Online-Journals.org

iJET International Journal of Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 19 No. 8 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijet.v19i08.52023

PAPER

Intelligent Support for Low Literacy Adults: The European Portuguese iRead4Skills Corpus

Maria Leonor Reis(✉), Sílvia Barbosa, Michell Moutinho, Ricardo Monteiro, Susana Correia, Raquel Amaro

NOVA CLUNL, Universidade NOVA de Lisboa, Lisbon, Portugal

mariareis@fcsh.unl.pt

ABSTRACT

This paper presents the Portuguese dataset of the *iRead4Skills project (Dataset 1: corpora by complexity level for FR, PT, and SP* – v.2.0), a representative sample of written European Portuguese for automatic complexity assessment that addresses a gap in existing resources for Portuguese. The corpus was created within the framework of the iRead4Skills project, which encompasses Portuguese, French, and Spanish. The project aims to develop an intelligent system to evaluate text complexity while recommending appropriate reading materials to native adult learners with low literacy skills. The corpus compilation involved a manual selection of text samples across various textual genres and document types, covering a wide range of existing written materials and focusing on the reading needs and reading habits of the target audience—low literacy adults enrolled in vocational education and training centres or adult learning (AL) centres. The collected texts were categorised into the three distinct levels of complexity targeted and defined by the project: very easy, easy, and plain levels. Texts of higher complexity were also included, resulting in the creation of four distinct sub-corpora. The resulting Portuguese dataset consists of 2,186 texts and 942,818 tokens and serves as the foundational source for training and testing the project's complexity analysis systems. This paper presents a comprehensive overview of the compilation process of the corpus, encompassing its methodological design and the challenges faced. Although some existing Portuguese corpora were used for complexity studies and tool development, these primarily consist of texts classified according to CERF levels and retrieved from didactic materials designed for L2 teaching/learning or texts produced by L2 learners. The corpus presented in this paper introduces a new resource that addresses a significant gap in materials needed to inform and support studies and applications related to text complexity. The resulting dataset provides a novel and important language resource for European Portuguese, with several applications including research on linguistic complexity, development of automatic text complexity and readability assessment systems, and educational purposes.

KEYWORDS

classified corpus, text complexity, adult learning (AL), low literacy

Reis, M.L., Barbosa, S., Moutinho, M., Monteiro, R., Correia, S., Amaro, R. (2024). Intelligent Support for Low Literacy Adults: The European Portuguese iRead4Skills Corpus. *International Journal of Emerging Technologies in Learning (iJET)*, 19(8), pp. 61–81. https://doi.org/10.3991/ijet.v19i08.52023

Article submitted 2024-08-01. Revision uploaded 2024-09-18. Final acceptance 2024-09-23.

1 INTRODUCTION

Reading skills greatly influence how we gather information and comprehend the world. They are indispensable for acquiring knowledge across all facets of life [1] [2]. Individuals with limited literacy skills face heightened challenges in acquiring and retaining skills necessary for navigating a rapidly evolving job market [3]. Moreover, adult learning presents challenges, exacerbated by poor reading skills and the lack of appropriate reading materials designed for this demographic. This shortcoming further diminishes adults' motivation to learn, as experts described, underscoring how some adult learners feel patronized upon realizing that their learning materials resemble those of their children or grandchildren [4]. This is especially relevant in the context of formal and informal education [5], such as adult learning (AL) and vocational educational training (VET), and in practical and empirical work contexts.

While several European Portuguese corpora exist, they do not meet the project's requirements due to either being intended for non-native speakers (such as [6]) or being restricted access.

The target audience for the iRead4Skills project[1] comprises native adult speakers with low literacy levels, namely with poor reading skills. The project's primary objectives are to promote the development of reading skills by creating an open-access, web-based iRead4Skills system. This system will include an automated complexity assessment tool to evaluate text difficulty and recommend reading materials suitable for the user's reading proficiency level. As its general goal, the system aims to improve reading literacy, bridge skill gaps, and facilitate access to diverse information and culture. It serves as a resource for trainers in AL and VET centres, enabling them to select and/or adapt texts to match the abilities of their learners. The iRead4Skills project combines an interdisciplinary team covering fields such as ICT, linguistics, economics, and education, and bringing together governmental entities, tech companies, and research institutions to provide ground breaking research and innovation.

The iRead4Skills project acknowledges the multifaceted nature of training this demographic, comprehending fundamental reading skills and motivational factors. Consequently, the development and testing of the system are informed by direct input from AL and VET trainers and trainees. This collaborative effort spans from initial surveys to discern reading needs and preferences to annotating and classifying texts based on complexity descriptors[2] informed by end-users' sensitivity to complexity.

The project aims to contribute to (i) upskilling and reskilling adults with low literacy skills; (ii) developing their flexibility and adapting abilities to stay apace with the job market and to accommodate new skills related to technology; (iii) influencing the creation of new educational strategies related to formal education, answering to individual and social needs, and also boosting adults' motivation and participation; and (iv) influencing innovative and inclusive education systems, using digital technologies to provide quality training.

The design and compilation of corpora were fundamental to fulfilling the project goals. Although the project addresses three target languages, Portuguese, Spanish, and French, we will only consider the Portuguese corpus compilation process. The European Portuguese iRead4Skills corpus addresses a gap in data concerning texts suitable for native adult speakers with poor reading skills, covering different real

[1] iRead4Skills is a Horizon Europe project responding to the topic 'Conditions for the successful development of skills matched to needs,' HORIZON-CL2-2022-TRANSFORMATIONS-01-07. https://iread4skills.com/

[2] Complexity descriptors are characterizations of textual complexity based on several linguistics and extralinguistic dimensions (cf. [7]).

written materials—both digital and in print—that cover distinct topics, genres, and communication situations.

In this paper, we address the following aspects of this new resource: In Section two, we present the steps undertaken to prepare the corpus compilation, namely the surveys conducted and the clarification of what is understood by textual complexity. In Section three, we detail the methodology supporting the development of the iRead4Skills Dataset 1, which guided us through the creation of the corpus, covering the definition of the relevant text typology and metadata, the text selection and organisation phase, and the validation of the data in terms of the text classification. In Section four, we present an overview of the current results, focusing on understanding what was gained in the process and discussing the outcomes and challenges of the corpus compilation process. Finally, in Section five, we present the corpus contributions to ongoing research, AL and VET activities, and open questions for further development.

2 TEXTUAL COMPLEXITY

Textual complexity can be looked upon from many different approaches and perspectives, such as i) quantitative approaches, which propose readability formulas [8] [9] based on descriptions of different types of texts' features that would impact text readability and comprehension [10]; ii) qualitative approaches, which, besides focusing on complexity features by themselves, also highlighted how these features would impact the reader's comprehension [11]. Therefore, the notion of textual complexity may vary considerably among authors. While some establish it as only depending on linguistic features (i.e., the number of features belonging to text in a specific language that can be analysed and studied separately from the reader's difficulties with it [12] [13]), others see it as also depending on variants apart from linguistic features (e.g., concepts used in the text or the overall context of communication) [14] [15]. These approaches, however, tend to disregard the characteristics of specific groups, as demonstrated in [16] [17] [18], and impact the performance of complexity analysis systems [19] [20] [21], to name a few.

