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:
D9.1 Evaluation plan

Dissemination Level:
PUBLIC

Lead Beneficiary:
UGOT

Project Coordinator:
UCL

Contributors:
All Partners

Reviewers:
Mina Vasalou (UCL), Ylva Hård af Segerstad (UGOT),
Leona Bunting (UGOT), Manolis Mavrikis (UCL)

Revision	Preparation date	Period covered	Project start date	Project duration
V2	February 2020	Month 13-36	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 content

1 EXECUTIVE SUMMARY

The aim of the iRead Project is to enhance the technology previously developed during another EU FP7 Project, iLearnRW, elevating it to a flexible, scalable and cost-effective cloud-based infrastructure offering personalised learning services. This deliverable describes the evaluation studies to be performed in the final two years of the project, focusing on both the efficacy of specific technologies designed to support children's reading, and the embedding and scaling up in everyday practice. However, as this is a rather long period, and some studies will create output that serves as input for subsequent studies, it is foreseen than this deliverable will need to be updated after submission.

2 Introduction

2.1 Goals

This deliverable describes the evaluation approach for the iRead project. Addressing our readers, it is important to first set out the contextual dimensions of our research. Our project has two overarching research goals, each of which requiring a different methodological strategy. *First*, we seek to evaluate the efficacy of specific technologies designed to support children's reading. This research goal requires methodologies that will allow us to isolate and study relationships between design features, learning processes and outcomes. *Second*, we seek to embed and scale up (within, or across schools) the iRead technology in everyday practice across a number of diverse teaching contexts e.g. schools, language centres, special education needs sessions etc. This research goal requires tight collaborations with teachers that involve training through Continuous Professional Development (CPD), while collaboratively designing new ways of embedding the iRead technologies into teaching activities. Thus, unlike the control and precision required for the first goal, the second one requires flexibility and recognition of the different learning approaches employed in each context, resulting in potentially different configurations of technology.

2.2 Evolution of our strategy

In the DOA we originally proposed addressing the two goals above through a single, shared methodology. We additionally conceived of efficacy as an outcome of the combined use of the iRead products and services, i.e., the e-Reader Amigo app and text classification, the Navigo games, the learning analytics, as well as CPD activities. Since then, we have recognised the tension at play between the methodological control required to understand efficacy, and the flexibility required to support integration in practice leading us to rethink the single approach. Moreover, we have recognised that our scholarly contributions will have more impact if they rely on isolating the unique dimensions of each technology (e-Reader and games) to show how they support specific learning processes involved in reading acquisition. We have also acknowledged that for some of the learner groups involved in iRead (children learning English as a foreign language, beginning readers, and children with dyslexia) education research has produced more evidence than for others. This requires us to take both exploratory and confirmatory approaches in evaluating the same design dimension, depending on the learner group. Finally, in proposing research methodologies, we have had to ensure we address credibility markers associated with each academic community we intend to speak to, so that we are able to publish across the interdisciplinary communities represented in our consortium e.g. second language acquisition, technology-enhanced learning, typical and atypical reading.

2.3 The Monitoring and Evaluation Approach

Taking this together, we have decided to take a different approach from what was originally defined in the DOA. Instead of performing the series of pilot studies with an integrated 'system' as previously described in our project plan, we have decided to roughly follow the monitoring and evaluation (M&E) approach described by Patricio, Miguel, Ximena, and Marcos (2010) for the evaluation of ICT-supported education programs in schools. In this approach, there is a distinction between the *intervention* and the *implementation* phase.

The M&E approach assumes that the program to be monitored and evaluated has already demonstrated the efficacy of the intervention and implementation processes. During the *intervention*, the efficacy of the technology is evaluated to determine how it supports student learning. In moving toward the *implementation* phase, the teachers and students have to develop the necessary autonomy for using an ICT-supported pedagogical model, and

researchers may help them to adapt and modify it to fit their context. This involves activities such as teacher training, hands-on experiences and class observations in accordance with an intervention plan. This process leads up to teachers and students applying the model in their work context, supporting scaling and integration.

iRead consists of several products and services:

- the Amigo e-Reader (directed at students in class)
- the Navigo games (directed at students in and outside class)
- the text classification (underlying service to provide appropriate texts to students)
- the learning analytics and CPD (directed at teachers)

Our evaluation approach for iRead will consist of the following three tasks:

- **Task 1 (preparing for implementation phase):** from January to June/July 2019 the first version of the continuing professional development - CPD (consisting of learning designs, a teacher manual and a training programme) will be developed through a collaborative process with teachers. This phase will be described in more detail in Deliverable 9.2. It will support the delivery of the implementation phase. This task will focus on activities on teachers supported by the learning analytics.
- **Task 2 (intervention phase):** from April-June/July 2019, both controlled quantitative and qualitative methods studies are performed to understand efficacy of the unique qualities of the different components of iRead. As previously mentioned, the approaches taken may differ between target learner groups in order to comply with previous research and methods in the relevant fields, while we will use methodologies that isolate specific relationships. In this task, we study the Amigo e-Reader, the Navigo games, as well as the text classification.
- **Task 3 (implementation phase):** from September 2019 onwards, several studies will be performed in line with the monitoring and evaluation approach as described by Rodriguez et al. (2010). However, while they mainly focused on the correlation between appropriation and efficacy through a large-scale study, we will focus on understanding appropriation and scaling in a qualitative way. This phase will start with an intervention where the researchers will help the teachers and students to develop the necessary autonomy, based on the outcomes of Tasks 1 and 2. We will end with the implementation, where the researchers leave it to the teachers to implement the components in their own context, but will monitor their use. In this phase, all products and services developed in iRead are integrated.

Task 1 is the CPD, which will be described in Deliverable 9.2. The remainder of the present deliverable will thus describe Tasks 2 and 3.

3 Task 2: Focused evaluations

3.1 e-Reader highlights studies

3.1.1 Description of iRead app dimension considered

As described in D7.1, the Amigo e-Reader relies on the child's user model to deliver two personalised learning features:

- ***Explicit instruction***: a pre-reading explicit instruction activity that explicitly teaches a language rule
- ***Textual enhancement***: Personalised text highlights of a language rule within the book the child is reading to bring attention it

It is the efficacy of this design feature that will be evaluated with EFL, typically and atypically developing readers. It is noted that the EFL study takes a confirmatory approach building on available empirical findings, whereas the typically and atypically developing readers study takes an exploratory approach reflecting the scarcity of research on this topic. Moreover, given the breadth of the reading domain models underpinning our technology, each study focuses on a domain model level and category that is pertinent to learner group in question: EFL on morphology and typical and atypical early reading on GPC and morphology.

