

Horizon Europe



iRead4Skills Dataset 2: annotated  
corpora by level of complexity  
for FR, PT and SP



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.

**Control****Information:**

Settings	Value
<b>Deliverable No.</b>	<b>D3.7</b>
Document Title:	iRead4Skills Dataset 2: annotated corpora by level of complexity for FR, PT and SP
Project Title:	iRead4Skills
Document Author(s):	<p>Ricardo Monteiro, NOVA University Lisbon  Raquel Amaro, NOVA University Lisbon  Thomas François, UCLouvain</p> <p>French corpus:  Alice Pintard (CENTAL, Université Catholique de Louvain), Thomas François (CENTAL, Université Catholique de Louvain) (creators);  Justine Nagant de Deuxchaisnes (CENTAL, Université Catholique de Louvain) (contributor).</p> <p>Portuguese corpus:  Sílvia Barbosa (CLUNL - Universidade Nova de Lisboa), Maria Leonor Reis (CLUNL - Universidade Nova de Lisboa), Michell Moutinho (CLUNL - Universidade Nova de Lisboa), Ricardo Monteiro (CLUNL - Universidade Nova de Lisboa), Raquel Amaro (CLUNL - Universidade Nova de Lisboa), Susana Correia (CLUNL - Universidade Nova de Lisboa).</p> <p>Spanish corpus:  Sandra Rodríguez Rey (CITIUS-Universidade de Santiago de Compostela), Keran Mu (Universitat Autònoma de Barcelona), Marcos Garcia González (CITIUS-Universidade de Santiago de Compostela), André Bernárdez Braña (CITIUS-Universidade de Santiago de Compostela), Xavier Blanco Escoda (Universitat Autònoma de Barcelona).</p>
Reviewer(s):	Raquel Amaro, NOVA University Lisbon
Doc. Version:	Version 2.0
Sensitivity:	<b>Public</b> <a href="https://doi.org/10.5281/zenodo.12821882">https://doi.org/10.5281/zenodo.12821882</a>
Date:	30/11/2024

**Document Location:** The latest version of this controlled document is stored in OneDrive\fcsh.unl.pt\iRead4Skills\Project\WorkPackages\WP3\ D3.7 - Annotated Data set\FINAL\_DELIVERABLE\_NOV2024

## Table of Contents

<b>1. Introduction .....</b>	<b>1</b>
<b>2. Classification and annotation tasks.....</b>	<b>1</b>
2.1 Dataset design .....	1
2.2 Task for trainers .....	2
2.3 Task for students .....	4
<b>3. Data description.....</b>	<b>4</b>
3.1 Format .....	5
3.2 Qualitative and quantitative description .....	6
3.2.1 Trainers' data .....	6
3.2.2 Students' data .....	9
3.2.3 Annotations .....	10
3.2.4 Majority calculation .....	11
3.3 Inter-annotator agreement.....	13
<b>4. Samples .....</b>	<b>14</b>
Trainers' Annotations .....	16
Students' Annotations .....	38

## I. Introduction

This report provides a detailed overview of **Dataset 2: annotated corpora by level of complexity for FR, PT and SP**. It is organized into three major sections - 2. *Classification and annotation tasks*, 3. *Data description*, and 4. *Sample results* - and provides a detailed account on the classification and annotation processes undertaken, the composition of the dataset, including data format, qualitative and quantitative analyses and inter-annotators agreement, as well as a sample of annotated texts extracted from the dataset.

The **Dataset 2: annotated corpora by level of complexity for FR, PT and SP** is a collection of texts categorized by complexity level and annotated for complexity features, presented in xlsx format. These corpora were compiled and annotated under the scope of the project *iRead4Skills – Intelligent Reading Improvement System for Fundamental and Transversal Skills Development*, funded by the European Commission (grant number: 1010094837). The project aims to enhance reading skills within the adult population by creating an intelligent system that assesses text complexity and recommends suitable reading materials to adults with low literacy skills, contributing to reducing skills gaps and facilitating access to information and culture (<https://iread4skills.com>).

This dataset is the result of specifically devised classification and annotation tasks, in which selected texts were organized and distributed to trainers in Adult Learning (AL) and Vocational Education Training (VET) Centres, as well as to adult students in AL and VET centres. This task was conducted via the Qualtrics platform.

The goal of this task was gathering input on complexity phenomena and on the perception of texts' complexity from both these target populations, based on the texts compiled and classified in the **iRead4Skills Dataset 1: corpora by level of complexity for FR, PT and SP** (<https://doi.org/10.5281/zenodo.10055909>).

In order to illustrate how the task was conducted by the participants, a sample of annotated texts is presented, for each language under the project's scope.

The full dataset is available under creative CC BY-NC-ND 4.0 license in Zenodo, <https://doi.org/10.5281/zenodo.12821882>.

## 2. Classification and annotation tasks

### 2.1 Dataset design

The **Dataset 2: annotated corpora by level of complexity for FR, PT and SP** is derived from the **iRead4Skills Dataset 1: corpora by level of complexity for FR, PT and SP** (<https://doi.org/10.5281/zenodo.10055909>), which comprises written texts of various genres and complexity levels. From this collection, a subset of texts was selected for classification and annotation. This classification and annotation task aimed to provide additional data and test sets for the complexity analysis systems for the three languages of the project: French (FR), Portuguese (PT), and Spanish (SP).

Approximately 20% of the texts in each of the language corpora (FR, PT, and SP) were selected, taking into account the diversity of topics/domains, genres, and the reading preferences of the target audience of the iRead4Skills project, as depicted in Table I. This percentage amounted to the total of 462 texts per language, which were divided by level of complexity, resulting in the following distribution:

- 140 Very Easy texts
- 140 Easy texts
- 140 Plain texts
- 42 More Complex texts.

Preferences' percentage	topic	genres/sub genres/document types
60-40	current issues	editorials, news, reportage, interview, opinion articles
40-30	health and well-being	magazines, self-help books, essays
30-25	society	magazines, interview, news, reportage
30-20	history	historic novels, history books, encyclopedias
30-25	traveling	chronicles, travel reports, travel/tourist guides
30-20	cooking	cookbooks
25-20	sports	news, reportage, interviews
30-15	romance and love	novel, epic, drama, ...
25-20	family	magazines, novels, self-help books, essays, diaries, chronicles, ...

**Table I:** Reading preferences of low-literacy adults retrieved from the iRead4Skills D2.1 Reading skills survey<sup>1</sup>

## 2.2 Task for trainers

Trainers were asked to classify the texts according to the complexity levels of the project (iRead4Skills - Complexity Levels<sup>2</sup>), here informally defined as:

- **Very Easy** (everyone can understand the text or most of the text).
- **Easy** (a person with less than the 9th year of schooling can understand the text or most of the text)
- **Plain** (a person with the 9th year of schooling can understand the text the first time he/she reads it)
- **More Complex** (a person with the 9th year of schooling cannot understand the text the first time he/she reads it)

It should be mentioned that, for French, we also asked trainers to further refine their judgement, using a Likert scale (from 1 to 5). In other words, after having decided on the gross level (e.g. Easy), trainers had to estimate whether the given text was more representative of the beginning of the level (1 or 2), typical of the level (3), or already close to the next level (4 or 5). In case the trainers forgot to answer this question, we assigned a score of 2,5.

<sup>1</sup> <https://zenodo.org/records/10179536>

<sup>2</sup> Monteiro, R., Amaro, R., Correia, S., Pintard, A., Gauchola, R., Moutinho, M., & Blanco Escoda, X. (2023). iRead4Skills - Complexity Levels (V1.0). Zenodo. <https://doi.org/10.5281/zenodo.10459090>

Furthermore, trainers were also asked to annotate the parts of the texts considered complex according to various type of features, at word-level and at sentence-level (e.g., word order, sentence composition, etc.). The annotating categories were:

*Lexical/word-related features*

- unknown word
- word too technical/specialized or archaic
- complex derived word
- points to a previous reference that is not obvious
- word (other)

*Syntactic/sentence-level features*

- unusual word order
- too much embedded secondary information
- too many connectors in the same sentence
- sentence (other)
- other (please specify)

To better distribute and execute this task, different batches containing texts to classify and annotate were constituted. Each batch contained texts from all levels of complexity. For the trainers' task, the goal was to have different trainers validating the same set. This not only provides different judgments on the sets but makes inter annotator agreement calculations possible. The table below shows the distribution of sets per annotator for PT and SP planned.

<b><i>number of texts per trainer</i></b>	<b><i>number of data sets &amp; trainers' groups</i></b>	<b><i>total number of trainers</i></b>
60 texts (15 Plain texts) + (5 +complex texts) (15 Easy texts) + (5 Plain texts) (15 Very Easy texts) + (5 Easy texts)	8	24

**Table 2:** *Distribution of texts per set & trainers' groups for PT and SP*

The sets were divided in three parts in Qualtrics, and, in each part, the texts were shown randomly to the annotator. The sets were divided into three parts in Qualtrics, and, in each part, the texts were shown randomly to the annotator.

