8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)
20-21 September 2019, Graz, Austria

Overview of LARA: A Learning and Reading Assistant

Elham Akhlaghi[1], *Branislav Bédi*[2], *Matthias Butterweck*[3], *Cathy Chua*[3], *Johanna Gerlach*[4]
Hanieh Habibi[4], *Junta Ikeda*[3], *Manny Rayner*[4], *Sabina Sestigiani*[5], *Ghil'ad Zuckermann*[6]

[1]Ferdowsi University of Mashhad, Iran; [2]The Árni Magnússon Institute for Icelandic Studies, Iceland; [3]Independent scholar; [4]FTI/TIM, University of Geneva, Switzerland; [5]Swinburne University, Australia; [6]University of Adelaide, Australia

elham.akhlaghi@mail.um.ac.ir, branislav.bedi@arnastofnun.is, matthias@butterweck.de, cathyc@pioneerbooks.com.au, Johanna.Gerlach@unige.ch, hanieh.habibi@unige.ch, ikedaj_91@hotmail.com, Emmanuel.Rayner@unige.ch, ssestigiani@swin.edu.au, ghilad.zuckermann@adelaide.edu.au

Abstract

We present an overview of LARA (Learning and Reading Assistant), a set of tools currently being developed in the context of a collaborative open project for building and using online CALL content. LARA offers a range of options for semi-automatically transforming text into a hypertext version designed to give support to non-native readers. Functionality includes construction of a personalised concordance based on the learner's reading history, addition of recorded audio files, and insertion of links to translations and online linguistic resources. We present initial evaluations of LARA content developed for Icelandic, Farsi and Italian, and briefly describe content created in several more languages. We conclude by noting ethical issues that arise and outlining plans for further development of LARA.
Index Terms: CALL, reading, hypertext, open source

1. Introduction and motivation

A key problem when creating human language applications is scalability. Techniques for building many kinds of speech and language applications are now well described in the literature, but the question is how to produce them quickly in large numbers; the standard answer is machine learning, but ML is not always the right tool. A complementary approach is crowdsourcing. If the tools needed to construct applications can be made easy enough to use, it is possible to recruit a large workforce and distribute the task. The most prominent example is Amazon's Alexa, where upwards of 100,000 'skills' have now been built and deployed.

This paper describes another platform of the same general kind. LARA (Learning and Reading Assistant; https://www.unige.ch/callector/text-content/) is a collaborative open project, initiated during Q3 2018, whose goal is to create resources via crowdsourcing techniques that help people read L2 texts in foreign/archaic languages. It does this by providing tools that make it easy to transform plain text documents into hypertext versions that give non-native readers various kinds of help.

There are now several mainstream platforms, most obviously the Amazon Kindle, which support non-native readers by allowing them to look up words in a dictionary and providing TTS support for at least some texts. Related but less well-known cases are LingQ (https://www.lingq.com/en/) and White Rabbit Press (https://play.google.com/store/apps/details?id=com.whiterabbitpress.jgr). It is not clear, however, that bilingual lexicons and audio are really the things which benefit the intermediate-level student most. Having audio is a definite plus, but the benefit of lexicon support is in contrast less obvious. It is useful to beginners; but an intermediate-level reader who already has a basic grasp of L2 grammar and vocabulary can guess the approximate meaning of most new words if they see a few examples, and learning the word from context internalises it more effectively than looking it up in a dictionary [1, 2, 3, 4].

When we started investigating the idea of annotating text to support students using the reading strategy, we decided to make this the central question: how could we help learners learn intuitively, by making it easier for them to examine words in context? In contrast to previous work, we confront the issues directly. The learner receives a personalised version of the text they are reading, marked up in such a way that they can immediately compare all occurrences of any word *in their own reading progress*. Concretely, the screen is divided into two halves, with the text on the left. When the student clicks on a word, the right-hand side shows a personalised concordance. Figure 1 illustrates, showing a reading progress where the student has read *Peter Rabbit* and the first three chapters of *Alice in Wonderland*. In addition to the personalised concordance, LARA also offers conventional support for the reader. As the figure shows, this includes optionally linking words and sentences to translations and audio recordings.

