Horizon Europe

D3.2 Baselines for complexity lexicons definition

(Report)

https://iread4skills.com/

Funded by
the European Union

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

Document Control

Information:

Settings	Value
Deliverable No.	**3.2**
Document Title:	Baselines for complexity lexicons definition
Author(s):	Xavier Blanco; Raquel Amaro; Thomas François
Reviewer(s):	Marcos Garcia
Sensitivity:	**Public**
Date:	31/10/2023

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

TABLE OF CONTENTS

0. Glossary

Lemma: a wordform considered as the citation form together with all the inflected forms (for instance, the infinitive form of a verb as 'work') associated to a given part-of-speech, but with no meaning differentiation (*work*$_1$=to have a job and *work*$_2$=to operate or function). Usually, corresponds to an entry in a dictionary.

Lexeme: wordform or inflectional paradigm of wordforms (for instance, the infinitive form of a verb as 'be') that constitute the association of a meaning, a form and a bunch of combinatorial properties. Informally, it corresponds to a meaning of usage dictionaries.

Phraseme: phrase that is not formed freely by the speaker, but taken from the lexical flow. It constitutes, like the lexeme, the association of a meaning, a form and a content, but it consists of more than one form word.

Vocable: set of lexemes that share a form and a certain semantic content. Informally, it usually corresponds to an entry in a dictionary of use.

Vocabulary: finite list of words (usually lemmas or lexemes) that occur in a specific dataset (for instance, newspaper corpus) or in a specific domain (for instance, Biology).

Wordform: linguistic sign that presents, from the syntactic point of view, a certain autonomy of insertion in the discourse and, from the morphological point of view, a certain internal cohesion. Informally, it corresponds to what, at least in Romance languages, is called "word" when referring to the number of words in a writing.

1. Introduction

Reading skills are essential to acquire technical and scientific knowledge. This is especially relevant in formal education and training contexts, as Adult Learning (AL) and Vocational Educational Training (VET), and in work contexts, such as when companies provide written instructions to their workers. People with low literacy skills are less able to acquire and maintain transversal and durable skills needed to stay apace with the changing job market and to lead meaningful and complete lives. However, promoting reading habits and skills in adults is quite challenging due to the lack of dedicated and/or adequate reading materials.

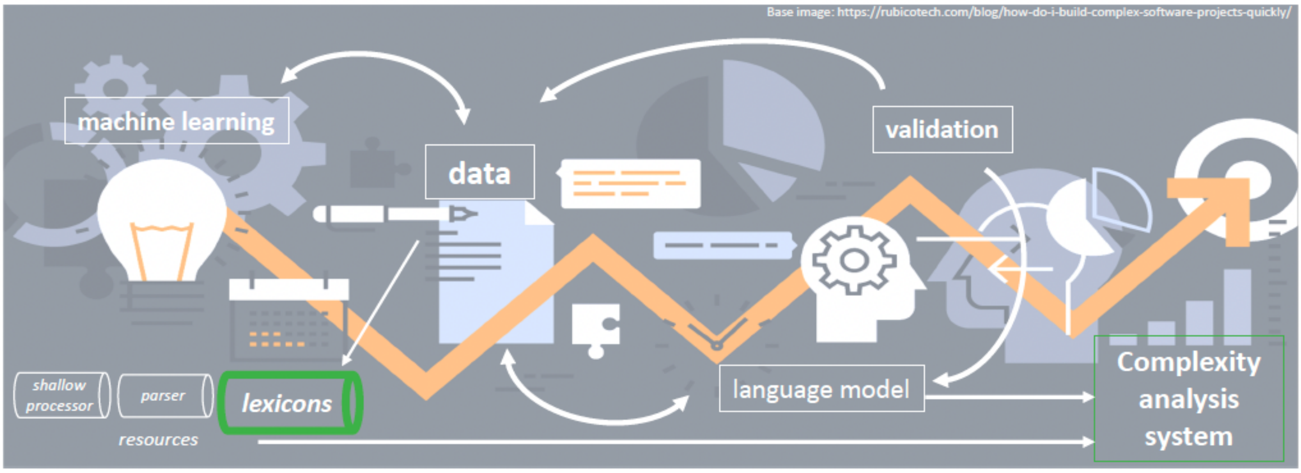
Supporting the adoption of innovation in adult training, the iRead4Skills project aims to promote the development of reading skills through an innovative intelligent system that evaluates texts complexity and suggests reading materials adequate to the user reading level, which can also be used by trainers in the creation or adaptation of texts with the appropriate level of complexity for their individual students.

