# The AEREST Reading Database

Marcello Ferro Istituto di Linguistica Computazionale ILC-CNR Pisa, Italy marcello.ferro@ilc.cnr.it Sara Giulivi Scuola Professionale della Svizzera Italiana SUPSI Locarno, Switzerland sara.giulivi@supsi.ch

Claudia Cappa Istituto di Fisiologia Clinica IFC-CNR Pisa, Italy claudia.cappa@cnr.it

#### Abstract

Aerest is a reading assessment protocol for the concurrent evaluation of a child's decoding and comprehension skills. Reading data complying with the Aerest protocol were automatically collected and structured with the *ReadLet* web-based platform in a pilot study, to form the Aerest Reading Database. The content, structure and potential of the database are described here, together with the main directions of current and future developments.

Aerest è un protocollo di valutazione della lettura che misura in parallelo la capacità di decodifica e quella di comprensione del testo. Il protocollo è stato applicato in uno studio pilota i cui dati sono stati raccolti attraverso la piattaforma web ReadLet. L'articolo descrive il contenuto, la strutture e le potenzialità del data set risultante, insieme a future direzioni di sviluppo.

#### 1 Introduction

In the PISA 2000 report (OECD, 2003), a distinction is introduced between the concept of "reading literacy" as opposed to "reading", the latter being restricted to the ability of decoding or reading aloud, the former including a much wider and more complex range of cognitive and meta-cognitive competencies: decoding, vocabulary, grammar, mastery of larger linguistic and textual structures and features, knowledge about the world, but also use of appropriate strategies necessary to process a text (p. 23). In the PISA 2019 report (OECD, 2019) "reading literacy" is defined as "an individual's capacity to understand, use, evaluate, reflect on and engage with texts in order to achieve one's goals, develop one's knowledge and potential, and participate in society", and as the "range of cognitive and linguistic competencies, from basic decoding to knowledge of words, grammar and the larger linguistic and textual structures needed for comprehension, as well as integration of meaning with one's knowledge about the world" (p.28). Achieving reading literacy is crucial for an individuals' participation in society and ultimately for their realization in academic context, in workplace or, more generally, in life.

To achieve reading literacy, pupils need first and foremost to be able to read accurately, understand what they read, and do this in a reasonably small amount of time. This multifaceted ability is defined here as "reading efficiency". Efficient reading implies on its turn, in the subject, the development of deep comprehension skills. As a matter of fact, comprehension is a complex construct that requires coordination and processing of several cognitive abilities at word, sentence, and text level (Perfetti et al., 2005; Padovani, 2006), including, but not limited to, building coherent semantic representations of what is being read (Nation and Snowling, 2000), making lexical and semantic inferences, using reading strategies, activating metacognitive control (Carretti et al., 2002).

When it comes to assessment, the above described complexity is not given due consideration and is, among other aspects, at the basis of the inadequacy of most protocols currently available. The latter often measure comprehension performance (in a way the "product" of reading comprehension) without considering the underlying processes, or treat those processes as if they were independent, not in interaction with one another. In addition, reading comprehension tests often tend to be used interchangeably, while they actually

Copyright @2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

measure different skills or processes and are not really comparable to one another (Colenbrander et al., 2017; Keenan et al., 2008; Cutting and Scarborough, 2006; Calet et al., 2020; Joshi, 2019). Finally, most currently available reading assessment tools fail to focus on reading efficiency, as they normally measure decoding and reading comprehension separately. This leads to failure in the identification of kids having difficulties in integrating the above mentioned abilities.

The AEREST protocol for reading assessment was designed and developed to fill this gap, by testing student skills in three tasks: reading aloud, silent reading, and listening comprehension. In the last two conditions, the student's comprehension of the text being read is assessed through a questionnaire. Only in the reading aloud condition, the text can also contain non-words.

In 2019, AEREST was tested in schools located in Southern Tuscany (Italy) and in the Canton of Ticino (Switzerland), involving a total of 433 children, from the  $3^{rd}$  grade of the Italian primary school through to the first grade of the Italian middle school ( $6^{th}$  grade). The protocol was automatically administered using a prototype version of *ReadLet* (Ferro et al., 2018a; Ferro et al., 2018b), a web-based platform that records large streams of time-aligned, multimodal reading data.