3.1.2 The impact of explicit instruction and text enhancement on EFL reading

In the field of second language acquisition (SLA) research, exposure to comprehensible input (i.e., language to which the learner is exposed) is generally considered to be a necessary condition for second language (L2) learning to occur. However, L2 learners usually do not process all the information that is made available in the input (Corder, 1967). Only a subset of the input to which learners are exposed gets processed and then learned, and attention is regarded as a key cognitive process determining what subset of the input gets selected for subsequent processing (Robinson, 2003; Schmidt, 2001). As a result, a principal question in instructed SLA research and practice is how learners' attention can be directed to linguistic features during L2 learning activities.

In the context of reading, researchers have suggested that one way to draw learner attention to language is by the means of textual enhancement. The aim of textual enhancement techniques is to make linguistic features salient in the input (Sharwood Smith, 1991, 1993), and thereby draw learners' attention to linguistic constructions that may otherwise remain unnoticed and thus unlearned (Leow, 1997, 1999, 2001; Robinson, 1995; Schmidt, 1990) while keeping learners' primary attention on meaning. Enhancing written input typically involves some kind of textual modification, such as underlining, highlighting, **boldfacing**, *italicization*, CAPITALIZATION, colouring or using different types of fonts (Sharwood Smith, 1991, 1993). The efficacy of textual enhancement has been examined in a large number of studies; a meta-analysis of 16 studies (Lee & Huang, 2008) found a small but positive impact of textual enhancement on L2 grammatical development. Most of the existing research, however, has been short-term (Han, Park, & Combs, 2008; Lee & Huang, 2008) and has focused on the acquisition of syntax and inflectional morphology by adult populations. Little is known about the *longitudinal* effects of textual enhancement on L2 development in derivational morphology by young learners, despite the fact that morphological knowledge is a reliable predictor of reading skills (Geva & Ramirez, 2015), and most second language instruction targets young learners. We

therefore plan to perform three studies in the EFL population using the E-Reader developed in iRead.

- The aim of the first study on highlighting is to help fill these gaps, looking at efficacy of **textual enhancement** (highlighting) for young L2 learners.

- We also plan to conduct a second study on highlighting to examine whether any effects of **textual enhancement** might be enhanced by the presence of **explicit instruction** integrated in the reader. Leow and Martin (2017) have recently suggested that textual enhancement is more likely to facilitate learning if it is used together with other attention-getting tools (e.g., feedback, explicit instruction), given that a combination of techniques is likely to lead to a higher level of processing.

- Finally, in a third study, we plan to investigate whether any effects of the **explicit instruction** are enhanced, when it is coupled with **textual enhancement**. Previous SLA research has generated plenty of evidence demonstrating that explicit instruction fosters second language learning. For example, meta-analyses carried out by Goo, Grañena, Yilmaz, and Novella (2015), Norris and Ortega (2000), and Spada and Tomita (2010) showed that explicit instruction led to greater development than implicit instruction. So far, however, little research has looked into whether the impact of explicit instruction can be further enhanced by textual enhancement.

Research question(s)

- **Study 1:** To what extent does textual enhancement lead to development in the knowledge of L2 derivational morphology?
- **Study 2:** To what extent does explicit instruction influence development in the knowledge of L2 derivational morphology, when textual enhancement is kept constant?
- **Study 3:** To what extent does textual enhancement influence development in the knowledge of L2 derivational morphology, when explicit instruction is kept constant?

Research design

Each of the three studies will employ a pre-test – post-test longitudinal design with 10 treatment sessions between the pre- and post-test. In Study 1, the participants will be randomly assigned to two groups, an experimental group and a control group. During the treatment, the experimental group will read texts with the target morphological constructions highlighted, whereas the control group will read texts without highlighting. In Study 2, the participants will be randomly assigned to two comparison groups. In the treatment sessions, both groups will read texts where the target constructions are highlighted, but one group will receive explicit instruction within the reader on the target feature, whereas the other group will not be exposed to explicit instruction. In Study 3, participants will also be randomly assigned to two comparison groups. This time, both groups will receive explicit instruction in the treatment sessions, but only one of the groups will be exposed to textual enhancement.

Participants

In each of the three studies, there will be 60 young learners. They will be at B1 proficiency level according to the Common European Framework of Reference (CEFR). In each study, learners will have the same first language.

Instruments

Treatment texts

During each treatment session, participants will read two texts. That is, altogether they will be exposed to 20 texts during the treatment. Using the tool Text inspector (http://www.englishprofile.org/wordlists/text-inspector), we will choose texts in which 95% of the vocabulary items are categorised as being at CEFR B1 or lower level. This will make it likely that B1 participants will achieve an acceptable level of comprehension of the texts (Laufer & Ravenhorst-Kalovski, 2010), and thereby have attentional capacity available to notice (Schmidt, 1990, 2001) the targeted morphological forms.

The target morphological constructions will be selected based on corpus analysis of the available texts. We will select constructions that are frequent in the texts and learners are unlikely to have full knowledge of (this will be confirmed through piloting with populations similar to the study participants).

Assessments

Proficiency test

In all three studies, participants will be administered the Oxford Placement Test to determine their proficiency level. Only students at CEFR level B1 will take part in the experiment, given that the reading texts available in iRead are suitable for this level.

Pre- and post-test measures

The pre-test and post-test will comprise equivalent versions of a series of morphological tasks (Spencer et al., 2015). Two tests of morphological structure (Carlisle, 2000), a *decomposition* and *derivation* task, will assess the extent to which students are aware of the relationships between base and derived forms. The decomposition task will require students decompose derived words to complete sentences (e.g., The word is *teacher*. The sentence is: Some young people are keen to _____). The derivation task will ask learners to produce a derived word to finish a sentence (e.g., The word is *drive*. The sentence is: My father is a _____.) In addition, three tests of derivational suffix choice will be employed. One of the derivational suffix choice tasks will include real words, and will require participants to select a correct option from four derivationally related options to complete a sentence (e.g., “How did you share the _____ ? (a) inform, (b) informing, (c) informed, (d) information.”) Another derivational suffix choice task will have the same format, but will use non-words (e.g., “My friend taught me how to ____ tennis. (a) daddling, (b) daddles, (c) daddle, (d) daddled.” In the third and last derivational suffix choice task, learners will be presented with an improbable stem and the will be asked to choose one of four sentences where the stem is correctly used (e.g., “Dogless. (a) The dogless can run fast. (b) He was in the dogless. (c) When he got a new puppy, he was no longer dogless. (d) He did not try to dogless.”)

