Beyond Philology No. 19/1, 2022
ISSN 1732-1220, eISSN 2451-1498

https://doi.org/10.26881/bp.2022.1.09

Using the LARA platform to crowdsource a multilingual, multimodal Little Prince

ELHAM AKHLAGHI, ANNA BĄCZKOWSKA, BRANISLAV BÉDI, HAKEEM BEEDAR, CATHY CHUA, CATIA CUCCHIARINI, HANIEH HABIBI, IVANA HORVÁTHOVÁ, CHRISTÈLE MAIZONNIAUX, NEASA NÍ CHIARÁIN, NATHALIE PATERSON, CHADI RAHEB, MANNY RAYNER, CHUNLIN YAO

Received 28.02.2022,
received in revised form 24.11.2022,
accepted 25.11.2022.

Abstract

We describe an ongoing project, in which an informally organised international consortium is using the open source LARA platform to create multimodal annotated editions of Antoine de Saint-Exupéry's *Le petit prince* in multiple languages, so far French, English, Italian, Icelandic, Irish, Japanese, Polish, Farsi and Mandarin. LARA versions of the book include integrated audio and translations and an automatically generated lemma-based concordance, and are freely available online. We describe the methods used to construct the various versions. In some cases, work for a given language was simply divided by type, typically with one person adding translations and another recording audio. In other languages, we experimented with crowdsourcing methods, splitting the text into chapter-sized units and using the LARA platform to distribute these to multiple annotators, then com-

bining the results at the end. Finally, we report an initial classroom study, where the French version was used by intermediate-level Australian students of French.

Keywords

Computer Assisted Language Learning, multimedia, crowdsourcing, English, Farsi, French, Icelandic, Irish, Italian, Japanese, Mandarin, Polish

Wykorzystanie crowdsourcingu w wielojęzycznym i multimodalnym opracowaniu „Małego Księcia" na platformie LARA

Abstrakt

Artykuł opisuje projekt, w którym nieformalnie zorganizowane międzynarodowe konsorcjum wykorzystuje platformę open source LARA do stworzenia multimodalnych, anotowanych tłumaczeń „Le petit prince" Antoine'a de Saint-Exupéry'ego. Tłumaczenia te obejmują kilka języków, w tym język francuski, angielski, włoski, islandzki, irlandzki, japoński, polski, farsi i mandaryński. Na platformie LARA dostępne są przekłady ww. książki, zintegrowane z nimi nagrania audio, a także automatycznie generowane konkordancje oparte na lematach. W artykule opisane zostały metody użyte do skonstruowania poszczególnych wersji językowych. Praca nad tworzeniem przekładów organizowana była w dwojaki sposób. Dla niektórych języków zadania dzielono na dwie osoby, z których jedna dodawała tłumaczenia, a druga nagrywała do nich dźwięk. W innych przypadkach eksperymentowano z metodami crowdsourcingowymi, dzieląc tekst na rozdziały i wykorzystując platformę LARA do przydzielania fragmentów tekstu wielu anotatorom, a następnie łącząc wyniki ich pracy w jeden tekst. Ponadto, w artykule opisano wstępne badania dotyczące wykorzystania platformy LARA w kontekście edukacyjnym, gdzie australijscy studenci na średnim poziomie zaawansowania językowego korzystali z przekładu „Le petit prince" w toku nauki języka francuskiego.

Słowa kluczowe

Computer Assisted Language Learning, multimedia, crowdsourcing, j. angielski, j. perski, j.francuski, j. islandzki, j. irlandzki, j.włoski, j. japoński, j. mandaryński, j. polski

1. Introduction

In this paper, we present a case study using the Learning And Reading Assistant (LARA; Akhlaghi et al 2019; https://www.unige.ch/callector/lara/), an open platform that has been under development by an international consortium since 2018. The purpose of LARA is to enable easy conversion of plain texts into a multimedia form designed to support development of reading skills, adding embedded audio, translations and other annotations, and typically addresses high beginner and intermediate level learners. Related functionality is offered by several other currently available platforms, high-profile examples including LingQ[1] and LearningWithTexts[2].

The overall goal of the project we describe here is to create LARA editions, in many languages, of Antoine de Saint-Exupéry's classic short novel *Le petit prince* ("The Little Prince"). We selected this text for several reasons. The book is out of copyright, has been translated into nearly four hundred languages, and is widely appreciated by both children and adults. The grammar and vocabulary are easy enough to be approachable at an intermediate level, but hard enough to stretch the reader; similarly, the length, about 15,000 words, is short enough to be manageable but long enough to give a sense of achievement if the student can reach the end. These positive factors combine to make the book hugely popular as an intermediate level reader.