For the iRead4Skills project, textual complexity is the interplay between linguistic features and extralinguistic knowledge. In this perspective, whether a text is deemed complex depends on its content and/or language, but it also depends on extralinguistic knowledge such as concepts or contexts of information. Within this framework, compiling the European Portuguese iRead4Skills corpus required careful consideration of several factors before collecting the written materials or texts. Firstly, it was informed by results on reading needs and preferences from the target audience, the learners in AL and VET centres. These were collected through a specifically designed survey focused on reading skills [22] [23]. Secondly, it required defining the relevant textual complexity levels for the project, determined through objective and comprehensive descriptors validated by AL and VET trainers. The compilation of the corpus and the validation of the descriptors were performed in tandem, allowing us to achieve a better equilibrium between what we aimed to describe and what we found in the compiled materials. This approach facilitated ongoing feedback between the text compilation process and the fine-tuning and initial validation of the descriptors used to establish the complexity levels.

The definition of the iRead4Skills complexity levels is further discussed in the next subsection, as the complexity levels played a crucial role in the compilation of the corpus.

2.1 Definition of the complexity levels

In the iRead4Skills project, we defined intra-language and cross-linguistic descriptors to build a text complexity analysis and classification framework. We considered linguistic and paralinguistic properties of texts, creating three distinct levels of complexity with hereditary components (i.e., the descriptors validated to a level 1 are also considered for levels 1 + 1 and 1 + 2). The three levels obtained can be described in lay terms as follows[3]:

- **Very easy:** Level concerning texts that are fully or almost fully understood by everyone, including people with very low schooling (i.e., that did not finish primary school (ca. 6th year)) and almost no reading experience.
- **Easy:** Level concerning texts that are fully or almost fully understood by people with low schooling (i.e., that completed primary school but do not have more than the 9th year) and with poor reading experience.
- **Plain:** Level concerning texts that are understood the first time they are read by people who completed the 9th year and have a functional-to-average reading experience.

In addition to the complexity levels deemed pertinent to the project, we used a higher complexity level (+Complex), addressing texts beyond the project's scope to serve as a reference. The collection process considered this aspect, ensuring a balanced and usable outcome for the system under development. The three complexity levels targeted are further described below. More comprehensive information on this process, including the full set of descriptors associated with each level, can be consulted in [24].

i. Very easy level

Short (approximately 50 words) **and simple texts** to perform familiar tasks OR short and simple texts introducing new information (e.g., didactic texts).

Typically, simple, everyday concepts. It is assumed that the speaker has limited access to communication domains. Ideally, the topic is previously presented.

Basic communication contexts: Basic day-to-day (transport schedules/lists, menus/general instructions, item price information); **family/personal communications; simple/basic information.**

Absence of figures of speech.

Basic lexicon (active lexicon): Known words and simple expressions memorized = used in everyday matters. (e.g., transport, food, family, work). Frequent and concrete main and copulative verbs and frequent and concrete nouns, that is, concepts/ideas with a higher level of concreteness than abstraction. Rare affixation; except for frequent affixes such as PT-mente.

Short periods, with simple conjunctions and in direct order (**Subject-Verb-Object**). Rare auxiliary verbs (except for copulative verbs) and few anaphoric references: the referential chain is complete and does not occur in an elliptical way (e.g., *O João não gosta dela porque ela não é simpática*). 'João doesn't like her because she's not nice' vs. *O João não gosta dela por não [-] ser simpática.* 'João doesn't like her for [-] not being nice'. **Coordination structures** (noun phrase and noun phrase, adjective and adjective): copulative, disjunctive, and adversative conjunctions are admitted. Some frequently used subordination structures (subordinate temporal adverbials, for example), except for those less frequent (reduced adverbials of infinitive), are admitted.

[3] While labelling the complexity levels, we took a transparency factor into account. Simplicity and transparent names that characterize the texts and not the readers may impact the use of the system by the target audience.

Occurrence of some periphrastic constructions, more usual and therefore more decipherable (e.g., PT *estar/começar/andar* + *a* + Infinitive (~ *to be/start* + *-ing* Verb form); PT *deixar/acabar* + *de* + Infinitive; *ir* + *Infinitive*) (~ *to stop* + *-ing* Verb form). No compound tenses.

Use of indicative verb tenses (PT except for the simple future (compound tense - *ir* 'go' + Infinitive - or present + adverb)). Personal Infinitive and Gerund are admitted.

Simple temporal location. Temporal cohesion is given through temporal adverbs or connectors (*today, tomorrow, before, after* ...) and not by verb tenses (e.g., 'She left before John arrived.' vs. 'John arrived and she had left').

ii. Easy level

Short texts (approximately 150 words) that are interesting for the reader to inform themselves or in moments of leisure or to carry out tasks.

Some presence of abstract concepts (such as feelings, states of mind, religiosity, qualities, and defects, etc.).

Concepts related to the reader's personal and professional experiences.

Usual communication contexts: work context (specific instructions); media (news of interest, e.g., sports); commercial communication (ads).

Commonly used figures of speech (e.g., *os preços dispararam,* 'the prices skyrocketed').

Basic lexicon (active lexicon) and expanded **passive lexicon** with frequent words. Main and copulative verbs and frequent nouns in different domains where the reader routinely interacts or is interested in.

Some affixation, frequent and productive prefixes, and suffixes (e.g., *-agem, -ção, -eiro, -mente, -ade, -íssimo, -inha, -ice*).

Short periods, with coordinated conjunctions and most of the subordinate conjunctions, both in direct order (subject-verb object) and in other possibilities. Admits subject relative subordinate clauses but not object clauses (e.g., *O rapaz que abraçou a sua mãe.* 'The boy who hugged his mother.' *vs. O rapaz que a mãe abraçou.* 'The boy who his mother hugged.'). Admits subordinates with indicative, subjunctive, and infinitive.

Verbs in simple tenses, including simple future (but not frequently). Some periphrastic constructions, such as the passive voice (especially in the Indicative). Compound tenses are present (e.g., *Pretérito Mais que Perfeito Composto* (*tinha* + Past Participle).

More complex temporal reference in linear sequence. Temporal cohesion can be given via verb tenses. The reader can link different parts of the text and make a global sense of them.

iii. Plain level

Texts of different sizes (approximately 250 words) **and on varied topics** of interest to the reader for information or leisure.

Varied concepts. Readers can step out of their comfort zone. More contact with the online world. The reader can infer at a more complex level (e.g., infer opinions from opinion texts), including at a multimodal level, with texts in less common formats (e.g., infographics).

Various communication contexts: Leisure (stories; travel diaries, fiction); professional (theoretical articles); media (reportage, opinion articles); online (forums) communication contexts.

Varied lexicon to express subjects in any of the communication domains. **Less passive lexicon,** due to diverse contact. Presence of polysemic words. Occurrence of frequent foreign words (e.g., *timing, hobby, show*).

Nouns that express both concrete and abstract concepts to describe situations, reactions, emotions, thoughts, etc. Some frequent domain-specific verbs and nouns, describing trendy or situations known to the reader. Main and copulative verbs, frequent, and some domain specific.