The rest of the paper is structured as follows. §2 describes how LARA content is created, §3 describes initial evaluation exercises using Icelandic, Farsi and Italian content, and §4 describes other LARA content we have built. §5 summarises our position on ethical issues and says how to obtain LARA. The final section outlines further directions.

2. Creating content

The process of creating a piece of LARA content consists of three steps. First, the source corpus is marked up to support construction of the concordance and the other resources. Markup currently contains elements for breaking text first into pages and then into segments (segments are typically but not always sentences), marking multi-words and compound words, tagging inflected words by their associated lemmas, and optionally including HTML and CSS formatting. It is also possible to mark a passage as "plain text", i.e. to be left unchanged by LARA. Figure 2 illustrates. The most laborious part of the process, adding the lemma tags, can be performed semi-automatically for the thirty-odd languages currently supported by TreeTagger [5]. TreeTagger's error rate varies widely depending on lan-

10.21437/SLaTE.2019-19

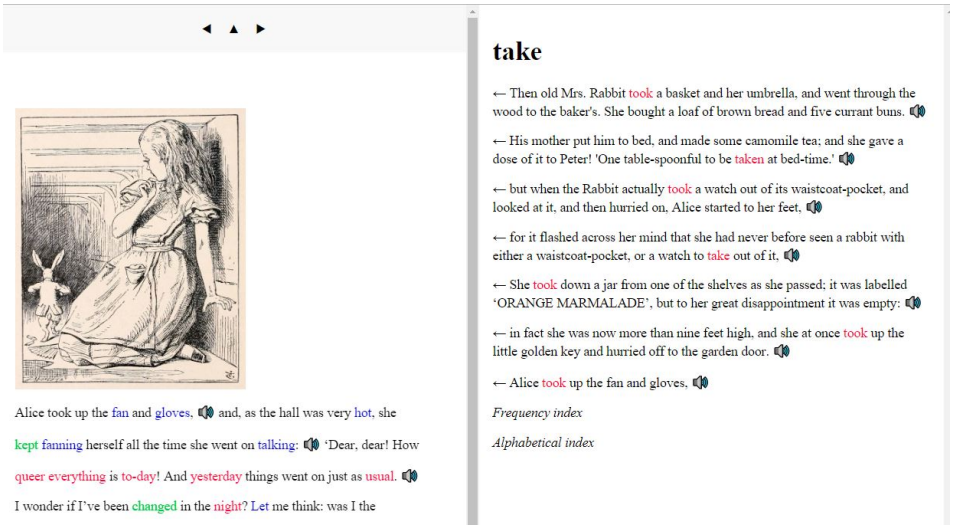

Figure 1: *Example constructed using current LARA prototype (available online here) showing a page from the personalised reading progress. The learner has so far read* Peter Rabbit *followed by the first three chapters of* Alice in Wonderland*. The left hand side shows the marked-up text, where the student has just clicked on the word "took". The right hand side displays occurrences of different inflected forms of "take" in both source texts. Colours show how many times words have occurred: red means the word has only occurred once, green two or three times, blue four or five times, black more than five times. The back-arrow at the start of each line on the right is a link to the point in the text where the example occurs. Hovering the mouse over a word plays an audio file and shows a translation for that word; hovering over a loudspeaker icon shows a translation for the preceding segment, and clicking plays an audio file. Most of the above functionality is optional and can be turned off if desired.*

guage and genre, but is typically in low single digits. To be able to handle languages not covered by TreeTagger, we plan soon to add support for other taggers, in particular for Icelandic [6] and Farsi.

```
<page>
/* Peter Rabbit, 2019-05-07 */
@Once upon a time@ there were#be# four
little Rabbits#rabbit#, and their
names#name# were#be#--
          Flopsy#Flopsy#,
       Mopsy#Mopsy#,
   Cotton-tail#Cottontail#,
and Peter#Peter#.||

They lived#live# with their Mother in
a sand-|bank, underneath the root of a
very big fir-|tree.||
<img src="01VeryBigFirTree.jpg" width="271"
height="317"/>
<page>
```