The main target audience of the iRead4Skills project are, thus, low literacy skills adults, which include adult native speakers, integrated in a dynamic linguistic community, but also other second language (L2) speakers, not necessarily in L2 formal classes.

To achieve the main goals of the project, we focused on three levels of complexity, defined according to relevant reference documents that cover adult skills (PIACC 2016, ANQEP 2021) and language proficiency levels (Common European Framework of Reference for languages (CEFR) 2020) and that explicitly describe reading skills. The levels of complexity defined for the project are object of a specific task (see D3.1 Complexity levels), but can be succinctly characterized as follows:

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

As expected, and as described in the literature from its beginning to date (from Lively & Pressley (1923) to Pirali et al. (2022), to name some), the words used in the texts are an essential factor of complexity. An automatic complexity analysis requires, thus, information on the lexicon used and/or expected at each level.



Base image: https://rubicotech.com/blog/how-do-i-build-complex-software-projects-quickly/

However, complexity analysis targeting adult native speakers may require lexical resources that somewhat differ from existent resources related to lexical complexity directed to L2 learners, as the passive knowledge and the needs from native speakers are expected to be different.

This report discusses the baselines for defining complexity lexicons within the iRead4Skills project, considering the major methods in place, the possible update and extension of existing resources for the target languages (French, Portuguese, and Spanish) and, whenever possible, the specific issues related to the identification of complex words for native speakers with low-reading skills.

2. Methods for defining complexity lexicons

For the purposes of this report, the methods for determining complexity lexicons, i.e., lexicons that are useful for complexity analysis of texts, can be divided in three major categories: corpus-based methods, methods that rely on automatic inference and manual methods that use expert knowledge.

2.1 Corpus-based

Corpus-based approaches offer a robust method for compiling lexicons to aid in assessing text complexity and readability, since these are based on authentic linguistic data. Corpus-based lexicon compilation is a data-driven approach that relies on large, authentic collections of texts to extract not only word lists, but also linguistic information such as word frequencies, word co-occurrence data, and vocabulary richness/lexical diversity.

Within these approaches, lexicons can be created considering, for instance, word frequency. It has been well established for a long time that word frequency is an important variable in word recall, recognition, and processing (Brysbaert, Buchmeier, et al., 2011). The facilitating effect of frequency on word recognition has been empirically confirmed by Howes & Solomon (1951) and, more recently, by Monsell (1991) and Brysbaert et al. (2000). This frequency effect can generally be explained by the fact that common words in the mental lexicon are easier to access than less common words (due to a lower threshold or a higher base activation threshold), or because the search strategy in the mental lexicon is frequency-based (Brysbaert et al., 2000). In this view, complexity lexicons are defined by identifying the most common and least common words in a specific corpus that is supposed to represent the language. Among these lists, we distinguish between educational resources and resources based on adult content.

The first pedagogical list was built by Thorndike (1921), based on a huge corpus – for the time – of more than 4 million words. A later notable list is the “"American Heritage Word Frequency Book”" by Carroll et al. (1971), built from a 5-million words corpus based on texts used in American schools. Besides the raw frequencies of words, the authors also introduced the notion of frequency dispersion, a variant of the entropy metrics, aimed at assessing the distribution of a word across different topics or, on the contrary, its specificity to a limited number of contexts. More recently, Zeno et al. (1995) published the “Educator’s Word Frequency Guide”, based on a corpus of 17 million words extracted from schoolbooks.

In parallel, other researchers developed frequency lists without a clear pedagogical aim. A very famous example is the list of Kucera and Francis (1967), built from a small corpus of 1 million words, composed of texts intended for adults. Later, other lists were developed, mostly for English, such as CELEX (Baayen et al., 1993), the BNC list (Leech et al., 2001) or SUBTLEX (Brysbaert and New, 2009), that was derived for different languages, including French, Spanish, and Portuguese. Complexity lexicon based on the notion of frequency have been further refined using corpus in which the difficulty level of the texts is known – either through a manual or automatically classification. From these corpora, lists of frequent *vs.* rare words in each level are extracted and it is also possible to compute the frequency distribution of words across the difficulty levels. This is the concept implemented in the CEFRLex project (François et al., 2014). Depending on the type of unit considered, e.g., wordforms or lemmas, the resulting lists can be of different dimensions. The results obtained this way allow us to classify texts according to the words they contain: simpler texts are expected to contain more frequent words, whereas more complex texts are expected to contain more infrequent words. In addition, texts of a given level are expected to have a high percentage of the vocabulary defined for that level, which include the vocabulary defined for lower levels.