Occurrence of most affixations (*-ite, -itude, -ume -vel, -oso, -ismo*), except for erudite and less frequent affixes. The presence of frequent compound words, which can have their meanings retrieved from each component. The reader can infer the meaning of derived words and some more frequent and non-idiomatic compound nouns.

Longer periods, with simple and compound sentences and a greater variety of conjunctions and syntactic order.

Some high-frequency irregular verbs are admitted. Presence of modal verbs, with uses and meanings in common expressions and in unusual contexts. Indicative, subjunctive imperative, and conditional moods, both in the active voice and in the passive voice. Passives with *-se* (e.g., *O trabalho faz-se bem,* 'The work does itself well').

Complex temporal reference in non-linear sequence. The reader can infer information (albeit basic) that is not explicit in the text.

As explained later, a thorough and objective description of the levels and their key features was crucial for guiding the informed collection and classification of the texts that make up the corpus.

3 IREAD4SKILLS DATASET 1 V2.0 METHODOLOGY

The corpus presented here is an open dataset[4] that reflects the complexity levels relevant for European Portuguese adult native speakers with poor reading skills and their needs and expectations.

As stated before, textual complexity arises from multiple factors, from simple metrics such as the number of syllables per word or number of words per sentence to subjective elements such as the level of concreteness of concepts used, impacting different ways and dimensions of comprehension. Thus, in corpus compilation, while accounting for the multiple linguistic and extralinguistic properties of the texts, such as topic and communication purpose, it was also considered adult readers' preferences and needs to ensure that the sample covered relevant communication situations and topic preferences and/or trends. Also, being a corpus for specific purposes, namely training the automatic complexity analysis tools, the idea was to produce a robust set of diverse materials that handles a wide sample of potential texts the system will have to analyze.

Regarding text length, although relevant for the very easy level, as reflected in the level description and as demonstrated by several complexity analyses [25], texts of different dimensions were collected for all the levels. This way, we ensured that the short text feature was one of many to be considered relevant in the automated analysis. As also stated before, a sample of texts of a complexity level higher than the ones foreseen for the project was compiled to provide a specific upper limit to the analysis.

The first compilation of texts within each level followed a superficial and intuitive reading of the texts but also contemplated a cross-check of the properties stated in

[4] The data files are open according to the CC BY-NC-ND 4.0 license and accessible through Zenodo (https://zenodo.org/records/10889888). The access is granted for the inspection of research results and to ensure research results reproducibility, but the texts cannot be published freely or for any purposes not compliant with the CC BY-NC-ND 4.0 license, as part of the texts in the corpus may still be under copyright.

the complexity descriptors previously defined. It is important to note that these two steps were carried out in tandem, allowing us to grasp the true nature of texts rather than solely relying on pre-defined expectations.

The overall work schema is represented in Figure 1 and will be further discussed in the following sections.

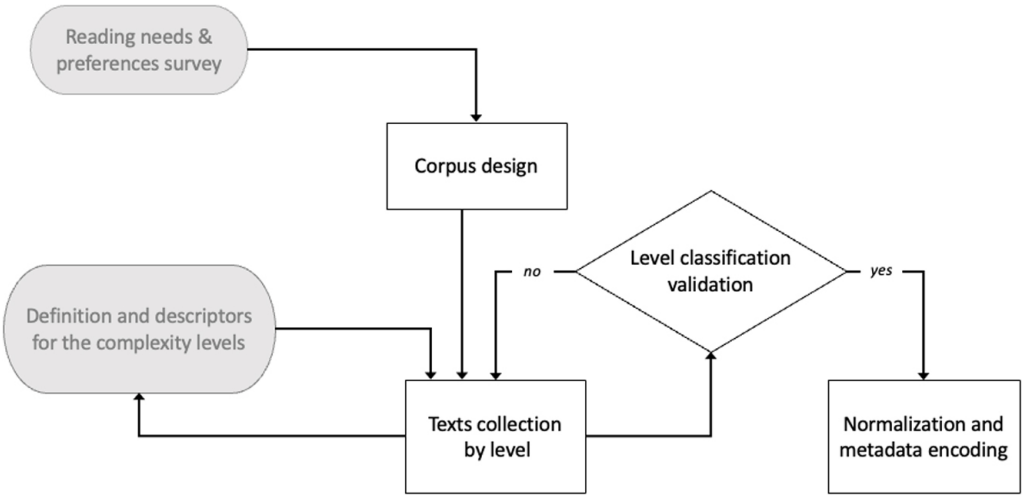


Fig. 1. iRead4Skills EP corpus compilation general workflow

The compiled data will provide the basis for the training and test sets for the complexity analysis systems. The collected corpus will be further validated and annotated by end-users, originating forthcoming versions and a second derived dataset.

3.1 Corpus general design

As part of the iRead4Skills project, we assembled a corpus for complexity analysis of texts, considering adult native speakers with low literacy skills. Our focus is on providing a versatile resource that includes texts in digital and/or paper formats, sourced from European Portuguese authors or through European Portuguese translations. The text dimension varies from 200 to 500 word excerpts, with an ideal distribution of 10 texts per subtype (i.e., further subcategories within general text genre and communication type categories). The initial goal envisaged 2000 texts, totaling between 500,000 and 600,000 words. Ultimately, besides providing the basis for developing automatic complexity analysis tools, the corpus will be a valuable resource for trainers, learners, and researchers interested in reading skills, text complexity, and skill development.

Corpus categories: text typology. In crafting our corpus, we were guided by reading preference surveys and existing materials to ensure robust results. We aimed to cover several communication domains and to assure relevance across different contexts. Based on well-known genders in corpora typologies [26] [27] and on the reference frameworks consulted to establish the complexity levels of the project [28] [29] [30] [31], we began by defining 11 major categories that reflected the goal and medium of communication and, for each, relevant sub-categories, to achieve a diversity of text genres and types, topics, and textual properties. As reflected in Table 1, this typology was organized according to two main aspects: communication goal and medium intended to consider relevant situations in which reading is involved,

and document type, which besides considering different text formats also reflects usual topics, purposes, and contents (e.g., cooking recipes talk about food, plants and animals used as ingredients, specific household tools, etc.).

Table 1. Text typology representative of the data universe

Goal & Medium	Document
Personal communication	note
	E-mail/personal letter
	list/agenda (groceries, tasks, …)
	diary
Institutional/professional communication (w/data anonymisation)	letter/e-mail
	report
	instructions
	task lists/agenda
	minute
	press release
	internal release/newsletter
	web page (about us/mission…)

The domain and content of the materials were also considered in this typology, as reflected in the document types defined for non-fiction books, for instance.

The ideal aim for each subcategory (document) was to find 10 text excerpts corresponding to the three levels (very easy, easy, and lain). This allowed us to check the compilation process intuitively and independently.

Text selection criteria. Before delving into the text collection process, several additional criteria were defined, namely:

1. To avoid artificially simplified texts, the materials should consist of authentic and contemporary texts.
2. The sources for collection should be legal websites of bookshops, book publishers, blog pages, national political websites (such as the parliament or the Portuguese political parties' websites), and newspaper or magazine websites.
3. To avoid extra optical character recognition (OCR) work, digital format should be the primary choice.