For French, we initially used batches of about 32 texts but soon received negative feedback from the annotators about the exaggerated number of texts. We therefore decided to reduce batch size to 16 or 17 texts. The distribution of text is operated randomly, ensuring more or less the same amount of text per difficult level (except for More Complex).

<b><i>number of texts per data sets</i></b>	<b><i>number of data sets &amp; trainers' groups</i></b>	<b><i>total number of trainers</i></b>
17 texts or		17

16 texts	28	
----------	----	--

**Table 3:** *Distribution of texts per set & trainers' groups for FR*

### 2.3 Task for students

A similar task was conducted with students from AL centers. However, smaller sets were considered to avoid fatiguing this population, which could result in skewed results.

In this student's task, each set of texts was divided by level of complexity (Very Easy, Easy, Plain). The students who participate in the task are matched with the set of texts adequate to their literacy level, in order to validate their reading comprehension of the texts. Each set contains texts from a given level, plus one text of the level immediately above. The texts of this higher level serve to provide us with a control task, helping us assess the validity of the task. The table below shows the distribution of text sets per student for PT and SP.

<b>number of texts per student &amp; level</b>	<b>number of data sets per level</b>	<b>total number of students</b>
5 (4 texts of the students' level + 1 text of the level above)	18 (6 sets per level)	54 (18 students from each level)

**Table 4:** *Distribution of texts per set, level and students' groups for PT and SP*

Additionally, the students are also requested to annotate words and sequences of words in the text that they did not understand. Different categories and tags were used in this specific annotation task, to avoid overwhelming the students. The categories used were:

- difficult word
- difficult part of the text

For French, the same annotation design was used. Unfortunately, due to various factors (poor connection with the VET and AL centers, motivational levers less efficient than for the PT and SP contexts, calendar problems, etc.), the number of recruited students was smaller. annotation of the texts. The table below shows the distribution of text sets for FR:

<b>number of texts per student &amp; level</b>	<b>number of data sets per level</b>	<b>total number of students</b>
5 (4 texts of the students' level + 1 text of the level above)	Very Easy: 5 Easy: 6 Plain: 7	19

**Table 5:** *Distribution of texts per set, level and students' groups for FR*

## 3. Data description

This section provides a description of the data resulting from the two tasks detailed above – Task for trainers and Task for students, including details on data formats, quantitative and qualitative information, and inter-annotator agreement. The descriptions and analysis of the data are organized by language: French, Portuguese, Spanish.

### 3.1 Format

The classification and annotation tasks were conducted using the Qualtrics platform. The data generated from these tasks was exported in CSV format and subsequently processed into XLSX format. The data was organized as a matrix, with rows displaying the input of each annotator and columns containing various details about the classified/annotated files. The data processing step was carried out separately for the results of the annotation and classification tasks, resulting in two different files. The final data format and organization is presented in the following subsections.

The complete results and datasets are in XLSX format in Appendixes I to III. These are organized by language in pairs of two files, with one file concerning the results from the classification (trainers)/validation (students) task and one file concerning the results from the annotation task.

The table below provides the description of each variable.

The data in each Excel file is transparently provided and organized. Each row contains the input from a single annotator, while the columns correspond to the variables at play, as presented in Table 5 below.

Column name	Data
<b>Annotator's ID</b>	The randomly generated ID code for each annotator, together with information on the dataset assigned to them.
<b>Progress</b>	Information on the completion of the task (for each text).
<b>Duration (seconds)</b>	Time used in the completion of the task (for each text).
<b>File Name</b> <b>N1 = Very Easy</b> <b>N2 = Easy</b> <b>N3 = Plain</b> <b>N4 = More Complex</b>	File internal identification, providing its iRead4Skills classification.
<b>Text</b>	The content of the file, i.e. the text itself.
<b>Annotated Level</b>	Level assigned by the annotator (trainer).
<b>Proficiency SubLevel (Likert Scale - 1 to 5)</b>	SubLevel assigned by the annotator (trainer) for FR data.
<b>Corresponding CEFR Level</b>	CEFR level closest to the iRead4Skills
<b>Additional Info</b>	Observations made by the trainers/students
<b>Annotated Term</b>	Word or set of words selected for annotation
<b>Term Label</b>	Annotation assigned to the Annotated Term (difficult word, word order, etc.)
<b>TermIndex</b>	Position of the annotated term in the text
<b>Annotator's Proficiency Level</b>	Level of AL/VET of the student
<b>Text adequate for user</b>	Validation of the text by the students

The complete datasets are available under creative CC BY-NC-ND 4.0 license in Zenodo, <https://doi.org/10.5281/zenodo.12821882>.



As visible in the Annexes, the columns Proficiency Level, Proficiency SubLevel, Proficiency File Level, and Proficiency Level CEFR are exclusive to the classification task, whereas the columns Annotated Term, Term Index, and Term Label are exclusive to the annotation task. The columns Additional Info and Annotator's ID are also common to both files.

The configuration of the XLSX files for both students' and trainers' tasks is nearly identical. However, in the classification task, the students were not asked to classify the text. Instead, the texts were presented to them based on their predetermined proficiency levels. Therefore, the students were not directly classifying the texts but rather indicating whether they understood them or not. This indirectly classified the texts as either corresponding to their proficiency level or not.

As shown in the table above, the *Text adequate for user* column indicates whether the students understood the text presented to them during the task or not (using a Boolean value, in the native language). The *Annotator's Proficiency Level* column refers to the reading proficiency level of the student. These proficiency levels correspond to levels assigned to the students by their educational institutions. In collaboration with the AL and VET centers and the trainers, a mapping of these external levels to the project's levels was carried out. The iRead4Skills level assigned to the text is indicated in the *File Name* column through the prefixes N1 (= Very Easy); N2 (= Easy); N3 (= Plain) and N4 (= More Complex).

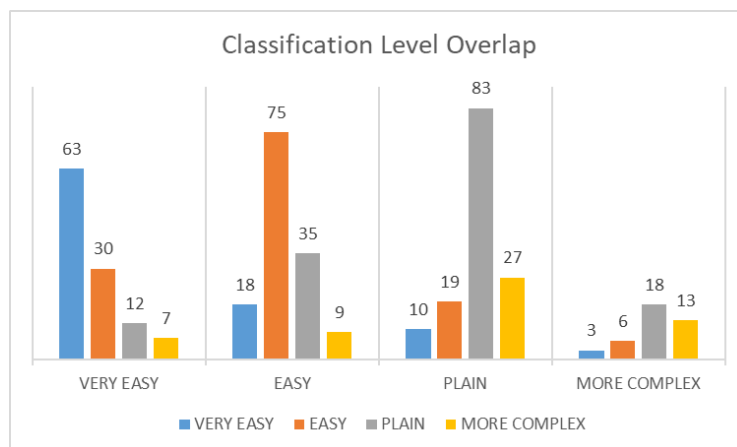
### 3.2 Qualitative and quantitative description

This subsection presents information on the data collected, organized by the different annotators/respondents (Trainers and Students) and the languages covered in the project. It presents quantitative details on the types and amounts of data collected, along with analyses of the results, such as annotators' agreement and majority-based calculations.

#### 3.2.1 Trainers' data

##### *Portuguese*

Trainers were asked to classify the texts according to the complexity levels (Very Easy, Easy, Plain, More Complex). The overlap of the results from this classification and the levels in the iRead4Skills PT corpus are presented in the Figure below.



**Figure 6:** *Texts' classification results per level – Portuguese*

The results indicate alignment with the iRead4Skills levels. The majority of texts at each level were classified as belonging to the corresponding level: 63 texts as Very Easy, 75 as Easy and 83 as Plain. Texts in the More Complex category appear to be an outlier, as the majority of them that were classified as belonging to a different level (13 texts as More Complex and 18 texts as Plain). However, it is important to interpret this critically, as the More Complex level contained fewer texts to be classified than the others and primarily served as a threshold level for the Plain level.

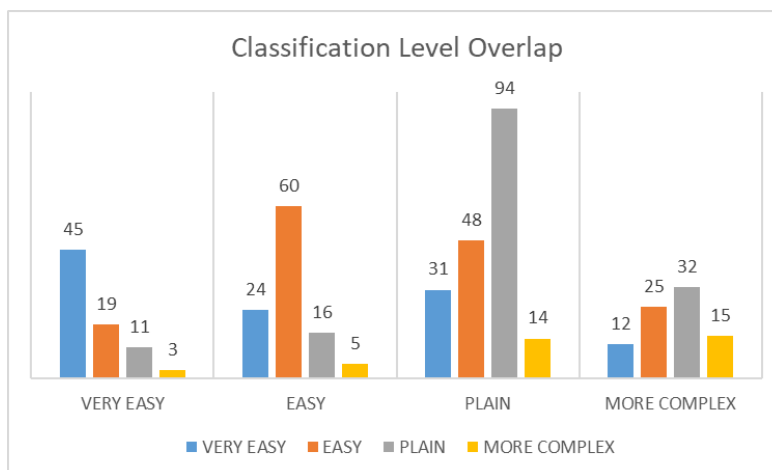
At the relevant levels (Very Easy, Easy, and Plain), a significant proportion of texts were classified as belonging to other levels: 44% in the Very Easy level; 45% in the Easy level, and 40% in the Plain level. For the Very Easy level, this suggests an underestimation of difficulty of these texts. In the other levels, the results are more evenly distributed, although there is notable tendency for texts to be classified into higher complexity levels.