### 2 ReadLet

The ReadLet platform monitors and records a user's behaviour during the execution of various reading tasks. It includes a central repository and a set of web applications, background services for pre- and post-processing analysis and query tools. The ReadLet endpoint is an ordinary tablet running a web application which is responsible for the administration of the reading protocol. The ReadLet app overrides most of the actions taken by a tablet to respond to typical touch events on the screen (tapping, scrolling etc.), which is needed to allow a reader to slide across the text displayed on the touchscreen as one would normally do on a printed text on paper.

The child is asked to read a short story displayed on the tablet screen either silently or aloud, and to finger-point to the text while reading. The story is displayed on the tablet one page at a time and the child is free to flip the pages back and forth. During each reading session, the audio stream is recorded along with the time-stamped touch events caused by the interaction of the user with the touchscreen. At the end of a session, all data are sent to the central repository, ready for post-processing and for further analysis. In the listening task, ReadLet provides an audio-player playing a pre-recorded story. As the user finishes reading or listening, a multiple-choice questionnaire is presented one question at a time. In answering each question, the reader/listener can get back to the full text or play back the audio-player, and search for relevant information.

Captured data are recorded, anonymized, and encrypted locally by the application, and sent to a remote server: i) the user information along with the session settings; ii) the text disposition and layout on the screen; iii) the audio stream (i.e. the user's voice while reading aloud), iv) the time-stamped finger interaction during the reading task and in filling the questionnaire; v) the timing of the answers to each question, along with possible self-corrections. ReadLet is equipped with tools for the automated linguistic analysis of texts. The tools, together with a finger-trackingto-text alignment module, make it possible to capture the user finger-tracking behaviour (e.g. forward tracking, regressions, tracking pauses) and the time spent on the text for different text unit levels (page, paragraph, sentence, token, syllable, morpheme, n-gram, letter) and different linguistic levels (e.g. morphological, lexical, syntactic). Furthermore, the ReadLet speech-to-text alignment module (currently under development) will allow the automatic assessment of decoding accuracy during reading-aloud sessions, by analysing hesitations, reading errors, and self-corrections.

### **3** The AEREST protocol

As already mentioned, the AEREST protocol was created to provide teachers and education professionals with an accurate, non-invasive, childfriendly assessment tool that could identify the full range of students with low reading efficiency. Unlike current protocols, that usually fail to identify students who do well in the single abilities underlying reading when assessed one at a time, but struggle in the integration of those abilities, the AEREST protocol allows identification of all children manifesting difficulties, in so doing favoring access to specifically tailored enhancement training programs for all those who may need them. The AEREST assessment protocol includes three tasks: 1. Reading comprehension; 2. Listening comprehension; 3. Decoding.

### 3.1 Reading comprehension

In order to carry out this task, subjects are provided with a tablet, displaying a story that contains narrative as well as descriptive parts. The texts used for comprehension assessment are based on existing stories written by well-known authors and modified by adding or cutting out text, in order to achieve two main objectives.

The first objective is to obtain a balanced mixture of narrative and descriptive text. In our opinion, this reflects more closely the kind of texts we normally encounter in life, which are hardly ever barely descriptive or barely narrative. Keeping this separation (as most reading assessment tools actually do) would lead, in our opinion, to a less ecological way of assessing reading comprehension.

The second objective is to obtain a text that would allow assessment of all (or most of) the cognitive processes involved in reading comprehension (this is usually not found in other assessment tools currently available). This is made possible through 15 comprehension questions that engage subjects in:

- 1. retrieving the general content of the text;
- 2. identifying specific information in the text; (who/what/where/when/...). Usually 4 questions out of 15 concerns this kind of information;
- 3. identifying temporal relations;
- 4. identifying cause-effect and sequential relations;
- 5. making inferences of different kinds;
- 6. retrieving information from syntactic structure (for example understanding if some event in the story has actually happened or not, based on the verb tenses used by the author);
- 7. forming mental representations (in general, subjects are prompted with 4 different images of a character or situation in the story and are asked to determine which image corresponds to what they have read);
- 8. spotting incongruities and errors;
- 9. retrieving word meaning from context;
- 10. identifying text register and style;
- 11. identifying text type.

For each question, the subject can choose among four different answers, out of which only one is correct.

