Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 323–331
Marseille, 11–16 May 2020

Constructing Multimodal Language Learner Texts Using LARA: Experiences with Nine Languages

**Elham Akhlaghi[1], Branislav Bédi[2], Fatih Bektaş[3], Harald Berthelsen[4], Matthias Butterweck[5]
Cathy Chua[6], Catia Cucchiarini[7], Gülşen Eryiğit[3], Johanna Gerlach[8], Hanieh Habibi[8]
Neasa Ní Chiaráin[4], Manny Rayner[8], Steinþór Steingrímsson[2], Helmer Strik[7]**

[1]Ferdowsi University of Mashhad, Iran; [2]The Árni Magnússon Institute for Icelandic Studies, Iceland;
[3]Istanbul Technical University (ITU), Turkey, [4]Trinity College, Dublin, Ireland;
[5]Independent scholar; [6]University of Adelaide, Australia; [7]Centre for Language and Speech Technology (CLST),
Radboud University Nijmegen, The Netherlands; [8]FTI/TIM, University of Geneva, Switzerland
elham.akhlaghi@mail.um.ac.ir, branislav.bedi@arnastofnun.is, bektas18@itu.edu.tr, berthelh@tcd.ie,
matthias@butterweck.de, Cathy.Chua@adelaide.edu.au, c.cucchiarini@let.ru.nl, gulsenc@itu.edu.tr,
Johanna.Gerlach@unige.ch, hanieh.habibi@unige.ch, Neasa.NiChiarain@tcd.ie, Emmanuel.Rayner@unige.ch,
steinthor@arnastofnun.is, w.strik@let.ru.nl

Abstract

LARA (Learning and Reading Assistant) is an open source platform whose purpose is to support easy conversion of plain texts into multimodal online versions suitable for use by language learners. This involves semi-automatically tagging the text, adding other annotations and recording audio. The platform is suitable for creating texts in multiple languages via crowdsourcing techniques that can be used for teaching a language via reading and listening. We present results of initial experiments by various collaborators where we measure the time required to produce substantial LARA resources, up to the length of short novels, in Dutch, English, Farsi, French, German, Icelandic, Irish, Swedish and Turkish. The first results are encouraging. Although there are some startup problems, the conversion task seems manageable for the languages tested so far. The resulting enriched texts are posted online and are freely available in both source and compiled form.

