

Project Title:

INFRASTRUCTURE AND INTEGRATED TOOLS FOR PERSONALIZED LEARNING OF
READING SKILL

Project Acronym:**Grant Agreement number:**

731724 — iRead H2020-ICT-2016-2017/H2020-ICT-2016-1

Subject:

D5.1 User-Model Driven Content Classification Metrics

Dissemination Level:

PUBLIC

Lead Beneficiary:

UOI

Project Coordinator:

UCL

Contributors:

KNOW, UCL, PICKATALE


Revision	Preparation date	Period covered	Project start date	Project duration
V2	30/09/2018	Month 4-21	01/01/2017	48 Months
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No 731724				

Table of Contents

1.	Executive Summary	4
2.	Background: Content classification systems	5
3.	Content classification in iRead: Description of the system	11
3.1	Generic content classification metrics	12
3.1.1	Quantitative metrics	12
3.1.2	Linguistic metrics: syntactic complexity	13
3.2	User-model driven metrics for personalised content selection	16
3.2.1	Phonology	17
3.2.2	Word recognition and orthography	18
3.2.3	Morphology	18
3.2.4	Morphosyntax and syntax	19
3.3	Weighting user-driven metrics	22
4.	Methodology: training and testing the content classification system	24
4.1	Training: Preliminary Training Setup (September 2018)	24
4.2	Training: Full training (November-December 2018)	25
4.3	Testing the generic content classification	25
4.3.1	Participants	26
4.3.2	Materials and Procedure	26
4.4	Final testing of generic and user-driven content classification (Task 9.11)	26
5.	Conclusions	28
	References	29

List of Tables

Table 1 Comparison across different systems, from Toyama, Hiebert & Pearson, (2017:145)	10
Table 2 Syntactic metrics in order of increasing difficulty	15
Table 3 Phonology categories.....	17
Table 4 Word recognition, Orthography	18
Table 5 Morphology	19
Table 6 Morphosyntax.....	19
Table 7 Syntax.....	21
Table 8 Weights of linguistic levels.....	23
Table 9 Difficulty levels.....	24

Table 10 Texts given to each group of students.....	26
---	----

List of images

Image 1 The iRead content classification system	12
---	----

1. Executive Summary

This deliverable describes the rationale for, and the outline of the content classification module (CCM) that will be incorporated in the IREAD software. Specifically, a description of the developed metrics for classifying texts based on their “difficulty” will be provided, where difficulty is related to the language features that appear within texts. Given that the same text might be easy or hard for a reader based on the language features he has mastered or not, personalization of the content classification is achieved by taking into consideration the user’s model and child’s competence on the language features. The design of the CCM aims to provide individualized teaching assistance to different student groups (i.e., novice readers, children with dyslexia and L2 learners) by enabling a teacher or parent to classify texts with respect to the degree of appropriateness for a particular child, based on his/her profile, as well as to search for appropriate content for a particular child.

Text classification will be determined based on text’s readability and linguistic complexity, in the sense that the readability of a text increases as linguistic complexity decreases and vice versa.

2. Background: Content classification systems

2.1. Introduction

A child learning to read and/or write will practice with several pieces of text. However, not all text is appropriate to be used in the learning process (either for reading or for writing) of a particular child. Selecting teaching material that is appropriate for a given student group is crucial to supporting the effectiveness of a teaching intervention. Any material given to students, especially texts, needs to correspond to their learning needs and language skills, in order to provide them with the opportunity to improve their learning and enhance motivation and self-confidence (Grabe and Stoller, 2002). A text that is too easy may lead to demotivation, while a text that is too difficult may hinder self-confidence. An appropriate text needs to be challenging enough to stimulate students cognitively, but it also needs to include information that can be handled easily to assist comprehension and foster students' sense of accomplishment and self-confidence. Importantly, text classification is necessary because classroom assessments are also used to determine students' eligibility for an intervention, in addition to what a student gets to read in and out of the classroom (Toyama, Hiebert & Pearson, 2017). Reading assessments are also being used for summative evaluation of student reading growth or effectiveness of interventions (Deeney & Shim, 2016; Hall, 2006; Leslie & Caldwell, 2015; Mellard et al., 2009; Paris, 2002; Spector, 2005). So, if classroom assessments cannot meet specific requirements, then test users cannot attribute the differential performance of students on various passages to students' capacity to handle various levels of complex text. The selection of appropriate reading material for particular students is often a challenging and laborious process, which involves classifying texts with respect to their readability and linguistic complexity.

Several factors contribute to a text's complexity/ difficulty, such as text's linguistic complexity, reader's familiarity with the topic, word difficulty, sentence length, concreteness of ideas and concepts, among other factors. Specifically, Lipson & Wixson (2003) suggested that the factors which affect text readability include not only **quantitative features**, such as the number of syllables in the words, the number of words in the sentences, word frequency, word length, sentence length, text length, but also **qualitative features**, such as vocabulary and sentence structure, text organization, in addition to the amount of background knowledge that is required of readers (Chall, Bixsax, Conard, & Harris-Sharples, 1996). In a more detailed account, Hess & Bigham (2004) determined the following factors that contribute to text difficulty: word difficulty

and sentence structure, text structure, discourse style (e.g. satire or humor), genre, background knowledge, degree of familiarity with text topic, level of reasoning required, organization and layout of text and text length.

Thus, another perspective on text complexity is that text at any level can be rendered more or less difficult because of factors in the reading process in addition to properties of the text, namely comprehension and those related to reader, task, and context. With respect to comprehension for instance, an average sixth-grade student might have more trouble extracting the underlying meaning of an Aesop fable (which can be considered as a second-grade level reading) than selecting the main idea for a tenth-grade science passage about black holes in space. Context also plays an important role. Two passages of equivalent linguistic complexity can receive different instructional support in a classroom (e.g., one read independently and the second deconstructed in a small group of peers). In that case, comprehension results might differ among students. Reader factors, such as topical knowledge, linguistic sophistication, or interest also play an important role; several studies have shown that deep knowledge or interest in a topic might well overcome, or at least compensate for, a great deal of linguistically complex language in a text (Alexander, Kulikowich, & Jetton, 1994; Valencia, Pearson, & Wixson, 2011; Valencia, Wixson, & Pearson, 2014).