Comprehension tests

After each text, participants will answer comprehension questions based on the text they have read to ensure that they process the texts for meaning. Also, this will make it possible to assess whether textual enhancement, in the presence or absence of explicit instruction, affects text comprehension.

Procedure

Each study will span approximately 11 weeks. First, participants will be administered the Oxford Placement Test to identify CEFR B1 students. Then, the students who have been selected to

participate in the study will complete the pre-test. Next, the 10 treatment sessions will follow. The treatment sessions will take place once a week, and last about 30 minutes. During each treatment session, participants will read two texts in their respective conditions, each text being followed by comprehension questions. Once the last treatment session has been completed, the post-test will be administered to the students.

3.1.3 The impact of explicit instruction and text enhancement on typical and atypical reading

Existing reading schemes, for example in the UK (e.g., Read Write Inc), typically provide a pre-reading activity for each book and focus on drawing the child's attention to specific, targeted GPCs that will be repeated throughout the story book to consolidate learning. For instance, the focus may be on split digraphs, or vowel and consonant alternatives (e.g., the alternative graphemes 'ou', 'ow' that make the same sound). In a similar vein, the development of the E-Reader has included an explicit instruction, as well as text enhancement, i.e. a feature to highlight targeted GPCs within the text, as described in 3.1.1.

At this point, there is a lack of evidence in the literature about how this learning feature - highlighting features within a word - aids the reading process. For children with dyslexia, with known phonological difficulties, that often struggle to identify and manipulate phonemes within words, it is possible that increasing awareness to certain phonemes would be beneficial. Research has shown that e-books that have the function to change the font size, line spacing, and background colour are beneficial to readers with dyslexia (Machan & Alexio, 2016; McCarthy & Swierenga, 2010). Word and sentence highlighting alongside audio (text-to-speech) has also been shown to increase readability for poor readers (Gerbier et al., 2018) and children with a diagnosis of dyslexia (Ikeshita, Yamaguchi, Morioka & Yamazoe, 2018). Further, in a small-scale study, Cramer and colleagues (2016) report that colour-coding spelling and reading rules helps children to learn letter-sound relations. It is conceivable that focusing attention to features within a word (making the target feature more salient) will increase recognition and speed of retrieval of the phoneme. Utilising this feature can demonstrate to the reader how to pronounce certain words. Yet this is still to be explicitly tested.

A useful starting point to explore the role of the highlighting function and pre-reading instruction, which aim to draw attention towards specific target features within words, will be to explore reading performance with and without these functions in the e-reader. This comparison approach has been used by researchers that are interested in how different fonts impact on the reading process for children with dyslexia (see Kuster, van Weerdenburg, Gompel, & Bosman, 2017; Marinus et al., 2016). An exploratory study seems the first logical step before considering running an intervention study. Moreover, the findings may provide a rationale for presets in the e-reader (i.e., whether the pre-reading instruction should be shown in combination with the highlighter feature, or not).

It may also be possible to incorporate eye-tracking into this study (with the UCL team) to determine readability of texts (i.e., frequency of fixations) when the highlight function is used versus absent.

Research question(s)

There will be three research questions for this study:

1. Does **explicit instruction** improve reading accuracy of targeted lexical items?
2. Does **textual enhancement** (highlights) improve reading accuracy of targeted lexical items?

3. How does **explicit instruction** and **textual enhancement** impact on reading fluency and text comprehension?

Research design

Two comparable studies will be conducted using the same research design: one with beginner readers and the other with children with dyslexia. The studies will employ an experimental design, taking a within-participants approach. All children will take part in each of the three conditions (reading with no alterations to the text; reading with the pre-reading instruction; and reading with the pre-reading instruction plus the highlight. The order that the condition (task) is presented in will be counterbalanced across participants.

Participants

All children will be recruited from mainstream schools. We will aim for between 40-50 children in each study.

Beginner readers will be recruited from Years 2 and 3 (UK), while children with dyslexia will be recruited from Years 4, 5 and 6 (UK). The 'dyslexia' group will include children with a diagnosis as well as those that do not have a formal diagnosis but are receiving additional support outside the classroom for literacy difficulties. The decision to include those receiving additional support was made given that not all children will receive a diagnosis in primary school in the UK – it is becoming increasingly difficult, but many children in these schools still present with significant reading difficulties.

Procedure and Data Collection methods

Both studies will be researcher-led and testing will need to take place on a 1:1 basis within the school. Two sessions will be scheduled. This first testing session is expected to take 20 minute; while the second testing session may be expected to take between 40-50 minutes, with breaks factored in.

A screening assessment will be administered, first of all to determine the child's reading level. In the UK context, the York Assessment of Reading for Comprehension (YARC) Early Reading materials will be used, which assess vowel and consonant knowledge. A measure of non-verbal ability will be conducted using the Raven's Matrices task. An informal assessment of morphology will also take place (assessment to be confirmed). Comparable assessments will be used for other languages.

The features to be focused on in the e-reader are still to be decided upon but could cover split digraphs, more complex vowel alternatives, and/or morphology. The second session will involve reading books that contain these features. Features will be carefully analysed to make sure the number of occurrences is consistent and that the words that they appear in are matched to the same lexical level. Three books per feature will be supplied and the order of presentation will be counterbalanced and split across three conditions so that the child readers a book with: (1) no E-reader functions, (2) explicit instruction only, and (3) explicit instruction plus text enhancement. The researcher will take a running record while the child reads the book aloud, recording any errors made and the time taken. Comparisons across the three conditions will be made with respect to the research questions to determine the impact of the explicit instruction and text enhancement features on reading accuracy and fluency.

If the eye-tracker can be used (still to be confirmed), eye movements would be recorded while the child reads the text on the e-reader display. Fixations and saccades would be analysed around the target features.