However, after having trouble finding texts of the first levels (namely for the very easy level), we also resorted to printed sources. Whenever possible, author and topic diversification were considered. We also focused on contemporary Portuguese, although with no specific temporal restrictions.

3.2 Texts collection

The text gathering, processing, and treatment followed a rigorous process, as shown in Figure 2. This process was applied consistently across all subcategories multiple times until successful results were achieved.

Fig. 2. Text collection process overview

Firstly, we searched for accessible text excerpts in online bookstores or other legal repositories. These samples needed to be at least 200 words long and were selected based on the complexity level they were intended to represent. Specifically, the initial collection process was guided by the desired complexity level and text subcategory rather than the other way around. For example, we looked for Very Easy texts within the *cookbook* genre.

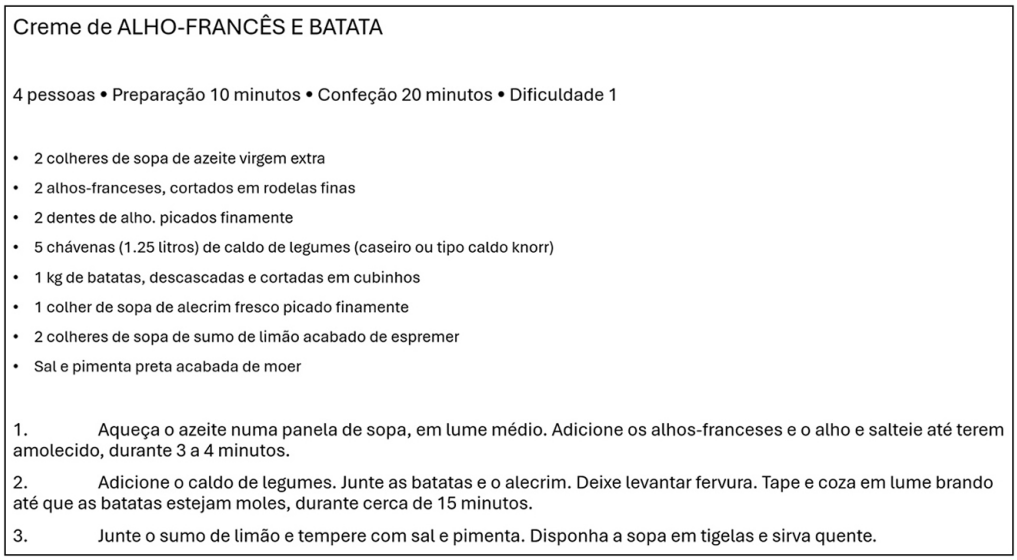
After the first examination and subsequent collection, the retrieved excerpts were kept in PDF format and later converted into TXT format. Whenever necessary, printed texts were collected, digitized into PDF format, and then converted into TXT format. We decided to keep the collected excerpts in both formats to guarantee access and traceability, ensuring the retention of digital proofs of the originals in case the original URL is removed or becomes unusable.

During the collection process, an initial superficial and intuitive assessment of the complexity level was conducted, followed by the complexity descriptors application. Two experts revised the texts and their initial assessments. Often, discrepancies arose between the features present in the collected samples and the descriptors associated with the assigned complexity level.

3.3 Data validation procedures

Given the high level of subjectivity of the classification task, all collected data was validated by a second expert team, not involved in the collection process. Only the texts whose classification and complexity assessment were validated by the second team were included in the corpus. This process was repeated, and whenever

necessary, new excerpts were retrieved until the desired number of texts for each subcategory and each level of complexity was reached. The descriptors used to establish each level were previously validated by a focus group with trainers from AL and VET centres.



Creme de ALHO-FRANCÊS E BATATA

4 pessoas • Preparação 10 minutos • Confeção 20 minutos • Dificuldade 1

- 2 colheres de sopa de azeite virgem extra
- 2 alhos-franceses, cortados em rodelas finas
- 2 dentes de alho. picados finamente
- 5 chávenas (1.25 litros) de caldo de legumes (caseiro ou tipo caldo knorr)
- 1 kg de batatas, descascadas e cortadas em cubinhos
- 1 colher de sopa de alecrim fresco picado finamente
- 2 colheres de sopa de sumo de limão acabado de espremer
- Sal e pimenta preta acabada de moer

1. Aqueça o azeite numa panela de sopa, em lume médio. Adicione os alhos-franceses e o alho e salteie até terem amolecido, durante 3 a 4 minutos.
2. Adicione o caldo de legumes. Junte as batatas e o alecrim. Deixe levantar fervura. Tape e coza em lume brando até que as batatas estejam moles, durante cerca de 15 minutos.
3. Junte o sumo de limão e tempere com sal e pimenta. Disponha a sopa em tigelas e sirva quente.

Fig. 3. Example of a text initially considered Very Easy but moved to Plain after expert validation

Initially, the text in Figure 3 was considered Very Easy because it was a typical soup recipe with ingredients familiar to any native Portuguese-speaking adult. However, during validation by other team members, the text was deemed inappropriate for the very easy level and was subsequently moved to a higher complexity level due to the following factors:

i. Used in a communicative context with specific instructions;
ii. Contained specific verbs such as “saltear” (“sauté”), “dispor” (“arrange”), and “adicionar” (“add”);
iii. Verbs in imperative such as “aqueça” (“warm up”), “junte” (“join”), or “tape” (“cover”);
iv. Domain-specific expressions like “levanter fervura” (“bring to a boil”);
v. Suffixes such as “-inho” and “-mente.”

The experts involved in the collection and validation teams were native Portuguese speakers with training in linguistics and/or language sciences at MA and PhD levels. They had specific knowledge of the project’s overall goals, including its research and innovation objectives, as well as a thorough understanding of the complexity levels and corresponding descriptors. The classifiers achieved a 100% agreement rate on including the texts in the dataset.

3.4 Normalization and metadata encoding

All collected material was organized in structured repositories, and comprehensive text metadata were registered. This was deemed necessary to facilitate the organization of textual information, ensure data traceability in case the original texts are removed from the internet, and guarantee data preservation. Texts were

normalised concerning formatting issues (e.g., odd line breaks) and inexistent word forms (e.g., parts of URLs, numeric codes, etc.).

The TXT files were numbered, and their relevant metadata were registered in a separate Excel file. Depending on the type of text and its characteristics, some fields in the metadata file were not filled in. Table 2 presents the metadata encoded.

Table 2. Set of metadata considered for the iRead4Skills EP corpus

Metadata	New File Information
Title	A demora no momento de pagar a conta
Author's name	A Lupa de Alguém
Translator's name	
Language	Portuguese
Publisher	
Place of publishing	
Date of publishing	07/09/2011
Document identification	
ISBN (International Standard Book Number)	
ISSN (International Standard Serial Number)	
URL (if online and available)	https://a-lupa-de-alguem.blogs.sapo.pt/219509.html
DOI (Digital Object Identifier)	
Copyright information	
Access Date	02/05/2023
Level of difficulty (L1, L2, L3)	N1 very easy
Purpose and mode	Personal communication
Document	Diary/personal journal
Area and content	
Number of extracted pages	
Number of words	105
ID BD_PDF / TXT	N1_Pess_diar_04

This file enables us to monitor several key aspects that may be relevant for other studies using this corpus. For example, it includes information on selected authors, publication dates, ISBNs, and the URLs from which the excerpts were collected, among other details. In future analyses of this corpus, it can provide valuable data for exploring correlations between linguistic complexity features and extralinguistic variables.