Before starting the task, kids are told that they have no time limit. Subjects are instructed to read the story silently from beginning to end, always pointing their finger to the text being read. Once they reach the end of the story, they are prompted with 15 comprehension questions. These are displayed, one at a time, on the bottom part of the screen, while the text is available in the top part. They can re-read the text, or chunks of it, as many times as they want, by scrolling up and down the text on the screen.

Analysing the responses to the comprehension questions, built as described above, allows to understand which of the processes underlying comprehension are leveraged by the subject and which ones are not efficient and need support through specific, personalised training.

In order to consider comprehension abilities independent of decoding skills (that may be weaker in some subjects, for example in kids with dyslexia) the listening comprehension test described underneath was included in the protocol.

### 3.2 Listening comprehension

As with the reading comprehension task, subjects are given a tablet and headphones for story listening. After hearing the whole story for the first time, kids start answering comprehension questions one by one, upon hearing them through their headphones and reading them on the tablet's screen. In order to reduce the child's working memory load, some of the questions are asked only after the text passage containing the relevant information is heard for the second time.

### 3.3 Reading aloud

In this task, children are asked to read aloud stories with a similar narrative structure. At the end of each story, one of the story characters (typically with some kind of supernatural powers: an alien, a witch, ecc.) starts speaking an unknown language, which consists of non-words following the phonology and morpho-syntax of Italian, and some Italian function words. We include here an example of text used for this task.

E come se stesse leggendo su quel vetro, rivelò a Lucilla la ricetta della segretis-

sima pozione: "Prendi una sirta mellusa e gafala in un tulo. Spisola una rifa e lubica una buva. Non zudugnare e non tapire le vughe. Quita le puggie, zuba i mumini e ralla un tifurno."

The administrator takes notes on the subject's errors, hesitations and self-corrections throughout the task. Meanwhile, the subject's performance is also recorded by the tablet. In addition, as for the reading comprehension task, children are instructed to always finger-point to the text being read. The child's reading score is then calculated taking off 1 point for each spelling error, 0.5 point for each self-correction. No points or fractions of point are sub-tracted for hesitations, as they already have an impact on reading time.

# 4 Data structure

Data are stored at different levels. Texts are pre-processed with NLP tools (Dell'Orletta et al., 2011) for text tokenization, POS tagging, dependency parsing, readability analysis, syllabification, *n*-gram splitting, and, finally, frequency information by means of a reference corpus.

Session settings are stored to include metadata such as the administrator identifier, user information (a unique identifier, child's affiliation and grade level, possible annotations), the text being read and its layout (e.g. margins, font size and family, letter and line spacing), task type (i.e. silent reading, reading aloud, or listening comprehension).

At the end of each session, all recorded data are sent to a remote server. Basic data include information about the tablet (e.g. the user agent string, the screen resolution), time-stamps of the beginning and end of the reading task and of questionnaire answering. More detailed data include the disposition of the text on the tablet screen (i.e. coordinates of the bounding box of each letter), touchscreen events (i.e. event type, time-stamp, and finger coordinates), the audio stream (sampled at 48KHz stereo and compressed in MP3 format at 128kbps), answers to the questionnaire and their timing.

Post-processing tools enrich stored data offline. A finger-tracking-to-text alignment algorithm binds touchscreen events over time to the text layout at the character level. This is done by creating two black and white images and performing a convolution operation over them: the first image represents the text disposition on the screen, where each line is rendered as a filled black rectangle on a white background; the second represents the user finger-tracking over time, where each segment between a touch-begin and a touch-end event is rendered as a black rectangle on a white background. During the execution of the convolution operation, the vertical and horizontal offsets which maximize the overlapping of the black areas within the two images indicate the optimal alignment to be taken into account. Such binding allows for subsequent modelling and evaluation of the reading dynamic, as well as for measurement of the reading time at different levels of granularity: from single letters and syllables through to sentences, and whole pages or documents.

# 5 Collected Data

In 2019, the AEREST protocol was administered to a total of 433 students. A total of 12 narrative texts was used, one for each of the four grade levels and the three assessment tasks. Details of participants and texts are reported respectively in Tables 1 and 2.