3.2 Game feedback study

3.2.1 Description of iRead app dimension considered

As detailed in Deliverable 6.1 the Navigo games provide both outcome feedback (feedback about performance) and elaborative feedback (feedback to improve understanding). Additionally, for those games that can be played using a trial and error strategy, when this is detected, rules have been implemented to provide the child with the correct response. This study will explore the impact of elaborative feedback for the three focal iRead learner groups.

3.2.2 The impact of elaborative game feedback on reading

In 2012, Connolly et al. (2012) conducted a systematic review of the games-based learning literature by identifying 129 papers evidencing the impact of games on learning. They reported there was evidence for "learning and behavioural outcomes including knowledge acquisition, perceptual and cognitive, behavioural, affective, motivational, physiological and social outcomes." Most of the papers they reviewed applied a quasi-experimental design comparing children playing the game with a control group. In iRead we could also take a similar comparative and confirmatory approach like this past work has done, by measuring children's improvements in reading skills before and after the game. However, as discussed in the introduction, we seek to move away from a 'black box' approach to technology to look at these improvements from a process perspective, focusing on a specific instructional dimension in each app in how it supported learning. Our view also connects with the perspective of Wouters and van Oostendorp (2013) who consider the relationship between well-designed instructional support (including particular dimensions, such as modelling, modality of explanations and feedback) and learning as a critical area for research.

Instructional design in games includes the provision of feedback, enabling connections between gameplay and initial instructional objectives, and informing the learner about their next step (Johnson et al 2017). Feedback reduces cognitive load and brings attention to part of the task; it signals a gap between performance and the learning aim; and it provides information for correcting inappropriate task strategies (Shute, 2008; Johnson et al., 2017). Johnson et al. group the learning game feedback types into *outcome feedback* (relating to task-level feedback) and *elaborative* or *explanatory feedback* (relating to all levels of feedback). Outcome feedback includes information about the response correctness, error location and performance measures (e.g. via a numerical scoring system). Elaborative feedback includes specific task/topic information, corrective strategies, why a response is (in)correct or hints/prompts. These feedback types are not mutually exclusive – e.g. a game score could be combined with guidance on how to improve that score next time, but the specific use of elaborative feedback has been shown to be very effective in learning achievement (Shute, 2008; Johnson et al 2017). There are three key characteristics to how this information is designed:

- *Content* captures the level of information designed into the feedback, how complex and specific it is (Shute, 2008; Johnson et al., 2017; Narcis, 2008). Feedback is more effective when it is specific rather than vague, and less complex and lengthy (Shute, 2008). The information embedded in the feedback can vary. It can include support to understand the task, knowledge about the concepts covered in the task, flagging up specific errors, providing strategies to process the task, and giving support in developing meta-cognitive skills (Narciss, 2008; Benton et al., 2018; Hattie and Timperley, 2007).
- *Timing* reflects when the feedback is presented. Immediate feedback follows directly after an item response and delayed feedback comes at the end of a task (Van der Kleij et al 2012; Johnson et al., 2017).

- *Modality* captures whether the feedback is verbal or visual, building on multimedia theory's proposition that people learn visually and orally (Johnson et al., 2017; Mayer and Moreno, 2003). Johnson et al (2017) draw on past research to argue that in primarily visual tasks, such as games, feedback presented orally is better processed (also Mayer and Moreno, 2003).

In evaluating game feedback types, Moreno (2004) found that novice college students learned more (in the context of botany) when provided with outcome-elaborative feedback than just outcome feedback. Mayer and Johnson (2010) replicated these results in the context of electronic circuitry with the same profile of learners. Moreno (2004) suggests elaborative feedback may reduce novice learners' cognitive load as they do not then spend time searching for a plausible explanation for their result. However the existing literature does not clearly outline how these findings would apply to early or struggling learners, who are considered novice learners in a large number of domains. We believe that young children's elaborative feedback needs careful design to reflect their current levels of cognitive development, prior knowledge, and metacognitive capabilities within the specific domain. Given the little research in young children's interaction with game feedback, we first engaged in a literature review to identify the key questions and gaps. This yielded the following questions:

RQ1: What is the impact of prior knowledge on children's ability to recover from errors within the game?

RQ2: What is the interaction between prior knowledge and feedback design?

RQ3: What do children perceive within game feedback and how does this influence their subsequent choices within the game?

3.2.3 Research design

3.2.3.1 Participants

Participants belonging to the three main iRead learner groups will be recruited: beginning readers (aged 5-7), children with dyslexia (aged 10-12), children learning English as a foreign language (aged 10-12). Children will be recruited through connections with participating iRead schools.

3.2.3.2 Procedure

The study will take place over a period of 2-4 weeks (depending on the RQs each partner will pursue). The study will be researcher-led i.e. the iRead research fellows will run the studies. Children will be taken out of the class during times arranged with the school. They will play the game in groups of 5-10, depending on the school's preference, once a week for 30 minutes. Each child will have a designated tablet, which they will use for the duration of the study. The game will be initialised taking into account the domain model pre-sets associated with the learning profile of the child.

3.2.3.3 Data Collection Methods

This will be a mixed methods study that takes a pragmatic position using methods that will answer the specific questions we want to address:

- RQs 1 and 2: Video data will be collected, namely to identify children's response correctness, the feedback type they received if they were wrong (outcome or elaborative), and the

impact of the feedback on their subsequent game response. This will be analysed statistically, comparing the child's correction of their response after the provision of feedback.

- RQ3: recordings of children's game play interactions will be made that will be used to conduct interviews with them using the recordings as a context to discuss their perception of feedback. We will examine how children interact with the feedback they are given and how it motivates their learning. E.g. do they notice it? Are they confused when they see it? Do they understand the feedback? Taking a purposeful approach to sampling, we will choose children (who participated in RQ1/2) with different literacy and language levels in order to understand the impact of feedback on different learners.

3.3 Content Classification study

3.3.1 Description of iRead app dimension considered

Text classification is a smart classification of the text based on the preselected metrics. The Text Classification (TC) component of iRead will provide users with the opportunity to select reading material that is appropriate both to the general characteristics of each student's age group, *and* to his/her individual strengths and weaknesses. The TC component is structured in two main sectors: (i) the *generic classification* subsystem, which will classify texts based on a number of pre-defined age groups, by considering quantitative (word length, sentence length, frequency etc.) and structural (i.e. syntactic complexity) linguistic criteria; (ii) the *user-driven content selection subsystem*, which will consider each students' skills relative to a number of linguistic areas that are recorded in his/her individual user profile in order to select the most appropriate material for that particular student (see Deliverable 5.1 for a detailed description of the TC component).