In addition to wordform frequencies, depending on the corpora used, for instance, a reference corpus for a given language, a set of words forming the core/basic/fundamental vocabulary for that language can be compiled. The high occurrence of 'core vocabulary' in a text indicates text simplicity and, conversely, high occurrence of words not in the 'core vocabulary' in a text points to text complexity.

Determining the lexical frequencies of the corpus (considering the sub corpora per level) requires some form of normalization (François et al. 2014, Durlich & François 2018). Normalized frequency per million words can be obtained by computing the raw frequencies by level, which are then weighted by a dispersion index. This allows for reducing the effect of low frequency words occurring in a small number of texts but with an unusually high frequency (context-specific effect).

Finally, word co-occurrence data concern collocations and multiword expressions (MWE) identification. This means extracting from corpus sets of two or more words that co-occur with statistical relevance, in different relative position and distance. Co-occurrence data can reflect different degrees of fixedness and/or idiomatic meaning (Mel'čuk 1998, Sinclair 1991, Fonseca et al. 2017) and allow us to identify different phenomena such as collocations, i.e., co-occurrences between two or more words that tend to be more frequent than expected based on the frequency of each element in a corpus, nominal compounds, idioms, formulae, proverbs, light verb constructions, etc (some typologies are presented in Sag et al. (2002) or Cowie (1994, 2001)). Co-occurrence (distributional) data can also help distinguishing between difference senses of polysemic wordforms. Word sense disambiguation is still a challenge even for the most advanced NLP systems.

The most relevant qualities of corpus-based approaches are that these can provide objective, data-driven assessments of text complexity, reducing subjectivity in evaluating texts, and that these allow for customization, given that the corpora can be tailored to specific domains or genres, allowing for more accurate assessments of complexity in specific texts. On the other hand, corpus-based approaches require adequate (representative and up to date) data and adaptability, as lexicons have to be adapted and updated as language evolves, ensuring that the complexity assessments remain relevant.

2.2 Expert grading

We will refer, in this section of the deliverable, not so much to the globality of the methods based on expert opinion, but to the application of this methodology within the framework of the iRead4Skills project (focusing on work developed for the Spanish lexicon), its possible contributions, but also the difficulties encountered, and the limitations of the results obtained.

Every speaker (including every learner) of a given language (in the case at hand, Spanish) has the strong intuition that there are more common "words" (they can refer to them as more useful, necessary, basic, etc.) than others. This intuition is triggered by the exposition of the neural circuits of an individual brain to repeated linguistic input and can be referred to as subjective frequency of words or familiarity (Gernsbacher, 1984) (as opposed to the objective frequency computed from a corpus). Interestingly, Connine et al. (1990) has demonstrated that both objective frequency and familiarity effects coexist and that while the familiarity with frequent words remains relatively stable from one individual to the next – which justifies its approximation by objective frequency -, for low-frequency words it varies greatly from one person to another. Consequently, familiarity would be a better predictor of response time to various classical tasks than frequency in the case of uncommon words.

Based on such intuition, every language teaching/learning method will try to initially (e.g., CEFR, level A1) present those "words and phrases" that are thought to be essential for the learner. Although variations can be observed depending on the target audience (age of the learner, diatopic variety under study, level of training, etc.) and the methodological-theoretical assumptions of the learning document (communicative approach, marked priority —or not— to the oral language, academic or extra-academic approach, etc.) there will be a first core of almost common lexical material. The differences between the lexical material presented will become more and more important as the level progresses (B2, C1 and C2).

The mere selection of the lexicon and its delimitation by levels (we are not even referring here to the presentation in linguistic or paralinguistic contexts, but only to a first identification of the forms that should be included) raises theoretical and empirical problems that are not easy to solve. First of all, it is important to note that belonging to a certain level of learning (we use, by default, the CEFR scale) is not a linguistic characteristic of the lexical unit, which only presents semantic properties, formal or combinatorial; that is, properties linked to the language system, not to discourse.

Nothing prevents, however, from associating a CEFR label with a lemma as the value of a category of lexicographic information, which does not have to correspond to a linguistic property (categories of lexicographic information can refer, in fact, to very diverse aspects, including the organization of the lexical database itself).