	Italy		Switzerland	
Grade	N	Age	Ν	Age
3	78 (13)	8.6 (0.4)	22 (4)	8.8 (0.4)
4	71 (14)	9.6 (0.3)	21 (2)	9.7 (0.5)
5	94 (25)	10.6 (0.4)	23 (2)	10.7 (0.4)
6	54 (6)	11.5 (0.4)	70(2)	11.9 (0.4)
TOT	297 (58)	10.0 (1.1)	136 (10)	10.9 (1.3)

Table 1: Sample size (number of children with disorders between brackets) and mean age (standard deviation between brackets) of the participants involved in the study, across grades (from the  $3^{rd}$  to the  $6^{th}$  grade level) and countries (Italy and Switzerland).

	silent	aloud		listening
Grade	words	words	nonwords	words
3	588	177	53	572
4	750	180	74	527
5	951	216	80	941
6	711	352	83	734

Table 2: Number of tokens in the texts administered during the study, across grades (from the  $3^{rd}$  to the  $6^{th}$  grade level) and decoding conditions (silent reading, reading aloud, and listening).

#### 6 Results and discussion

Tablets proved to be easy to use and well accepted devices, extremely instrumental and accurate for data collection with toddlers and older children (Frank et al., 2016; Semmelmann et al., 2016). Tablet data confirmed high standards of ecological validity, and a high correspondence with data collected with other, more traditional tools (e.g. eye-tracking, see Lio et al. (2019)), and protocols. Within the present work, the collected data allowed for the evaluation of the decoding and comprehension skills of the children involved in the study. For each grade level, Aerest decoding performance, expressed in syllables per second, was shown to be in line with more classical reading assessment reports (Cornoldi et al., 2010), for both words and non-words. Furthermore, the use of the finger tracking allowed for the validation of the correlation of the time spent on each word with basic features such as frequency and length: statistical analysis with linear mixed-effect models shows a highly significant correlation (p < 0.0001), thus confirming the reliability of the adopted technique.

Decoding and comprehension performance scores are shown in Fig. 1. Data are normalized for each grade level group, so that all data groups can be overlapped on the same plot. Indeed, data belonging to each group was divided by the median value of control children only. In this way data can be graphically compared, being a value of 0.5 equal to half the mean performance of control children, a value of 1 equal to average behaviour, and a value of 2 indicates a double outperforming with respect of the average performance.

### 7 Conclusions and future work

The AEREST protocol was shown to be effective in characterizing the decoding and comprehension performance of children of late primary school and early middle school in text reading tasks. Results are clear and encouraging, opening the way to further, more detailed, dynamic, and multimodal analysis. Completion of the current AEREST protocol with a second battery of tests is foreseen in the near future. This will provide schools with two different test batteries, to be used for assessment at the beginning and end of school year, for adequate monitoring of pupils' reading and reading comprehension skills. A version of the protocol conceived for clinical context is also foreseen, as

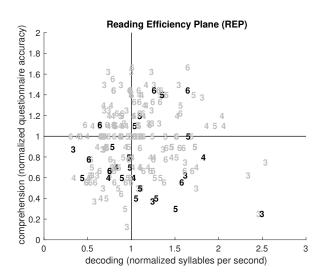


Figure 1: Reading Efficiency Plane for the reading comprehension task (silent reading and comprehension questions). The decoding performance on silent reading (expressed as the normalized syllables per second) is shown in the horizontal axis, while the comprehension performance (expressed as the normalized questionnaire accuracy) is shown in the vertical axis. For each grade level group (from the 3<sup>rd</sup> to the 6<sup>th</sup> grade level), the two measures are normalized on the basis of the performance of children with typical reading development. Each child is represented by a digit marker indicating the grade level. Typically and atypically developing readers are shown respectively in gray and black.

well as translation and adaptation of the protocol to languages other than Italian.

The collected data will be assembled in a multimodal linguistic resource and made freely available to the scientific community.

#### Acknowledgments

This work was supported by the Swiss grant "AEREST: An Ecological Reading Efficiency Screening Tool" (2017-2020) funded by the Department of Teaching and Learning of the University of Applied Sciences and Arts of Southern Switzerland (SUPSI), and by the Italian project "(Bio-)computational models of language usage" (2018-) funded by the Italian National Research Council (DUS.AD016.075.004, ILC-CNR).