Notably, the level of text difficulty may differ across student groups. For instance, although factors such as the child's age, the size of her vocabulary and the syntactic complexity of the text might be enough to determine the degree of text appropriateness for a child without learning difficulties, the same is not true for children with learning difficulties. In that case, many more factors must be combined to determine the degree of text appropriateness and render it as (un)suitable for a child. Therefore, a personalized profile is necessary for each user in order to identify the appropriateness of the content. Specifically, each child's profile will specify possible error-types, and texts rich in words/structures will be sensitive to these error types that are likely to cause more problems to the child during reading/writing. Text classification will thus be a major component of the on-line recourse bank that will be supported by iREAD. Content classification and, consequently, profile parameterized text searching is a valuable feature for a user with a reading/writing disability. In that case, we will be able to sort our texts not only based on their size but based on a decreasing order of suitability/appropriateness following the individuals' profile. The implementation of this component is a language dependent task and will be supported for English, Greek, German and Spanish.

The present document is organized as follows: Firstly, in Section 2.2.1, the notions of text complexity and reading difficulty are described in a brief literature

review. Then, the best-known readability tests that have been used for the purposes of content classification are presented in Section 2.2.2. A description of the methodology used for the development of the content classification module is given in Section 3. Descriptions of the user model as well as some basic definitions and metrics relevant to classification of words and texts are presented. In Section 5, the basic techniques and the resources that we used in order to implement the module are listed. Section 6 presents a demo application that was implemented for the purposes of the first annual review of the project. We conclude in Section 7.

2.2. Linguistic complexity and reading difficulty

Content classification can be used for individualized teaching assistance for different student groups (i.e., novice readers, children with dyslexia and L2 learners). This will enable teachers or parents to classify texts with respect to the degree of content appropriateness, based on each child's profile. Text classification can be made based on their readability, which is determined by the linguistic complexity of a text.

2.2.1. What is linguistic complexity?

Complexity is one of the most debated notions in the linguistic literature. Blache (2011) differentiates between **local complexity** (i.e., structural complexity/ difficulty), which involves processing aspects and cognitive load, and **global complexity**, which refers to the language as a system rather than the complexity of a given construction. Similarly, Miestamo (2008) distinguishes between global and local linguistic complexity, referring to the complexity of a language or language variety and the complexity of a particular linguistic domain respectively. Local complexity therefore includes phonological complexity (e.g. size of phonemic inventory, incidence of marked phonemes, phonotactic restrictions, maximum complexity of consonant clusters), morphological complexity (e.g. extent of allomorphy use and morphophonemic processes), syntactic complexity (e.g. level of clausal embedding and recursion), semantic and lexical complexity (e.g. extensive occurrence of homonymy and polysemy, type/token ratios), pragmatic complexity (e.g. degree of pragmatic inferencing) (see Szmrecsanyi & Kortmann 2012 for a review).

In order to address complex reading skills for the languages under investigation, the developed domain models incorporate different *linguistic levels* including phonology, morphology and syntax. Each linguistic level is represented by a number of phenomena or structures, called *language categories*, each of which includes a set of specific instances, the *features*. Specifically, linguistic difficulty has been instantiated by

number values that denote a scaling of the features that belong to a single category with respect to their relative difficulty or linguistic complexity. For instance, in a category with 10 features, these features would be ranked (ordered) from the easiest to the hardest and each one will be given a number value, which reflects this order. In this category, difficulty levels would range from 1 to 10 or lower, as each level of the scale may correspond to more than one feature (i.e. several features may be placed on the same level of difficulty).

2.2.2 Linguistic complexity and text complexity

The complexity of a particular text is the result of combinations and interactions of a variety of factors. These may include linguistic complexity factors, topic familiarity, word difficulty, sentence length, concreteness of ideas and concepts and others. The first qualitative classification tools were developed since the early 1920's, aiming to analyze text complexity, focusing on the description of text features that impact on their readability. Specifically, the first readability formulas aimed to match reading materials with specific readers, to select appropriate teaching materials for the classroom. Readability formulas were built based on the assumption that reading difficulty is determined by specific text features, which can be entered into an equation that will produce a numerical estimate of readability for a particular text. Then each level of readability could be mapped onto a specific age or educational level, enabling the selection of appropriate reading material for an individual reader of a particular age or educational background.

Klare (1984) described the four most commonly used readability formulas using two independent variables of complexity, syntactic and semantic: the Flesch Reading Ease Index (Flesch, 1948), the Fry Index (Fry, 1968), the Dale-Chall Readability Formula (Chall & Dale, 1995), and the Flesch-Kincaid Grade Level (GL) Score (Kincaid, Fishburne, Rogers, & Chissom, 1975). In these formulas, syntactic complexity is measured in terms of sentence length, while semantic difficulty is differently measured either in terms of word length measured in number of syllables (Flesch, Flesch-Kincaid and Fry) or in terms of mean word frequency (Dale-Chall).

An alternative measure of text readability was introduced by Stenner et al. (1988), which was adopted in elementary and middle schools in the U.S. The Lexille Framework for Reading uses two linguistic variables as proxies for complexity: syntactic complexity in terms of sentence length (e.g., Ravid, Dromi, & Kotler, 2010), and semantic complexity in terms of word frequency, which is established through occurrence counts in a large corpus of texts that constitute representative reading materials for students

from kindergarten through college. However, this measure does not take into account a number of significant linguistic factors that affect syntactic complexity, such as embedding, recursion, locality, as well as factors that affect semantic complexity, such as polysemy, concreteness etc.