These cases where there was no overlap were further analyzed. It was observed that texts in these situations were typically classified as belonging to adjacent levels, usually one level higher on the complexity scale, and rarely to a level significantly beyond that.

Results suggests that the iRead4Skills assessment of the texts' complexity levels was adequate and that is supported by the empirical knowledge and experience of trainers.

## *Spanish*

Also in the Spanish case, the data from the trainers' task corroborated the iRead4Skills proposed levels for the majority of texts.



**Figure 7:** Text's level classification results per level – Spanish

For each level, the majority of texts were classified as belonging to that same level. The More Complex level was the only exception, with most of texts being overestimated in terms of complexity. However, as previously noted, this level primarily served as a threshold for the preceding one, and therefore does not invalidate the data for the other levels.

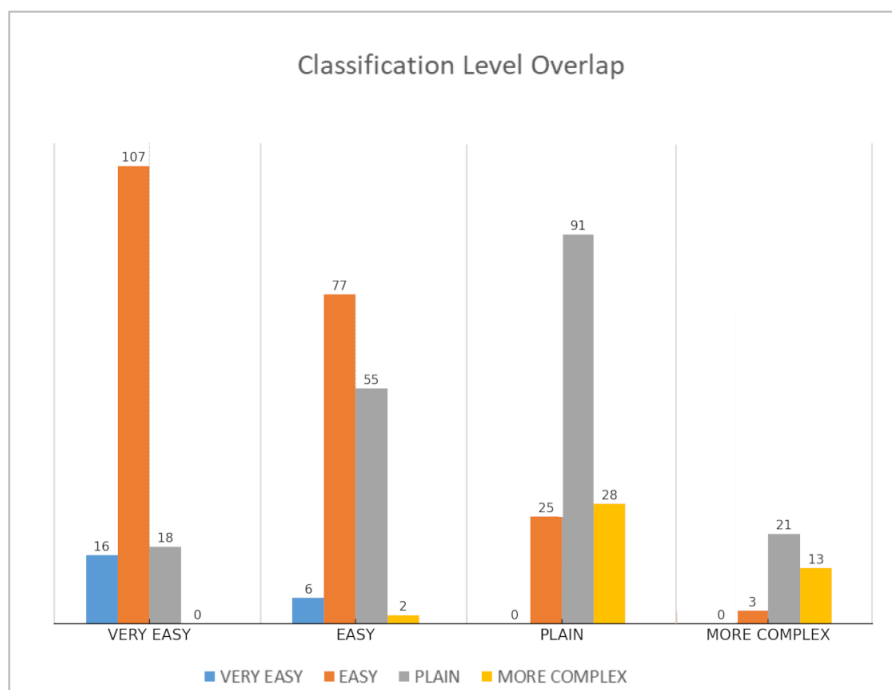
As in the Portuguese case, a significant proportion of texts were classified as belonging to other levels: in the Very Easy level, 42% of the texts were considered not to belong to this level, while in Easy and Plain levels, 43% and 49% of the texts, respectively, were classified as belonging to other levels. In Spanish, however, the tendency for the Easy and Plain levels is for texts to be classified into lower complexity levels.

Results, again, support that the iRead4Skills assessment of the texts' complexity levels was adequate.

## French

For French, it should be remembered that the initial classification was not carried out by humans, but by a deep learning algorithm capable of identifying the difficulty of texts according to the CEFR framework, called DMesure (Yancey et al., 2021)<sup>3</sup>. This automatic tool seems to have underestimated the difficulty of the texts that were retained, since 107 texts that he had considered Very Easy were recalibrated to Easy. For the two median levels, DMesure's assessment and that of the trainers seem to be broadly in line, although there is still a slight tendency to underestimate the Easy level. Conversely, for the More Complex level, we can see that it is the trainers who underestimate the difficulty of texts. This raises the question of the extent to which our annotators were not subject to a central tendency bias (tending towards average scores). Nevertheless, for the rest of the project, we'll be relying more on the trainers' scores than on DMesure's.

<sup>3</sup> Yancey, K., Pintard, A., & Francois, T. (2021). Investigating readability of French as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e linguaggio*, 20(2), 229-258.



**Figure 8:** Text's level classification results per level – French

### 3.2.2 Students' data

In the students' classification task, students, based on their assigned reading proficiency levels, were asked whether they understood the majority of the text or not. Unlike the trainers' task, it did not require students to reassign levels to the texts that were not understood.

#### Portuguese

The table below presents the results of the students' classification task. It details the percentage of overlap between the iRead4Skills classification for each level and the students' responses, reflecting the proportion of texts considered adequate by the target users. These results indicate that the texts initially proposed for each level are, for the most part, appropriate for the reading proficiency of the students.

Level	Overlap with the iRead4Skills level
Very Easy	77,6%
Easy	65,8%
Plain	77,6%
More complex	100%

**Table 6:** Percentage of texts overlapping with the iRead4Skills levels, based on student's answers

## Spanish

Level	Overlap with the iRead4Skills level
Very easy	66,7%
Easy	73,3%
Plain	80,2%
More complex	100%

**Table 7:** Percentage of texts overlapping with the iRead4Skills levels, based on student's answers

The table above presents the results of the students' classification task. It provides the percentage of overlap between the initial classification for each level and the students' responses, representing the proportion of texts deemed adequate to the target users. As the data indicates, there is quite a significant overlap between the proposed levels of the texts and the students' assessment.

## French

For French, as we did not have the manually assigned levels for texts but those of the DMesure tool, we considered more informative to compare the difficulty levels assigned by the teachers with the comprehension's statement of the students. It appears that 86.7% of the texts assigned to readers at the "Very Easy" level were correctly understood. For the "Easy" level, 75.6% of the texts were considered adequate by the students of this level, whereas 77% of the "Plain" texts were correctly understood by students of this level. Although this shows that the teachers' annotations are not perfect, these percentages remain reasonable, even for applicative purposes.

Level	Overlap with the initial level
Very Easy	86.7%
Easy	75.6%
Plain	77.1%
More complex	NA

**Table 8:** Percentage of texts overlapping with the teacher's levels, based on student's answers

### 3.2.3 Annotations

A complementary task was conducted with both the student and trainer populations – the annotation task. As explained earlier, respondents were asked (but not required) to:

- mark parts of the texts that they did not understand (students),
- annotate parts of the texts considered complex according to various features, at word-level and at sentence-level (e.g., word order, sentence composition, etc.).

The table below presents the quantitative results of this task for the three languages.

	<b>Portuguese</b>	<b>Spanish</b>	<b>French</b>
Student's annotations	103	73	95
Trainer's annotations	344	402	462

**Table 9:** *Number of texts with annotations per language dataset*

The table shows the number of texts with one or more annotations in each language dataset. Considering the total number of texts presented for this task, the number of texts annotated seems relatively satisfactory. However, the collected annotations are not sufficient to be directly used for machine learning purposes.

Nonetheless, further analysis of these data is expected to be valuable for developing and fine tuning the writing assistant tools.

### 3.2.4 Majority calculation

To process and use the results from the trainer and students' tasks, the degree of consensus among the different classifications was computed. This degree of consensus (or majority agreement) took each classification from every annotator, as well as the initial classification of the texts, as input values. The calculation considered the following categories:

- Very strong: All annotators and the initial file level are in complete agreement.
- Strong: There is one disagreement among the annotators (including the initial level).
- Weak: There are two disagreements among the annotators (including the initial level).
- Very weak: There are more than two disagreements among the annotators (including the initial level).

Since only a few texts in the students' task were classified by more than one person, this calculation was applied only to the data resulting from the trainers' task.

The analysis was performed for the results of the different languages.

#### Portuguese

<b>Majority</b>	<b>Number</b>	<b>Percentage</b>
Very weak	171	40%
Weak	179	39,72%
Strong	71	16,59%
Very strong	16	3,74%
<b>Total</b>	<b>428</b>	

**Table 10:** *Majority per type - PT*

	<b>Very Easy</b>	<b>Easy</b>	<b>Plain</b>	<b>More Complex</b>
Very weak	33,3%	41,1%	45,3%	40%
Weak	35,1%	42,6%	32,1%	52,9%
Strong	20,8%	15,6%	19,8%	7,1%
Very strong	10,8%	0,7%	2,8%	0%

**Table 11:** *Distribution of majority type per level - PT*

The tables above indicate low consensus among annotators. Almost 80% of the majority types are Very weak or Weak. The data also shows that texts classified as "Very Easy" tend to have better agreement, whereas texts from other levels exhibit higher rates of disagreement. This may suggest greater variability in complexity at higher levels than at lower ones, making it more challenging to reach consensus on higher complexity levels.