A special thanks goes to all schools that took part in the study, in particular: Ist. Comprensivo of Manciano-Capalbio (Grosseto, Italy), elementary school of Novaggio, (Ticino Switzerland), lower secondary school of Bedigliora (Ticino, Switzerland).

### References

- Nuria Calet, Rocío López-Reyes, and Gracia Jiménez-Fernández. 2020. Do reading comprehension assessment tests result in the same reading profile? a study of spanish primary school children. *Journal of Research in Reading*, 43:98–115.
- Barbara Carretti, Cesare Cornoldi, and Rossana De Beni. 2002. Il disturbo specifico di comprensione del testo scritto. In S. Vicari and M.C. Caselli, editors, *I disturbi dello sviluppo: neuropsicologia clinica e ipotesi riabilitative*, pages 169–189. Il Mulino, Bologna.
- Danielle Colenbrander, Lyndsey Nickels, and Saskia Kohnen. 2017. Similar but different: differences in comprehension diagnosis on the neale analysis of reading ability and the york assessment of reading for comprehension. *Journal of Research in Reading*, 40(4):403–419.
- Cesare Cornoldi, Patrizio E. Tressoldi, and Nicoletta Perini. 2010. Valutare la rapidità e la correttezza della lettura di brani. nuove norme e alcune chiarificazioni per l'uso delle prove mt. *Dislessia*, 7:89– 101.
- Laurie E. Cutting and Hollis S. Scarborough. 2006. Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific studies of reading*, 10(3):277–299.
- Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.
- Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Franco Alberto Cardillo, and Vito Pirrelli. 2018a. ReadLet: an ICT platform for the assessment of reading efficiency in early graders. page 61, Edmonton, Alberta (Canada), 25-29 September, 2018. 11th International Conference on the Mental Lexicon.
- Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Ouaphae Nahli, Franco Alberto Cardillo, and Vito Pirrelli. 2018b. Readlet: Reading for understanding. In 2018 IEEE 5th International Congress on Information Science and Technology (CiSt), pages 1–6.
- Michael C. Frank, Elise Sugarman, Alexandra C. Horowitz, Molly L. Lewis, and Daniel Yurovsky. 2016. Using tablets to collect data from young children. *Journal of Cognition and Development*, 17(1):1–17.

- R. Malatesha Joshi. 2019. Componential model of reading (cmr): Implications for assessment and instruction of literacy problems. In D. A. Kilpatrick, R. M. Joshi, and R. K. Wagner, editors, *Reading development and difficulties*, pages 3–18. Springer, Dordrecht (The Netherlands).
- Janice M. Keenan, Rebecca S. Betjemann, and Richard K. Olson. 2008. Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3):281–300.
- Guillaume Lio, Roberta Fadda, Giuseppe Doneddu, Jean-René Duhamel, and Angela Sirigu. 2019. Digit-tracking as a new tactile interface for visual perception analysis. *Nature Communications*, 10(5392):1–13.
- Kate Nation and Maggie J. Snowling. 2000. Factors influencing syntactic awareness skills in normal readers and poor comprehenders. *Applied psycholinguistics*, 21(2):229–241.
- OECD. 2003. Learners for life. student approaches to learning. results from PISA 2000. https: //doi.org/10.1787/9789264103917-en, OECD Publishing, Paris.
- OECD. 2019. Assessment and analytical framework. https://doi.org/10.1787/b25efab8-en, OECD Publishing, Paris.
- Roberto Padovani. 2006. La comprensione del testo scritto in età scolare. una rassegna sullo sviluppo normale e atipico. *Psicologia clinica dello sviluppo*, x(3):369–398.
- Charles A. Perfetti, Nicole Landi, and Jane Oakhill. 2005. The acquisition of reading comprehension skill. In M. J. Snowling and C. Hulme, editors, *The science of reading: a handbook*, chapter 13, pages 227–247. Blackwell, Oxford.
- Kilian Semmelmann, Marisa Nordt, Katharina Sommer, Rebecka Röhnke, Luzie Mount, Helen Prüfer, Sophia Terwiel, Tobias W Meissner, Kami Koldewyn, and Sarah Weigelt. 2016. U Can Touch This: How Tablets Can Be Used to Study Cognitive Development. *Frontiers in psychology*, 7:1021, jul.