2.2.1 Method proposal

It is a challenge for the lexicographer to assign, given a list of words presented as entries in a glossary or dictionary, a CEFR level for each entry. In this section, we present the *modus operandi* that we have followed[1] to, first, select and, subsequently, classify the 2,500 words (uniwords) that are accepted to constitute the lexical competence (at least the passive lexical competence) of a learner who has reached level B1 in a Romance language: 600 for level A1, another 600 for level A2 and 1,200 more to complete level B2 (some estimates are smaller: 500-500-1000, but remember that we target passive vocabulary since our object of study is written comprehension, not production). Let us note that in more advanced stages of learning, the disproportion between active vocabulary and passive vocabulary tend to become much greater in favor of the latter.

This method benefits from (but does not require) having large-coverage dictionaries (preferably in electronic format) with certain semantic information (at least syntactic-semantic features: *Human, Animal, Vegetable, Abstract, Locative, Temporal*; if possible also some other feature that could be useful (for example, *Body Part, Collective*) and, in some cases of large ambiguity, more precision in the form of certain syntactic-semantic class.

[1] Our method is partially but deeply inspired by the works of Mylène Garrigues and her notion of "plausibility", cf. for instance Garrigues (1992). On the application to an electronic Spanish dictionary, cf. Blanco (2001).

We start from a Spanish dictionary of about 60,000 disambiguated lemmas (which correspond, in principle, to lexemes)[2]. Each participating expert is asked to project a tripartition[3] on this lexicon, labeling each entry as:

- 1 — lexeme considered necessary for a proficient speaker of the language in question.
- 3 — lexeme considered particularly marked: diaphasic (e.g., form typical of literary language), diachrony (e.g., form in disuse or archaizing), diatopic (regionally marked form), diaspeciality (form closely linked to a precise field of knowledge, non-trivial terminology), dianormative (a commonly used form that is however erroneous from a normative point of view), diaintegrative (rare or foreign words), etc.
- 2 — label 2 marks lexemes that the expert cannot attribute to either 1 or 2.
- 0 — a label 0 is available for those entries that the expert wishes to mark as excluded for some reason (not recognized as forms of Spanish, etc.).

This classification must be carried out with some speed. We calculate no more than five or six minutes for every hundred lemmas (even greater speed is possible). Otherwise, the expert begins to take into account too much metalinguistic considerations, which leads to heterogeneous and, therefore, unusable results. Let us note that we are trying to reflect word familiarity and linguistic competence, not specialized lexicographic knowledge. Experts do not have to be professional linguists, but nothing prevents them from being so.

In this experience we have carried out, lexemes marked as 1 correspond, for the entire lexicon, to 30% of the Spanish dictionary, that is, about 20,000. It is interesting to note that this corresponds to the expected vocabulary breadth of the C2 level. Everything seems to indicate that the expert native speaker is encoding C2 as the first level, which is not surprising, since that is his level of internalized functioning. Halfway in terms of level (but not in terms of number of lexemes), we have B1.

From this list, the expert is asked again for a tripartition:

- a — a language learner (in early stages) cannot do without this lexeme without risking having greater difficulties of expression or understanding)
- c — a language learner (in early stages) can do without it because you have more common synonyms, hypernyms or hyponyms, paraphrases, etc.)
- b — There is reasonable doubt between both categories (hesitation).

Level *a* corresponds to just over 10% of the 20,000 forms previously marked as 1. That is, few lexemes are marked as “essential” (note that we are very close to 2,500 — levels A1 to B1).

Of the approx. 2,500 lexemes that would constitute the lexicon from A1 to B1, the proportion by parts of speech is at this moment (we give approximate percentages):

A (adjectives) (394)	16%
ADV (adverbs) (75)	3%
CONJ (conjunctions) (11)	0.4%
ART (articles) (2)	0,08%
INT (interjections) (9)	0.4%

[2] Since lemma is a lexicographic category and lexeme is a lexicological category, it is appropriate to continue making the difference even if we reach the ideal 1 lemma = 1 lexeme.

[3] Tripartition has a cognitive basis: "yes", "no", "neither yes nor no".

N (names) (1460)	59%
PREP (prepositions) (19)	0.8%
PRO (pronouns) (25)	1%
V (verbs) (427)	17%
XI (residual) (6)[4]	0,24%

Once this base of approx. 2,500 A1-A2-B1 lexemes has been established, the labeling phase will be carried out by Spanish learners of approximately B2-C1 level. They will be asked to also project a tripartition: 1 "it is basic vocabulary from my point of view", 3 "it is part of my linguistic competence, but it is not basic vocabulary", 2 "I hesitate", 0 "it is not at all part of my competence".

Let us note that the learners must be more advanced than B1, to avoid reflecting too partial a competence: if they have obtained, but not consolidated their B1. To the extent that they classify but do not select the lexemes, there is no danger that their extra "competence" will interfere: they will not be able to label as 1 a B2 or C1 lexeme that would no longer be part of the 2,500.