1 https://www.lingq.com/
2 https://sourceforge.net/projects/lwt/

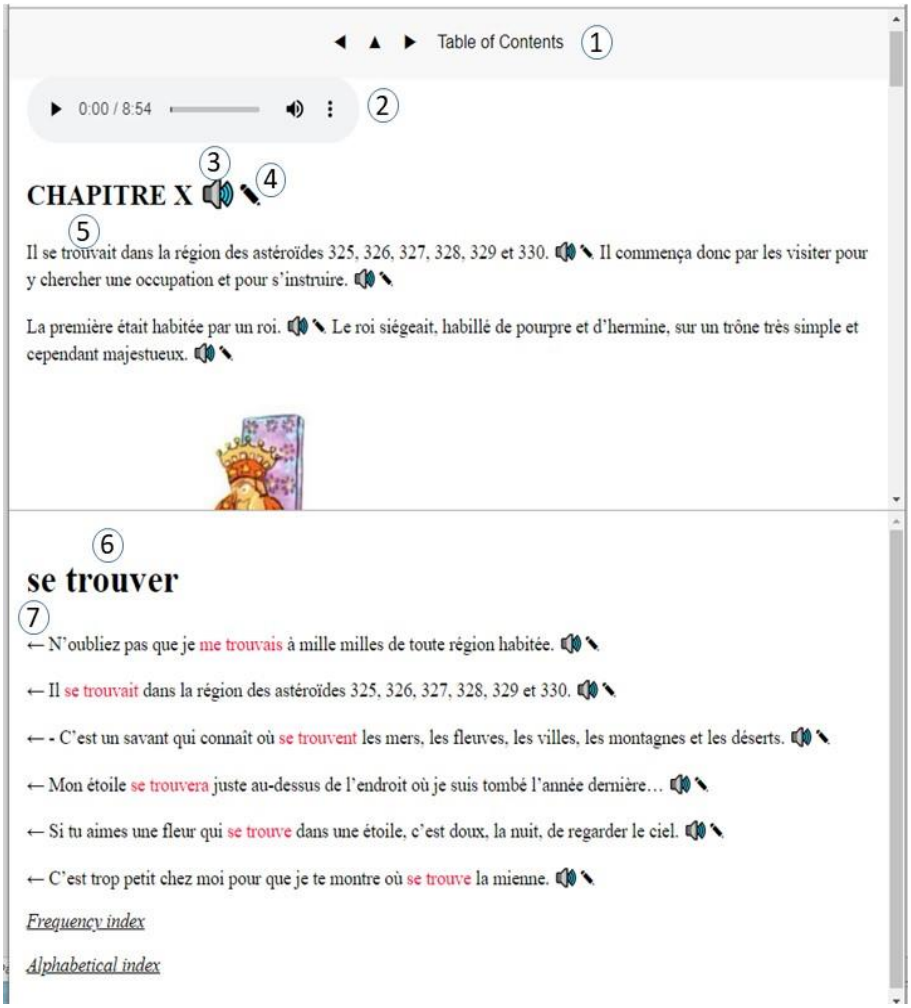

Figure 1

A sample page from the French LARA edition of Le petit prince. The text is in the upper part of the screen. Controls at the top (1) support navigation; each page contains a control (2) that plays audio for the complete page. Audio for a sentence can be played using the loud-speaker icons (3), and hovering over a pencil icon (4) shows a pop-up translation for the sentence. Clicking on a word plays audio for it and also shows a concordance; here, the user has just clicked on the multiword se trouvait ("found himself") (5), bringing up the concord-ance for the lemma se trouver ("to find oneself") (6). Clicking on a backarrow in the concordance (7) moves to the place in the main text where that sentence occurs.

The specific languages we have used are English, Farsi, French, Icelandic, Italian, Irish, Japanese, Mandarin and Polish: in some cases, work is complete, in others it is still in progress. We have experimented with a number of methods for dividing up the work required to create the LARA versions. For English, French, Icelandic, Italian, Irish and Mandarin, we simply assigned responsibility for the whole of each type of work to a single person. Thus, for example, one person added the translations for that language, one recorded the embedded audio, and so on. In Polish, Farsi and Japanese, we crowdsourced at the level of the text; we divided it into chapter-length pieces and assigned each piece to a different person, combining the results at the end. For Polish, this was done manually. For Farsi and Japanese, we have gone further and added infrastructure to the LARA platform to make the process automatic, with the platform itself taking responsibility for distributing the pieces of text and recombining the annotations added by the crowd-workers.

The rest of the paper is organised as follows. Section 2 presents background on the LARA platform, and Section 3 describes the new LARA crowdsourcing component. Section 4 describes the work we carried out to build the different versions of The Little Prince; in several cases (Farsi, Italian, Japanese, Mandarin and Polish), this involved extending the platform to provide better support for the language in question. Section 5 presents an initial study where the French version was trialed in the context of an intermediate-level course. The final section concludes and outlines further directions.

2. The LARA platform

Learning And Reading Assistant (LARA; https://www.unige.ch/callector/lara/; Akhlaghi et al 2019) is an open source project whose goal is to develop easily accessible tools that support conversion of texts into an annotated multimedia form designed to support non-native learners. LARA has its roots in enetCollect (Lyding et al 2019; https://enetcollect.eurac.edu), a European COST network that links together about 200 researchers intere-

sted in the intersection of crowdsourcing and CALL. Initial funding came from a Swiss National Science Foundation grant to the University of Geneva, under the SNSF's COST program. The project has since expanded to involve individuals and groups in over a dozen countries. After Switzerland, still the de facto centre, the most active groups are the ones in Iceland, Iran, Ireland, Australia, Holland and Poland. Publications and links to examples of LARA content can be found on the LARA site.

The LARA software platform consists of three levels. The lowest level, the *core engine*, was the first to be developed; it consists of a set of Python scripts which carry out the central functions of converting text and other resources (recorded audio, translations, etc) into multimedia form. It is possible to use the core engine on its own by invoking it from the command line. In practice, however, most people access LARA through the *portal* (https://lara-portal.unige.ch/) which provides an easy-to-use wizard style interface, implemented in PHP. The third layer, in development, is the *social network layer*. As the name suggests, this will provide a top level which organises LARA users and content in a social network inspired way, with home pages for each user and piece of content and the expected links between them. In this paper, since we are focusing on the content creation process, we will mostly consider relevant parts of LARA's portal level functionality.

Figure 1 above shows an example of LARA content, a page taken from the French edition of *Le petit prince*, and illustrates the main types of annotations supported by LARA: audio, translations and multi-words. We now briefly describe how these annotations are created. Full details can be found in the online documentation (Rayner et al, 2021).

2.1. Portal top level

The portal top level provides easy access to the other types of functionality offered. The greater part of this involves the content creation and editing process, relevant aspects of which are described in sections 2.2-2.5. §3 describes the new crowdsour-

cing mechanism, which is being used for the first time in the project presented here.

2.2. Lemma tagging

The first step in the annotation process is to process the text so as to tag each word with its associated lemma. This can be done by hand, but for most of the languages we work with[3] is in practice performed by a tagger/lemmatiser, followed by manual post-editing of the automatically added tags. The LARA portal has several tagger/lemmatisers integrated so that the initial automatic tagging step can be invoked with a single click. Our experience shows that the performance of modern tagger/lemmatiser packages for this kind of task varies a great deal. In good cases (well-resourced languages like English and French; texts similar to those used to train the resources) error rates are in low single digits, and an experienced editor can carry out the post-editing task at a rate of several thousand words per hour (Akhlaghi et al, 2020). In less favorable cases, some of which are described in this paper, error rates can be a good deal higher. An example of tagged LARA text is shown in Figure 2.

```
CHAPITRE X||
Il se trouvait#trouver# dans la#le# région des#du# asté-
roïdes#astéroïde# 325, 326, 327, 328, 329 et 330.|| Il
commença#commencer# donc par les#le# visiter pour y cher-
cher une#un# occupation et pour s'#se#|instruire.||
La#le# première#premier# était#être# habitée#habiter# par
un roi.|| Le roi siégeait#siéger#, habillé#habiller# de
pourpre et d'#de#|hermine, sur un trône très simple et
cependant majestueux.||
```

Figure 2

Example of tagged LARA text automatically produced by an integrated tagger/lemmatiser. Lemma tags are delimited with # … #, segments with ||, compound words with |.

[3] LARA documents have been created for some languages where no tagger/lemmatiser exists.

2.3. Audio recording

Audio recording is performed using the LiteDevTools online recording tool (LDT; Akhlaghi et al, 2019). Once the lemma tagging step is complete, the LARA portal automatically uploads recording scripts for segments and words to LDT, where they can be accessed by the voice talent. The simple and intuitive interface is shown in Figure 3.

Audio can also be created using an integrated TTS engine; the portal currently supports ReadSpeaker, Google TTS and ABAIR for Irish (ABAIR 2021). Comparison between human and TTS audio suggests that the gap between them is now quite narrow. In a recent empirical study we carried out, subjects preferred TTS to some of the human voices used (Akhlaghi et al 2021).