## Spanish

For Spanish, the majority results were as follows:

<b>Majority</b>	<b>Number</b>	<b>Percentage</b>
Very weak	113	24,9%
Weak	135	29,7%
Strong	175	38,6%
Very strong	31	6,8%
<b>Total</b>	454	

**Table 12:** *Majority per type - SP*

	<b>Very Easy</b>	<b>Easy</b>	<b>Plain</b>	<b>More Complex</b>
Very weak	17,3%	40,2%	26,1%	12,2%
Weak	25%	31,8%	27,8%	33%
Strong	46,1%	18,7%	38,9%	53%
Very strong	11,6%	9,3%	7,2%	1,8%

**Table 13:** *Percentage of type of majority per level*

In general, the data show a relatively even distribution of consensus among annotators, with the agreement category "Strong" having the highest level of occurrence. The data further indicates that texts categorized as "Very easy", "Plain", and "More Complex" tend to have a strong agreement, whereas "Easy" texts exhibit higher rates of Very weak agreement.

## French

For the French case, the majority results were as follows:

Majority	Number	Percentage
Very weak	6	1.3%
Weak	286	61.9%
Strong	155	33.5%
Very strong	15	3.3%
<b>Total</b>	<b>462</b>	

**Table 14:** Majority per type

	Very easy	Easy	Plain	More Complex
Very weak	4.3%	0.6%	0%	0%
Weak	62.6%	64.6%	60%	56.1%
Strong	32.3%	32.9%	34.5%	36.6%
Very strong	0.9%	1.9%	5.5%	7.3%

**Table 15:** Distribution of majority type per level

The figures for French are roughly similar to those of the other languages, except that there is much less “Very weak” agreement (1.3% compared to 24.9% and 40%), but more “Weak” agreement (61.9% vs. 29.7% and 39.72%). The proportion of “Strong” and “Very strong” agreements (36.8%) is slightly lower than in Spanish (45.4%), but higher than in Portuguese (20.4%).

In addition to providing informing on the properties of the *iRead4Skills Dataset 2: annotated corpora by level of complexity for FR, PT and SP*, the majority calculation, along with data from the texts' classification task, will be further used to determine the project's golden standard corpora.

### 3.3 Inter-annotator agreement

The inter-annotator agreement analysis used Cohen's Kappa and Fleiss' Kappa to evaluate the agreement between annotators, considering the trainers' classification tasks performed for the three languages.

Cohen's Kappa is calculated for a pair of annotators within a batch and measures the agreement between the annotations of two evaluators. This allows a detailed comparison of how consistently two specific annotators classify the same data.

Fleiss' Kappa, on the other hand, evaluates the agreement among multiple annotators for a single batch. It provides an overall measure of consistency across all annotators, helping to assess the general reliability of the classifications in each form.

Two types of Cohen's Kappa were calculated: linear and quadratic. The linear version is more sensitive to small discrepancies, while the quadratic penalizes disagreements more severely. The reported value is the average kappa on all batches and for all pairs of annotators within each batch.



	Cohen's Kappa				Fleiss' Kappa
	Linear weights		Quadratic weights		
	average	standard deviation	average	standard deviation	
French	0,080	0,22	0,31	0.25	0,010
Portuguese	0,241	0,180	0,241	0.222	0,199
Spanish	0,259	0,145	0,390	0.165	0,049

**Table 16:** Cohen's Kappa and Fleiss' Kappa results per language

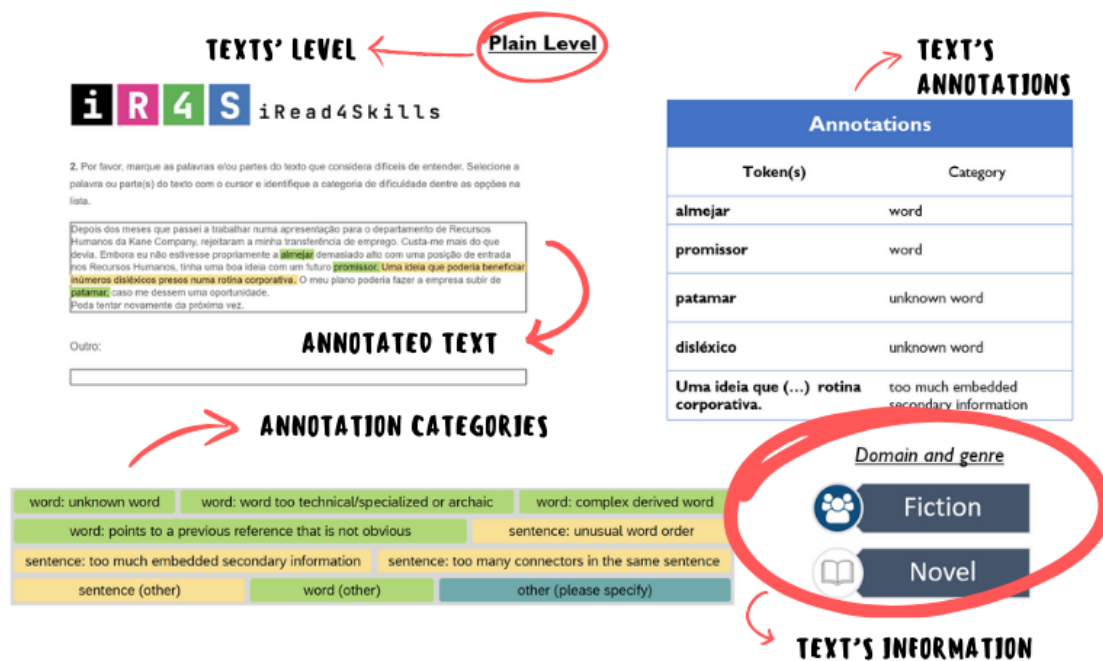
According to Landis & Koch (1977)<sup>4</sup>, these results indicate between Slight and Fair agreement<sup>5</sup>. These findings align with the difficulty and subjectivity of the task and reflect the range of variables that influence one's perception of what makes a text complex and difficult to read. Nonetheless, inter-annotator agreement provides measurable and relevant information about the compiled dataset and should be considered alongside the other quantitative and qualitative information for evaluating the dataset usability and its suitability for different goals. It should also be mentioned that although agreement is not optimal, disagreement will be averaged when we create the gold standard, meaning that the resulting difficulty score for each text will be a more robust estimation than a single annotator judgement.

## 4. Samples

The following pages provide also some samples of the results of the tasks described above in a more human-accessible format, with particular focus on the annotation tasks. The texts are organized by complexity level and language. Samples from both the students and trainers' tasks are presented. Figure 9 shows what the elements in the pages allude to.

<sup>4</sup> Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>.

<sup>5</sup> Kappa Level of Agreement: >0,8: Almost perfect; >0,6: Substantial; >0,4: Moderate; >0,2: Fair; >0; <0 No agreement, from <https://datatab.net/statistics-calculator/reliability-analysis/fleiss-kappa-calculator>.



**Figure 9:** Sample elements

## Trainers' Annotations

## Very Easy Level



2. Por favor, marque as palavras e/ou partes do texto que considera difíceis de entender. Selecione a palavra ou parte(s) do texto com o cursor e identifique a categoria de dificuldade dentre as opções na lista.

Reparo desde pequena que os adultos vivem muito em casais.  
O amor constrói. Gostamos de alguém, mesmo quando estamos parados durante o tempo de dormir, é como fazer prédios ou cozinhar para mesas de mil lugares.  
Mas amar é um trabalho bom. A minha mãe diz.

Outro:

3. Se achar relevante, deixe-nos os seus comentários ou sugestões. Obrigado.

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
O amor constrói	points to a previous reference that is not obvious
Gostamos de alguém (...) lugares.	too many connectors in the same sentence

### Domain and genre



Fiction



Short story

## Very Easy Level



2. Por favor, marque as palavras e/ou partes do texto que considera difíceis de entender. Selecione a palavra ou parte(s) do texto com o cursor e identifique a categoria de dificuldade dentre as opções na lista.

Quase todos partilham a opinião de que os jovens dos **países do Sul da Europa** gostam de vestuário. Em Portugal, **as indústrias nacionais são a têxtil e a do calçado**; portanto, só nos fica bem gastar uma boa parte do nosso rendimento em roupas e sapatos. Somos, de facto, dos melhores da Europa neste exercício. Segundo um estudo feito pela **Eurostat (Instituto de Estatística da União Europeia)**, são os jovens dos países do Sul quem mais se preocupa em vestir roupa de marca.

Outro:

3. Se achar relevante, deixe-nos os seus comentários ou sugestões. Obrigado.

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious		sentence: unusual word order
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
<b>Países do Sul da Europa</b>	points to a previous reference that is not obvious
<b>as indústrias nacionais são a têxtil e a do calçado</b>	word too technical/specialized or archaic
<b>Eurostat (Instituto de Estatística da União Europeia)</b>	unknown word

### Domain and genre

	Social Media
	News

## Easy Level



2. Por favor, marque as palavras e/ou partes do texto que considera difíceis de entender. Selecione a palavra ou parte(s) do texto com o cursor e identifique a categoria de dificuldade dentre as opções na lista.