4 RESULTS

In this section, we present the collected data. We start by showcasing the discrepancy between the ideal criteria, which specified the number of texts required for each category, and the actual number of texts collected.

Table 3. Total number of collected text for each category

Typology	Ideal Number of Texts	Number of Collected Texts
personal communication	120	200
institutional/professional communication	210	246
social media	300	465
commercial communication/dissemination	210	260
book non-fiction	330	403
book fiction	180	407
book didactic	120	419
academic	210	176
political	90	91
legal	90	85
religious	120	168
Total	1980	2920

The second column of Table 3 depicts the number of texts ideally envisaged. The third column shows the final number of collected texts. The absolute number of texts collected, in general, was higher than the ideal ones (2920 > 1980). However, this did not occur for the category Academic, namely in the very easy level (cf. Appendix 1). Initially, finding texts corresponding to the very easy level was challenging, especially for specific text subcategories. For instance, sometimes, when texts were available, they were often too short, which led us to gather additional documents for some subcategories to achieve the expected number of tokens.

Consequently, we had to compile more texts for other levels or subcategories to balance the quantity of texts in some categories. This also happened when the team members disagreed with an initial classification and a collection process of new materials had to be initiated.

After selecting the sources, collecting, processing, and organizing the texts according to the level (Very Easy, Easy, Plain, and +Complex) and category/subcategory, the iRead4Skills Dataset 1 (v.2.0) consists of a total of 2,915 texts with 802,125 tokens. The texts were distributed as depicted in Table 4, finishing with 2920 texts and 802125 tokens.

Table 4. Text distribution by complexity level

Levels	Very Easy	Easy	Plain	More Complex	TOTAL
N. texts	776	658	752	734	2920
N. tokens	109567	136493	257069	298996	802125

The first line in Table 4 shows the number of TXT files, while the second line presents the number of tokens collected for each level.

The validation process also led to the reclassification and reorganization of some excerpts. Some texts initially classified at a certain level were either reclassified to a different level, such as +Complex, or removed from the dataset. This led to some subcategories of some levels (mainly +Complex level) ending up with more texts than

the others. The corpus distribution by category is presented in the following graphs for the three main levels established within the project (see Figure 4). The complete data on tokens and text distribution per level, category, and subcategory is provided in Appendix 2.

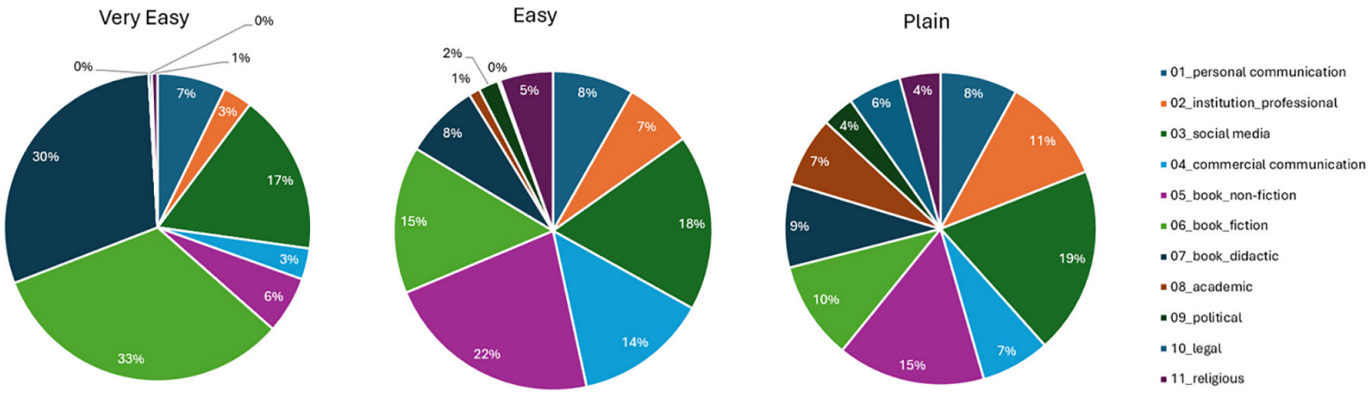


Fig. 4. Token distribution (in percentage) per category and complexity level

4.1 Corpus limitations

As expected, the lower levels of complexity have a more unbalanced distribution by category. Authentic texts with very low levels of complexity, such as Very Easy and Easy, are scarce because many written texts across various genres are not typically designed for readers with reading difficulties.

The compilation of the corpus can be viewed as a subjective task. Even with the use of objective descriptions for complexity levels, some texts may be perceived as misclassified due to individual judgement.

4.2 Discussion

In this paper, we presented the design and compilation process of the Portuguese corpus for the iRead4Skills project, which fills a gap in existing Portuguese language resources. We explain the reasoning behind the definition of levels, the determination of data categories and subcategories of sampled texts, and the process of collecting and selecting the text excerpts for the corpus. In this subsection, we discuss some challenges encountered in this process and how we addressed them to obtain a diversified corpus capable of meeting the project's objectives.

The corpus compilation presented some major challenges related to the difficulty of finding authentic texts with low levels of complexity, which are reflected in the corpus's final format. For example, the variations in the number of tokens in texts of easier levels highlight the difficulty in finding sufficiently lengthy samples for the lower complexity levels compared to those for higher levels. To sustain a solid methodology that allows us to represent reality instead of resorting to the artificial creation of such samples, we assume that some genres, such as academic texts, and others used in functional and essential parts of daily life situations, are difficult to collect and represent. This way, the unbalance of some categories and subcategories in the different levels of complexity is a direct reflex of the universe samples.

The compilation followed a well-informed methodology for both data collection and preservation. Nonetheless, the validation process was demanding. We attest that

textual complexity is quite subjective, depending on individual human judgement, even among experts with similar backgrounds and a clear understanding of the complexity and the corpus's specific purposes. The proposed methodology aimed to address this issue and ensure higher levels of objectivity and consistent assessment. Defining and using objective descriptions of complexity levels, including specific trends for each level and examples of lexical, morphological, syntactic, and semantic phenomena to include or avoid, was crucial in refining the team's judgments. However, on several occasions, it was not possible to reach consensus, resulting in several collected samples being discarded. As machines tend to be more systematic, automatic complexity analysis can help reduce this task's high level of subjectivity.

Existing corpora used for automatic complexity analysis are typically composed of data from i) textbooks, i.e., texts deemed suitable for students at specific proficiency levels, whether native speakers (L1 students) [19] or language learners (L2 students) [32]. ii) Graded texts for specific readers, i.e., texts transformed or adapted to better suit readers with low literacy or proficiency [33]. iii) Texts classified by non-expert participants in specific tasks or environments [34].

