White paper on Linguistic Infrastructure

White paper on Linguistic Infrastructure

As part of iRead's 4 year project to develop personalised learning technologies to support reading skills, we have developed a **Linguistic Infrastructure.** The Linguistic Infrastructure is composed of **Dictionaries** and **Language Models.** We want to share the Linguistic Infrastructure with literacy edtech developers to **a) help improve the quality of these technologies** and **b) save companies time and money**. Ultimately, as a consortium of research institutions and industry, the aim of the iRead project is to help improve literacy rates around the world.

Dictionaries provide linguistic information about words including:

- Phonics features (i.e. individual sounds, digraphs, sight words etc)
- Chunking and syllabification
- Grammar (i.e. parts of speech, negations, modal verbs, suffixes, etc)

Language Models capture the journey of a student learning to read and contain language features that need to be mastered. The journey of a 6 year old British child learning to read English is different from the journey of an 8 year old Spanish child learning to read English, and our Language Models reflect these differences. They contain:

- Language features
- Difficulty (compared to other features in the same category)
- Prerequisites that should be mastered before introducing a new language feature

1. Dictionaries

Dictionaries are provided for the English, Greek, German and Spanish languages. The English dictionary contains 12317 lemmas, the Greek dictionary contains 8419 lemmas, the German dictionary contains 12203 and the Spanish dictionary contains 23134 lemmas. Dictionaries can be accessed as a JSON file or through an online dictionary visualiser.

For each language, more than one dictionaries may be available. The dictionaries may differ in the number of words contained in each of them, as well as in the detailed material provided for each word. In the iRead project, two dictionaries are provided for each language, namely a game-dictionary (containing words selected and curated in order to support the literacy games in iRead) and a reader-dictionary (covering a larger set of words). The reader-dictionaries are an extension of the game-dictionaries, that is, all lemmas found in a game-dictionary are also included in the corresponding reader-dictionary. Apart from size, there are the following differences:

Lemmas in the game-dictionary have audio-files, while no audio-files exist for the additional lemmas of the reader-dictionaries. The game-dictionaries have been checked for correctness through an extensive Quality Assurance process.

A dictionary lemma contains linguistic information that is included in any dictionary. **What makes iRead dictionaries special is the information associating each lemma with features in the corresponding language Language Model**. For example, the word apple, is associated with three linguistic features: æ-a, p-pp, l-le.

The lemma attributes contained in iRead's dictionaries are listed below. Attribute "Feature occurrences" links each lemma to a set of related features and also provides information about the position in which the corresponding features appear in a word.

Attribute	Definition	Comment
Id	An integer identifier	
Name	The word	
Phonetic	The phonetic transcription of the word in IPA	
Part of speech	The part of speech of the word (e.g. verb, noun, article, etc)	
syllables	Syllabification of the word	
Grapheme Phoneme Correspondence (GPC)	Pairs of graphemes (letters) to phonemes	Not generated for abbreviations
cv-form	Transcription of the word's syllables in consonant-vowel form	Sound based for Spanish, letter based for other languages
Number of characters	The character-length of the word	
Number of phonemes	The number of GPC pairs	May differ for Spanish
Number of syllables	The number of syllables	
Word frequency in child language	Frequency of the word in child appropriate corpus of texts	
Audio file	The name of the audio file for the word	
Prefix	The prefix of the word	
Prefix type	Indicates if adding the prefix alters the root-word (e.g. add, visual, drop, etc)	Not present in Spanish and German
Suffix	The suffix of the word	In Greek suffix is always the inflectional suffix
Suffix type	Indicates if adding the suffix alters the root-word (e.g. add, visual, drop, etc)	Not present in Spanish and German
Stem	The dictionary lemma corresponding to the root word	Not defined if the root word is not in the dictionary
Feature occurrences	Information about the features the word contains. For each feature's occurrence, the starting and ending position in the word are provided	

Word type info	Additional information about the word (inflectional information etc)	Depends on each language

2. Language Models

The following Language Models are available. Each Language Model has a different set of attributes depending on the user characteristics:

- English (novice reader) – primary years 1-3 [279 features, lemmas per feature vary from a minimum of 3 to a maximum of 4655]
- Greek (novice reader) – primary years 1-3 [466 features, lemmas per feature vary from a minimum of 1 to a maximum of 5703]
- German (novice reader) – primary years 1-3 [316 features, lemmas per feature vary from a minimum of 1 to a maximum of 7594]
- Spanish (novice reader) – primary years 1-3 [326 features, lemmas per feature vary from a minimum of 1 to a maximum of 19907]
- English as a foreign language (Mother language: Greek) [279 features, lemmas per feature vary from a minimum of 3 to a maximum of 4655]
- English as a foreign language (Mother language: Spanish) [279 features, lemmas per feature vary from a minimum of 3 to a maximum of 4655]
- English as a foreign language (Mother language: Romanian) [279 features, lemmas per feature vary from a minimum of 3 to a maximum of 4655]
- English as a foreign language (Mother language: Swedish) [279 features, lemmas per feature vary from a minimum of 3 to a maximum of 4655]

Structure

A Linguistic Language Model contains information about language acquisition through different type of information (lexical, morphosyntactic etc.). In order to address complex reading skills for the languages under investigation, the developed Language Models incorporate the following linguistic levels: **phonology**, **morphology, morphosyntax** and **syntax**.