Todo o país em risco muito elevado de exposição à radiação UV

Para as regiões com risco muito elevado, o IPMA aconselha a utilização de óculos de sol com filtro UV, chapéu, t-shirt, guarda-sol, protetor solar e que se evite a exposição das crianças ao sol.

Todos os distritos de Portugal continental e os arquipélagos da Madeira e Açores apresentam esta quinta-feira um risco muito elevado de exposição à radiação ultravioleta (UV), segundo o Instituto Português do Mar e da Atmosfera (IPMA).

Na sexta-feira, Portugal continental e os arquipélagos da Madeira e Açores vão manter-se com níveis muito elevados e elevados de exposição a esta radiação.

A escala de radiação ultravioleta tem cinco níveis, entre risco extremo e baixo.

Para as regiões com risco muito elevado, o IPMA aconselha a utilização de óculos de sol com filtro UV, chapéu, t-shirt, guarda-sol, protetor solar e que se evite a exposição das crianças ao sol.

O IPMA recomenda para as regiões com risco elevado o uso de óculos de sol com filtro UV, chapéu, t-shirt e protetor solar.

Outro:

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

## Annotations

Token(s)	Category
<b>Para as regiões (...) crianças ao sol</b>	other (please specify) – “sentence must be removed because it repeats”
<b>Radiações ultravioleta (...) Atmosfera (IPMA).</b>	points to a previous reference that is not obvious
<b>Eurostat (Instituto de Estatística da União Europeia)</b>	unknown word

## Domain and genre

	Social Media
	Weather

## Easy Level



2. Por favor, marque as palavras e/ou partes do texto que considera difíceis de entender. Selecione a palavra ou parte(s) do texto com o cursor e identifique a categoria de dificuldade dentre as opções na lista.

Declaração Universal dos Direitos do Homem

Artigo 1º

Todos os seres humanos nascem livres e iguais em dignidade e em direitos.

Dotados de razão e de consciência, devem agir uns para com os outros em espírito de fraternidade.

Outro:

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious		sentence: unusual word order
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
<b>Declaração Universal dos Direitos do Homem</b>	points to a previous reference that is not obvious
<b>livres e iguais (...) espírito de fraternidade.</b>	too many connectors in the same sentence

### Domain and genre



Political Communication



Motion

## Plain Level



2. Por favor, marque as palavras e/ou partes do texto que considera difíceis de entender. Selecione a palavra ou parte(s) do texto com o cursor e identifique a categoria de dificuldade dentre as opções na lista.

Depois dos meses que passei a trabalhar numa apresentação para o departamento de Recursos Humanos da Kane Company, rejeitaram a minha transferência de emprego. Custa-me mais do que devia. Embora eu não estivesse propriamente a **almejar** demasiado alto com uma posição de entrada nos Recursos Humanos, tinha uma boa ideia com um futuro **promissor**. **Uma ideia que poderia beneficiar inúmeros disléxicos presos numa rotina corporativa.** O meu plano poderia fazer a empresa subir de **patamar**, caso me dessem uma oportunidade. Poda tentar novamente da próxima vez.

Outro:

Annotations	
Token(s)	Category
<b>almejar</b>	word
<b>promissor</b>	word
<b>patamar</b>	unknown word
<b>disléxico</b>	unknown word
<b>Uma ideia que (...) rotina corporativa.</b>	too much embedded secondary information

## Domain and genre

word: unknown word

word: word too technical/specialized or archaic

word: complex derived word

word: points to a previous reference that is not obvious

sentence: unusual word order

sentence: too much embedded secondary information

sentence: too many connectors in the same sentence

sentence (other)

word (other)

other (please specify)



Fiction



Novel



## Plain Level



2. Por favor, marque as palavras e/ou partes do texto que considera difíceis de entender. Selecione a palavra ou parte(s) do texto com o cursor e identifique a categoria de dificuldade dentre as opções na lista.

O meu nome é Andre Agassi. Sou casado com Stefanie Graf. Temos dois filhos, um menino e uma menina, de cinco e três anos de idade. Vivemos em Las Vegas, no estado do Nevada, mas, de momento, estamos instalados numa suíte do hotel Four Seasons de Nova Iorque, porque estou a participar no Open dos Estados Unidos de 2006. O meu último Open dos Estados Unidos. Na verdade, a última competição na qual participarei na vida. Ganho a vida a jogar ténis, **embora odeie o ténis; odeio-o com uma paixão secreta e sombria, sempre o odiei.**

Outro:

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious		sentence: unusual word order
sentence: too much embedded secondary information		sentence: too many connectors in the same sentence
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
<b>embora odeie o ténis (...) sempre o odiei.</b>	too much embedded secondary information

### Domain and genre



Non-fiction



Autobiography

## Very Easy Level



2. Marque las palabras o las partes del texto que crea que dificultan la comprensión del texto.

Seleccione la palabra o las partes del texto con el cursor e identifique la categoría de dificultad eligiéndola de la lista. Tenga en cuenta que las categorías se refieren al nivel del texto (no a su propio nivel como lector competente: ¿considera esta palabra "desconocida", "demasiado técnica" etc. para el nivel en cuestión).

Confesión  
¿Sabéis? A vosotros no os voy a mentir.  
Encontrar al amor de tu vida con 22  
es una bendición,  
pero sobre todo una putada.  
El destino se alía con el presente  
para decirte textualmente:  
Pero tú, niñato, ¿de qué vas?  
con todo lo que te queda por vivir.  
Y tienes que darle la razón.  
Hasta que aparece ella,  
repito, ella,  
y entonces aprendes que la libertad  
no era lo que tanto habías defendido,  
que la libertad también está  
en los desayunos por la mañana,  
en madrugar para irte de viaje,  
en **trasnochar** sin alcohol,  
en llegar de currar y que te abracen.  
Vamos, en todas esas cosas que tanto critiqué  
y que podéis encontrar en cualquier película  
mala de esas de Antena 3.  
Ahí también estaba la libertad,  
y las dudas,  
las dudas siempre están,  
pero en estos casos  
son el Imperio otomano  
y tú eres ese niño pequeño  
al que su madre ha dejado solo en la cola del súper.  
No puede ser ella, no puede ser ella.  
Me lo repito en bucle  
porque es la única manera de autoconvencerme:  
repetir la mentira,  
hasta que ya no recuerde  
cuál era la verdad.

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
<b>trasnochar</b>	word too technical/specialized or archaic

### Domain and genre



## Very Easy Level



2. Marque las palabras o las partes del texto que crea que dificultan la comprensión del texto.

Seleccione la palabra o las partes del texto con el cursor e identifique la categoría de dificultad eligiéndola de la lista. Tenga en cuenta que las categorías se refieren al nivel del texto (no a su propio nivel como lector competente: ¿considera esta palabra "desconocida", "demasiado técnica" etc. para el nivel en cuestión).

Es temprano, alrededor de las siete, y os aseguro que pocas veces me he despertado tan pronto en esta ciudad. Hoy he hecho un esfuerzo porque por una vez quería disfrutar a toda costa del amanecer sobre Barcelona desde aquí arriba. El mejor sitio para hacerlo es el restaurante bar Mirablau. Un clásico, nada especial desde un punto de vista **gastronómico**, pero con un ventanal que seguramente ofrece una de las mejores panorámicas de Barcelona. El camarero, **somnoliento** y gruñón, refunfuña algo para sus adentros mientras anota mi comanda. Típico de España. ¿O es sólo que no me ha entendido?

Le he pedido «un cortado y un sándwich mixto»; a lo mejor no ha querido entenderme. Pruebo en catalán: «Un tallat i un biquini».

Al final me lo trae «Aquí tens», me ladra cuando vuelve y me tira la taza y el plato encima de la mesa: 'Aquí tienes'. Busco el asiento más apartado de él.

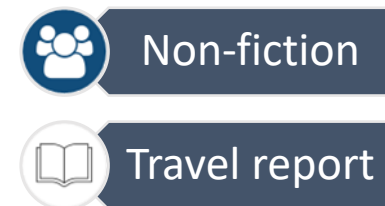
En el fondo me gusta esta **hosquedad**, este malhumor descarado que brota de las entrañas y se exhibe sin ningún **recato**. Porque mi abuelo era precisamente así.

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
<b>gastronómico</b>	word too technical/specialized or archaic
<b>somnoliento</b>	word too technical/specialized or archaic
<b>hosquedad</b>	word too technical/specialized or archaic
<b>recato</b>	word too technical/specialized or archaic

### Domain and genre



## Easy Level



2. Marque las palabras o las partes del texto que crea que dificultan la comprensión del texto.

Seleccione la palabra o las partes del texto con el cursor e identifique la categoría de dificultad eligiéndola de la lista. Tenga en cuenta que las categorías se refieren al nivel del texto (no a su propio nivel como lector competente: ¿considera esta palabra "desconocida", "demasiado técnica" etc. para el nivel en cuestión).