The same problem stands for all traditional readability measures presented so far. Specifically, these formulas tend to overlook important variables that determine linguistic complexity of a text (described in 2.1.), such as discourse characteristics, density of information, inferential requirements, textual cohesion, rhetorical structure, text genre, complexity of ideas etc., as well as reader-related variables, such as motivation, cultural background and general world knowledge (McNamara, Louwerse, and Graesser, 2002). Given that these critical variables for comprehension have not been taken into account, these formulas have often been considered “overly simplistic” (Sawyer, 1991) with regards to the complexity of what is being assessed. It is also interesting to note that the grammar used in a text is not considered an aspect of the text’s complexity according to the presented models although “grammatical knowledge can also aid reading comprehension and interpretation” (Common Core State Standards Initiative, 2010b, p. 29). Grammar is thus acknowledged as relevant, but its exclusion from the model for measuring text complexity, except as an aspect of a quantitative measure, nevertheless suggests a low priority placed on its consideration.

Recently, new quantitative tools (see Table 1) have been developed that use sophisticated statistical methods and multiple measures to determine text complexity, such as the Coh-Metrix (Graesser, McNamara, & Kulikowich, 2011), TextEvaluator (TE, formerly known as Source-Rater; Sheehan, Kostin, Napolitano, & Flor, 2014), and Reading Maturity Metric (RMM; Landauer, Kireyev, & Panaccione, 2011) (see Toyama, Hiebert & Pearson, 2017; for a review).

Table 1 Comparison across different systems, from Toyama, Hiebert & Pearson, (2017:145)**Table 1.** Four Analytical Tools of Text Complexity Used in the Study.

Analytical Tools (developer)	Unit	Linguistic/Text Variables		
		Word Level	Sentence Level	Discourse/ Text Level
Traditional Two-Factor	Flesch-Kincaid Grade Level (Kincaid et al., 1975)	• Word length	• Sentence length	
	Lexile (MetaMetrics)	• Word frequency	• Sentence length	
Newer- Multi-Factor	Reading Maturity Metric (RMM) (Pearson Education)	• Word Maturity • Word length	• Sentence length • Punctuation • Coherence	• Coherence • Order of info • Paragraph complexity
	TextEvaluator (TE) (ETS)	• Word unfamiliarity ¹ • Word concreteness ¹ • Academic vocabulary ¹	• Syntactic complexity ¹	• Lexical cohesion ¹ • Interactive style ¹ • Narrativity ¹ • Argumentation ¹

¹ A component derived from multiple variables based on principal component analysis.

Specifically, the RMM (Landauer et al., 2011) measures a range of text structure features and vocabulary, by using a mathematical model of human language that simulates the development of word meanings as learners' exposure to language increases. The RMM provides an overall text complexity score in grade-level units and, additionally, identifies the 10 most difficult words in a given text. The TextEvaluator system differs from other systems since it does not only provide a single, holistic score of overall complexity, but it takes into account other contributing components such as, part-of-speech tags and syntactic parses, unlike many competing systems which tend to only incorporate two or three basic features, such as average sentence length and average word frequency. Specifically, text complexity is based on eight dimensions: (a) academic vocabulary, (b) syntactic complexity, (c) word concreteness, (d) word unfamiliarity, (e) interactive/conversational style, (f) degree of narrativity, (g) lexical cohesion, and (h) argumentation (Sheehan et al., 2014; Napolitano, Sheehan & Mundkowsky, 2015). These eight components, tend to have higher correlations and can account for over 60% of variation in text difficulty across a wide range of passages as judged by human experts (in line with Nelson, Perfetti, Liben & Liben, 2012).

3. Content classification in iRead: Description of the system

Content classification is a smart classification of the text based on the preselected metrics. To make the process effective, the latest AI technologies to automate these tasks were used. IRead is going to use automated text classification Application Programming Interfaces (APIs) of two parties -- DFKI Munderline and Knowble 360 AI -- which work on training and test principles. The Content Classification System will apply generic content classification that will utilise general quantitative (Section 3.1.1) and linguistic (syntactic) metrics (Section 3.1.2). Based on these features, Ai is being trained on tagged/categorised content, which consequently will allow the algorithm to automatically classify it. The content classifier system will be trained using 100 books/examples per each difficulty level.

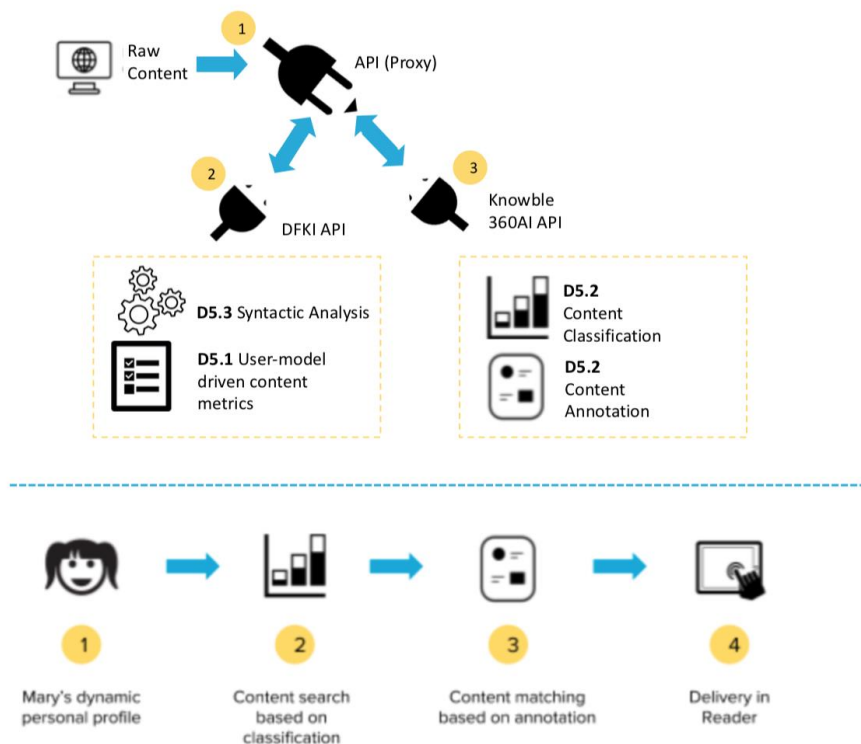


Image 1 The iRead content classification system

This is an overview of what **the system** will do (see Picture 1):

1. Content is being sent via proxy system to reach the APIs
2. Content is being analysed and annotated with all predefined language features via Munderline API
3. Content is being classified and annotated for the objective of content difficulty (grade). The use grade level is bundled into groups of 2 (e.g. 4-5; 5-6; 6-7 etc.), the difficulty level is indicated by the letter (e.g. A, B, C etc.)
4. The API can be used by IRead product to classify the text against age groups

3.1 Generic content classification metrics

3.1.1 Quantitative metrics

At basis of quantitative metrics, the Flesch-Kincaid Ease and Grade level are being used. Both provide readability scores that are based on the formulas depicted below.