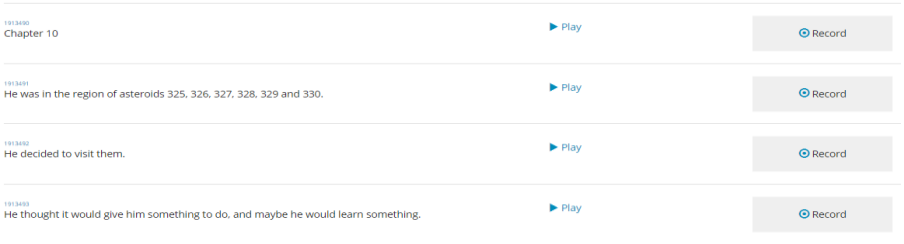

Figure 3

Recording audio with the online LDT tool. The voice talent clicks 'Record', reads the text on the left, and clicks again to stop. They can listen to the recorded audio using the 'Play' button and, if necessary, rerecord it. Items can be attempted in any order, and the voice talent is free to log out and return later. If the source text is edited in the LARA portal, the recording tasks are automatically updated to reflect this.

2.4. Translations

Sentence-level (more properly, segment-level) translations are entered using another simple portal interface, illustrated in Figure 4.

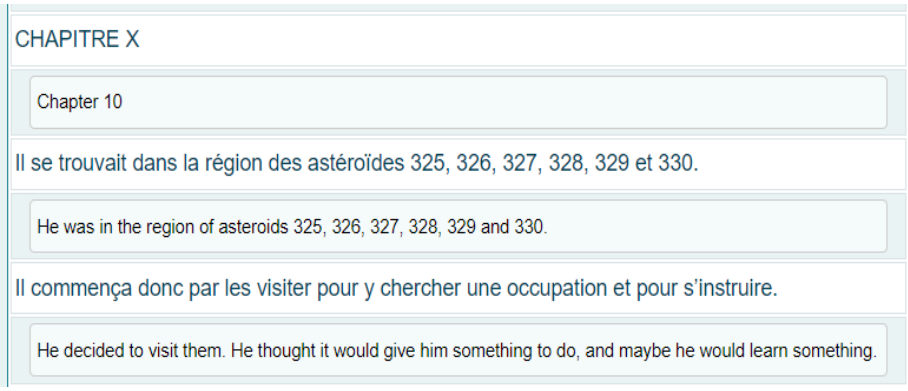

Figure 4

Segment-level translation interface. The portal presents the segments in the reading language, the annotator fills in the translations in the student language. Segments are presented in the order in which they occur in the text.

Word translations are created in two stages. First, the annotator fills in translations for word types, using an interface similar to the one used for segments. This is shown in Figure 5. Word type translations are shared between different texts as a common resource.

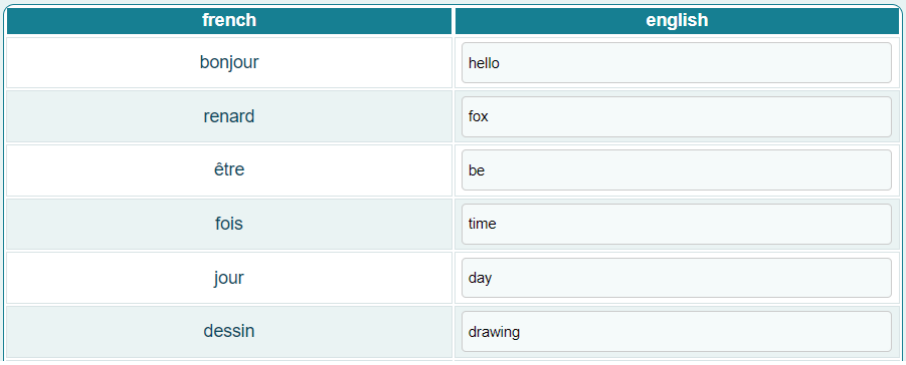

Figure 5

Word type translation interface. The reading language words are presented on the left, the annotator fills in the student language words on the right. Hovering over one of the words on the left shows a pop-up with a list of occurrences of the word in the current text.

The word type translations are then used to populate a set of pages ("word token translations") which allow the annotator to change the translation of a word in a specific context. An example of a word token translation page is shown in Figure 6.

Figure 6

Word token translation interface. Lines are presented in groups of three. The first line is a sentence from the text, the third is the sentence translation, and the annotator fills in the word glosses in the middle. Initial values are set from the type translations. Multiword expressions are highlighted in red.

2.5. Multi-word expressions

Multi-word expressions (MWEs) are common in most languages. We will use this term in a broad sense, to refer to any sequence of words, for example a phrasal verb, which intuitively should be treated as a unit. An occurrence of an MWE often, though by no means always, consists of a contiguous set of words. Thus, for example, the continuous sequence "just about" in "I had **just about** enough water for eight days" is treated as an MWE, and similarly the discontiguous sequence "looked … like" in "They **looked** exactly **like** his own flower" is treated as a form of the MWE "look like".

LARA provides a systematic mechanism for annotating MWEs. Each language includes an MWE lexicon, which in general has been compiled partly from LARA texts and partly from

sources on the web. An entry in the MWE lexicon is a list of words, each of which can be either a surface word (the word must appear exactly as listed in the MWE entry), or a lemma (the word has the named lemma tag). In the notation used, lemmas are distinguished by adding asterisks around them. Thus, in the two examples above, the lexicon entry for "just about" is

```
just about
```

while the lexicon entry for "look like" is

```
*look* like
```

Other features make it possible to define lexicon entries for MWEs like reflexive verbs, by introducing the possibility of defining a lexicon token to represent a class of words. Thus, for example, in the French MWE lexicon, the token se is defined to represent the class of reflexive pronouns, which means that a lexicon entry like

```
se *coucher*
```

("oneself lie-down", "go to bed") matches any of the sequences *me couche* (1st person singular), *vous couchez* (2nd person plural), etc. A full description of the MWE lexicon notation can be found in the online documentation but is somewhat beside the point here, since the annotator does not see the lexicon infrastructure. All they see is a list of possible MWE occurrences, presented using the interface shown in Figure 7. Their task is to mark each occurrence as genuine or spurious.