El derecho a la movilidad es una necesidad de imprescindible cobertura por las administraciones públicas en la sociedad moderna, que se concreta en la exigencia de mantenimiento de una red de transporte público de calidad.

En este sentido, la Unión Europea reconoce de manera específica el derecho de los ciudadanos a un transporte seguro, eficaz y de calidad en el Libro Blanco de los Transportes.

En línea con esta política comunitaria, el Gobierno español ha establecido recientemente la Estrategia Española de **Movilidad Sostenible**, aprobada por el Consejo de Ministros, de 30 de abril de 2009, donde se da carta de naturaleza a la **potenciación** del transporte público como una obligación de las Administraciones Públicas, promoviendo todo tipo de actuaciones encaminadas a la generación de una alternativa de movilidad al transporte privado que pueda considerarse realmente sostenible.

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

Annotations	
Token(s)	Category
<b>Movilidad Sostenible</b>	word too technical/specialized or archaic
<b>potenciación</b>	word too technical/specialized or archaic

## Domain and genre



## Easy Level



2. Marque las palabras o las partes del texto que crea que dificultan la comprensión del texto.

Seleccione la palabra o las partes del texto con el cursor e identifique la categoría de dificultad eligiéndola de la lista. Tenga en cuenta que las categorías se refieren al nivel del texto (no a su propio nivel como lector competente: ¿considera esta palabra "desconocida", "demasiado técnica" etc. para el nivel en cuestión).

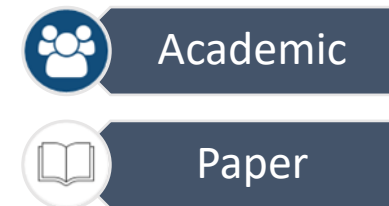
A lo largo de este estudio, hemos ofrecido los datos históricos que nos permiten deducir que el periodo de la primera mitad del s. XX fue clave para la institucionalización de la enseñanza del español para no nativos, con los primeros cursos de español para extranjeros tanto en España como en México. También lo fue para la consolidación de los departamentos y cátedras de español en universidades de prestigio españolas, del resto de Europa y de América, sin los que sería impensable la labor de tantos hispanistas como a lo largo de las décadas siguientes han estudiado la lengua y cultura hispánicas y han formado a nuevas generaciones que, desde otras lenguas maternas, han conseguido prender la llama del interés por ellas hasta el punto de dedicar a ello su profesión. El tema tratado es tan amplio y sus bifurcaciones tan interesantes, que por fuerza este estudio tiene claras limitaciones, pues conscientemente no profundiza en las figuras o instituciones citadas, sino que se centra en el hilo conductor que las relaciona.

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
También lo fue (...) su profesión.	unusual word order
cátedras	word too technical/specialized or archaic

### Domain and genre



## Plain Level



2. Marque las palabras o las partes del texto que crea que dificultan la comprensión del texto.

Seleccione la palabra o las partes del texto con el cursor e identifique la categoría de dificultad eligiéndola de la lista. Tenga en cuenta que las categorías se refieren al nivel del texto (no a su propio nivel como lector competente: ¿considera esta palabra "desconocida", "demasiado técnica" etc. para el nivel en cuestión).

(Sale CLARÍN, gracioso.)  
 Di dos, y no me dejes  
 en la posada a mi cuando te quejes; que si dos hemos sido  
 los que de nuestra patria hemos salido  
 a probar aventuras,  
 dos los que entre desdichas y locuras aquí habemos llegado,  
 y dos los que del monte hemos rodado, ¿no es razón que yo sienta  
 meterme en el pesar, y no en la cuenta? ROSAURA:  
 No quise darte parte  
 en mis quejas, Clarín, por no quitarte, llorando tu desvelo, el derecho que tienes al consuelo; que tanto  
 gusto había  
 en quejarse, un filósofo decía, que, a trueco de quejarse, habían las desdichas de buscarse. CLARÍN:  
 El filósofo era  
 un borracho barbón.  
 ¡Oh, quién le diera  
 más de mil bofetadas!  
 Quejarse después de muy bien dadas. Mas ¿qué haremos, señora,  
 a pie, solos, perdidos y a esta hora  
 en un desierto monte,  
 cuando se parte el sol a otro horizonte? ROSAURA:  
 ¡Quién ha visto sucesos tan extraños! Mas si la vista no padece engaños que hace la fantasía,  
 a la medrosa luz que aún tiene el día, me parece que veo  
 un edificio.

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
Di dos, y no me (...) no en la cuenta?	unusual word order
No quise (...) borracho barbón.	unusual word order
Oh, quién le	word too technical/specialized or archaic
más de mil bofetadas!	word too technical/specialized or archaic
Mas si (...) veo un edificio	unusual word order

### Domain and genre



## Plain Level



2. Marque las palabras o las partes del texto que crea que dificultan la comprensión del texto.

Seleccione la palabra o las partes del texto con el cursor e identifique la categoría de dificultad eligiéndola de la lista. Tenga en cuenta que las categorías se refieren al nivel del texto (no a su propio nivel como lector competente: ¿considera esta palabra "desconocida", "demasiado técnica" etc. para el nivel en cuestión).

3. Energía, trabajo y cambios mecánicos

### INTRODUCCIÓN

Hasta aquí hemos conseguido una explicación unitaria, común, a la existencia de distintos tipos de movimiento de los objetos y hemos obtenido un principio de conservación totalmente general en la naturaleza: la conservación de la cantidad de movimiento. No obstante, existen preguntas y situaciones de gran interés a las que la Dinámica sola no puede dar respuesta. Unas están relacionadas con el cambio en el movimiento de los objetos y otras se refieren a cambios que no pueden expresarse en términos de velocidades y aceleraciones.

Imaginemos, por ejemplo, la siguiente situación familiar: dos péndulos exactamente iguales hechos con bolas de acero cuelgan verticalmente de modo que las dos esferas se encuentran juntas. Separamos una de ellas ligeramente de su posición inicial manteniendo el hilo tenso y la soltamos de manera que choque frontalmente con la que estaba en reposo. ¿Qué es lo que observamos? Si realizamos la experiencia comprobaremos que la bola que estaba en movimiento se queda en reposo y la que estaba en reposo empieza a moverse hasta que alcanza prácticamente la misma altura desde la que se había soltado la primera.

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

## Annotations

Token(s)	Category
Hasta aquí hemos (...) común.	points to a previous reference that is not obvious
Dinámica	word too technical/specialized or archaic
péndulos	word too technical/specialized or archaic

## Domain and genre



## Very Easy Level

Casserole de chocolat chaud à l'ancienne

Ingrédients pour 4 personnes

- 190g Chocolat noir
- 1l Lait
- 10cl Crème liquide
- 1 cuil. à soupe Cacao

Etapes de préparation

**Hachez** le chocolat au couteau. Versez le lait et la crème dans une casserole, ajoutez le cacao **tamisé** et faites chauffer doucement. Fouettez puis, hors du feu, **incorporez** le chocolat haché et fouettez à nouveau pour bien faire fondre le chocolat. Servez bien chaud.

Si vous avez indiqué "Autre", donnez une indication sur le type de difficulté identifié :

Commentaire ou suggestion pour ce texte (optionnel) :

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
<b>Hachez</b>	Unknown word
<b>tamisé</b>	Unknown word
<b>incorporez</b>	Unknown word

### Domain and genre

	Didactic
	Recipe



## Very Easy Level

Le coq est mort

Le coq est mort,  
Le coq est mort,  
Le coq est mort.

Il ne dira plus cocodi, cocoda  
Il ne dira plus cocodi, cocoda,  
cocodicodi, codicoda  
cocodicodi, codicoda

Si vous avez indiqué "Autre", donnez une indication sur le type de difficulté identifié :

Commentaire ou suggestion pour ce texte (optionnel) :



word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
<b>cocodi</b>	Unknown word
<b>cocoda</b>	Unknown word

### Domain and genre



Lyric



Fiction

## Very Easy Level

Mardi passé, je me suis levée tôt.  
J'ai pris un énorme sac et je suis partie au marché.  
J'arrive près de l'arrêt du tram et... le tram démarre. Zut!  
Alors, je repars à l'appart car je me dis: « Pas de nuages, pas de pluie! Le vélo sera plus rapide. »

Arrivée au marché, j'achète un tas de légumes:  
des tomates, des pommes de terre, des carottes, des haricots et du basilic.  
Je me dis: « Ce sera un super potage pour la famille! »

J'ai repris le vélo et... un chiot me suit, il va vite, il va me rattraper, il hurle: «Rrrwaf! Rrrwaf! » comme un malade.  
Il va attaquer le vélo. Alors je m'arrête net. Catastrophe!