Compared to these corpora, the iRead4Skills Portuguese corpus offers several advantages:

- It follows a specific classification methodology based on objective descriptors and is carried out by an expert team, unlike textbooks, where texts are often selected intuitively by different authors with varying sensitivities and goals.
- Unlike those in graded reader sets, it consists of real, non-adapted texts.
- It is of suitable size for machine training, unlike the corpora collected in user studies or experiments.

Moreover, not all the corpus types mentioned above are available for Portuguese. In fact, corpora used in European Portuguese readability and complexity analysis are quite small (e.g., Camões corpus of 114 texts [35], composed of excerpts collected from Portuguese books, news, and articles used in Portuguese language classes) and focused on language learners (e.g., c500 corpus [35], Camões CEFR A1-C1 corpus used in [36]), or are learners' corpora, composed of texts written by learners in language assessment exams (e.g., COPLE2 corpus with 1,634 texts [6]), thus far from being representative of the universe we aim to analyze. The iRead4Skills Portuguese corpus is therefore the first solid, diverse, and large dataset of its kind.

In times to come, as more materials may become available, the current corpus can be redesigned or supplemented to fulfil different criteria and, more importantly, serve various purposes (for instance, leverage the relevance of text length, topic, or genre in automatic complexity analysis).

In addition, by providing texts of varying levels of complexity, the current corpus can have useful and practical implications for educators, developers of literacy tools, and policymakers. For example, teachers and trainers now have a set of texts from multiple communication domains (e.g., professional communication, legal and political domain, etc.) to practice reading and comprehension skills adapted to the learners' needs. Texts with increased difficulty can be explored progressively, with learners of different reading skills and needs. Also, writing activities for specific daily life activities can now be proposed to learners in VET and AL centres based on the texts provided by our corpus. Authentic texts from varied complexity levels in our corpus can also be included in textbooks dedicated to adult learners and classroom materials and activities developed by the trainers in AL and VET centres. As AL difficulties can range from basic reading skills to motivational factors, this corpus addresses both by covering the real needs of end-users and providing information

on reading skills issues, along with materials for individuals, skill development institutions, and employers. Furthermore, by incorporating reading materials from a wide range of domains, this corpus enables adult learners to develop their reading skills across different areas of life. As a result, adult readers will engage with not only general information but also literature and culture.

Finally, the whole process of corpus compilation allowed us to verify the existence of a gap in ideally universally accessible texts (i.e., Very Easy, Easy, and Plain texts) in several functional areas of society, such as transportation, health services, and social security, with potentially profound impacts not only on the motivation of low-literacy individuals but also on their professional, social, and personal achievements.

5 FINAL REMARKS AND FUTURE WORK

The Portuguese corpus in the iRead4Skills Dataset 1 (v2.0) consists of a solid sample of authentic texts, reflecting a diverse collection of texts written by diverse people for diverse people, used in different activities, and covering a wide range of topics. Focusing on the relevant complexity levels, it presents a balanced data set covering four levels of text complexity that illustrates the phenomena that can be the basis of the difficulty of understanding authentic texts for native low-literacy adults, but also the need for this type of assessment and resources for trainers in AL centres.

Besides filling in a gap in data concerning texts suitable for native adult speakers with poor reading skills, and covering different real written materials—both digital and in print—that cover distinct topics, genres, and communication situations, the compiled corpus will allow the development of the iRead4Skills Portuguese complexity analysis system and has also allowed for the creation of the basic lexicons for the complexity levels targeted [37]. The iRead4Skills system will be able to automatically and immediately analyze texts and suggest readings according to their complexity level, as well as assist in text assessment and simplification. The system can assist low literacy adults in autonomously selecting readings based on their preferences and needs.

Planned extensions of this dataset include the classification validation by non-experts, comprising our two major target users' groups: low literacy adults and AL trainers, which is an ongoing process, as well as annotation of complex phenomena and parts of the texts of a sub-part of the corpus.

Additionally, it can be used by trainers and content creators to develop or adapt texts to the appropriate level of complexity for their target audiences.

Encompassing such a diverse set of authentic texts, including materials from different domains and genres, the iRead4Skills Dataset 1: corpora by complexity level for FR, PT, and SP (v.2.0) also provides a knowledge foundation that can be immediately used by AL and VET trainers as teaching materials.

The iRead4Skills Dataset 1: corpora by complexity level for FR, PT, and SP (v.2.0) provides new and relevant language resources for complexity studies and the development of automatic text complexity classification and analysis systems and can be accessed at https://zenodo.org/records/10889888 for research purposes under the CC BY-NC-ND 4.0 license.

6 ACKNOWLEDGEMENTS

This study is supported by (1) the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZONCL2- 2022-TRANSFORMATIONS-01-07, DOI:10.3030/101094837). Views and opinions expressed are those of the authors

only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. (2) Portuguese national funding through the FCT—Portuguese Foundation for Science and Technology, I.P., as part of the projects UIDB/LIN/03213/2020, 10.54499/UIDB/03213/2020 and UIDP/LIN/03213/2020, 10.54499/UIDP/03213/2020—Linguistics Research Centre of NOVA University Lisbon (CLUNL).

7 REFERENCES

[1] R. Desjardins, "Rewards to skill supply, skill demand and skill match-mismatch: Studies using the adult literacy and lifeskills survey," *Lund Economic Studies,* no. 176, 2014.

[2] M. Xu, K. Lv, and X. Bi, "Computer network assisted teaching of college English reading," *International Journal of Emerging Technologies in Learning (iJET),* vol. 11, no. 8, pp. 47–53, 2016. https://doi.org/10.3991/ijet.v11i08.6048

[3] OECD, "Skills matter: Additional results from the survey of adult skills," OECD Publishing, 2019. https://doi.org/10.1787/605ec8b3-en

[4] C. Gomes and J. P. hosted by Frazão, *Da Capa à Contracapa* [Audio Podcast], Fundação Francisco Manuel dos Santos, 2024. [Online] Available: https://ffms.pt/pt-pt/ffms-play/da-capa-contracapa-podcast/como-ainda-se-vive-sem-saber-ler-e-escrever-em-portugal#resumo

[5] M. Conde and A. Hernández-García, "Data driven education in personal learning environments – What about learning beyond the institution?" *International Journal of Learning Analyctics and Artificial Intelligence for Education,* vol. 1, no. 1, pp. 43–57, 2019. https://doi.org/10.3991/ijai.v1i1.11041

[6] A. Mendes, S. Antunes, M. Janssen, and A. Gonçalves, "The COPLE2 corpus: A learner corpus for Portuguese," *in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., 2016, pp. 3207–3214.

[7] R. Monteiro *et al*., "iRead4Skills – Complexity Levels," *iRead4Skills* (V1.0), 2023. https://doi.org/10.5281/zenodo.10459090

[8] J. P. Kincaid, R. P. J. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas Automated Readability Index, Fog Count and Flesch Reading Ease Formula for Navy enlisted personnel," *Institute for Simulation and Training,* vol. 56, 1975. [Online]. Available: https://stars.library.ucf.edu/istlibrary/56

[9] I. Lorge, "The Lorge and Flesch readability formulae: A correction," *School and Society*, vol. 67, pp. 141–142, 1948.