- The term **linguistic level** refers to the level of analysis each language structure belongs to. For example, phonemes and syllables refer to the phonological level of analysis.
- Each linguistic level is represented by a number of phenomena or structures, called **language categories**, like phonemes and syllabification, prefixes or suffixes, embedding or passive voice.
- Each linguistic level includes a set of specific instances, the features. **Language features** refer to the specific instances of each category included in Language Models. For instance, the category phonemes of the Greek Language Model includes the Phonemes /p/, /b/, /d/, /g/, /s/ and so on in word-initial and word-internal position, while the category of Syllabification includes different syllable structure combinations like CV-CV, CVC-CV, VC-CV, etc.

The following table provides an illustration of how the information stored in the English Language Models is structured.

Linguistic Level	Language Category	Example feature
Decoding words (phonology)	Phonics	‘s’ as in sad
	Blends	'sl' as in 'slap'
Chunking words (phonology)	Syllables	Chunking 2-syllable words
Recognising words (word recognition)	Common sight words	Frequent words 1 - Reception ('a', 'and', etc.)
	Confusing letters	‘d’ and ‘b’
Using prefixes and suffixes (morphology)	Prefixes	'mono', 'multi', etc.
	Suffixes	Quickly, loudly
Using suffixes and grammar	Adverbs	More slowly
	Verb tenses	's' as in 'he plays'
	Plurals	es' as in 'the buses'
	Comparative adjectives	'-er' as in 'warmer'
	Superlative adjectives	'the fastest'
Syntax	Adjectives	nice' as in 'a nice dress'
	Pronouns	each other' as in 'they like each other'
	Determiners	this' as in 'this chai over there'
	Proper nouns	Sam' as in 'Sam is back'
	Noun with no determiner	cats' as in 'Cats like fish'
	Articles	the' as in 'the cat'
	Passives	The cat was chased'
	Complex sentences: coordinations	Coordination in 'or', 'and', 'but', ...
	Complex sentences: subordination	after me' as in 'he came in after me'
	Negations	not' as in 'I do not/don't know him'
	Prepositions	down', 'in', 'in front of', 'on', 'above', ...
	Determiners	A few', 'few', 'fewer', 'fewest'
	Adverbs	apparently' as in 'They have apparently arrived'
	Wh- questions	What is the man chasing?'
	Yes/no questions	Do you like milk?'
	Modal verbs	can', 'may', 'might', 'could'

Attributes

The Language Models provide (a) ratings of difficulty, indicating each feature’s difficulty or complexity relative to the rest of the features within the same category, as well as (b) progression schemes, in the form of prerequisites, indicating the order of teaching of those features. Specifically, rich texts in words/linguistic structures are likely to cause a barrier in reading. Therefore, all language features included in the models are classified as relatively easier or harder based on an initial taxonomy of features per category formulated by empirical data. The table below displays the attributes for each linguistic feature in a Language Model.

Attribute	Description
Id	An integer identifier; consecutive and starting from one (1)
Linguistic level	The linguistic level of the feature (e.g.: Phonology, Morphology, Morphosyntax, Syntax)
Category	The (linguistic) category of the feature (e.g. GPC and Clusters under Phonology, Prefixes and Derivational Suffixes under Morphology)
Feature type (optional)	The (linguistic) type of the feature (e.g. Consonant and Vowel under GPC-Phonology)
Description	A short description of the feature
Human readable linguistic level	The corresponding attributes in a 'human readable' fashion, for UI usage and in the corresponding language
Human readable category	
Human readable feature type	
Human readable description	
Examples (optional)	Examples of words or phrases that contain the feature
Exceptions (optional)	Examples of words or phrases that are exceptions of the feature's definition
Difficulty (across category)	An integer value, denoting the difficulty of a feature compared to other features of the same category
Prerequisites	List of features that should be mastered before working the feature
Frequency	The frequency of the feature measured in child-appropriate corpus of texts
Group (optional)	Names of groups that contain the feature
Word level complexity (optional)	Indicates the suggested 'complexity' of words that a user could practice for the feature
Initial competence (optional)	Indicates the percentage of initial mastery for the feature; used for EFL models, where some features are highly relevant to the first language and are typically mastered before second language acquisition

3. Research Basis and Methodology

A number of steps were taken to create the Language Models and dictionaries:

- Reviewed literature in first and second language reading and atypical reading development to identify language features, difficulty levels and progression
- Constructed a corpus of 551 children's texts and books to conduct a frequency analysis. This informed further choices on which language features to include in the final Language Models

- The corpus of children's texts and books was used to construct two dictionaries of words – game-dictionary and reader-dictionary. These words were automatically tagged. Within the game-dictionary they were subsequently assessed by human raters to ensure accuracy.
- For English as a second language reading, we worked with academic experts in Swedish, Romanian, Greek and Spanish to identify which features had been mastered in children's first language to narrow the focus of the Language Model

In what follows, we describe in detail the rationale of the language features and the process we followed to create the Language Models.