Le gros sac glisse du vélo sur le sol.  
Les pommes de terre s'échappent sur la rue.  
Les tomates s'écrasent sur la rue.  
Les haricots s'étaient sur la rue.  
Le pot de basilic éclate sur la rue.  
Je ramasse les légumes abîmés vite, vite car le tram sonne, le bus rôle, les voitures trépignent.  
Je suis fâchée: Je me dis: « Mardi, pas de potage pour ma famille plutôt une bonne purée! »

Si vous avez indiqué "Autre", donnez une indication sur le type de difficulté identifié :

Commentaire ou suggestion pour ce texte (optionnel) :

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
l'appart	Unknown word
rôle	Unknown word
trépignent	Unknown word

### Domain and genre

	Personal Communication
	Diary

## Easy Level

Mercredi 12/04/23 Bonjour,  
Il y a **plusieurs** problèmes avec le matériel.  
D'abord, il faut de nouvelles chaussures de sécurité pour toute l'équipe. Ensuite, le camion est en panne et on ne peut plus faire de **livraison**.  
Merci de régler ces problèmes rapidement !  
Mickaël SANTORO Le chef d'équipe

Si vous avez indiqué "Autre", donnez une indication sur le type de difficulté identifié :

Commentaire ou suggestion pour ce texte (optionnel) :

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

## Annotations

Token(s)	Category
<b>plusieurs</b>	Unknown word
<b>livraison</b>	Unknown word

## Domain and genre


Institutional/Professional Communication


Letter

## Easy Level

Je me décris

Je m'appelle Thérèse. J'ai presque 70 ans. J'ai les cheveux blonds et gris. De plus en plus gris et blancs. Ils sont courts. J'ai les yeux bleus. Je mesure **1m65**. Je ne porte pas de lunettes. Je porte un pantalon gris et une chemise orange.

Réponds aux questions :

- Comment t'appelles-tu ? Quel âge as-tu ?
- De quelle couleur sont tes cheveux ?
- Sont-ils longs ou courts ?
- De quelle couleur sont tes yeux ?
- Portes-tu des lunettes ?
- Combien mesures-tu ?
- Que portes-tu comme vêtements ?

Si vous avez indiqué "Autre", donnez une indication sur le type de difficulté identifié :

Commentaire ou suggestion pour ce texte (optionnel) :

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

## Annotations

Token(s)	Category
<b>1m65</b>	Unknown word

## Domain and genre



Didactic



Manual

## Easy Level

Bienvenue à Gardanne

Gardanne est une ville située dans le département des Bouches du Rhône, dans le sud de la France.

C'est une ville avec beaucoup d'histoire.

Autrefois, les hommes de Gardanne travaillaient à la mine de Gardanne. Celle-ci n'existe plus aujourd'hui.

A Gardanne, de nombreuses activités et balades sont possibles : Randonnées, visites, pêche.

Vous pourrez aussi découvrir nos événements festifs :

Le Marché de Noël, la Semaine Provençale, la Fête de la Saint-Jean et les Journées du Patrimoine.

Bonne visite à Gardanne.

Si vous avez indiqué "Autre", donnez une indication sur le type de difficulté identifié :



Commentaire ou suggestion pour ce texte (optionnel) :

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

### Annotations

Token(s)	Category
<b>Gardanne</b>	Unknown word
<b>Autrefois</b>	Unknown word
<b>Provençale</b>	Unknown word
<b>Saint-Jean</b>	Unknown word
<b>Journées du Patrimoine</b>	Unknown word

### Domain and genre

	Non-fiction
	Travel guide

## Plain Level

Ça y est, je suis rentrée dans le XXI<sup>e</sup> siècle, je suis connectée à Internet. Je **surfe** je navigue, enfin pour l'instant je rame. Ça a commencé quand j'ai acheté l'ordinateur.

- Monsieur, je voudrais un **Mac** parce que PC, ça veut dire **plante** constamment.

- Mac ou PC, c'est pareil, dans trois mois, votre matériel sera **obsolète**, j'arrive.

- Faut peut-être mieux que j'attende trois mois ?

- C'est pareil madame, avec l'informatique, tout va vite, tout va très très vite.

C'est vrai que ça va vite, en cinq minutes, j'ai dépensé 1398 euros.

En plus mon ordinateur, j'essaie de faire tout ce qu'il me dit mais lui il fait rien de ce que je veux **déjà quand il me parle**, je comprends rien :

« Vous avez mal éteint l'ordinateur, nous allons le **reconfigurer** »

Qui ça nous ? Ils sont plusieurs là-dedans ?

« L'application ayant servi à créer ce document est introuvable. »

**Si lui il la trouve pas** comment je la trouve moi ?

« Une erreur système est **survenue inopinément** »

**Genre** t'as **une erreur système qui se promène** : « Je suis une erreur système, qu'est ce que je vais faire ? Tiens, je vais **survenir inopinément** »

« Veuillez libérer de la mémoire. »

Je demande pas mieux moi. **Mémoire, par ordre de sa majesté, je vous libère** ! Où elle est la touche mémoire ? Ya pas de touche

mémoire. Tu sais ce que ça veut dire PC ? P'tit con. Il est très poli

mon ordinateur, **j'ai beau l'insulter** il continue de me vouvoyer. **Poli**

**mais mauvais caractère** des fois il se **braque** y'a plus aucune

touche qui marche : « **Bad Command, invalid Response** »

Quand il parle anglais, c'est qu'il est très énervé.

Si vous avez indiqué "Autre", donnez une indication sur le type de difficulté identifié :

Commentaire ou suggestion pour ce texte (optionnel) :

## Domain and genre



Fiction



Drama

word: unknown word    word: word too technical/specialized or archaic    word: complex derived word

word: points to a previous reference that is not obvious

sentence: unusual word order

sentence: too much embedded secondary information

sentence: too many connectors in the same sentence

sentence (other)

word (other)

other (please specify)

## Annotations

Token(s)	Category
<b>surfe</b>	Unknown word
<b>Mac</b>	Unknown word
<b>plante</b>	Unknown word
<b>obsolète</b>	Unknown word
<b>Déjà quand il me parle</b>	Unknown word
<b>reconfigurer</b>	Unknown word
<b>Si lui il la trouve pas</b>	Unknown word
<b>Survenue inopinément</b>	Unknown word
<b>Genre</b>	Unknown word
<b>Une erreur (..) promène</b>	Unknown word
<b>Mémoire, (...) libère.</b>	Unknown word
<b>J'ai beau l'insulter</b>	Unknown word
<b>Poli mais mauvais caractère</b>	Unknown word
<b>braque</b>	Unknown word
<b>Bad (...) Response</b>	Unknown word

## Plain Level

Une **sexagénaire** suédoise a eu la surprise de recevoir par la poste le portefeuille qu'elle avait perdu quarante ans plus tôt avec tout l'argent qu'il contenait à l'époque.  
 Gulli Wihlborg avait 18 ans lorsqu'elle **égara** son porte-feuille en faisant du vélo dans Trelleborg durant l'été 1963. Il contenait 45,54 couronnes (l'équivalent de 412 couronnes aujourd'hui, soit 45 euros), la moitié de son loyer d'alors, a-t-elle expliqué.  
 Le paquet est arrivé à son domicile de Malmö, où elle habite depuis vingt-cinq ans, avec une **note manuscrite** anonyme disant : « Chère Gulli, il ne faut jamais désespérer. Voici le portefeuille perdu dans la rue Ostersjogatan il y a plusieurs années. Salutations de Trelleborg. »  
 « C'est assez **inouï** », a déclaré Gulli Wihlborg au journal local Trelleborgs Allehanda. **L'expéditeur** reste inconnu.

Si vous avez indiqué "Autre", donnez une indication sur le type de difficulté identifié :

Commentaire ou suggestion pour ce texte (optionnel) :

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

## Annotations

Token(s)	Category
<b>sexagénaire</b>	Unknown word
<b>égara</b>	Unknown word
<b>note manuscrite</b>	Unknown word
<b>inouï</b>	Unknown word
<b>L'expéditeur</b>	Unknown word

## Domain and genre



## Plain Level

Quelques conseils pour réduire nos quantités d'emballages

- Préférons les produits vendus en **vrac**, ou **à la découpe**, dans les commerces de proximité (boucherie, fromagerie, etc.).
- Refusons les produits suremballés.
- Prévoyons un panier, un sac ou un filet, pour faire les courses, et refusons les sacs jetables.
- Choisissons des produits concentrés (produits d'entretien, lessives, etc.) ; ils nécessitent moins d'emballages, et sont aussi efficaces. Ou encore mieux, faire ses produits d'entretien soi-même.
- Adaptons nos achats à nos besoins : six petits yaourts demandent plus d'emballages qu'un seul grand pot de quantité **équivalente. Proportionnellement**, les grands **conditionnements** sont moins chers et moins polluants.
- Choisissons des emballages réutilisables ou **consignés**.
- Buons l'eau du robinet ; elle est de bonne qualité, ne nécessite aucun emballage et coûte 500 fois moins cher que la plupart des eaux en bouteille !
- **Dressons** une liste de courses avant de nous rendre au magasin : cela nous permettra de moins nous laisser tenter par des achats imprévus.