[10] E. H. Hiebert and P. D. Pearson, "The state of the field," *The Elementary School Journal,* vol. 115, no. 2, 2014. https://doi.org/10.1086/678297

[11] T. A. Harley, *Psychology of Language: From Data to Theory*. London: Taylor and Francis, 2013.

[12] J. W. Cunningham and H. A. Mesmer, "Quantitative measurement of text difficulty: What's the use?" *The Elementary School Journal,*" vol. 115, no. 2, 2014. https://doi.org/10.1086/678292

[13] O. Dahl, "Definitions of complexity," in *Proceedings of the Colloquium on Complexity, Accuracy and Fluency in Second Language Use, Learning & Teaching,* 2007, pp. 41–46.

[14] CCSS/ELA, "National Governors Association Center for Best Practices & Council of Chief State School Officers," *Common Core State Standards*, 2010.

[15] A. Tourimpampa, A. Drigas, A. Economou, and P. Roussos, "Perception and text comprehension. It's a matter of perception!" *International Journal of Emerging Technologies in Learning (iJET),* vol. 13, no. 7, pp. 228–242, 2018. https://doi.org/10.3991/ijet.v13i07.7909

[16] M. Heilman, K. Collins-Thompson, J. Callan, and Eskenazi, "Combining lexical and grammatical features to improve readability measures for first and second language texts," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, 2007, pp. 460–467.

[17] H. A. Mesmer, J. Cunningham, and E. Hiebert, "Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future," *Reading Research Quarterly,* vol. 47, no. 3, pp. 235–258, 2012. https://doi.org/10.1002/rrq.019

[18] M. Flor, B. B. Klebanov, and K. Sheehn, "Lexical tightness and text complexity," in *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, 2013, pp. 29–38.

[19] K. Collins-Thompson and J. P. Callan, "A language modeling approach to predicting reading difficulty," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* 2004, pp. 193–200.

[20] S. Vajjala, D. Meurers, A. Eitel, and K. Scheiter, "Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts," in *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 2016, pp. 38–48.

[21] V. Yaneva, I. Temnikova, and R. Mitkov, "Accessible texts for autism: An eye-tracking study," in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, 2015, pp. 49–57. https://doi.org/10.1145/2700648.2809852

[22] S. Correia, R. Amaro, and R. Gauchola, "iRead4Skills – Reading skills survey," *iRead4Skills,* (V1.0), 2023. https://doi.org/10.5281/zenodo.10179536

[23] F. Clément, L. Hauret, and R. Amaro, "iRead4Skills – Overall skills and gaps survey," *iRead4Skills,* (V1.0), 2023. https://doi.org/10.5281/zenodo.10181135

[24] X. Blanco Escoda, R. Amaro, T. François, and M. Garcia, "iRead4Skills – Baselines for complexity lexicons definition," *iRead4Skills* (V1.0), 2023. https://doi.org/10.5281/zenodo.10069793

[25] S. R. Goldman and C. D. Lee, "Text complexity: State of the art and the conundrums it raises," *The Elementary School Journal,* vol. 115, 2014. https://doi.org/10.1086/678298

[26] Lee, David, Genres, and Registers, "Text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle," *Language Learning & Technology*, vol. 5, no. 3, pp. 37–72, 2001.

[27] D. Biber, *Variation Across Speech and Writing.* New York, NY: Cambridge University Press, 1988. https://doi.org/10.1017/CBO9780511621024

[28] ALTE. "The ALTE can do project (1992–2002)," ALTE, 2002. Available at: https://www.alte.org/Materials

[29] M. J. Alves and S. Lameira, "Referencial de Competências-chave de Educação e Formação de Adultos – Nível Básico," ANQEP, 2021.

[30] CoE – Council of Europe "Common European framework of reference for languages: Learning, teaching, assessment – Companion volume," Council of Europe Publishing, Strasbourg, 2020. Available at: https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

[31] OECD, "The survey of adult skills reader's companion," OECD Publishing, 2013.

[32] T. François and C. Fairon, "An 'AI readability' formula for French as a foreign language," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* J. Tsujii, J. Henderson, and M. Pasca, Eds., 2012, pp. 466–477.

[33] S. Vajjala and W. D. Meurers, "On improving the accuracy of readability classification using insights from second language acquisition," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP,* 2012.

[34] S. Gooding, Y. Berzak, T. Mak, and M. Sharifi, "Predicting text readability from scrolling interactions," in *Proceedings of the 25th Conference on Computational Natural Language Learning*, A. Bisazza and O. Abend, Eds., 2021, pp. 380–390. https://doi.org/10.18653/v1/2021.conll-1.30

[35] R. Santos, J. Rodrigues, A. Branco, and R. Vaz, "Neural text categorization with transformers for learning Portuguese as a second language," in *Progress in Artificial Intelligence*, G. Marreiros, F. S. Melo, N. Lau, H. Lopes Cardoso, and L. P. Reis, Eds., 2021, pp. 715–726. https://doi.org/10.1007/978-3-030-86230-5_56

[36] E. Ribeiro, N. Mamede, and J. Baptista, "Automatic text readability assessment in European Portuguese," in *Proceedings of the 16th International Conference on Computational Processing of Portuguese, EPIA 2021*, in Lecture Notes in Computer Science, P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, Eds., vol. 12981, Springer, Cham, 2024, pp. 97–107.

[37] R. Wilkens *et al.*, "iRead4Skills – Basic lexicons per complexity level," *iRead4Skills* (V1.0), 2024. https://doi.org/10.5281/zenodo.10889986

8 APPENDIX

8.1 Appendix 1

Goal & Medium	Document	Very Easy		Easy		Plain		+Complex	
		Texts	Tokens	Texts	Tokens	Texts	Tokens	Texts	Tokens
Personal communication	note	20	1526	8	463	9	714	2	173
	E-mail/personal letter	18	2586	12	1789	13	7978	10	6479
	list/agenda (groceries, tasks, ...)	29	3575	14	8204	14	9183	10	4989
	diary	4	1986	15	3898	12	8851	10	4760
Institutional/ professional communication	letter/e-mail	16	1521	17	2900	13	2070	8	1892
	report	1	72	2	343	16	7489	11	4739
	instructions	6	648	6	1182	13	4762	10	5623
	minute	0	0	4	563	21	9752	6	3250
	press release	2	475	10	3233	10	3754	10	6349
	internal release/newsletter	5	630	9	2516	8	3722	10	5414
	web page	3	730	6	1412	13	5307	10	5613
Social media (newspapers, magazines)	editorial	6	929	17	2416	12	4308	10	4859
	news	26	4666	15	3439	10	6128	10	4262
	reportage	3	227	21	3922	9	3906	10	5486
	interview	44	9756	7	1179	13	7350	11	5690
	opinion article	10	1892	10	2809	13	9334	11	6607
	scientific dissemination article	5	912	13	3389	10	5677	10	4724
	profile (brief presentation of a notorious person)	18	2675	16	7397	12	7020	10	6080
	horoscope	5	485	13	1596	6	3319	8	3127
	obituary	6	410	14	2728	10	11394	7	13763
	weather report	5	674	9	2270	10	6132	10	3739

(Continued)