Flesch-Kincaid Ease

$$206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right)$$

The score is able to indicate how easy it is to read this piece of content and what level of education is required - the higher the score, the easier it is to read this text (Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975).

Flesch-Kincaid Reading Level

$$0.39 \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left(\frac{\text{total syllables}}{\text{total words}} \right) - 15.59$$

The score of this formula helps to identify which grade level the piece of content belongs to and is equivalent to the US grade level of education (Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975).

The following quantitative metrics will be employed by the system:

- Total number of paragraphs
- Total number of sentences
- Total number of words
- Total number of syllables
- Average sentence length
- Average number of syllables

3.1.2 Linguistic metrics: syntactic complexity

As reviewed above (Section 2.2), for most existing systems of measuring text complexity, syntactic complexity is viewed in a narrow sense, essentially solely in terms of sentence length. Although sentence length is clearly a factor which influences the complexity of a text, sentence length alone is too crude a measure: two sentences of equal length may impact text complexity in dramatically different ways, depending on which linguistic phenomena are instantiated in each sentence. To give a simple example, combining matrix sentences with each other (parataxis) is not syntactically identical to combining dependent sentences with (a) matrix one(s) (hypotaxis); within the class of dependent

(i.e. embedded) sentences, the different categories (complement vs. adjunct clauses; different types of complement clauses) are not syntactically equally complex.

Such considerations suggest that in measuring syntactic complexity sentence length should be complemented with metrics of syntactic phenomena contained in sentences. In order to better model syntactic complexity, we thus chose a number of phenomena, all of which are contained in the domain model, which impact the complexity of the sentence in which they occur. We grouped together phenomena which have a similar impact, and we assigned a value on a scale (from 0 to 9) to each grouping of phenomena, indicating their difficulty based on information available in the literature. In setting these difficulty values, we considered evidence coming from language acquisition and language processing research, so that structures that are acquired later or are more difficult to process were given a higher value (references). The table that follows presents the phenomena included, as well as the numerical value on the aforementioned scale assigned to each one. Syntactic phenomena included in the model which are, however, not computationally exploitable (i.e. will not be identified by the syntactic parser) were not included. Thus, at this stage their impact is essentially impossible to compute reliably.

i. Binding phenomena

- Reflexive pronouns (e.g. *himself, ourselves*), Personal (object) pronouns (e.g. *him, us*): 0,1
- Reciprocal pronouns (e.g. *each other*): 0,2

ii. Coordination

- *or, and, nor, but, or, yet, so, and nor, but nor, or nor, neither, no more, only*: 0,3

iii. Negation & Modals

- **Negation**: *do not, don't, am not, is not, did not, have not, haven't, had not, hadn't, should not, shouldn't, would not, wouldn't, could not, couldn't, not*: 0,4
- **Modals**; *can/may/might/could, should/must/ought to/have to*: 0,4

iv. Conjunctions

- *either...or, not only...but (also), neither...nor, both...and, whether...or, just as...so, the...the, as...as, as much...as, no sooner...than, rather..than*: 0,5

v. Embedded clauses

- Complement clauses, introduced by the complementizers *that*, *who*, *what*, *whether*: 0,6
- Adverbial clauses (e.g. temporal clauses ... *while I was reading the book*): 0,7
- Relative clauses:
 - *that*-RCs, subject extracted, right-branching: 0,8
 - object extracted: 0,8
 - RCs with a relative pronoun: 0,8

vi. Passive structures

- Short passive (*was read*): 0,9
- Long passive (*was read by John*): 0,9

vii. Wh-questions

- Subject extracted *who* questions: 1
- Wh-questions object extracted *what* questions: 1
- Wh- questions object extracted *who* questions: 1
- *Which*-NP questions: 1
- Adjunct questions: 1

Table 2 Syntactic metrics in order of increasing difficulty

ID	DM Category	Factor (metric)	weight
183	Binding	Reciprocal pronouns: each other	0,2
184	Binding	Personal (object) pronouns	0,1
185	Binding	Reflexive pronouns	0,1
186	Coordination	or, and, nor, but, or, yet, so, and nor, but nor, or nor, neither, no more	0,3
187	Coordination	either...or, not only...but (also), neither...nor, both...and, whether...or, ... etc.	0,5
198	Embedding	Adverbial clauses	0,7
199	Embedding	that-RCs, right-branching	0,8
200	Embedding	that-RCs, centre embedded	0,8
201	Embedding	that-RCs, subject extracted	0,8

202	Embedding	that-RCs, object extracted	0,8
203	Embedding	RCs with a relative pronoun	0,8
205	Embedding	Complement clauses: Complementizers (that compl.)	0,6
209	Negation	do not, don't, am not, is not, did not, have not, haven't, had not, hadn't, etc.	0,4
211	Passive	Short passive	0,9
212	Passive	Long passive	0,9
225	Wh- questions	object extracted 'what' questions	1
226	Wh- questions	Which-NP Questions	1
227	Wh- questions	Subject extracted 'who' questions	1
228	Wh- questions	object extracted 'who' questions	1
229	Wh- questions	adjunct questions	1
230	Wh- questions	subject extracted 'what' questions	1
232	Modals	Predictive: will/would/shall	0,4
233	Modals	Possibility: can/may/might/could	0,4
234	Modals	Necessity: should/must/(ought to/have to)	0,4