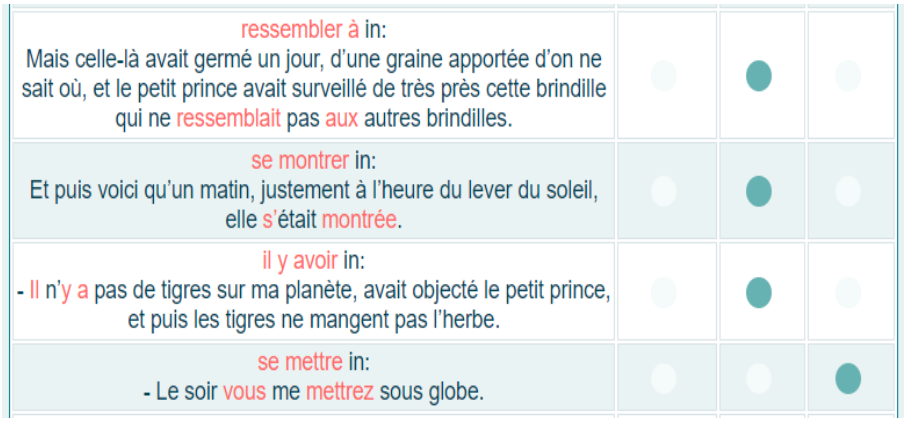

Figure 7

Interface for marking occurrences of multiword expressions. The candidates are ordered by the number of words skipped, with contiguous occurrences first. In the screenshot above, the first three candidates are marked as genuine occurrences of MWEs, the fourth as spurious.

3. Adding support for crowdsourcing in LARA

The functionality offered by the LARA crowdsourcing interface is straightforward and intuitive. When creating a LARA project, the developer may optionally divide the text into a number of sections by inserting separator tags. If they have done this, they can then use a dashboard control to *crowdsource* the project. The effect is that a set of subprojects are created, one for each section. Each subproject has the same settings as the original project, except that it contains only the relevant portion of the corpus.

The crowdsourced subprojects are made visible under the *Available Tasks* tab of the crowdsourcing dashboard, which is accessible from the portal top level. Any user who wishes can *pick up* a subproject from the Available Tasks tab. This transfers the subproject to its new owner, who is then free to add any of the types of annotations described under §§2.2-2.5 above. However, the owner of the original project (the *task requester*) can still view the crowdsourced subtask. They can also, at any time, use another dashboard control to *collect* the current versions of

the subprojects. As the name suggests, this means that the texts and annotations currently present in the subprojects are combined back into the parent project, so that the task requester can obtain an up-to-date snapshot of the project as a whole. The task requester and the crowdworker can communicate through the crowdsourcing dashboard by updating the message thread attached to each subproject. It is possible for the crowdworker to *give back* a task they no longer wish to do, and for the task requester to *take back* a task if they feel insufficient progress is being made.

A detailed description of the crowdsourcing functionality can be found in the online documentation[4].

Table 1

Status of "LARA Little Prince" projects by chapter for the nine languages currently finished or under development, as of May 18 2021. Languages are ordered by completeness.

Language	Chs	Tagging corrected	Audio		Translations	
			Sents	Words	Sents	Words
French	27	All	All (Human)	All (TTS)	All	All
English	27	All	All (Human)	All (TTS)	All	All
Italian	27	All	All (Human)	All (Human)	All	No
Farsi	27	All	26 chs (Human)	All (Human and TTS)	25 chs	22 chs
Icelandic	27	All	All (TTS)	All (TTS)	No	No
Polish	18	18 chs	18 chs (Human)	18 chs (Human)	18 chs	18 chs

[4] https://www.issco.unige.ch/en/research/projects/callector/LARADoc/build/html/crowdsourcing.html

Japanese	27	3 chs	All (TTS)	All (TTS)	9 chs	1 ch
Irish	2	2 chs	2 chs (TTS)	2 chs (TTS)	2 chs	2 chs
Mandarin	1	1 ch	1 ch (Human)	1 ch (Human)	1 ch	1 ch

4. Creating LARA versions of The Little Prince in different languages

In this section, we describe how we created LARA versions of *The Little Prince* in English, Farsi, French, Icelandic, Italian, Irish, Japanese, Mandarin and Polish. The work involved was of two kinds. In all cases, texts were annotated using the methods described in Sections 2 and 3 to add embedded audio and translations, and tag occurrences of multiword expressions. For some languages, we first needed to extend the platform's support for the language in question, most often by adding or improving automatic tagging capabilities. Links to the LARA versions, in many cases still ongoing, are posted on the LARA examples page[5]; a summary of the current state of play, as of May 2022 is presented in Table 1 above. We group the languages by the type of work involved.

4.1. French, English, Icelandic and Irish

These four languages have been used extensively since the start of the LARA project, and no further work was required on infrastructure. English and French perform tagging and lemmatization using TreeTagger with the appropriate packages (Schmidt 1994); Icelandic uses a combination of the Greynir, Nefnir and ABLtagger packages (Steingrímsson 2019, Akhlaghi et al 2020). In all four languages, work was organised so that one person was responsible for post-editing tagging, adding translations, and marking MWEs. For French and English, audio was recor-

[5] https://www.unige.ch/callector/lara-content

ded by a child/young teen (12-year-old boy for French; 15-year-old girl for English) paid at the local babysitter rate. For Icelandic and Irish, we used TTS engines, ReadSpeaker for Icelandic and ABAIR for Irish.

4.2. Italian

For Italian, we also used TreeTagger for tagging and lemmatization, but some work was needed to handle postverbal clitics. Italian orthography affixes these directly to the verb (*masticarla* = *masticar-la* = "chew it"; *spiegargliele* = *spiegar-glie-le* = "explain them to him"). However, we found that TreeTagger only tags words of this kind with the lemma for the verb, discarding the clitics. We corrected by adding post-processing code which operates on words that have been tagged as verbs of the relevant types and end in affixes which match legal sequences of postverbal clitics.

This time, annotation work was divided up so that correction of tagging, and marking of MWEs, was performed by two people, one native and one near-native in Italian, while a third person added translations. As with French and English, audio was recorded by a young teen voice, this time a 14-year-old girl.

4.3. Polish

When we started work on the Polish edition, we soon determined that performance for the TreeTagger Polish package was very poor. As this is also the case for the Russian and Slovak packages, it may be that the TreeTagger architecture is not well suited to the highly inflected Slavic languages. We looked around for alternatives, and found that the combination of the Morfeusz2 lemmatiser (Woliński 2017) and Concraft tagger was recommended as giving state-of-the-art performance. As in the case of the Italian TreeTagger package, integrating the Morfeusz2/Concraft combination was not entirely trivial. For complex reasons having to do with the grammar of Polish, past tense verb affixes are analysed by Morfeusz2 as separate morphemes,

but this approach is not appropriate for LARA, which needs to analyse a verb as a single word. After some discussion with the Morfeusz team, who were helpful in explaining the issues, we found that we could solve the problem in post-processing by joining together verbs and their associated past tense morphemes.

Departing from the recipe we had used for the five languages above, we experimented with a new one for Polish. The text was manually divided into chapters and distributed among the members of a course which the Polish member of the consortium was teaching. Each student was responsible for all aspects of annotation for their chapter. At the end, a script was used to recombine the annotations from the multiple users into a single LARA text, following which we did a little post-editing to add basic markup for headings, images and division into pages.

This experiment was not entirely successful as there were only enough students to assign 20 out of 27 chapters, and some students did not adequately complete their assignments. The majority, however, did their tasks well, giving us valuable information to suggest that the crowdsourcing method was a step in the right direction. This motivated the development of the crowdsourcing infrastructure previously described in §3.