2.2.2 Operational remarks

The important obstacle posed by the so-called "polysemy" (in reality, the fact of working on vocables instead of working on lexemes) can be considered resolved empirically and provisionally if we take into account that, up to level B1 (but especially in A1 and A2), only the most common lexeme of each word is considered and known, with some exceptions that should be specifically addressed.

Regarding phraseology, it is probably most advisable to follow, at first and until a more appropriate treatment can be adopted, a double strategy based on taking into consideration some unavoidable phrasemes (for example, some conjunctions and adverbs) and a tripartition between a) non-compositional phrases (which can block comprehension), b) quasi-compositional phrases (which tend to be understood, at least partially and, especially, by speakers of similar languages) and c) collocations (which are formally very varied but have the advantage to have a reduced number of meanings —at least the most common ones).

Derivation (authentic composition is residual in Spanish) should also be discussed, which can multiply the number of lexemes available as passive vocabulary with a reduced investment on the part of the learner.

2.3 Automatic inference

Automatic inference approaches to the definition of complexity lexicons concern methods using specific rules, knowledge, or standards to infer if a word should be considered as simple or complex or to assign it to a given complexity level.

Inference can be based on various word characteristics, including imean number of letters, phonemes, syllables, word frequency, sound-script correspondences, orthographic neighbors, imageability, age of acquisition, etc. It can also refer to more sophisticated measures, combined with NLP tools and methods, usually using strutured resources such as computational lexicons (wordnets), distributional lexicons, etc. However, the major innovation brought by automatic inference aproaches is the use of language models "crucial to predict the level of difficulty of a word by combining and weighting the different predictors over large amounts of data." (Gala et al. 2013: 147).

Relevant work has been developed within this approach, especially considering L2 lexicons, as vocabulary is an essential part of language learning and is considered in all reference documents (CEFR, PIACC, etc.). For

[4] Normally form-words that only appear within the framework of phraseological units.

instance, Kidwell et al (2011) have developed a complex statistical method that automatically estimates word acquisition age on a corpus of educational texts. Brooke et al. (2012) produced a graded lexicon using a method inspired by the automatic design of polarity lexicons. More recently, Gala et al. (2013) and François et al. (2016) combined various word features within a support vector machine classifier to assess the complexity of synonyms.

Closer to our goals is the fact that advanced L2 reading is linked to the ability to recognize a large set of words, above 10 000 (Grabe 2014). Pintard & François (2020) presents work on automatic inference of CEFR levels for words, combining expert knowledge, reflected in already existing resources, with frequency information extracted from non-structured corpora. Considering the Reference Level Descriptions (RLD) for French as a gold standard for the mapping of words to CEFR levels, the authors infer a statistical model from this pedagogical information that is able to transform the word frequency distribution from graded corpora into the adequate CEFR levels. In this approach, the knowledge from the French RLD is leveraged to train a mathematical function, based on machine learning algorithms, able to transform any lexical distribution into a CEFR level.

Besides making use of the resources resulting from the approaches previously described, Pintard & François (2020) also reports on the into the advantages and shortcomings of the frequency *vs.* expert knowledge information. From the issues discussed, at least three are also relevant for the definition of lexicons for native speakers:

- Lexical coverage: RLD semantic organization reflects expert knowledge on semantic domains, as well as lexical availability, resulting in more comprehensive lists available words.
- Representativeness and relevance of the topics covered: data-driven results reflect corpus composition. Often the materials compiled avoid relevant topics while other topics are overrepresented. Expert knowledge can easily compensate that, although it was clear that some topics have denser and more detailed nets of words (e.g., food vs. human body).
- POS coverage: lists provided by experts with the goal of augmenting learners' vocabulary often focus on content words (i.e., words conveying denotational meaning, typically nouns, verbs and adjectives), as these are the ones describing the entities, situations and events particular to speficic topics. Grammatical or function words (i.e., words which have grammatical function but light or no meaning, as determiners or conjunctions) are considered by experts as relating to other part of the L2 teaching/learning process focusing on grammar and are not considered in the vocabulary lists. Frequency information, however, constrasts with this as pronouns, determiners, prepositions, etc., are highly frequent.

3. iRead4Skills mixed approach

As described above, the three major types of approaches mentioned complement each other. Manually built resources are richer in terms of some specific information but tend to be shorter and more-time consuming to build. Data-driven and automatic inferred resources are quicker to get but require representative and balanced corpora as well as other structured source resources.