3.2 User-model driven metrics for personalised content selection

In this task we develop, implement and evaluate personalized metrics for content selection that will be selected based on each user's individual needs. The content classification component will make use of information drawn from the individual user profile and each child's performance on the linguistic features included in the domain model. In order to achieve this, the features of the English domain model were used as linguistic factors/metrics that will determine the selection of content that is appropriate for each user's specific learning profile. However, as the linguistic content of the English DM is too fine-grained, including more than 279 language features in 5 linguistic levels (phonology, morphology, morphosyntax/syntax, orthography and word recognition) and 28 language categories, these features were grouped into more general linguistic

factors that will serve as metrics, in order to minimize complexity of computation and maximize functionality. After the factors were formed, the linguistic levels and categories and the language factors within each level and category were differently balanced, so that weights were assigned to each factor based on each user group's learning needs. For instance, phonology was assigned a higher weight for the dyslexia group, as dyslexia has been found to significantly affect phonological skills. Lexical features (i.e. irregular verb or noun formations), on the other hand, were assigned higher weight for the EFL group, due to EFL limited vocabulary compared to native speakers. This means that three different sets of user-driven metrics were formulated, one for each user group (novice readers, dyslexia and EFL). The formation process of the user-specific metrics along with a justification of their weighting is described in the following sections.

3.2.1 Phonology

The linguistic level of Phonology includes 3 language categories and 165 features: Grapheme-Phoneme Correspondence (120 features), Clusters (40 features) and Syllabification (5 features). Grouping in phonology was based on phonemes' similarities, difficulty level (included in English DM) and frequency (included in English DM). There are 3 metrics in the Grapheme-Phoneme correspondence category (GPC), 2 in clusters and 1 in syllabification (see Table 3). More specifically, as for GPC, single consonants and single vowels were grouped together in one metric. A different grouping was made for double consonants, and for phoneme alterations (graphemes that have different phoneme correspondence, e.g. letter <c> in the word *city*, which is pronounced as ['sɪtɪ] or letter <u> in the word *bus*, which is pronounced as [bʌs]) and exceptions (e.g. letter <y>, that is pronounced as /i/ in the word *sunny* ['sʌni] but as /j/ in the word *you*), as well as split digraphs (e.g. /aɪ/ -i_e in the word *bike*, pronounced as [baɪk]). All features in the Clusters category are used as one metric and features under Syllabification as another.

Table 3 Phonology categories

Category	Metric N	Features
GPC	1	Single phonemes (consonants+vowels) (features 1,7,8, 10, 11, 13, 18, 20, 22, 23, 78, 80, 82, 86, 88,89, 93, 96, 99, 102, 104, 106 - 108)

	2	Double consonants (features 2, 3, 9, 12, 14, 79, 81, 83, 84, 87, 94, 95, 97, 100, 109, 110)
	3	Alterations and exceptions (consonant alterations, consonant exceptions, vowel alterations, vowel exceptions, split digraphs; features 4-6, 15-17, 19, 21, 24-77, 85, 90-92, 98, 101, 103, 105, 11-120)
Clusters	4	clusters (features 121- 160)
Syllabification	5	Syllables (features 161- 165)

3.2.2 Word recognition and orthography

Orthography is part of the English DM with an emphasis on letter similarity; for instance, letters <d> and <p> <q> are the most commonly reversed letters in the English language due to their shape and sound (Brendler and Lachmann 2001). These features are used as one metric, while all features in the Word Recognition level were included in another. This level includes 400 words which are organised in features (each feature contained a list of words) based on the order in which they tend to be taught, and their word length.

Table 4 Word recognition, Orthography

Linguistic Level	Metric N	Features
Word recognition	6	word recognition (frequency) (features 166-175)
Orthography	7	confusing letters (features 176-179)

3.2.3 Morphology

Within the linguistic level of morphology, in a similar way to phonology, the domain model features were used as 2 metrics (see Table 5). The first metric includes prefixes like *re-*, *pre-*, *ex-* etc. (e.g. *re-write*). Derivational suffixes were also used as a single factor,

including all grammatical categories (suffixes of nouns, e.g. *pay-ment*, verbs, e.g. *popular-ise*, *short-en*, adjectives, e.g. *industri-al*, *hero-ic*, and adverbs, e.g. *quick-ly*).

Table 5 Morphology

Category	Metric N	Features
Prefixes	8	Prefixes (features 235-243)
derivational suffixes	9	Derivational suffixes (nouns, verbs, adjectives and adverbs; features 244-253)

3.2.4 Morphosyntax and syntax

The linguistic level of Morphosyntax includes 10 categories (see Table 6) and their grouping was based on feature difficulty level and the particular characteristics of each category. The first metric includes adverbs (Comparative adverb: e.g. *more slowly*), whilst in the second includes agreement in the use of auxiliaries in the present and past tense (e.g. *I am*, *s/he is*, *I do*, *you did*). Inflectional suffixes of verbs and nouns are grouped together in one metric, while Irregular forms (past, plural and comparative/superlatives) are grouped in a separate metric. In addition, adjectives are used as one metric, while all irregular formations are included in the Lexical metric.

Table 6 Morphosyntax

Category	Metric N	Features
Adverbs	10	Comparative adverb (more+adverb) (features 254)
Agreement	11	auxiliary 'be', 'have', 'do', present and past (features 255-257)
Inflectional Suffixes	12	Present tense, regular past, aspect, plural (features 258-269)
Adjectives	13	Comparative + superlative (features 271-276)

Lexical	14	irregulars (comparatives/superlatives, plural,past) (features 270, 277-279)
---------	----	--