4.4. Farsi

Farsi (Persian) is an Indo-European language written with an Arabic script. We perform tagging and lemmatization using the hazm package.[6] A noteworthy property of Farsi is the unusually high frequency of phrasal verbs. This meant that an early priority in developing Farsi resources for LARA was to compile a substantial lexicon of phrasal verbs compatible with the LARA MWE processing module (cf. §2.5).

The Farsi version was produced using the automatic crowdsourcing method described in §3. The tagging phase was first carried out by one of the authors, a Farsi native speaker with

[6] https://pypi.org/project/hazm/0.3/

long-time experience of using LARA, after which the text was divided into 27 chapters and put on the crowdsourcing platform. Students enrolled in a course taught by the author who performed the tagging were invited to pick up chapters and add human audio and translations. They were told that completion of the assignment within a three-week window would be rewarded with a 10 % bonus credit on the course.

16 students completed their chapter and 10 more returned a partial result. Missing word audio, the most common missing part, was added using the ReadSpeaker integration of the Ariana TTS engine.[7]

This exercise was the first large-scale test of the crowdsourcing mechanism from §3. It performed correctly, with no obvious technical faults, but several of the students complained that the portal was not user-friendly enough for casual users who had no previous experience with it.

4.5. Mandarin

Chinese, the first non-Indo-European language of the ones we used, posed special problems at the level of infrastructure. The highly analytical nature of Chinese grammar means that lemmatization as such is not required. This is replaced by the well-known Chinese segmentation problem. Since standard Chinese orthography does not mark spaces between words, the first task in text processing is typically to segment the input. After experimenting with a couple of alternatives, we decided to use the popular Jieba package[8], which was easy to integrate. It was also important to provide word glosses in pinyin (Roman alphabet), the pedagogical value of which is well attested. There are many sites on the web which provide accurate pinyin conversion tools. We integrated the one from chineseconverter.com.[9]

[7] http://www.farsireader.com/english/

[8] https://github.com/fxsjy/jieba

[9] https://www.chineseconverter.com/en/convert/chinese-to-pinyin

The above issues make the Chinese workflow slightly different from the one used for the other languages. Instead of correcting an automatically tagged version of the text, the annotator starts by correcting an automatically segmented version; thus they edit segment boundaries rather than lemma tags. Figure 8 illustrates.

As of May 2022, one sample chapter has been produced for Mandarin.

当|我|还|只有|六岁|的|时候|，在|一本|描写|原始森林|的|名叫|《真实|的|故事|》的|书|中|，看到|了|一副|精彩|的|插画|，画|的|是|一条|蟒蛇|正在|吞食|一只|大|野兽|。||页头|上|就是|那副 画|的|摹本|。||

这本书|中|写道|："这些|蟒蛇|把|它们|的|猎获物|不|加|咀嚼|地|囫囵|吞下|，尔后|就|不能|再|动弹|了|；它们|就|在|长长的|六个月|的|睡眠|中|消化|这些|食物|。"||

当时|，我|对|丛林中|的|奇遇|想得|很多|，于是|，我|也|用|彩色|铅笔|画出|了|我的|第一副|图画|。||我的|第一号|作品|。||它|是|这样|的|：||

我|把|我的|这副杰作|拿给|大人|看|，我|问|他们|我的画|是不是|叫|他们|害怕|。||

他们|回答|我|说|："一顶|帽子|有|什么|可怕|的|？"||

Figure 8

Example of hand-editing Chinese segmentation produced by the Jieba package. As in the case of Figure 2, segment boundaries are marked with double bars, ||, and word boundaries with single bars, |. Red is used to show the changes made by the editor.

4.6. Japanese

Similar to Chinese, Japanese as a non-Indo-European language posed significant issues for LARA. Like Chinese, Japanese is normally written without spaces between words, and segmentation is thus again a substantial problem. Unlike Chinese, however, Japanese is primarily a synthetic language, and its morphosyntax demonstrates inflection of agglutinative roots.

At the moment, LARA performs segmentation, lemmatization and tagging using the Google Cloud Natural Language Morphology and Dependency Trees package for Japanese[10]. Unfortunately our experience with the current version of the package on various texts shows a high error rate, typically more than two errors per sentence. Errors are frequent both in segmentation and in lemmatization. As an example of the former, the common word あいだ (aida, literally "interval" but with the sense of "for") in the typical context ６か月のあいだ (rokkagetsu no aida, "For six months") is segmented as two lexemes, あい (ai "love") and だ (da, copula). With regard to lemmatization, verbs are typically not tagged with the dictionary form: thus, for example, 出して (dashite, "leaving") is tagged as 出し (dashi) instead of the expected 出す (dasu). Some verbs are, however, correctly lemmatised; in particular, the irregular して (shite, doing) is always lemmatised correctly as する (suru). We suspect that these errors may be teething problems in a new package, perhaps caused by very insufficient training data. The task is further complicated by the fact that pedagogical considerations do not always coincide with standard Japanese linguistic conventions. For example, while morphosyntactically and phonologically 本には (hon-ni-wa, "in the book") is realised as a single, inflected unit in Japanese, it is more useful for a learner to see it split up into a content word and two polysemous particles (Masuda 2018: 63–96). For all these reasons, tagged Japanese text so far requires extensive post-editing.

The Japanese version is using the automatic crowdsourcing method described in §3. The workflow is similar to that in the Chinese version, though crowd-workers, for the reasons described above, need to revise both segmentation and tagging. Detailed instructions to the crowd-workers are posted online[11]. Crowd-workers found it natural to divide up work within the

[10] https://cloud.google.com/natural-language/docs/morphology

[11] https://www.issco.unige.ch/en/research/projects/callector/LARAJapaneseLPDoc/build/html/index.html

chapters, with one crowd-worker correcting the segmentation and tagging, handing it to a second crowd-worker who adds the sentence translations, and then taking it back to add the word translations. The LARA platform's crowdsourcing functionality efficiently supports this flexible division of labor. Unfortunately, however, although all technical aspects appear to work correctly, it has so far proved extremely difficult to attract crowd-workers, and only two chapters have been completed.

5. Using the LARA Little Prince in the classroom

Since July 2020, links to the different versions of the LARA Little Prince have been publicly posted. Although our impression is that they have been well received,[12] we wished to get more formal student feedback. With this in mind, we organised the utilization of the French LARA version of *Le petit prince* as part of a low intermediate level French course held at Flinders University, South Australia, using the French LARA version of *Le petit prince*.

5.1. Context/Background

A set of three sessions was devoted to reading the first three chapters in a class of 16 (11F, 5M) A2-B1 level students of French as a foreign language. The students had completed three semesters of study and had spent around 400 hours learning French. Their majors included international relations, law, psychology and education.