3.3.2 The impact of text classification on reading

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). This renders text classification a very useful tool that gives students and teachers the opportunity to make the most out of their teaching/learning experience.

However, in order for text classification to provide users with the best possible learning assistance, it needs to consider a number of linguistic criteria that compose the notion of text complexity, as well as factors that relate to each user's individual needs. Most existing text classification systems consider mainly quantitative criteria like word, sentence and text length, while linguistic complexity as it is manifested in the structural properties of phrases and sentences are usually ignored (see Deliverable 5.1 for a description of existing systems). Additionally, factors that relate to a student's particular difficulties are not generally incorporated in text classification formulas. These two areas constitute the innovative contribution of the iRead TC component. However, since this kind of factor consideration in text classification has not been tried out before, this pilot aims to evaluate the extent to which the predefined quantitative and syntactic criteria, as well as the user-driven linguistic factors lead to an accurate selection of reading material for particular users.

3.3.3 Research questions

The aim of this pilot study is to evaluate and validate the way texts are automatically 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. To achieve this end, one general research question was formed for this study:

RQ1: Does the TC component select reading material that is appropriate for each user group?

This research question will be approached by addressing the following more specific questions:
- Do students rate the texts that are selected for them by the TC component as appropriate?
- Does the TC component produce similar results for both user groups (novice readers, dyslexia)? – Are there any group-related effects (differences between user groups)?
- From those texts that are rated as not appropriate by students, how many are rated as too easy and how many as too hard?
- Which factors are responsible for the rating of texts as not appropriate (too easy or too hard)? This is actually a way of fine-tuning the classification metrics.

3.3.4 Research design

The performance of the generic text classification subsystem, as well as the user-driven metrics will be validated with text ratings collected from school-aged children.

The generic text classification system will classify texts based on the following difficulty level classification:

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

The user-driven text selection system will then use personalized information of student performance on the features of domain models to select most appropriate texts for each student. Therefore, the texts given to the students during the iRead pilot will be pre-selected based on their individual needs as considered by the overall text classification (TC) component. The final testing of the TC system will be completed by asking students to rate the texts selected for them as easy, hard or appropriate. The following sections provide a description of the methodology for the study to determine how well the automatic classification coincides with children's ratings.

3.3.4.1 Participants

Sixty (60) English-speaking school-aged children will participate in this study, 30 per reader group:

- 6-7 years old children with no learning difficulties (Group A, novice readers)
- 10-year-old students with dyslexia (Group B, dyslexia)

The groups are selected in such a way to enable the testing of all three difficulty levels defined for the text classification component, as well as for the validation of the user-driven classification

component, which will be used to select appropriate texts for students whose skills fall behind those of their typical age norm (i.e. dyslexia). The children will be students of state and independent (private) schools in the London area who have already been participating in the project previously.

3.3.4.2 Materials

The English text library compiled for the project includes close to 300 books and texts. This will form the material for the current study.

3.3.4.3 Procedure of data collection and analysis

Students will be involved in reading sessions, where they will be presented with texts selected through the TC component for them. After reading each text, students will be asked to decide how difficult they thought each text 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, so as to determine whether there the texts selected through the TC component are found appropriate (i.e. "OK") by the students. Ratings for 10 texts will be collected from each student in order to establish a statistically valid evaluation of the TC component.

After the completion of data collection, children's ratings will be analysed as following:

- the number and proportion (%) of texts that were rated as appropriate by students will be calculated,
- a quantitative analysis of the texts rated as non-appropriate (too easy or too hard) will be carried out to determine whether these are mostly found as too easy or too hard (calculation of % of non-appropriate texts that were *too easy* and those that were *too hard*), so as to determine if there is a pattern in the way the TC fails in selecting appropriate texts,
- a quantitative comparison between user groups,
- a qualitative analysis of texts that were rated as non-appropriate so as to determine the particular features that lead to their misclassification by the TC component.

In analysing students' ratings, a threshold of rating at least 80% of texts selected by the TC component as appropriate is set in order to validate the TC performance.

4 Task 3: Monitoring and evaluation

In this phase, we foresee two types of use of the iRead tools: in schools as part of the curricular teaching, and unsupervised by children at home. The first study focuses mainly on appropriation, while the second investigates dissemination opportunities outside the school context. We will perform two studies to address these types of use.

4.1 School use

Harrison et al. (2007) argue that when designing technological artefacts, the study of the local, situated practices of the users should be the focal point, since meaning is created in the context and situation, often in collaboration between the people, the artefact and the environment as well as the resources available where the artefact is used. Because of the situated nature of use, many technologies are not used in the ways designers had originally envisioned. However, according to Dix (2007) 'these improvisations and adaptations around technology are not a sign of failure, things the designer forgot, but rather show that the technology has been domesticated, that the users understand and are comfortable enough with the technology to use it in their own ways' (p. 27). This phenomenon is called appropriation. In addition to the situatedness of the technology, Dix also points to two additional advantages of appropriation:

dynamics and ownership. Over time, environments and the people in them change, and most likely so does their use of technology. A design for use must therefore be a design for change [2]. Appropriation can also create a feeling of ownership, as the users feel that they are in control or find new ways to achieve their goals.

The Model of Technology Appropriation (Carroll, 2006; Fidock & Carrol, 2006) was developed to show the appropriation process from Technology as Designed to Technology in Use (Figure 1).