The linguistic level of syntax contains 14 categories including function words (e.g. clitics and articles), prepositions, negative particles, embedded constructions, passives, and discourse anaphors. An extensive literature review on typical reading development and language acquisition (as well as reading difficulties) was conducted prior to selecting and grouping syntactic features into metrics for content classification and selection, while features that were hard to computationally identify were excluded from the list of metrics. All categories and features form 16 metrics (see Table 7), depending on their difficulty level and syntactic characteristics. Group 1 includes adjectives in both attributive (e.g. *Maria has a **nice** dress*) and predicative position (e.g. *Maria's dress is nice*). All pronouns (reciprocal, personal and reflexive pronouns, features 183-185), are grouped together in metric 16, while coordination is included in metric 17 (e.g. conjunctions like **and**, **or**). Discourse Anaphors include personal pronouns, proper names and bare nominals (e.g. *I am sick*, *Pencils are made of wood*) as well as demonstrative pronouns and possessive nominals/ nominals with a genitive as a determiner (*John's book*), definite nominals, indefinite nominals (e.g. *this chair*, *a plate*), which are all included in metric 18. As for embedding, it is divided in 3 metrics depending on the feature's syntactic characteristics: metric 19 includes adverbial clauses (e.g. temporal, causal and conditional clauses), metric 20 includes relative clauses (RCs), and metric 21 includes complement clauses. Metric 22 includes negation (e.g. *do not*, *couldn't*) and metric 23 noun phrases with two or more premodifiers (e.g. *soft gentle flow*). All passive verbs are classified together in metric 24, whether they are followed by a *by*-phrase or not (e.g. *The ball was kicked*, *The woman was kissed by the boy*). All prepositional phrases are grouped under metric 25 and all quantifiers under metric 26. With respect to questions, we included simple *yes/no* questions (e.g., *Did you kiss him?* metric 28), as well as constituent (*wh*-) questions (including adjunct questions (e.g., *When did you kiss him?*), complex (e.g. *Which girl did you kiss?*) and simplex argument questions with animate (e.g., *Who did you kiss?*) or inanimate referents (e.g. *What did you kick?*)), which were all included in metric 27. Finally, all modal verbs were included in metric 29, while reduced relatives (feature 204), complement clauses with *that*-deletion (feature 208) and split infinitives (feature 207) were omitted, as they proved to be computationally unidentifiable.

Table 7 Syntax

Category	Metric N	Features
Adjectives (modifiers)	15	attributive, predicative, secondary predicate (3 features)
Binding	16	Reciprocal, Personal, Reflexive pronouns (features 183-185)
Coordination	17	Conjunctions (features 186-187)
Discourse anaphors	18	Personal pronouns, Proper Names, Bare NPs Demonstrative pronouns, Possessive NPs: NPs with genitive NPs as determiner, Definite NP, Indefinite NP (features 188-195)
Embedding	19	Adverbial clauses (features 196-198)
	20	Relative clauses: that-RCs (4 features)
	21	Complement clauses (features 205, 206)
Negation	22	(1 feature)
Noun phrases	23	two or more premodifiers (feature 210)
Passives	24	be + V+ed/PP (features 211-212)
Prepositional phrases	25	Prepositional phrases (functional, semi-lexical, Stranded prepositions) (features 213-216)
Quantifiers	26	All quantifiers (features 217-223)
Wh-Questions	27	object extracted 'what' questions, Which-NP Questions, Subject extracted 'who' questions, object extracted 'who' questions (features 225-230)
Yes-No Questions	28	Yes-No Questions (feature 231)
Modals	29	Predictive, Possibility, Necessity (features 232-

		233)
--	--	------

3.3 Weighting user-driven metrics

The weights of the metrics described in 3.2 were balanced per linguistic level, so that each of the 5 levels can contribute to the final selection not based on the number of features it includes, but based on its significance in determining a student's difficulty when dealing with written texts. Balancing was based on literature-based information on typical difficulties encountered by each group of users, leading to the formation of three distinct sets of content selection metrics, one for each group. So, for the group of **Novice Readers** the level of Phonology was assigned a lower weight compared to the other levels (10%), since typically developing children are expected to have acquired all phonological rules of their mother tongue long before the age of 7 (Carroll et al, 2011; Fowler, 1991; Ingram, 1974). On the other hand, children at that age keep enriching their syntax, as well as their vocabulary and spelling, which is why the levels of Syntax and Word recognition/Orthography were given a weight of 25% each, while the levels of Morphology and Morphosyntax contribute by 20% each. The weighting is somewhat different for the group of **Dyslexia**, which mainly affects children's phonological skills and graphophonemic awareness. Therefore, the levels of Phonology and Word recognition/Orthography were given a weight of 25% each, while Morphology and Syntax contribute by 15% each. Morphosyntax has a weight of 20%, as inflectional suffixes are often problematic in dyslexia as well (Carlisle, 2000). Finally, the levels that were considered most significant for the **EFL** group are Word recognition/Orthography and Morphosyntax. This is because EFL students have more limited vocabulary compared to native speakers (Gabb, 2000), which often hinders reading comprehension, while syntactic structures also tend to cause difficulties, which are milder or more severe depending on the learner's L1 (Day and Bamford, 1998). For this reason, Word recognition/Orthography contributes by 35% to the final content selection decision, with lower weights given to the rest of the linguistic levels.

Finally, it should be noted that all features within each linguistic level contribute equally to the level's final weight. The whole weighting system for the three groups is given in Table 8.

Table 8 Weights of linguistic levels

	Weight		
Linguistic Level	Dyslexia	Novice readers	EFL
Phonology	25%	10%	15%
Word recognition Orthography	25%	25%	35%
Morphology	15%	20%	20%
Morphosyntax	20%	20%	10%
Syntax	15%	25%	20%
<i>Total</i>	<i>100%</i>	<i>100%</i>	<i>100%</i>

4. Methodology: training and testing the content classification system

4.1 Training: Preliminary Training Setup (September 2018)

The text classification system will be trained in two phases: a phase of preliminary training, which will provide us with a training infrastructure. This includes implementing the extraction of all features for the algorithm. It will be completed by the end of September 2018 and consists of the following steps:

- Collecting pre-classified books for preliminary training setup: The set of pre-classified texts that has been obtained so far includes 205 texts written for various age groups, beginning from 4 years and up to 11 years of age. These texts will be used as an example to setup the infrastructure and test the extraction of features.
- Preliminary training will classify text based on difficulty level and age groups. The difficulty levels used can be found in the Table 9.
- Once the model has been set with the most optimal classification accuracy, the results need to be verified manually.