Models of reading are useful to identify the processes involved in skilled reading and this, in turn, can help us to understand why children may struggle to learn to read. The Simple View of Reading (SVoR; Hoover & Gough, 1990) highlights that learning to read requires both word recognition (decoding) and language comprehension skills.

In the early years of schooling, emphasis is placed on teaching phonics and developing decoding skills (i.e. phonological awareness: the ability to segment and blend sounds within words) so that children can access texts and subsequently learn to comprehend the books/material that they read. Research supports that children develop their reading skills by first learning grapheme-phoneme correspondences (GPCs; i.e. letter to sounds; Ehri, 2005). In addition, children are typically taught to identify syllables to develop their phonological awareness skills (Hatcher, Duff & Hulme, 2014). Syllabification can be an important strategy for beginner readers because it enables children to work with larger units than phonemes when reading, which may in turn increase fluency.

With instruction and practice, children should transition from decoding each phoneme (sound) within a word to whole word recognition ('sight words') - thus, a key aim in the reading process is to develop fluency, which further supports comprehension (Kim, Park, & Wagner, 2014). Moreover, whole word recognition is particularly important for children learning to read in English given that a number of words cannot be decoded phonetically (often labelled as 'irregular' or 'exception' words). It is considered important that children become familiar with high frequency words early on in the school system (Ehri, 2005). Therefore, developing a sight vocabulary alongside developing phonological recoding processes are both essential to becoming a skilled reader. In support of this, Share's (1995) self-teaching hypothesis argues that once readers have gained letter knowledge and adequate decoding skills, they are well-equipped to become more independent readers and to use prior knowledge and inference to support the development of fluency when reading.

Based on the research literature and to tap into the development of children's word recognition ability, the iRead Language Model has an explicit focus on supporting: decoding (using phonics knowledge to read vowels, consonants, digraphs, trigraphs and blends), chunking (syllabification), and recognising whole words (common high frequency words and recognising confusing letters). The Language Model follows a synthetic and analytic phonics rationale that recognises the smaller sound

representation in words and also the bigger letter units in teaching children how to decode words.

In addition to focusing on decoding abilities, other higher-level language skills that map onto reading comprehension were considered when devising the Language Model. Specifically, morphological awareness and syntactic processing.

Morphologically derived words (e.g., help - helpful) make up 40% of unfamiliar words that children encounter in text in their late school years (Nagy & Anderson, 1984; Nagy et al., 1993). Research with typically developing readers has found that inflected words (e.g., help - helped) are easier to learn than derived ones (Carlisle, 1995). This may be because the morphological changes of derived words are less predictable and reliable, compared to inflected words. Within the Language Model, thus, we included several derivational prefixes and suffixes to enhance children's learning.

Children with dyslexia have been known to present with difficulties with productive and receptive (morpho-) syntactic skills (Scarborough, 1990, 1991; Lyytinen et al., 2001). Moreover, primary-aged poor comprehenders have been shown to have difficulties with past tense formation (see Nation, Snowling, and Clarke, 2005; Joanisse et al., 2000), and verb agreement marking (Casalis et al., 2012; Cantiani et al., 2013; Joanisse, Manis, Keating, & Seidenberg, 2000; Rispens & Been, 2007). The importance of such skills are reflected in the curriculum, contribute to understanding word meaning, and aid text comprehension.

At the morphological level, the Language Model covers: derivational prefixes and suffixes, and inflectional suffixes (e.g. past tense, plurals). Within the syntactic level we included categories that relate to morpho-syntax supporting grammar (proper nouns, articles, prepositions, negative particles, embedded constructions, passives, complex sentences and so on). When selecting morphosyntactic categories for inclusion in the Language Model, we consulted published literature on what morphosyntactic features may influence reading comprehension. We also focused on the most frequent morphosyntactic features identified in a corpus analysis based on first language children's texts.

When it comes to the EFL Language Models, we requested experts to select those features from the overall EFL Language Model which were anticipated to cause reading difficulty considering the existing linguistics and literacy skills of a particular L1 group. Our rationale for this was that, by the time second language (L2) readers learn to read, they often have developed some literacy skills in their first language (L1). It has been shown that, once L2 readers have reached a certain level of proficiency, they can successfully use these existing L1 strategies and skills when processing L2 texts (Grabe, 2009). Importantly, while L1 reading strategies may indeed assist beginning L2 readers to cope with certain tasks, they may have adverse effects in other situations, leading to slower reading speed and issues with comprehension (Koda, 2007).
As shown above, a wide range of linguistic structures are assessed within Language Models across each language, going beyond the skills of decoding of sound-letter correspondence and word recognition to more complex reading skills, including morphological, syntactic and discourse processing of text.

White paper on Linguistic Infrastructure

In order to reliably select the appropriate linguistic phenomena for the new models, we first used evidence from published studies for the populations we are interested in the literature. We conducted extensive literature reviews not only to define the relevant language areas for reading development, but also to define which language areas are harder to develop while learning to read or while acquiring language in general.