The group was fairly homogeneous, motivated and enthusiastic. They were engaged in their studies and generally applied themselves diligently to assignments outside the classroom. Most of them spoke French spontaneously in class.

[12] In particular, a page for the French LARA edition (https://www.goodreads.com/book/show/54923317-le-petit-prince) appears on the popular Goodreads review site. As of November 2022 it has 18 ratings, with an average rating of 4.72.

At this level, students were entitled to 4 hours of weekly lessons, divided into two seminars of two hours each. The first seminar was mostly dedicated to reading and writing, with the introduction of new grammatical points. The second seminar focused on oral comprehension and expression, combining exercises of different types, some using authentic material, and others designed specifically for FFL. Songs and short dictations were also included.

5.2. Process/method

The entire course was based on a textbook (*Entre Nous 2*, Editions Maison des Langues), which provided core material and a point of focus for grammatical and thematic features. Numerous additional documents were offered relating to grammar, linguistic and lexical content, as well as to assist oral and writing skills. In support of reading, the teachers-researchers generally offered short texts to be prepared at home and subsequently discussed in class.

Introducing *The Little Prince* was part of an ongoing process of discovering literary texts; the fact that the book was familiar to many of the students was a facilitating element. Due to time constraints, only the first three chapters were used. The overall aims were to allow students to feel more motivated and more confident in their target language reading skills, help them understand the text, and make them aware of the poetry and the literary effects created by the text.

We expected that using the LARA platform would add value by giving students immediate access to word meanings (both translations and other occurrences of the word in the text, sometimes illustrating contextually different meanings), and also to audio. Audio was made available in human-recorded form for sentences and in TTS voice for words. TTS audio was produced by the ReadSpeaker engine, which at a word level gives quality judged comparable to human audio (Akhlaghi et al 2021).

The two teachers designed a worksheet based on each of the three chapters presented on the LARA platform, whilst maintaining the grammatical points programmed for the semester. Each of the worksheets included comprehension questions on the passage (oral or written) as well as questions relating to grammar or conjugation. The following activities were offered:

1. Activities aimed at working on and checking comprehension

A. An individual activity consisting of listening and reading simultaneously to be discussed in pairs to enable understanding of the chapter (Chapter 1)
B. An individual activity consisting of listening and reading simultaneously where each student individually initially answered a series of questions relating to the literal understanding of the passage. This was followed by a consideration at a deeper level, that is, first literal and then literary – of the text and its effects on the student-reader (Chapter 1 and Chapter 3)
C. A listening and reading activity designed to be done individually with the aim of presenting a brief oral summary of the chapter (Chapter 2 and Chapter 3)
D. Individual listening / reading activity to test comprehension through a multiple-choice response type activity (Chapter 2)

2. Activities aimed at working on pronunciation/prosody and at expanding the lexicon

A. Observing selected words with respect to 1) their pronunciation 2) their specific occurrences in the text. Students were also invited to observe the tenses or moods used in the case of verbs (Chapter 2)
B. Students recorded new vocabulary in their exercise book. (Chapter 2, Chapter 3)
C. An activity aimed at listening and imitating designed to observe and then reproduce the rhythm of the sentence created by the commas (Chapter 2, first paragraph) and

to observe and reproduce the intonation in the dialogue section (“Mais qu’est-ce que tu fais là?”; Chapter 2)

D. Pronunciation activities to work on the contrast between nasal sounds such as (ON [ɔ̃] and AN or open O [ɔ] and closed O [o]) vowels. This type of activity involved clicking on a particular word and repeating the sentence aloud (Chapter 3)

3. Activities aimed at raising awareness of the literary nature of the book

A. Analysis of the gradual unveiling of the Little Prince’s character, both from the textual and visual point of view, as well as a question asking the students to put themselves in the shoes of a child reader discovering the image in question of the Little Prince (Week 2, Chapter 2)

B. Activities based on supporting documents about the text and the author. These documents comprised an audio biography from the TV5 Monde platform, concise iconographic presentations of the text and the author’s life, an article about the Antoine de Saint-Exupéry Foundation, a tribute to the author by the French Air Force, and an interview with the son of Léon Werth, the person to whom the book is dedicated.[13]

5.3. Questionnaire and responses

In the week immediately following the conclusion of the experiment, the students were asked to fill out an English language questionnaire (see the appendix). The ten responses were analysed following Patton (2002). We organise the material under five headings:

[13] https://www.youtube.com/watch?app=desktop&v=Veq_K08R1H8&ab_channel=QuelleHistoire, https://ticsenfle.blogspot.com/2014/12/le-petit-prince-ressources-lecture.html, https://www.fondationdefrance.org/fr/fondation/fondation-antoine-de-saint-exupery-pour-la-jeunesse, https://perelafouine.com/la-patrouille-de-france-rendra-hommage-a-antoine-de-saint-exupery-et-son-petit-prince/, https://www.youtube.com/watch?v=O5Z-9tyEZ4I&ab_channel=SaintExuperyTV.

1. Previous knowledge of Le Petit Prince
2. Practising reading with LARA
3. Practising pronunciation with LARA
4. Work on vocabulary with LARA
5. General comments on the approach and platform, suggestions

Previous knowledge of Le Petit Prince
3 of the 10 students knew the book before the experiment.

Practising reading with LARA
Several learners highlighted how their comprehension was facilitated by the audio and translation annotations. Student feedback in this and following sections is presented as originally provided without correction.

> "LARA is a great way to read and at the same time listen to the correct pronunciation. It is great that the meaning of each word can easily be seen (the translation in English; that makes it much easier and more fun as I don't have to look up each unknown word in the dictionary" (student A)
> "[reading was] very good. Was helpful to understand pronunciation and learning new vocabulary and sayings" (student C)
> "I thought it was very useful because it exposed the class to a classic French story. Reading the text as a class also improved my reading confidence" (student D)
> "Listening to the platform, read the text and then reading it through as a class and individual helped my pronunciation and fluency of reading" (student F)
> "I thought it was useful that we could hear the words being spoken out loud as we read it. Additionally, the fact that we could hover over a word to discover its meaning in English was extremely useful, and even enabled me to learn new words" (student G)
> "It worked well. We were able to effectively cover everything in an easy understandable way" (student H)
> "I enjoyed being able to read along with the speaker and being able to find out what the individual words mean. Overall I think it was very successful" (student J)

Practising pronunciation with LARA

The students appreciated being able to listen and repeat the words and phrases, imitating what they heard:

> "It was very helpful to listen to the pronunciation of certain words and to repeat it" (student A)
> "it was good as we went through the correct pronunciation while we were reading (…) audio was great, LARA platform is really helpful, pronunciation tasks were good (…) the pronunciation is really helpful because you can hear the liaisons and different tenses" (student B)
> "Good, the fact that each word is provided with two pronunciations (boy and female) in a sentence but also just the word made it very helpful" (student C)
> "studying pronunciation helped me understand the way a word can change in sound with aspects like liaison. Being able to use a platform that also read the words with a click was also very helpful" (student D)
> "Being able to isolate individual sentences and words within the text is useful to improve pronunciation and recognition of liaisons between words for pronunciation (…) a strength is (…) being able to easily repeat sections for practise" (student F)
> "I thought it was useful that we could listen to each passage sentence or word being read aloud by a French speaker, which helped to properly understand how to pronounce the words I hadn't previously known" (student G)
> "this was a good approach allowing us to see many ways each word might be used, allowing us to properly understand how to pronounce it" (student H)
> "Amazing. I had never heard or used LARA before. The use of AI has helped with some pronunciation. I also like the capacity to repeat sections of the text" (student I)
> "I thought that it was very effective to be able to hear and repeat the words both on their own and in sentences (…) it helped to be able to listen and go word by word pronunciation wise" (student J)

Work on vocabulary with LARA

Students found the LARA platform helped with the expansion of their vocabulary, many mentioning the large number of new words and idioms they discovered:

"I learned new words and expressions (…) by reading the text, being able to see the pronunciation of unknown words and by answering questions on the worksheet regarding the text and also by looking at single words or expressions (…) I think it is a great way to learn vocabulary. It also helped me write down new words" (student A)
"I thought that this was really helpful as the LARA platform went through all the scenarios/tenses of certain words (…) I thought this went well, however, it would be more helpful if we translated the text more to get a better understanding" (student B)
"Very good. Went over new vocabulary and the platform also tells you the meaning of each word/phrase so you can easily understand" (student C)
"I think I have gained more French vocab since studying the text and also been exposed to more day-to-day French phrases" (student D)
"Studying the vocabulary used in the text and being able to isolate words to see how they are used in various formats is very helpful to build my vocabulary as well as my ability to use it" (student F)
"I thought it was useful to be able to translate each sentence, especially when the sentence was idiomatic" (student G)
"… gave us lots of opportunities to practise and perfect words (…) hovering over a word showed the translation and helped a lot (…) exact translation of paragraphs are sometimes unclear" (student H)
"This was a bit of a 'curve ball' for me as I (didn't) consider it a real learning objective. However, I now appreciate some of its value" (student I)

General comments on the approach and platform, suggestions
There were several positive comments about LARA's general flexibility/versatility and its ability to combine with relevant pedagogical activities:

"If class missed you can still go over on your own for comprehension and writing" (student C)
"Another strength is that we can access LARA from home to practise" (student D)
"A strength is accessing it from home" (student F)
"Flexibility and ease of repetition outside the class environment has allowed me to listen, read, speak and write more fluently" (student I)

> "When reading a book or text where there are several words I don't know, it can always be a bit difficult to remember all the new vocabulary. Anyways, by answering questions, reading and listening to the text (maybe several times), we really engage with the text and it makes understanding, remembering and learning easier and also interesting" (student A)
>
> "… was very helpful (…) also for forming sentences" (student C)
>
> "I think LARA is a great platform as it is very accessible, simple and provides a lot of helpful tools to break down French vocab, pronunciation and listening (…) I'd like to read more French classics on LARA as I find it very easy to navigate" (student D)

A few comments were, however, more critical of the pedagogical approach, LARA, or both:

> "LARA is great the way it is. More translation activities in class would be beneficial for a better understanding" (student B)
>
> "The use of the LARA platform could have been better (…) The approach is good for teaching pronunciation and vocabulary but it is difficult for people who struggle with reading and comprehension (…) I would have preferred if the chapters were read aloud in class rather than having to read them at home" (student E)
>
> "An improvement of the platform could be including the form of grammar when you click on a word" (student F)
>
> "Exact translations of paragraphs are sometimes unclear (…) may be have translations of paragraphs or sentences available" (student H)
>
> "The 'click' sound at the end of each sentence can be annoying" (student I)
>
> "Perhaps single subject options (1 or 2 pages) of a journal, contemporary news story etc. that can be played and discussed to create opinions and differing points of view" (student I)
>
> "A weakness is either listening to one sentence or the whole thing and not be able to listen paragraph by paragraph" (student J)

Summary of the questionnaire responses

The high proportion of positive responses suggests the approach using LARA was appreciated by students. In particular, the ability to hover over a word to discover its meaning in English and hear how it is pronounced was evidently considered very useful.

Several students indicated that they would have liked to be able to listen to full paragraphs when working with LARA and that more time needed to be spent in class on translation and ensuring that the content was properly understood by the group as a whole. The general favorable feedback invites continued use of LARA. Next year's course may include either an extended reading of the whole of *Le Petit Prince*, or allow students to discover another French literary work available on LARA.

6. Conclusions and further directions

We have given an overview of the "LARA Little Prince" project, an informal multinational collaboration aimed at producing annotated multimodal editions of *Le petit prince* in many languages using the open source LARA platform. Each language, in general, poses its own special problems; but we have found that these can be solved, and the project is making steady progress. The first few languages are now at the point where the online texts can be used as learning resources. As reported in §5, initial classroom feedback is encouraging.

A novel aspect of the project has been the introduction of a crowdsourcing method to divide up the tasks for a given language between multiple workers. The method was used for Polish and Farsi with partial success. About two-thirds of the chapters were completed in both languages, though the quality was variable. For Japanese, the crowdsourcing method was unfortunately not successful at all. The tasks were posted on an online Japanese forum where one of the authors was an active member, but despite verbal expressions of interest no one picked them up. The sample is small, but one cannot help remarking on the fact that the Polish and Farsi crowd-workers were students who performed the tasks in the context of courses where they would receive extra credit for completing their assignments, while the prospective Japanese crowd-workers had no similar incentive. It seems to be important to think more about issues of motivation and quality control.

We hope to explore these ideas further. People who would like to take part in the project, either in the existing languages or starting new ones, are very welcome to contact us.

Acknowledgements

Work at the University of Geneva was funded by the Swiss National Science Foundation under projects IZCOZ0_177065 and 100015_204503. We gratefully acknowledge ReadSpeaker's help in making their TTS software available to the LARA project.

Appendix: text of the student questionnaire

1) Did you read or study *Le Petit Prince* prior to this semester? At school? At home? In French or English?
2) What do you think of the way we practised *reading* in the first three chapters of Le Petit Prince class?
3) What do you think of the way we studied *pronunciation* in the first three chapters of Le Petit Prince?
4) What do you think of the way we worked on *vocabulary* in the first three chapters of Le Petit Prince?
5) What were the strengths/the weaknesses of this approach?
6) Do you think that using the LARA platform helped you? Why?
7) Did you have a look at the LARA recordings in other languages? If yes, which ones? What did you think of them?
8) Do you think we made good use of the LARA platform?
9) What else could we do or what could we do differently?

References

ABAIR: An Sintéiseoir Gaeilge – The Irish Language Synthesiser ABAIR (2021). http://www.abair.ie. Accessed 20 July 2022.