Table 9 Difficulty levels

Difficulty levels	Age	Year
0	4-5	Reception
A	5-6	Y1
A	6-7	Y2
B	7-9	Y3-Y4
C	9-11	Y5-Y6

Preliminary testing will produce indicative weights for each of the syntactic metrics defined for content classification. This was considered necessary as no available information on the impact each syntactic factor has on text difficulty can be found in the

literature, since no existing text classification system has included such detailed syntactic analysis of a text. Once preliminary weights have been set, a full training of the system will follow, where a much larger number of pre-classified texts will be used to fine-tune weights of syntactic metrics.

4.2 Training: Full training (November-December 2018)

After preliminary training has been completed, full training of the system will follow, which will involve the following steps:

- Collecting pre-classified books for full training (at least 100 examples of pre-classified texts per age-group).
- Full training will focus on improving the results of difficulty levels and age group classification.
- Once the model has been set with the most optimal classification accuracy, based on the dataset, the results need to be verified manually.

4.3 Testing the generic content classification

The testing of the generic content classification is focused on validation of the data accuracy. The metadata will be removed from the content and then analysed by the AI. Afterwards the results will be compared with the originally provided metadata. The results will be expressed in a F1 score as true positives, false positives, false negatives and true negatives.

The performance of the generic text classification performance will be validated with text ratings collected from school-aged children. The aim of this process is to validate the way texts are classified by the system, so that its output coincides with real user opinions regarding the difficulty of a given text for readers learning to read in their mother tongue.¹ During this process, only the generic content classification outputs can be tested, as testing the user-driven metrics requires use of the whole iRead system, which will occur during the iRead evaluation phase and will begin in January 2019 (see **Section 4.4** below). The following sections describe the methodology followed in the current testing procedure.

¹ Note that this testing process aims to validate classification results for novice readers, not EFL students.

4.3.1 Participants

Thirty (30) English-speaking school-aged children will participate in the testing, 10 of each of two age groups: 6 years old (Group A, 15 children) and 8 years old (Group B, 15 children). The age groups were selected in such a way to enable the testing of all three difficulty levels defined for the text classification component (see Section 4.1), so that texts of each level could be given to children below, within and above the age range it was written for. The children will be students of state and independent (private) schools that are already participating in the project around the London area.

4.3.2 Materials and Procedure

The students will be asked to read a number of texts selected from the sets of pre-selected texts that were used for training (see Section 4.1). Each age group will be given 5 texts from three different difficulty levels: Group A will be given 5 texts from difficulty level 0, 5 from level A and 5 from level B, while Group B will be given texts from Levels A, B and C (see Table 8), so as to establish that each group reads set of texts that are expected to be *easy*, a set of *appropriate* texts and a set of *hard* texts (Table 10).

Table 10 Texts given to each group of students

Age group	Level of texts rated		
	Easy	Appropriate	Hard
A (6 yrs)	Level 0 (4-5 yrs)	Level A (5-7 yrs)	Level B (7-9 yrs)
B (8 yrs)	Level A (5-7 yrs)	Level B (7-9 yrs)	Level C (9-11 yrs)

Students will be instructed to read the text and then decide how difficult they thought it was by rating it on a 5-level scale: too hard – hard – OK – easy – too easy. The students' responses will be analysed and compared to the level of classification given by the text classification system.

4.4 Final testing of generic and user-driven content classification (Task 9.11)

The effectiveness and usefulness of the user-model driven content classification component will be tested in a separate pilot of the iRead evaluation phase, which will

occur in Task 9.11. All the available/accessible materials provided by the publisher partners, libraries or the Internet will be deployed. Children and teachers will participate in the validation of the results. The results of the individual analysis, as well as the comparisons across case studies will provide firm conclusions to address the research goals stipulated in T9.11.

5. Conclusions

This deliverable describes the rationale for the content classification module (CCM) that will be incorporated in the iREAD software. As reviewed earlier, given that many factors (quantitative and qualitative) can affect text readability, it is not always easy to determine the degree of text appropriateness and render it as (un)suitable for a child. For instance, **quantitative features**, such as the number of syllables in the words, the number of words in the sentences, word frequency, word length, sentence length, text length, as well as **qualitative features** including vocabulary and sentence structure, text organization, in addition to the amount of background knowledge required from readers, play an important role. However, text complexity may also differ across different users. For instance, although factors like the aforementioned might be enough to determine the degree of text appropriateness for a child without learning difficulties, the same is not true for children with learning difficulties.

Since the overarching goal of the iREAD project is to enhance reading abilities across different student groups (i.e., novice readers, children with dyslexia and L2 learners), the development of a personalized content classifier which takes into account individual user's profile and child's competence on the language features is required. This technology will enable us to sort our texts not only based on their size but based on a decreasing order of suitability/appropriateness following the individuals' profile. The implementation of this component is a language dependent task and will be supported for English, Greek, German and Spanish.

The fact that preselected personalized metrics based on each user's individual needs will be used in the iRead content classification renders it quite innovative. Metrics will include not only quantitative features, such as total number of paragraphs, sentences, words, and syllables, average sentence length and number of syllables, but also, linguistic factors/metrics which will be integrated to determine texts' complexity based on each user group's learning needs. Therefore, the fact that different sets of user-driven metrics were formulated for each user group (novice readers, dyslexia and EFL) constitutes the system an innovative, personalised learning tool, which is expected to prove valuable to students and teachers.

References

Alexander, P.A., Kulikowich, J.M., Jetton, T.L. (1994). The role of subject-matter knowledge and interest in the processing of linear and nonlinear texts. *Rev. Educ. Res.* 64:201–52.

Blache, P. (2011). A Computational Model for Linguistic Complexity, In *proceedings of the First International Conference on Linguistics, Biology and Computer Science*.

Brendler, K., Lachmann, T. (2001). Letter reversals in the context of the Functional Coordination Deficit Model. In E. Sommerfeld, R. Kompass, & T. Lachmann (Eds.), *Proceedings of the International Society for Psychophysics*. Lengerich, Berlin: Pabst. 17: 308-313.