Figure 2: *Example of marked-up LARA text, showing page boundaries (<page>), segment boundaries (||), multi-words (@ ... @), compound words (|), lemma tags (# ... #), plain text (/* ... */) and HTML tags.*

Once the corpus is marked up, the second and third steps, adding audio recordings and translation, are straightforward. Audio recordings are efficiently produced using an online tool. The compiler extracts two scripts from the marked up corpus, one for segments and one for words, and the tool presents them to the voice talent, who can complete the items in any order and review recordings before submitting them. When the recording process is complete, the content creator downloads the audio files and associated metadata. A similar method is used to obtain translations of words and segments for each supported L1. For users who have installed LARA on their own machines, there is a simple GUI which wraps the above functionality and allows the operations to be carried out by simple button-presses. A web portal is currently undergoing initial testing, and should be available by the time of the SLaTE 2019 conference.

What cleverness there is in LARA is related to the process of combining together the different resources — annotated text, images, audio and translations — into a personalised hypertext concordance summarising the individual student's reading progress. Since each reader has their own concordance, and they are frequently updated, it is important to construct them efficiently. It is also desirable to allow content to be distributed over multiple servers, both since LARA is set up with crowdsourcing in mind and for the ethical reasons outlined in §5.

We use two main tricks. First, in order to avoid copying the multimedia files (images and recorded audio), which account for over 90% of the web space required, we require each LARA corpus resource, which can be anywhere on the web, to instantiate a uniform structure, with metadata listing the corpus, image and audio files; a master file lists the root URLs for the available resources. When constructing the concordance for a given reading history, the compiler only needs to download the corpus text and metadata for the resources referenced, using the metadata to insert links to the multimedia where it appears. The set of web data representing the learner's personalised concordance thus contains only text, and can be kept manageably small. Second, we cache intermediate data for each student's reading progress and only recompute when necessary. Thus we keep copies of downloaded corpus files, internalised forms of the files' con-

tent, generated pages, and the data used to build them; when the student advances to a new page P, it is only necessary to remake the contents of P and the concordance pages for words appearing on P. On a medium-range laptop, the personalised concordance can typically be updated in two to five seconds, the time required depending on the size of the pages.

Table 1: *Currently available online LARA content. "Lng" = language;"#Seg" = number of segments; "#Tok" = number of surface word tokens; "#Typ" = number of lemma types; "Link" = link to online LARA resource.*

Text	Lng	#Seg	#Tok	#Typ	Link
Evaluated					
Tina	IS	207	2743	466	☞
Choopan	FA	31	391	150	☞
Boz-Boz Ghandi	FA	75	539	191	☞
Ebne Sina	FA	35	257	94	☞
Arash	FA	34	329	128	☞
Molana	FA	37	482	221	☞
Ungaretti	IT	127	328	199	☞
Dante	IT	135	840	395	☞
Not yet evaluated					
Peter Rabbit	EN	41	966	356	☞
Hy. Ikita Neko	JP	88	967	225	☞
Wilhelm Busch	DE	73	674	281	☞
Alice in Won.	EN	1478	26868	2007	☞
Le petit prince	FR	1436	15421	1568	☞
Nibelungenlied	DE	11937	81853	2826	☞
Barngarlidhi M.	BJB	498	456	339	☞
Revivalistics	HE	69	129	124	☞

3. Initial evaluations

Table 1 shows nontrivial LARA content created so far. The unusual mix of languages reflects the project's distributed organisation. The LARA tools are mostly being developed by the core group at Geneva University, working under funding from the Swiss National Science Foundation, but content has been produced by multiple sites. The most active partners have been the Árni Magnússon Institute for Icelandic Studies (Icelandic), the Ferdowsi University of Mashhad (Farsi), Swinburne University (Italian), the University of Adelaide (Israeli Hebrew, Barngarla) and three independent scholars: Cathy Chua has produced content in English, Matt Butterweck in German and Middle High German, and Junta Ikeda in Japanese.