Si vous avez indiqué "Autre", donnez une indication sur le type de difficulté identifié :

Commentaire ou suggestion pour ce texte (optionnel) :

word: unknown word	word: word too technical/specialized or archaic	word: complex derived word
word: points to a previous reference that is not obvious	sentence: unusual word order	
sentence: too much embedded secondary information	sentence: too many connectors in the same sentence	
sentence (other)	word (other)	other (please specify)

## Annotations

Token(s)	Category
<b>vrac</b>	Unknown word
<b>à la découpe</b>	Unknown word
<b>équivalente. Proportionnellement</b>	Unknown word
<b>conditionnements</b>	Unknown word
<b>consignés</b>	Unknown word
<b>Dressons</b>	Unknown word

## Domain and genre



Institutional/Professional Communication



Fact sheet



## Students' Annotations

## Very Easy Level



b) Por favor, marque as palavras ou partes do texto que considera que fazem o texto difícil de compreender.

Selecione a palavra/parte do texto com o rato:

-----

BEM-VINDO, RODOLFO!

Rodolfo é o mais recente animal nascido no Jardim Zoológico de Lisboa. E não é um bicho qualquer! Trata-se de um veado-da-birmânia, uma espécie em perigo. Rodolfo nasceu no final de outubro, depois de cerca de oito meses na barriga da mãe. As pessoas caçam estes animais para os comer e também para ficar com as hastes que os machos desenvolvem todos os anos, e é por essa razão que esta espécie está em perigo.

Difficult word

Difficult part of text

### Annotations

Token(s)	Category
Zoológico	Difficult word
Veado-da-birmânia	Difficult word
espécie	Difficult word
hastes	Difficult word

### Domain and genre



Social Media



News

## Easy Level



b) Por favor, marque as palavras ou partes do texto que considera que fazem o texto difícil de compreender.

Selecione a palavra/parte do texto com o rato:

-----

Havia um templo no alto das montanhas.  
Rodeado por uma vasta floresta.  
E um lago profundo e parado.  
O templo já vira melhores dias.  
Mas, para os dois amigos que fizeram daquele antigo lugar a sua casa, isso não tinha importância.  
Durante o dia, eles subiam as montanhas.  
E **exploravam** as **densas** e **emaranhadas** florestas, na esperança de **vislumbrarem** as criaturas que ali viviam.  
À noite, observavam as estrelas e bebiam o chá quente que o Pequeno Dragão preparava com tanto cuidado.

Difficult word

Difficult part of text

### Annotations

Token(s)	Category
<b>exploravam</b>	Difficult word
<b>densas</b>	Difficult word
<b>emaranhadas</b>	Difficult word
<b>vislumbrarem</b>	Difficult word

### Domain and genre



Non-fiction



Travel report

## Plain Level



b) Por favor, marque as palavras ou partes do texto que considera que fazem o texto difícil de compreender.

Selecione a palavra/parte do texto com o rato:

### SOCIEDADE/ OBITUÁRIO

Morreu o político e advogado Adriano Moreira. Tinha 100 anos

O Este artigo tem mais de 6 meses

Antigo ministro do Ultramar, ex-presidente do CDS, advogado e professor universitário, Moreira foi um dos primeiros académicos portugueses a refletir sobre o colonialismo português no século XX.

23 out. 2022, 12:36

Rita Cipriano

Texto

Adriano Moreira morreu este domingo de manhã aos 100 anos, avançou o Diário de Notícias, que confirmou a notícia junto de fonte familiar. **A Agência Lusa** obteve confirmação junto de um membro do CDS, de que Moreira foi presidente entre 1986 e 1988.

**Condecorado** pelo Presidente da República em junho passado com a Grã-Cruz da Ordem de Camões, antes de completar 100 anos, Moreira destacou-se não só como político e **estadista**, mas também como professor universitário e pensador na área das Relações Internacionais.

Antigo ministro do Ultramar do Estado Novo, ex-presidente do CDS, advogado e professor universitário, Adriano Moreira nasceu a 6 de setembro de 1922 em Grijó, em Macedo de Cavaleiros. Mudou-se para Lisboa em criança, quando o pai, António José Moreira, foi nomeado subchefe da Polícia de Segurança Pública no porto de Lisboa. Morou no bairro de Campolide.

Difficult word

Difficult part of text

### Annotations

Token(s)	Category
<b>A Agência Lusa</b>	Difficult part of text
<b>Condecorado</b>	Difficult word
<b>estadista</b>	Difficult word

### Domain and genre



Social Media



Obituary

## Very Easy Level



b) Por favor, marca las palabras o las partes del texto que no has entendido.  
Selecciona la palabra o la parte del texto con el cursor.

-----

Habíamos pasado el fin de semana más romántico de nuestra corta historia juntos. Dos días completos en su casa de la playa. Dos despertares sin prisa. Dos desayunos con sus amplias **sobremesas**, dos largos paseos, incluso un baño helador en el mar. Nada de oscuridades y rincones. Nada de contraseñas o de sexo prohibido en el interior de un coche. Todo a la luz del día, como si por una vez fuéramos una pareja de verdad y sin nada que ocultar.

Patricia, su mujer, llamó solo una vez en todo ese tiempo. Fue un momento tenso en medio de tanta felicidad. Era sábado por la tarde y bebíamos unas copas de vino frente a la chimenea. Kerman cogió la llamada y se metió en su despacho sin cerrar la puerta, así que le oí mentir. Hablarle de la aburrida monotonía de su fin de semana **ermitaño** en la playa, «ya sabes, todo igual». Le dijo que estaba trabajando en la reforma del **granero**, terminando uno de los baños de la planta baja. ¿Qué tal ella por Madrid?, preguntó. Yo estaba desnuda sobre la alfombra, con la copa en la mano, casi aguantando la respiración.

Difficult word

Difficult part of text

### Annotations

Token(s)	Category
<b>sobremesas</b>	Difficult word
<b>ermitaño</b>	Difficult word
<b>granero</b>	Difficult word

### Domain and genre



Fiction



Novel

Easy Level



b) Por favor, marca las palabras o las partes del texto que no has entendido.  
Selecciona la palabra o la parte del texto con el cursor.

FÓSIL

Son los restos conservados de organismos desaparecidos hace mucho tiempo. Puede ser una parte dura inalterada (un diente o un hueso), la petrificación (de madera o hueso), partes blandas inalteradas o parcialmente alteradas (un mamut congelado).

Fago

Virus que parasita a una bacteria

Fagosomas

Nombre con el que se designan a las vacuolas cuando efectúan procesos digestivos en su interior. Si degradan estructuras propias de la célula, se llaman autofagosomas.

Fecundación

Fusión de dos núcleos gaméticos haploides; forman el núcleo de un cigoto diploide. 1. f. interna: los gametos se encuentran y fusionan en estructuras internas del organismo 2. f. externa: los gametos se encuentran y se fusionan externamente a los organismos involucrados en el proceso de la fecundación.

Fecundación cruzada

fusión de gametos formados por diferentes individuos; opuesto a autofecundación.

Fenotipo

Conjunto de caracteres que manifiestan los individuos de una especie en un ambiente determinado.

Feofitas

Filo del grupo de las algas. Sinónimo de algas pardas.

Fermentación

Reacción de degradación de compuestos orgánicos en ausencia de oxígeno; produce menos energía que los procesos aeróbicos.

Fibra muscular

Célula muscular; célula larga, cilíndrica, multinucleada, que contiene numerosas miofibrillas y se contrae cuando se estimula.

Difficult word

Difficult part of text

Annotations

Token(s)	Category
petrificación	Difficult word
Fago	Difficult word
Fagosomas	Difficult word
vacuolas	Difficult word
autofagosomas	Difficult word
gaméticos haploides	Difficult word
gametos	Difficult word
involucrados	Difficult word
Fenotipo	Difficult word
Feofitas	Difficult word
multinucleada	Difficult word

Domain and genre



Didactic



Glossary

## Plain Level



b) Por favor, marca las palabras o las partes del texto que no has entendido.  
Selecciona la palabra o la parte del texto con el cursor.

Abandonados (hijos)

Los menores de 18 años, o mayores con **discapacidad** igual o superior al 65 %, y hayan sido abandonados por sus padres, siempre que no se encuentren en régimen de acogimiento familiar, pueden ser beneficiarios de la asignación económica por hijo o menor acogido a cargo.

Accidente de Trabajo (AT)

Se entiende por accidente de trabajo toda lesión corporal que el trabajador sufra con ocasión o por consecuencia del trabajo que ejecute por cuenta ajena. En el trabajo por cuenta propia de los **Regímenes Especiales de trabajadores Autónomos y de trabajadores del Mar**, se entenderá como accidente de trabajo el ocurrido como consecuencia directa e inmediata del trabajo que realiza por su propia cuenta y que determina su inclusión en el campo de aplicación del régimen especial. La ley establece, además, diversas consideraciones y presunciones para determinar el concepto de accidente de trabajo.

Accidente no laboral

Lesión o alteración de la salud derivada de accidente siempre que éste no sea consecuencia del trabajo realizado.

Difficult word

Difficult part of text

### Annotations

Token(s)	Category
<b>discapacidad</b>	Difficult word
<b>Regímenes Especiales de trabajadores Autónomos y de trabajadores del Mar</b>	Difficult part of text

### Domain and genre



Didactic



Glossary