Considering the different available resources for each language and taking advantage from the expertise gathered in the iRead4Skills consortium, the compilation of the lexicons to be used in the project will follow a mixed approach. It will combine:

- corpus-driven methods (e.g., for Portuguese, to extract graded lexicons per CEFR levels and to extract core/fundamental general language vocabularies),

- automatic inference (e.g., for French using graded lexicons for CEFR levels and distributional data from iRead4Skills validated corpus),

- expert grading (e.g., for Spanish to define core vocabularies usable for complexity analysis and for automatic inference tools; for Portuguese, to validate and further adapt CEFR graded lexicons to native-speakers' adult case).

To attain the iRead4Skills project goals, we have to adapt and/or built the resources for the specific case of native speaker with low reading skills, trying to account for the differences between L1 and L2 oriented lexicons, as described in Pintard & François (2020), namely that the L2 lexicons represent word distribution in materials targeting L2 learners, but also that the L1 lexicons represent native distribution of words considering average L1 speakers with no reading difficulties. So, differences between active and passive lexicon of these speakers are in order, as well as specific phenomena affecting readability and not necessarily lexicon acquisition.

As also captured in previous work (Pintard & François 2020; Alfter et al. 2022), mixed approaches have the ability to uncover and compensate the shortcomings of approaches taken in isolation. For instance, expert grading approaches tend to focus on content words (nouns, verbs, adjectives), treated in more detail concerning semantic information and disambiguation, and to ignore function words (determiners, copula verbs, conjunctions, etc.) (although not always, cf. section 2.2), as these are transversal to semantic domains (food, housekeeping, school) and communicational contexts. Corpus-based and inference approaches can compensate this, as function words, besides being highly frequent, are also easily retrieved since these make up closed lists. In the opposite direction, expert graded lexicons and/or manual validation by experts allow us to account for regular phenomena, regardless of their occurrence in the data. For instance, expert knowledge can provide information on productive affixes for each language, augmenting in a sustained way the vocabularies considered.

Figure 1 presents the general schema for the definition of complexity lexicons in iRead4Skills project, combining the different approaches and accounting for the diversity of the available resources for each language. The idea is to allow for the use of different data sources, of different analysis and processing systems, and well as of different pipelines to reach the best outcome. Graded corpora considered as input include the corpora specifically built for the iRead4Skills project (see D3.7 Data set 1: corpora by level of complexity FR, PT and SP). The resulting lexicons will also be subject to the target-audience indirect validation, as adult learners will participate in a classification task for validating the corpora compiled, that contemplates the option of annotating words and expressions perceived as complex.

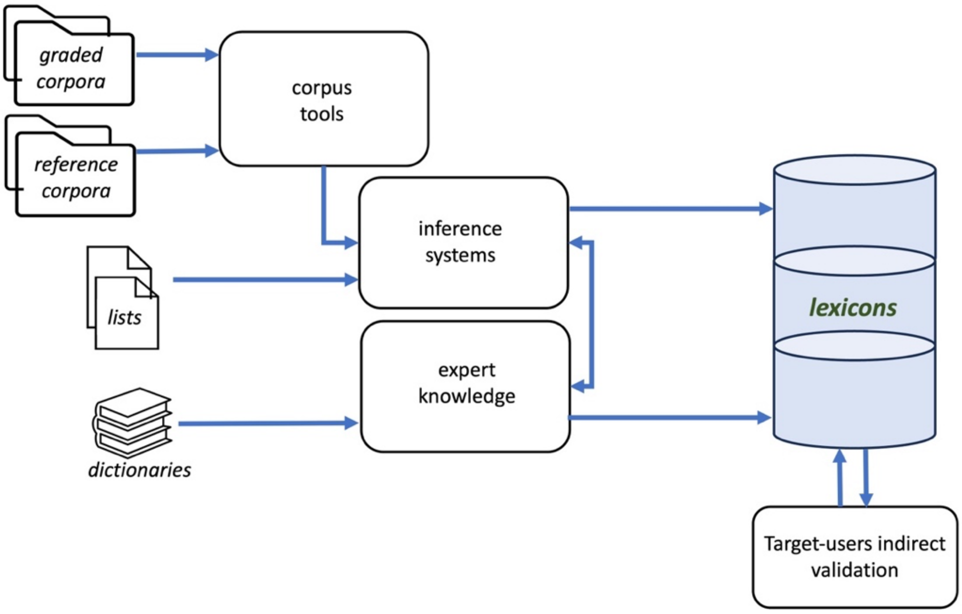


Figure 1: Method for complexity lexicons compilation

The resulting lexicons for the three languages covered in the project are expected to respond to the lexical features implemented by the complexity analysis (see D4.1 Report on the set of features for the ICA).