Caldwell, J., Leslie, L. (2010). Thinking aloud in expository text: Processes and outcomes. *Journal of Literacy Research*, 42: 308-410. doi:10.1080/1086296X.2010.504419

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing: An Interdisciplinary Journal*, 12: 169 –190.

Carroll, J.M., Bowyer-Crane, C., Duff, F.J., Snowling, M.J., Hulme, C. (2011). *Developing language and literacy: Effective intervention in the early years*. Malden, MA: John Wiley.

Chall, J. S., Bissett, G. L., Conard, S. S., Harris-Sharples, S. (1996). *Qualitative assessment of text difficulty: A practical guide for teachers and writers*. Cambridge, MA: Brookline Books.

Chall, J. S., Dale, E. (1995). *Readability revisited, the new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.

Day, R. R., Bamford, J. (1998). *Extensive reading in the second language classroom*. Cambridge: Cambridge University Press

Deeney, T. A., Shim, M. K. (2016). Teachers' and students' views of reading fluency: Issues of consequential validity in adopting one-minute reading fluency assessments. *Assessment for Effective Intervention*, 41(2): 1–18.

Flesch, R. F. (1948). A New Readability Yardstick. *Journal of Applied Psychology*, 32: 221-233.

Fowler, A. E. (1991). How early phonological development might set the stage for phoneme awareness. In S. Brady & D. Shankweiler (Eds.), *Phonological processes in literacy*. Hillsdale, NJ: Erlbaum

Gabb, I.(2000).From Talk to Print: Preparing Students to Read with Ease. *Field Notes*, (10); Retrieved on Nov1'2004 from ' <http://www.sabes.org/resources/Voll0>'

Grabe, W., Stoller, F. (2002). The nature of reading abilities. In W. Grabe & F. Stoller (Eds.), *Teaching and researching reading*, London, UK: Pearson: 9-39.

Graesser, A.C., McNamara, D.S., Kulikowich, J.M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40:223–234.

Hall, L.A. (2006). IRA Outstanding Dissertation Award for 2006: Anything but lazy: New understandings about struggling readers, teaching, and text. *Reading Research Quarterly*, 41: 424-426.

Hess, K. and Biggam, S. (2004). A Discussion of “Increasing Text Complexity”. Article produced in partnership with the New Hampshire, Rhode Island, and Vermont Departments of Education.

Ingram, D. (1974). Phonological rules in young children. *Journal of Child Language*, 1(1): 49-64.

Kincaid, J. P., Fishburn, R. P., Rogers, R. L., Chissom, B.S. (1975). Derivation of new readability formulas for navy enlisted personnel. *Technical Report Research Branch Report*, Millington, Tenn, Naval Air Station: 8-75.

Klare, G. R. (1984). Readability. *Handbook of reading research*, ed. P. D. Pearson. New York: Longman : 681-744.

Landauer, T.K., Kireyev, K., Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading*, 15(1): 92–108. doi:10.1080/10888438.2011.536130

Leslie, L. & Caldwell, J.S. (2015). *Content Area Reading Assessment: A Formative Measure of the Common Core State Standards*. Pearson.

Lipson, M.Y., Wixson, K.K. (2003). *Assessment and instruction of reading and writing difficulty*. Boston: Allyn & Bacon.

McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). Coh-Metrix: *Automated cohesion and coherence scores to predict text readability and facilitate comprehension (Tech. Rep.)*. Memphis, TN: Institute for Intelligent Systems, University of Memphi.

Mellard, D., McKnight, M., Woods, K. (2009). Response to intervention screening and progress monitoring practices in 41 local schools. *Learning Disabilities Research & Practice*, 24: 186–195.

Miestamo, M. (2008). Grammatical complexity in a cross - linguistic perspective. In: Miestamo, M., Sinnemäki, K. and Karlsson, F. (eds), *Language Complexity: Typology, Contact, Change*. Amsterdam: Benjamins: 23–41.

Napolitano, D., Sheehan, K., Mundkowsky, R. (2015). Online readability and text complexity analysis with TextEvaluator. In *Proceedings of NAACL 2015: Demonstrations*, Denver, Colorado: 96-100.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors.

Nelson, J., Perfetti, C., Liben, D., Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York, NY: Student Achievement Partners.

Paris, R., (2002). International Peacebuilding and the ‘Mission Civilisatrice. *Review of International Studies* 28(4): 637–656.

Ravid, D., Dromi, E., & Kotler, P. (2010). Linguistic complexity in school-age text production: Expository versus mathematical discourse. In M. A. Nippold & C. M. Scott (Eds), *Expository Discourse in Children, Adolescents, and Adults: Development and Disorders*, New York: Taylor & Francis: 123–150.

Sawyer, M. H. (1991). A review of research in revising instructional text. *Journal of Reading Behavior*, 33: 307– 333.

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The Text Evaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *Elementary School Journal*, 115(2): 184–209.

Spector, B. (2005) (ed.). *Fighting Corruption in Developing Countries*, Bloomfield, CT: Kumarian Press, Inc.

Stenner, A. J., Horabin, I., Smith, D.R., Smith, R. (1988). *The Lexile Framework*. Durham, NC: Metametrics, Inc.

Szmrecsanyi, B., & Kortmann, B. (2012). *Introduction: Linguistic complexity—Second Language Acquisition, indigenization, contact*, 6–34.

Toyama, Y., Hiebert, E. H., & Pearson, P. D. (2017). An analysis of the text complexity of leveled passages in four popular classroom reading assessments. *Educational Assessment*, 22(3): 193–170.

Valencia, S. W., Pearson, P. D., & Wixson, K. K. (2011). *Assessing and tracking progress in reading comprehension: The search for keystone elements in college and career readiness*. Princeton: Center for K-12 Assessment & Performance Management at ETS. Retrieved from http://www.k12center.org/publications/through_course.html

Valencia, S. W., Wixson, K. K., & Pearson, P. D. (2014). Putting text complexity in context: Refocusing on comprehension of complex text. *Elementary School Journal*, 115(2): 270–289. doi:10.1086/678296