iRead
Linguistic infrastructure

Asimina Vasalou, Laura Benton, Elisabeth Herbert, Nelly Joye, Andrea Revesz, Emma Sumner
UCL Institute of Education and Antonis Symvonis, Chrysanthi Raftopoulou, Dionysis Panagiotopoulos, National Technical University of Athens.

As part of Horizon 2020 iRead's 4 year project to develop personalised learning technologies to support reading skills, we have developed a Linguistic Infrastructure composed of Language Models and Dictionaries.

Language Models

Linguistic Language Models contain information about language acquisition through different types of information. To address complex reading skills for the languages under investigation, the developed Language Models incorporate the following linguistic levels: **phonology, morphology, morphosyntax** and **syntax.** An English language model was developed for primary years 1-3, older struggling readers (primary years 4-6) and children learning English as a Foreign Language (279 features, lemmas per feature vary from a minimum of 3 to a maximum of 4655).

'**Linguistic level**' refers to the level of analysis each language structure belongs to. e.g., phonemes and syllables refer to the phonological level of analysis.

Each linguistic level is represented by a number of **language categories**, e.g., phonemes and syllabification/ prefixes or suffixes/embedding or passive voice.

Each linguistic level includes a set of specific features. **Language features** refer to the specific instances of each category included in Language Models.

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

linguistic Level	language Category	example feature
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'

Attributes
The Language Models provide (a) ratings of difficulty, indicating each feature's complexity relative to other features within the same category, (b) progression schemes, in the form of prerequisites, indicating the order of teaching of those features.

Dictionary

The English dictionary contains 12317 lemmas. A dictionary lemma contains linguistic information included in any dictionary. **What makes the dictionaries special is the information associating each lemma with features in the corresponding Language Model**. For example, the word apple, is associated with three linguistic features: æ-a, p-pp, l-le. The lemma attributes contained in the dictionary are listed below. Attribute "feature occurrences" links each lemma to a set of related features and provide information about the position in which the corresponding features appear in a word.

attribute	identification
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	transcription of the word's syllables in consonant-vowel form
grapheme phoneme correspondence (GPC)	graphemes (letters) to phonemes
cv-form	transcription of the word's syllables in consonant-vowel form
number of characters	the character-length of the word
number of phonemes	the number of GPC pairs
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)
suffix	the suffix of the word
suffix type	indicates if adding the suffix alters the root-word (e.g., add, visual, drop, etc)
stem	the dictionary lemma corresponding to the root word
feature occurrences	information about the features the word contains. For each feature's occurrence, the starting and ending position in the word is provided
word type info	additional information about the word (inflectional information etc)

Several steps were taken to create the Language Models and dictionary:

- A review of the literature in first and second language reading and atypical reading development to identify language features as well as inform difficulty levels and progression.
- Construction of 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 the dictionary. Words were automatically tagged. They were subsequently assessed by humanraters to ensure accuracy and suitability for children.

Click here to read how the linguistic infrastructure has been used in the Navigo literacy game - iread-project.eu/2020/11/24/using-the-linguistic-infrastructure-to-develop-the-navigo-literacy-game

Horizon 2020 European Union Funding for Research & Innovation

A 4-year EC Funded project (Jan 2017-Jun 2021).This project has received funding from the European Union's Horizon 2020 innovation programme under grant agreement No 731724.