4. Available resources

The resources listed here comprise corpora, vocabulary lists, RDL and structured resources that can be used as sources to obtain the lexicons per complexity level for each language.

Resource type and title	language	link/reference
CETEMPúblico corpus	PT	https://www.linguateca.pt/cetempublico/
CHILDES FR corpus	FR	MacWhinney (2000)
CHILDES SP corpus	SP	MacWhinney (2000)
COMBINA-PT: Word Combinations in Portuguese Language list	PT	https://www.clul.ulisboa.pt/en/projeto/combina-pt-word-combinations-portuguese-language
COPLE2 corpus	PT	Mendes & Gonçalves (2016)
Corpus de Referência do Português Contemporâneo	PT	https://clul.ulisboa.pt/en/projeto/crpc-reference-corpus-contemporary-portuguese
Corpus of CEFR-graded exams	PT	Santos et al. (2021)
CREA - Corpus de Referencia del Español Actual, RAE	SP	https://corpus.rae.es/creanet.html, Real Academia Española: Banco de datos (CREA) [on line]. *Corpus de referencia del español actual.* <http://www.rae.es>
ELELex lexicon	SP	François et al. (2018)
FLELex lexicon	FR	Francois et al. (2014)
Lexique3 list	FR	New & Pallier (2019)
Multifunctional Computational Lexicon of Contemporary Portuguese	PT	Barreto & Amaro (2004), https://www.clul.ulisboa.pt/en/recurso/multifunctional-computational-lexicon-contemporary-portuguese
Portuguese Web corpus ptTenTen2020	PT	Kilgarriff (2014)
Reference Level Descriptors for FR	FR	Beacco et al. (2008)
Reference Level Descriptors for SP	SP	Instituto Cervantes https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/
Reference Level Descriptors for PT	PT	https://www.instituto-camoes.pt/activity/centro-virtual/referencial-camoes-ple
SUBTLEX-ESP list	SP	Cuetos et al., 2012

5. References

Alfter, D., Cardon, R. and François, T. (2022). A Dictionary-Based Study of Word Sense Difficulty. In *Proceedings of the 2nd READI Workshop @LREC2022*, pp. 17-24. https://aclanthology.org/2022.readi-1.3.pdf

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

Barreto, F. and Amaro, R. (2004). Multifunctional Computational Lexicon of Contemporary Portuguese: An Available Resource for Multitype Applications. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC'04), 1075-1078, http://www.lrec-conf.org/proceedings/lrec2004/pdf/303.pdf.

Beacco, J.-C., Lepage, S., Porquier, R., and Riba, P. (2008). *Niveau A2 pour le français: Un référentiel*. Didier.

Blanco, X. (2001). Dictionnaires électroniques et traduction automatique espagnol-français. *Langages* 143, pp. 49–70.

Brooke, J., Tsang, V., Jacob, D., Shein, F., & Hirst, G. (2012, June). Building readability lexicons with unannotated corpora. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations* (pp. 33-39).

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect : A review of recent developments and implications for the choice of frequency estimates in German. Experimental Psychology, 58(5), 412-424.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis : A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4), 977-990.

Brysbaert, M., Lange, M., & Wijnendaele, I. V. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition : Further evidence from the Dutch language. European Journal of Cognitive Psychology, 12(1), 65-85.

Carroll, J. B., Davies, P., & Richman, B. (1971). The American Heritage Word Frequency Book. Houghton Mifflin.

Connine, C., Mullennix, J., Shernoff, E. et Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. Journal of Experimental Psychology : Learning, Memory, and Cognition, 16(6):1084–1096.

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

Cowie, A. P. (1994). Phraseology. In *Encyclopedia of Language and Linguistics*, ed. Ronald E. Asher, 3168–3171. Oxford: Pergamon.

Cowie, A. P. (2001). Speech formulae in English: Problems of analysis and dictionary treatment. In *Making Senses: From Lexeme to Discourse. In Honour of Werner Abraham*, eds. Geart Van der Meer & Alice G. B. ter Meulen, 1–12. Groningen: Center for language and Cognition.

Cuetos, F. G. N., Maria Barbón, A. and Brysbaert, M. (2012). Subtlex-esp: Spanish word frequencies based on film subtitles. In *Psicológica*, vol. 33, núm. 2, 2012: 133-143. Universitat de València, https://www.redalyc.org/pdf/169/16923102001.pdf.