Goal & Medium		Document	Very Easy		Easy		Plain		+Complex	
			Texts	Tokens	Texts	Tokens	Texts	Tokens	Texts	Tokens
Commercial communication/ dissemination		add	22	2118	12	3679	8	7293	10	422
		posters/outdoors	13	264	9	129	6	134	10	208
		flyer/leaflet	4	431	10	2495	7	2428	10	1921
		menu	5	573	12	10542	10	4536	10	2678
		label	4	239	14	1823	11	2038	10	4539
		user manual/ instructions guide	4	630	11	3000	9	3888	10	4539
		medicine leaflet	2	130	7	1945	10	3581	10	4696
Book	non-fiction	(auto)biography	10	1963	13	8448	11	6197	15	10712
		chronicle	6	770	8	1480	6	3725	10	7475
		essay	0	0	6	1714	10	9409	7	6906
		diary	12	1986	15	3652	9	4084	11	6491
		preface/prologue	5	708	13	3522	10	5251	10	5287
		dedicatory	9	127	8	487	11	353	3	909
		self-help book	6	864	17	8448	11	4107	10	6467
		travel report/diary	4	637	10	2437	9	4122	10	7167
		memoirs	4	738	16	4528	11	5461	10	6182
		letters	2	241	6	1237	11	3335	13	7186
		travel/tourist guide	0	0	11	2493	14	5137	10	5108
	fiction	short story	106	31134	26	8483	9	14040	10	6203
		fable	23	4980	16	4055	4	1762	10	5935
		epic	0	0	8	2686	11	6911	15	7424
		novel	4	950	19	5890	14	6447	16	12272
		drama (play)	14	3964	16	4039	10	3242	10	6334
		lyric (poetry)	33	2625	11	868	10	1542	12	2023
	didactic	textbook	97	18294	19	4100	14	3926	10	3607
		encyclopedia/atlas	76	14130	34	7319	22	5785	10	3133
		cookbook	24	3295	5	1165	14	4750	10	3637
		glossary	44	4339	4	695	25	14417	11	7144
Academic		article/paper	0	0	4	1362	9	3247	14	7061
		report	0	0	1	95	8	3284	20	7355
		abstract/synthesis	0	0	0	0	18	4244	10	3242
		critical review	0	0	0	0	3	1245	17	8573
		thesis/PhD	0	0	1	114	4	1402	16	5381
		project proposal	0	0	1	257	6	1430	21	7206
		essay	0	0	1	136	10	9409	12	11467

(Continued)

Goal & Medium	Document	Very Easy		Easy		Plain		+Complex	
		Texts	Tokens	Texts	Tokens	Texts	Tokens	Texts	Tokens
Political	speech	0	0	3	576	13	4230	14	7363
	motion/report	0	0	1	1725	12	5056	18	9307
	program	0	0	4	1219	7	1848	19	8911
Legal	law	0	0	0	0	18	4972	10	4447
	contract	0	0	0	0	21	8998	5	3383
	notification letter/memorandum/ statutes/public notice	2	421	3	442	16	4454	10	5004
Religious	prayer	5	266	19	2025	16	2287	21	4008
	scriptures/psalms/ epistles	2	196	10	2215	11	3022	20	8348
	homilies	0	0	6	1897	15	5695	10	7323
	sermon/flyer/catechism	2	355	10	3254	11	3141	10	7024

8.2 Appendix 2

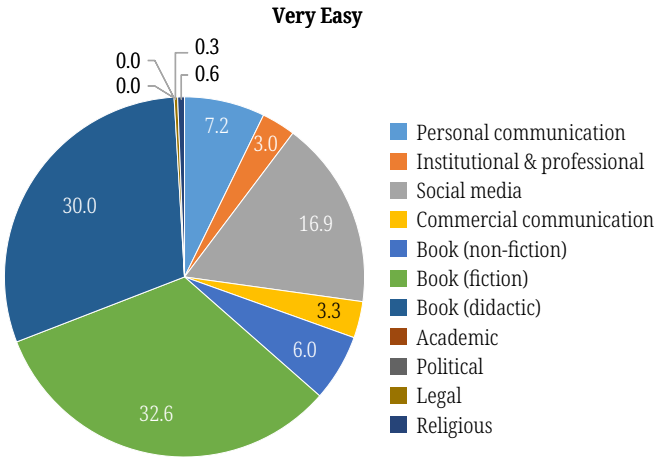


Fig. A1. Token distribution (in percentage) per category and complexity level [Very Easy]

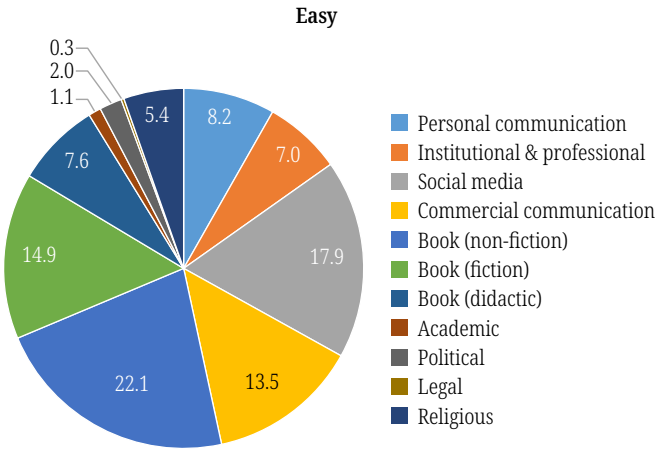


Fig. A2. Token distribution (in percentage) per category and complexity level [Easy]

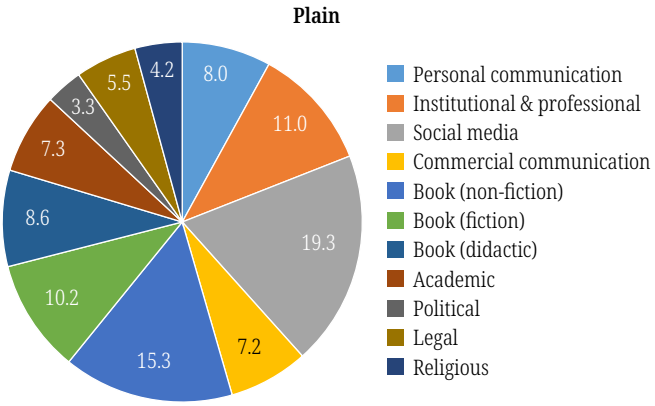


Fig. A3. Token distribution (in percentage) per category and complexity level [Plain]

9 AUTHORS

Maria Leonor Reis is with the NOVA CLUNL, Universidade NOVA de Lisboa, Lisbon, Portugal (E-mail: mariareis@fcsh.unl.pt).

Sílvia Barbosa is with the NOVA CLUNL, Universidade NOVA de Lisboa, Lisbon, Portugal.

Michell Moutinho is with the NOVA CLUNL, Universidade NOVA de Lisboa, Lisbon, Portugal.

Ricardo Monteiro is with the NOVA CLUNL, Universidade NOVA de Lisboa, Lisbon, Portugal.

Susana Correia is with the NOVA CLUNL, Universidade NOVA de Lisboa, Lisbon, Portugal.

Raquel Amaro is with the NOVA CLUNL, Universidade NOVA de Lisboa, Lisbon, Portugal.