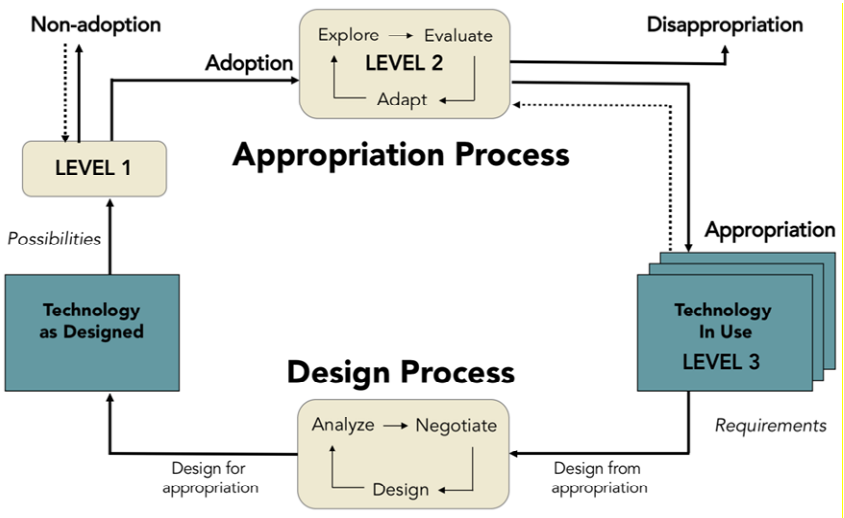

Figure 1 Model of Technology Appropriation

In this process, users evaluate the technology at three different levels:

- Level 1: At the first level they consider the technology **as originally designed**, based on its features and their own expectations of its value. At this level users may even reject the technology for a variety of reasons, such as a lack of technical skills, a lack of equipment etc.
- Level 2: At the second level, users **continuously explore and evaluate the technology as they use it and learn how it can support their practices**. As technology may enable or constrain certain activities, users change their activities to fit with the technology. In addition, depending on the malleability of the technology, users **try to adapt it to fit with current needs, or for completely new purposes**. During this stage, there are a number of influences that encourage or discourage continued appropriation. Three elements are needed to understand appropriation (Bruce et al. 2009):
    - The idealisation serves as an indication for the intentions of the developers as they often are important participants throughout the creation of the technology. Further, this element is used to identify how the technology is perceived or anticipated by the teachers.
    - The settings form an important part in how a technology emerges and considers aspects about the social context of use, i.e. cultural, institutional, and pedagogical aspects. This includes the goals and expectations of the participants, the institutional practice, constraints, and resources.
    - The realisation of the technology in each distinct context aims to study how the innovation is used, how the use changes and the reason for these changes.
- Level 3: If users do not reject the technology at level 2, the third level may be reached over time. At this level, the **technology has become an integral part of the users' practice**, and no further adaptations take place. However, this may change if the users' evaluation of the technology changes, in which case they return to the previous level where the technology may once again be rejected.

4.1.1 Research question(s)

The following main research question with underlying sub questions will be answered.

RQ1: What are the facilitative mechanisms to appropriation and how does this foster scaling?

- How does CPD and other modes of support within and across schools scaffold teachers in the process of appropriation?
- What challenges remain that cannot be addressed through these mechanisms?
- How does it inform our understanding of scaling within and across schools?

RQ2: How does the nature of appropriation develop over time?

- Is appropriation aligned with the conceptual underpinnings of technology (match between the idealisation and the realisation)?
- What contextual factors (settings) shape appropriation, for what ends, and who benefits?

4.1.2 Research design

With the goal of investigating how the iRead technologies may be scaled up, the aim of the appropriation study is to investigate how the technologies are being implemented and appropriated in the participating schools. The study takes the teachers' perspective, and focuses on how they choose to include the technology in their classrooms. The results will offer opportunities to un-pick the context of the appropriation, and shed light on how the technology scaled up, why it has become successful in some schools and why not in others. The appropriation study aims to investigate the main research question with underlying sub questions (cf. D9.1_evaluation plan).

All products and services of iRead (e-Reader, Navigo game, learning analytics and teacher tools), as well as the CPD were made available to the teachers. In this qualitative study, all teachers who were willing to use the iRead products and services in their teaching were first offered CPD, with the researchers helping them to set up their teaching in a way that fits their context. Once the teachers felt confident to start working more independently, the researchers stepped down, but continuously offer regular support to tackle problems to keep the teachers going.

4.1.2.1 Participants

All partners have found teachers who are willing to participate with their classes. The maximum number of teachers and children each partner was able to recruit for the study, was dependent on the number of tablets available and whether children can share tablets, and meant that there are several classes in the same school that can use the tablets at different times. Even teachers who were not willing to participate with their class were seen as participants, because we are interested in understanding their reasons for non-adoption (or lack of appropriation).

4.1.2.2 Data Collection

The appropriation study will run during the period November 2019 - June 2020, with a minimum period of 10 weeks per school. Given that each participant has a fixed number of tablets, it could

be necessary to restrict the time that each school use the tablets to be able to reach the number of students required for each participant. Participants will be asked to provide a justification for the length they decide on for the study.

An overview of the methods to be used in the appropriation study includes the following:

- **Descriptions of the schools** to provide contextual information.
- Semi-structured **interviews** with each teacher will be conducted using an interview guide (suggested duration 45 minutes/interview). The interviews must be *audio recorded* with the teachers' *informed consent* to allow for *transcription*.
- **Questionnaires**, e.g. using Google Forms, will be used to *supplement interviews* with the teachers.
- **Classroom observations** comprising *a case study* which will include (at least) *two contrasting schools* which differ from one another (e.g. differing in the amount of exposure to the Amigo Reader and the Navigo games, or how many of the teachers in the school use the apps in their teaching, for example All or almost all teachers, Half of the teachers, One teacher)
- **Log data** to *supplement the qualitative data* (logs and analyses will be provided by NTUA)

Depending on the resources available and duration planned for the study, participants may choose a *minimal study set-up* approach or a more *advanced study set-up* to conduct the appropriation study, described below.

4.1.2.3 Minimal study set-up

Minimal requirements for the appropriation study are that it is conducted during a period of 10 weeks minimum in each school. The following tools for data collection are included:

1. **Descriptions of the school** to provide contextual information.
2. **Interviews** with **each teacher** on **three occasions:**
    a. Prior to start using the apps in class - to understand expectations and settings. A suggestion is to conduct this interview in connection to the CPD phase.
    b. Halfway through the study - to understand and facilitate further use.
    c. At the end of the study (or when the teacher decides to stop using the apps, if this is before the end of the planned time period for the study) - to let the teachers reflect upon how they have appropriated the iRead technology
3. **Questionnaires** in between interviews, on two or more occasions (depending on length of each study):
    a. Questionnaire with each teacher **a quarter of the way through the study**
    b. Questionnaire with each teacher **three quarters of the way**

    The reason for using online questionnaires is to provide regularity of feedback on the teachers' progress of using the iRead apps in teaching, as well as to maintain contact between the researchers and the teachers. In this way, the questionnaires (including no more than a few questions each time) will supplement information gathered through the interviews. If the appropriation study is conducted during a period longer than the

minimum of 10 weeks, a recommendation is that the online questionnaire should be sent to each teacher every 2 months.