Durlich, L. and François, T. (2018). EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2018), 873– 879. https://aclanthology.org/L18-1140.pdf

Fonseca, A., Sadat, F., and Lareau, F. (2017). Combining Dependency Parsing and a Lexical Network Based on Lexical Functions for the Identification of Collocations. In *Computational and Corpus-Based Phraseology*, 447–461. https://doi.org/10.1007/978-3-319-69805-2_31.

François, T., Billami, M. B., Gala, N., & Bernhard, D. (2016). Bleu, contusion, ecchymose: tri automatique de synonymes en fonction de leur difficulté de lecture et compréhension. In *JEP-TALN-RECITAL 2016* (Vol. 2, pp. 15-28).

François, T., Gala, N., Watrin, P., and Fairon, C. (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, (LREC 2014), 3766-3773. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1108_Paper.pdfGala, N., François, T., and Fairon, C. (2013). Towards a french lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *Proceedings of eLex2013*, 132–151. http://eki.ee/elex2013/proceedings/eLex2013_10_Gala+Francois+Fairon.pdf

Garrigues, M. (1992). Dictionnaires hiérarchiques du français. In *Langue française* 96: 88–100.

Gernsbacher, M. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. Journal of Experimental Psychology : General, 113(2):256–281.

Howes, D. H., & Solomon, R. L. (1951). Visual Duration Threshold as a Function of Word-Probability. Journal of Experimental Psychology, 41(6), 401.

Kidwell, P., Lebanon, G., & Collins-Thompson, K. (2011). Statistical estimation of word acquisition with application to readability prediction. *Journal of the American Statistical Association*, *106*(493), 21-30.

Kilgarriff, J.M.P., J.S.T..W.P., A. (2014). PT TenTen: A corpus for portuguese lexicography. In T.d.L. T.B. Sardinha, T.B. São Bento Ferreira (ed.) *Working with Portuguese Corpora*, 111– 128. Bloomsbury Academic.

Kučera, H., & Francis, W. N. (1967). Computational Analysis of Present Day American English (1st edition). Brown University Press.

Leech, G., Rayson, P., & Wilson, A. (2001). Word Frequencies in Written and Spoken English : Based on the British National Corpus.

Lively B. A. & Pressey S. L. (1923). A method for measuring the vocabulary burden of textbooks. In *Educational administration and supervision*, 9(7), 389-398.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated a pproaches to lexical diversity assessment. Behavior Research Methods, 42(2), 381-392.

Mel'čuk, I. (1998). *Collocations and lexical functions. Phraseology. Theory, Analysis, and Applications*, ed. Anthony P. Cowie, 23–53. Oxford: Oxford University Press.

Mendes, A., & Antunes, S. (2016). Collocations in Portuguese: A corpus-based approach to lexical patterns. In B. V. Sanromán (Ed.), *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*, 141-166. Sanromán, Helsinki: Société Néophilologique.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. Basic Processes in Reading: Visual Word Recognition. Lawrence Erlbaum Associates Inc.,Hillsdale, NJ., 148-197.

New, B. and Pallier, C. (2019). *Manuel de Lexique 3*. http://lexique.org/_documentation/Manuel_Lexique.3.pdf

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

Pintard, A. and François, T. (2020). Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties* (READI), Marseille, France. ELRA, 85–92. https://aclanthology.org/2020.readi-1.13.pdf

Pirali, C., François, T. and Gala, N. (2022). PADDLe: a Platform to Identify Complex Words for Learners of French as a Foreign Language (FFL). In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties* (READI), Marseille, France. ELRA, 46–53. https://aclanthology.org/2022.readi-1.7.pdf

Sag, I., Baldwin, T., Bond, F. Copestake, A. and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of CICLING-2002*, ed. A. Gelbukh, 1–15.

Santos, R., Rodrigues, J., Branco, A. and Vaz, R. (2021). Neural Text Categorization with Transformers for Learning Portuguese as a Second Language. In Marreiros, G., Melo, F.S., Lau, N., Lopes Cardoso, H., Reis, L.P. (eds) *Progress in Artificial Intelligence. EPIA 2021*. Lecture Notes in Computer Science, vol 12981. Springer, Cham, 715-726. https://doi.org/10.1007/978-3-030-86230-5_56

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Thorndike, E. L. (1921). The Teacher's Word Book. Teachers College, Columbia University.

Zeno, S., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). The Educator's word frequency guide. Touchstone Applied Science Associates.