In this section, we will present initial evaluation exercises, carried out at the University of Iceland, Ferdowsi University of Mashhad and Swinburne sites respectively with three groups of students, which used the content in the upper half of the table. All three groups followed the same methodology. Initially, the students were given a ten-minute introduction to LARA and the relevant piece of content, then asked to use them on their own for about an hour, accessing the content through their own laptops. Finally, they spent twenty minutes filling in an anonymous questionnaire consisting of three sections: four demographic questions, seventeen questions on a five-point Likert scale, and eight open-ended questions. The same questionnaire was used for all three groups; the design was based on two examples from the literature which have already been widely copied [7, 8]. In the rest of this section, the people responsible for each group start by describing their learners and content, after which we present the results.

3.1. University of Iceland (Branislav Bédi)

Out of 177 students registered for the Icelandic Practical Diploma course at the University of Iceland, Reykjavik, 47 took part in the evaluation. They consisted of a mixed group of beginners and intermediate learners of Icelandic as a second language. Motivation varied. Some had specific goals in mind, such as taking an entry examination for a course, or to get a residence permit. Others simply wanted to improve. The text used for LARA, Esther Skriver's *Tína fer í frí*, is a children's book normally read by 6 to 8 year olds. The choice was based on the course level; the book somewhat stretched most of the students' vocabularies, but was not out of their reach. Reading a children's short story in classical printed book form is part of the normal course syllabus; using LARA to do the same thing with an interactive text offered several additional advantages. As well as giving support for vocabulary, grammar, and pronunciation, it also allowed the students to practise their listening abilities. This work is reported in more detail elsewhere [9].

3.2. Swinburne (Sabina Sestigiani)

Two groups comprise the cohort of language students of Italian from Swinburne University of Technology, Melbourne: Year 1 with an average of 50 students and Year 2 of about 20. The motivation is rather different from those of the Icelandic students. Swinburne students are taking on Italian by choice as a minor specialisation within their bachelor degree or as an elective and their motivation is therefore more a question of passion than purpose. Students might be interested in their Italian heritage, or attracted to Italian culture and lifestyle, or simply wish to learn a classical European language. The teaching style adopted for the courses relies on the appreciation of authentic literary texts and the enhancement of the students' oral skills in L2 through theatrical performances. The choice of two Italian poetry classics—Giuseppe Ungaretti's famous World War I poems and a few extracts from Dante's Inferno—was therefore dictated by the necessity to expose the students to the performativity of the voice in Italian literature. The intent was twofold: stimulate students' interest and fascination with the sounds of Italian poetry, and empower them to learn independently. The internationally celebrated poems—of which students had heard but never dared to read because they feared them to be inaccessible—could prove a potent motivation for learning.

3.3. FUM (Elham Akhlaghi and Hanieh Habibi)

Two groups of students used LARA contents; First, a group of 4 female and 3 male Russian students, aged 29 to 35. Second, a group of 12 male and 1 female Arab students, aged 20 to 47. All the students had Beginner and Intermediate level in Farsi. In the former group all had a Ph.D. degree in various fields of the Humanities and took part in a virtual Farsi course to use this language in their own field of academic research. The latter members aimed to start academic studying in Iran. Unlike the Icelandic and Italian classes, we had several short texts for Farsi, collected and simplified from famous children's stories. The reason for simplifying was that the students were only familiar with the simple past tense in Farsi and all texts had to be edited

in an appropriate way. One of the main characteristics of Farsi is the high frequency of occurrence of compound verbs with nouns and light verbs, and it can be hard for foreign students to distinguish similar verbs in a text. LARA helped here by showing all occurrences of each verb in a single concordance page. Having a completely different Arabic-based script also makes it difficult for learners to start reading Farsi texts. The option of listening to each word on mouseover made it easier for students to map the written alphabet to its phonetic form.