For instance, Google Forms may be used for the questionnaires, and the same questions will be used for each stage. The questionnaires may be distributed to the teachers through a means which suits each pilot lead best, e.g. via email notification, and it is open to each lead to decide which means to use for this.

4. **Log data will be** gathered throughout the study by NTUA.

Support visits and contact with the teachers must be provided ongoing by each participant, if needed.

Optionally, if the participating teachers agree, **Facebook or WhatsApp groups** could be set up "locally" for each participant (e.g. a iRead EFL Facebook group for the participating EFL teachers, a iRead Facebook group for the beginning readers, one for the dyslexia, etc.). The local groups on Facebook would mainly serve the following functions:

1. Easy access for the participating teachers to the (local) researchers (for technical support, questions and comments, etc.)
2. Forum for exchange between participating teachers and schools. Data will not be gathered *per se* from the groups.

4.1.2.4 Advanced study set-up

The advanced study set-up includes the same requirements as the minimal study set-up, with the addition of **classroom observations**. For the advanced set-up, at least *two contrasting schools* should be selected in which to make classroom observations to base the case study on. By January 2020, participants adopting the advanced study set-up must decide which schools to include in their case study. Observations in classrooms during the study will be conducted when the Amigo and Navigo apps are being used in teaching, with a minimum of two classroom observations per school on two separate occasions. Make sure that the observations are evenly spread out during the study.

4.1.2.5 Documentation

A timeline, protocols and guides to document the following have been provided to the researchers:

1. **Protocol for descriptions of the schools**
2. **Interview guides for the semi-structured teacher interviews**
    a. in connection to the CPD phase
    b. halfway through the study
    c. after the end of the study

   Consent forms are available under WP9 in Dropbox.
3. **Questionnaires (using Google Forms, for example)**
    a. a quarter of the way through the study
    b. three quarters of the way, a few weeks
4. **Protocol for observations in class**

The methods for data collection presented above, are intended to study the different levels of appropriation suggested in The Model of Technology Appropriation (Carroll, 2006; Fidock & Carrol, 2006) (Figure 1, above):

- Level 1: may not be possible since schools are already involved. Could be studied in case one needs to recruit more schools and/or teachers.
- Level 2:
    - The idealisation: our description of the use of iRead. The manuals we develop for teachers form the basis.
    - The settings: Interviews with teachers prior to use of (several teachers were already interviewed for WP3).
    - The realisation: Interviews, observations and questionnaires during the use of iRead
- Level 3: Interviews with teachers after the end of the evaluation. We may not reach this level for most teachers, but some may continue using iRead after the end of the study.

4.2 Unsupervised use

In the previous section, the study incorporating all components of iRead in an integrated way in schools was described. However, the iRead games can potentially also be used outside school by independent learners. To gain insights in the games' usage in such an unsupervised manner, we aim to perform one large-scale study for the Greek language model for novice readers.

4.2.1 Description of iRead app dimension considered

The game with the underlying language models for Greek novice readers is incorporated. If the user participation target is not reached for the Greek model, the English novice reader model will also be incorporated.

4.2.2 Research question(s)

There are four research questions for this study:

- How do the learners progress through the different features of the employed language model?
- What is the efficacy of the specific learning activities and game mechanisms?
- What is the degree of user engagement?

4.2.3 Participants

We aim to recruit 1000 online novice readers of Greek. The pilot will be advertised through the local teacher networks and by employing social media. Employment of specialized internet advertisements will be considered.

4.2.4 Data collection and analysis

We will gather the following data:

1. demographic data collected during the readers' registration
2. game usage data automatically logged by the Navigo game. The game usage data contains information about the literacy activities as well as the educational content presented to the reader and the readers' performance in these activities
3. A questionnaire to determine children's reading motivation, such as MRQ questionnaire (Wigfield & Guthrie, 1997) may also be employed. However, due to the unsupervised nature of the pilot, a high response ratio cannot be guaranteed.

Using the data gathered above, an exploratory analysis will be performed, where the data will be converted to appropriate format. Graphical data representations will be utilized for the presentation of results.

4.2.4.1 Procedure

The study will be advertised, as described above, and interested users will be asked to create an account to download and install the game. Then, the readers will use the game in an unsupervised manner for a longer period (September 2019-August 2020). By “unsupervised’, we mean that the study takes place outside the school environment, that is, the literacy activities (i.e., the language features and the relevant literacy content) presented to the readers are not predetermined based on a school curriculum or a specific, teacher generated and administered teaching plan, but rather they are determined in an automatic way by employing iRead’s adaptation/personalization capabilities. In this way, the reader’s course through the game depends on the language model employed and the reader’s performance on previously presented literacy activities.

5 Conclusion

This deliverable describes the evaluation plan for iRead, including the methodology and most of the measures used. Since this plan holds for studies to be performed under a long period of time, with some studies being performed more towards the end of the project, this document will have to be updated regularly.

6 References

Bruce, B.C., A. Rubin, and J. An, Situated Evaluation of Socio-Technical Systems, in Handbook of Research on Socio-Technical Design and Social Networking Systems, B. Whitworth and A.d. Moor, Editors. 2009, Information Science Reference: Hershey, PA. p. 685-698.

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

Carroll, J. (2004) Completing Design in Use: Closing the Appropriation Cycle, in ACIS. 2004. p. 337-347.

Carvalho, M.B., Bellotti, F., Berta, R., De Gloria, A., Islas Sedano,C., Baalsrud Hauge, J., Hu, J. and Rauterberg, M. 2015. An Activity Theory-Based Model for Serious Games Analysis and Conceptual Design. Computers & education 87, 166-181. http://dx.doi.org/https://doi.org/10.1016/j.compedu.2015.03.023

Connolly, T.M,. Boyle, E. A., MacArthur, E., Hainey, T., Boyle, J.M. (2012) A systematic literature review of empirical evidence on computer games and serious games, Computers & Education, Volume 59, Issue 2, Pages 661-686, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2012.03.004.

