). Overview of Annotation Creation: Processes and Tools. In: Pustejovsk, P. and, Ide, N. (eds.) *Handbook of Linguistic Annotation*. Springer. DOI 10.1007/978-94-024-0881-2, 167-191

Hovy, E. and Lavid, J. (2010). Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. In *International Journal of Translation*, Vol. 22, No. 1, Jan-Jun 2010.

Ide, N., Chiarcos, C., Stede M. and Cassidy, S. (2017). Designing Annotation Schemes: From Model to Representation. In: Pustejovsk, P. and, Ide, N. (eds.) *Handbook of Linguistic Annotation*. Springer. DOI 10.1007/978-94-024-0881-2, pp. 73-111

Neale, S., Silva, J. and Branco, A. (2015): A flexible interface tool for manual word sense annotation. In: Bunt, H. (ed.) *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (ISA-11). London, UK. http://www.aclweb.org/anthology/W/W15/W15-0208.pdf

OECD (2016), *The Survey of Adult Skills: Reader's Companion, Second Edition* (PIACC 2016), OECD Skills Studies, OECD Publishing, Paris, https://doi.org/10.1787/9789264258075-en.

Pirali, C., François, T., and Gala, N. (2022). PADDLe: a Platform to Identify Complex Words for Learners of French as a Foreign Language (FFL). In *Proceedings of the 2nd READI Workshop @ LREC2022*, Marseille, 24 June 2022. ELRA, pp. 46–53. https://aclanthology.org/2022.readi-1.7.pdf

Pustejovsky, J. , Harry, B. and Zaenen. A. (2017). Designing Annotation Schemes: From Theory to Model. In: Pustejovsk, P. and, Ide, N. (eds.) *Handbook of Linguistic Annotation*. Springer. DOI 10.1007/978-94-024-0881-2, 21-71

Stubbs, A. (2011). MAE and MAI: lightweight annotation and adjudication tools. In: *Proceedings of the 5th Linguistic Annotation Workshop* (LAWV), pp. 129–133. Association for Computational Linguistics., Portland, Oregon, USA http://www.aclweb.org/anthology/W11-0416.

Verhagen, M., Knippen, R., Mani, I. and Pustejovsky, J. (2006). Annotation of temporal relations with Tango. In: *Proceedings of the 5th Languange Resources and Evaluation Conference* (LREC 2006), pp. 2249–2252. European Language Resources Association, Genoa, Italy.