3.4. Results

Table 2: *Summary of questionnaire results from initial LARA evaluations, for three sites and five groups of questions. Numbers give average Likert scores for the question category and group, with 1 = least favourable and 5 = most favourable.*

	Iceland	Mashhad	Swinburne	Mean
# Subjects	47	23	17	87
Efficiency	4.12	3.90	4.53	4.24
Ease of use	4.20	3.85	4.63	4.22
Quality	4.19	3.92	4.71	4.23
General	4.09	3.91	4.45	4.16
Open	3.74	3.92	3.62	3.77
Mean	4.07	3.90	4.39	4.11

Full results of the study are posted at https://www.unige.ch/callector/lara-study-1/. Table 2 presents a summary, where we have grouped the questions from the second and third sections into five categories and given the average Likert score for each category and group. The students were evidently very happy with the issues addressed by the first four groups of questions (typical questions for each category: "Using this application increases my learning productivity"; "Using this application makes it easier for me to learn pronunciation"; "Compared to using books, using this application improves the quality of reading L2 texts"; "The application addresses my learning-related needs in this course"). They were more critical in the open-ended questions, where they were asked to suggest ways to make the app better. Common suggestions were that the minimalistic design could be improved and that there should be more support for tablets/mobile platforms.

4. Other content

We briefly describe the content in the bottom half of Table 1. The first two items are similar. *Peter Rabbit* [10] and *Hyakumankai Ikita Neko* [11] are well-known stories for younger children in English and Japanese respectively, both about a thousand words long. They have a complete set of LARA features, including audio and translations for both words and sentences. The third item consists of three illustrated poems in German by Wilhelm Busch; they show how it is possible to include complex formatting in LARA documents.

The next three items are full-length books, whose LARA versions still represent work in progress. *Alice in Wonderland* [12] has been fully tagged, using a tagger built from resources provided by the Python NLTK package [13] followed by manually cleaning. Audio is currently being recorded, and is about 60% complete. *Le petit prince* [14] is similar; here, the initial tagging was done using TreeTagger. *Das Nibelungenlied*, a 12,000 line poem in Middle High German, is the most ambitious LARA project so far attempted. It was tagged using the TreeTagger package for Middle High German and manually cleaned. Segment translations in High German were added from a publicly available source. There is embedded audio for one representative section.

The last two items represent an interesting idea we have recently begun investigating. Although the original purpose of LARA was to help students read L2 texts, it also seems useful for presenting mixtures of L1 and L2 text of the kind found in linguistics papers, language textbooks and similar. *Barngarlidhi Manoo* [15] is an 80 page alphabet book for the Australian aboriginal language Barngarla. *Revivalistics* is a four page extract from a forthcoming linguistics book [16]; we have chosen a passage in which the phonetic aspects of language are salient. In both of these examples, annotation makes use of "plain text" brackets (cf. Figure 2) so that the English text can be marked as to be left unchanged and passages in the various other languages as to be processed by LARA; audio was then recorded for each such phrase. The upshot is that the two texts could easily be transformed into versions where the reader can listen to any non-English phrase by hovering the mouse over it.

5. Ethical issues; obtaining LARA

We are only able to present a capsule summary of our position on ethical issue here, though they are a central part of the project; a longer discussion can be found in two recent papers [17, 18]. Very briefly, we consider that since the value of an internet community like the one we are establishing here is largely derived from the unpaid labour of the community's members, the founders incur a corresponding obligation towards these people. In particular, the community should both respect the members' intellectual property rights, and endeavour from the start to create a sustainable infrastructure which will preserve the work they have created. The history of online communities shows that these principles are often flagrantly disrespected.