Corder, S. (1967). The significance of learners' errors. International Review of Applied linguistics, 5, 161–170.

Cramer, E. S., Antle, A. N., & Fan, M. (2016). The Code of Many Colours: Evaluating the Effects of a Dynamic Colour-Coding Scheme on Children's Spelling in a Tangible Software System. Paper presented at the Proceedings of the 15th International Conference on Interaction Design and Children.

Dix, A. (2007) Designing for appropriation, in Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it. 2007, BCS Learning & Development Ltd. p. 27-30.

Fidock, J., Carroll, J. (2006) The model of technology appropriation: A lens for understanding systems integration in a Defence context, in ACIS 2006. 2006. p. 88.

Gerbier, E., Bailly, G., & Bosse, M. L. (2018). Audio–visual synchronization in reading while listening to texts: Effects on visual behavior and verbal learning. Computer Speech & Language, 47, 74-92.

Geva, E., & Ramírez, G. (2015). Focus on reading. Oxford: Oxford University Press.
Goo, J., Granena, G., Yilmaz, Y., & Novella, M. (2015). Implicit and explicit instruction in L2 learning: Norris & Ortega (2000) revisited and updated. In Rebuschat, P. (Ed.), Implicit and explicit learning of languages (pp. 443–482). Amsterdam: John Benjamins.

Han, Z., Park, E. S., & Combs, C. (2008). Textual enhancement of input: Issues and possibilities. Applied Linguistics, 29, 597–618.

Harrison, S., D. Tatar, and P. Sengers. (2007) The three paradigms of HCI. in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07). 2007. ACM Press.

Ikeshita, H., Yamaguchi, S., Morioka, T., & Yamazoe, T. (2018). Effects of highlighting text on reading ability of children with developmental dyslexia: a pilot study. International Journal of Emerging Technologies in Learning, 13(9), 239-251.

Johnson, C. I. Bailey, S.K.T and Van Buskirk, W. L. (2017). Designing Effective Feedback Messages in Serious Games and Simulations: A Research Review. In Instructional Techniques to Facilitate Learning and Motivation of Serious Games Pieter Wouters and Herre Van Oostendorp (eds.). Springer, 119-140.

Kuster, S. M., van Weerdenburg, M., Gompel, M., & Bosman, A. M. T. (2018). Dyslexie font does not benefit reading in children with or without dyslexia. Annals of Dyslexia, 68, 25-42.

Laufer, B., & Ravenhorst-Kalovski, G.C. (2010). Lexical threshold revisited: Lexical text coverage, learner's vocabulary size and reading comprehension. Reading in a Foreign Language, 22, 15–30.

Lee, S. K., & Huang, H. T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. Studies in Second Language Acquisition, 30, 307–331.

Leow, R. P. (1997). The effects of input enhancement and text length on adult L2 readers' comprehension and intake in second language acquisition. Applied Language Learning, 8, 151–182.

Leow, R. P. (1999). Attention, awareness, and focus on form research: A critical overview. In J. F. Lee & A. Valdman (Eds.), Form and meaning: Multiple perspectives (pp. 69–96). Boston: Heinle & Heinle.

Leow. R. P. (2001). Attention, awareness, and foreign language behavior. Language Learning, 51, 113–155.

Leow, R. P., & Martin, A. (2017). Enhancing the input to promote salience of the L2: A critical overview. In Gass, S. M., Spinner, P., & Beheney, J. (Eds.), Salience in second language acquisition (pp. 167–186). New York, NY: Routledge.

Machan, L., & Aleixo, P. (2016). E-readers as an alternative to coloured overlays for developmental dyslexia in adolescents. The Psychology of Education Review, 40(2), 33-38.

Marinus, E., Mostard, M., Segers, E., Schubert, T. M., Madelaine, A., & Wheldall, K. (2016). A special font for people with dyslexia: does it work and, if so, why? Dyslexia, 22, 233-244.

McCarthy, J. E., & Swierenga, S. J. (2010). What we know about dyslexia and Web accessibility: A research review. Universal Access in the Information Society, 9, 147-152.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. Instructional science, 32, 1: 99-113.

Mäkelä, S., Vellonen, V. (2018) Designing for appropriation: A DIY kit as an educator's tool in special education schools, International Journal of Human-Computer Studies, 118, 14-23, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2018.05.004.

Norris, J., & Ortega, L. (2000). Efficacy of L2 instruction: A research synthesis and quantitative meta-analysis. Language Learning, 50, 417-528.
Robinson, P. (1995). Attention, memory, and the “noticing” hypothesis. Language Learning, 45, 283–331.

Patricio, R., Miguel, N., Ximena, L., & Marcos, S. (2010). A Monitoring and Evaluation Scheme for an ICT-Supported Education Program in Schools. Journal of Educational Technology & Society, 13(2), 166-179.

Robinson, P. (2003). Attention and memory during SLA. In C. Doughty & M. H. Long (Eds.), The handbook of second language acquisition (pp. 631–678). Oxford: Blackwell.
Schmidt, R. (1990). The role of consciousness in second language learning. Applied Linguistics, 11, 129–158.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), Cognition and second language instruction (pp. 1–32). Cambridge: Cambridge University Press.Sharwood Smith, M. (1991). Speaking to many minds: On the relevance of different types of language information for the L2 learners. Second Language Research, 7, 118–132.

Sharwood Smith, M. (1993). Input enhancement in instructed SLA. Studies in Second Language Acquisition, 15, 165–179.

Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78, 1: 153-189.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. Language Learning, 60, 263-308.

Spencer, M., Muse, A., Wagner, R. K., Foorman, B., Petscher, Y., Schatschneider, C., Tighe, E. L., Bishop, M. D. (2015). Examining the Underlying Dimensions of Morphological Awareness and Vocabulary Knowledge. Reading and Writing, 28, 959-988.

Wigfield, A., & Guthrie, J. T. (1997). Relations of children's motivation for reading to the amount and breadth or their reading. Journal of Educational Psychology, 89(3), 420-432. doi:http://dx.doi.org/10.1037/0022-0663.89.3.420

Wouters. P. & Van Oostendorp, H.. 2013. A Meta-Analytic Review of the Role of Instructional Support in Game-Based Learning. Computers & Education 60, 1, 412-425. http://dx.doi.org/https://doi.org/10.1016/j.compedu.2012.07.018