Akhlaghi, Elham, Branislav Bédi, Matt Butterweck, Cathy Chua, Johanna Gerlach, Hanieh Habibi, Junta Ikeda, Manny Rayner,

Sabina Sestigiani, Ghil'ad Zuckermann (2019). “Overview of LARA: A learning and reading assistant”. In: *Proceedings of SLaTE 2019*, 99-103.

Akhlaghi, Elham, Branislav Bédi,, Fatih Bektaş, Harald Berthelsen, Matt Butterweck, Cathy Chua, Catia Cucchiarini, Gülşen Eryiğit, Johanna Gerlach, Hanieh Habibi, Neasa Ní Chiaráin, Manny Rayner, Steinþór Steingrímsson, Helmer Strik (2020). “Constructing multimodal language learner texts using LARA: Experiences with nine languages”. In: *Proceeding of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 323-331.

Akhlaghi, Elham, Anna Bączkowska, Harald Berthelsen, Branislav Bédi, Cathy Chua, Catia Cucchiarini, Hanieh Habibi, Ivana Horváthová, Pernille Hvalsøe, Roy Lotz, Christèle Maizonniaux, Neasa Ní Chiaráin, Manny Rayner, Nikos Tsourakis, Chunlin Yao (2021). “Assessing the quality of TTS audio in the LARA learning-by-reading platform”. In: Naouel Zoghlami, Cédric Brudermann, Muriel Grosbois, Linda Bradley, Sylvie Thouësny (eds.). *CALL and professionalisation: short papers from EUROCALL 2021*, 1-5.

Bédi, Branislav, Matt Butterweck, Cathy Chua, Johanna Gerlach, Birgitta Björg Guðmarsdóttir, Hanieh Habibi, Bjartur Örn Jónsson, Manny Rayner, Sigurður Vigfússon (2020). “LARA: An extensible open source platform for learning languages by reading”. In: Karen-Margrete Frederiksen, Sanne Larsen, Linda Bradley, Sylvie Thouësny (eds.). *CALL for widening participation: short papers from EUROCALL 2020*, 27-35.

Bédi, Branislav, Haraldur Bernharðsson, Cathy Chua, Birgitta Björg Guðmarsdóttir, Hanieh Habibi, Manny Rayner (2020). “Constructing an interactive Old Norse text with LARA”. In: Karen-Margrete Frederiksen, Sanne Larsen, Linda Bradley, Sylvie Thouësny (eds.). *CALL for widening participation: short papers from EUROCALL 2020*, 20-26.

Butterweck, Matt, Cathy Chua, Hanieh Habibi, Manny Rayner, Ghil’ad Zuckermann (2019). “Easy construction of multimedia online language textbooks and linguistics papers with LARA”. In: *Proceedings of ICERI 2019*, 7302-7310.

Lyding, Verena, Lionel Nicolas, Branislav Bédi, Karen Fort (2018). “Introducing the European NETwork for Combining Language Learning and Crowdsourcing Techniques (enetCollect)”. In: Peppi Taalas, Juha Jalkanen, Linda Bradley, Sylvie Thouësny (eds.). *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018*, 176-181.

Masuda, Kyoko (2018). *Cognitive Linguistics and Japanese Pedagogy: A Usage-based Approach to Language Learning and Instruction*. De Gruyter.

Rayner, Manny, Matt Butterweck, Hanieh Habibi, Cathy Chua (2021). Constructing LARA Content. Online documentation. https://www.issco.unige.ch/en/research/projects/callector/LARADoc/build/html/index.html

Schmid, Helmut (1994) "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Proceedings of the International Conference on New Methods in Language Processing*, 154-162.

Steingrímsson, Steinþór, Örvar Kárason, Hrafn Loftsson (2019). "Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019*, 1161-1168.

Elham Akhlaghi
ORCID iD: 0000-0002-4635-7780
Ferdowsi University of Mashhad
Azadi Square, Mashdad
Rhazavi Khorasan Province
Iran
elhamakhlaghi80@gmail.com

Anna Bączkowska
ORCID iD: 0000-0002-0147-2718
University of Gdańsk
Instytut Anglistyki i Amerykanistyki
ul. Wita Stwosza 51
80-308 Gdańsk
Poland
anna.baczkowska@ug.edu.pl

Branislav Bédi
ORCID iD: 0000-0001-7637-8737
The Árni Magnússon Institute
for Icelandic Studies

Skrifstofa Árnagarði við Suðurgötu
102 Reykjavík
Iceland
branislav.bedi@arnastofnun.is

Hakeem Beedar
Independent scholar
Adelaide
Australia
hbeedar@hotmail.com.au

Cathy Chua
ORCID iD: 0000-0002-5462-2441
Independent scholar
Adelaide
Australia
cathyc@pioneerbooks.com.au

Catia Cucchiarini
ORCID iD: 0000-0001-5908-0824
Centre for Language and Speech Technology (CLST)
Radboud University Nijmegen and
Union for the Dutch Language
Paleisstraat 9
2514 JA Den Haag
The Netherlands
c.cucchiarini@let.ru.nl

Hanieh Habibi
ORCID iD: 0000-0001-7602-2933
FTI/TIM
University of Geneva
40, Boulevard du Pont-d'Arve
1211 Geneve 4
Switzerland
hanieh.habibi@unige.ch

Ivana Horváthová
ORCID iD: 0000-0002-8388-7692
Constantine the Philosopher University
Department of English and
American Studies
Štefánikova tr. 67
Faculty of Arts, CPU
949 01 Nitra
Slovakia
ihorvathova@ukf.sk

Christèle Maizonniaux
ORCID iD: 0000-0003-2844-1820
Flinders University
College of Humanities, Arts and
Social Sciences
SSN3105, GPO Box 2100
Adelaide 5001
Australia
christele.maizonniaux@flinders.edu.au

Neasa Ní Chiaráin
ORCID iD: 0000-0002-4669-5667
Trinity College Dublin
Room 4074A Arts Building
College Green
Dublin
Ireland
neasa.nichiarain@tcd.ie

Nathalie Paterson
University of Adelaide
Faculty of Arts, Business, Law and
Economics (ABLE)
SA 5005 Adelaide
Australia
npa10716@bigpond.net.au

Chadi Raheb
ORCID iD: 0000-0002-5731-0443
University of Guilan

1841, Khalij Fars Highway
Rasht, Gullian
Iran
chadi.raheb@gmail.com

Manny Rayner
ORCID iD: 0000-0002-6002-2434
FTI/TIM
University of Geneva
40, Boulevard du Pont-d'Arve
1211 Geneve 4
Switzerland
emmanuel.rayner@unige.ch

Chunlin Yao
ORCID iD: 0000-0001-6006-5031
Tianjin Chengjian University,
No. 26 Jinjing Road, Xiqing
Tianjin
China
yao_chunlin@126.com