In LARA, we are attempting to follow the abstract guidelines laid out above. Most importantly, LARA software is all open source, and has been designed to be simple and portable. The codebase as of mid July 2019 consists of ∼7.5K lines of Python (the core code) and ∼5K lines of PHP, JavaScript and CSS (the online portal). It was originally developed at the University of Geneva, but is freely available from an open source repository; the online documentation gives details [19]. A non-Geneva person, Matt Butterweck, has already introduced substantial improvements and extensions. He has probably written about 25% of the Python code.

6. Summary and further directions

We have presented a brief overview of LARA. We are encouraged by progress to date. The first group of content constructors has used the platform to produce LARA resources in ten widely different languages; we have received many queries from other people interested in learning to do the same. The material is already being used in real language courses, and initial feedback from students has been extremely positive.

Our top priorities for the next few months revolve around stabilising the new LARA portal and making it generally available. We are also organising a two day workshop in late November, funded by the enetCollect COST network (https://enetcollect.net), where attendees will be able to get hands-on familiarity with LARA and other related tools. Details will be posted on the project home page before SLaTE 2019.

7. References

[1] R. J. Sternberg, "Most vocabulary is learned from context," *The nature of vocabulary acquisition*, vol. 89, p. 105, 1987.

[2] T. S. Paribakht and M. Wesche, "Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition," *Second language vocabulary acquisition: A rationale for pedagogy*, vol. 55, no. 4, pp. 174–200, 1997.

[3] J. H. Hulstijn, "Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity," in *Cognition and Second Language Instruction*, P. Robinson, Ed. Cambridge University Press, 2001.

[4] T. Johns, "Data-driven learning: The perpetual challenge," in *Teaching and learning by doing corpus analysis*. Brill Rodopi, 2002, pp. 105–117.

[5] H. Schmid, "Improvements in part-of-speech tagging with an application to German," in *Natural language processing using very large corpora*. Springer, 1999, pp. 13–25.

[6] S. Steingrímsson, Örvar Kárason, and H. Loftsson, "Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step," in *Proceedings of RANLP 2019*, Varna, Burgaria, in press.

[7] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989.

[8] D. Nesbitt, "Student evaluation of CALL tools during the design process," *Computer Assisted Language Learning*, vol. 26, no. 4, pp. 371–387, 2013.

[9] B. Bédi, C. Chua, H. Habibi, R. Martinez-Lopez, and M. Rayner, "Using LARA for learning Icelandic," in *Proc. EUROCALL 2019*, 2019.

[10] B. Potter, *The Tale of Peter Rabbit*. Frederick Warne & Co., 1904.

[11] Y. Sano, *Hyakumankai ikita neko*. Kodansha, 1977.

[12] L. Carroll, *Alice's Adventures in Wonderland*. Macmillan, 1865.

[13] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[14] A. de Saint-Exupéry, *Le petit prince: avec des aquarelles de l'auteur*. Gallimard, 1945.

[15] G. Zuckermann, *Barngarlidhi Manoo: Speaking Barngarla Together*, https://www.adelaide.edu.au/directory/ghilad.zuckermann?dsn=directory.file;field=data;id=41076;m=view and https://www.adelaide.edu.au/directory/ghilad.zuckermann?dsn=directory.file;field=data;id=41096;m=view, 2019.

[16] ——, *Revivalistics: From the Genesis of Israeli to Language Reclamation in Australia and Beyond*. New York: Oxford University Press, forthcoming.

[17] C. Chua and M. Rayner, "What do the founders of online communities owe to their users?" in *Proceedings of the enetCollect WG3/WG5 workshop*, Leiden, Holland, 2019, http://ceur-ws.org/Vol-2390/.

[18] C. Chua, H. Habibi, M. Rayner, and N. Tsourakis, "Decentralising power: how we are trying to keep CALLector ethical," in *Proceedings of the enetCollect WG3/WG5 workshop*, Leiden, Holland, 2019, http://ceur-ws.org/Vol-2390/.

[19] M. Rayner, H. Habibi, and M. Butterweck, *Constructing LARA content*, https://www.issco.unige.ch/en/research/projects/callector/LARADoc/build/html/index.html, 2019, online documentation.