Keywords: CALL, reading, tagging, open source

1. Introduction

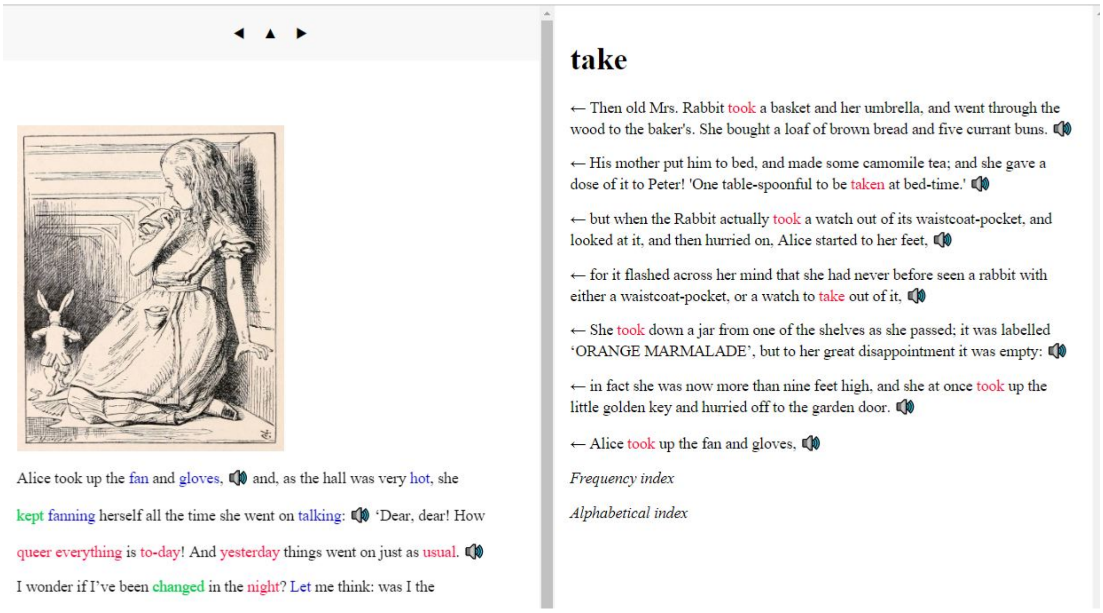
LARA (Learning and Reading Assistant; https://www.unige.ch/callector/lara/; (Akhlaghi et al., 2019)) is an open source project, initiated during Q3 2018. The general goal is to develop methods for creating online resources that support language learning through reading and and listening. It is uncontroversial that reading and listening practice are important for L2 learning (Grabe and Stoller, 2012), and several platforms now exist that facilitate conversion of texts into online versions tailored to the L2 learner; prominent examples include Learning With Texts[1], Alpheois[2] and Clilstore[3]. LARA, which uses a crowdsourcing/online community approach where content creators and content users interact in a shared online environment, adapts and extends many of the ideas used in these earlier platforms and adds new ones. In particular, LARA texts are organised so that, when the user accesses them through the online portal (Habibi, 2019), a personalised concordance is built up which associates each word in the text with previous occurrences in the learner's own reading history. Other core functionality includes linking words to audio recordings and translations. When the learner sees a word in a LARA text they are reading, they are thus always in a position to find out where they have seen the word before, what it sounds like, and what it means. An example is shown in Figure 1.

In previous papers, (Akhlaghi et al., 2019; Bédi et al., 2019), we have described LARA's functionality, shown examples of multimodal texts produced by LARA, and presented evidence that learners like it and find it useful. The next question is to determine more precisely how much work is involved in building a LARA resource: the goal of this paper is to provide an initial answer. In a series of experiments carried out in October and November 2019, we took substantial texts, up to the length of short novels, and tried to determine quantitatively what we needed to do to "LARA-ify" them. The process consisted of three main steps, of which only the first was obligatory: 1) manually correcting automatically produced markup added to the source text; 2) adding audio recordings and 3) adding translations. We report results for nine languages; some of them still pose problems, but for others, including widely spoken languages such as English and French, the conversion task already seemed manageable, with annotation effort on the order of thousands of words per hour for novel-length texts. The rest of the paper is organised as follows. §2. describes LARA in more detail. §3. and §4. summarise the relevant external resources and the texts used. §5., §6. and §7. present experiments, results and discussion. In the final section, we briefly outline subsequent work motivated by the experiments and carried out over the period December 2019–February 2020.

2. Overview of LARA

We expand on the sketch of LARA presented in the previous section, describing the LARA platform, the process of

[1] https://sourceforge.net/projects/lwt/
[2] http://alpheios.net/
[3] http://multidict.net/clilstore/



Alice took up the fan and gloves, 🔊 and, as the hall was very hot, she kept fanning herself all the time she went on talking: 🔊 'Dear, dear! How queer everything is to-day! And yesterday things went on just as usual. 🔊 I wonder if I've been changed in the night? Let me think: was I the

take

← Then old Mrs. Rabbit took a basket and her umbrella, and went through the wood to the baker's. She bought a loaf of brown bread and five currant buns. 🔊

← His mother put him to bed, and made some camomile tea; and she gave a dose of it to Peter! 'One table-spoonful to be taken at bed-time.' 🔊

← but when the Rabbit actually took a watch out of its waistcoat-pocket, and looked at it, and then hurried on, Alice started to her feet, 🔊

← for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, 🔊

← She took down a jar from one of the shelves as she passed; it was labelled 'ORANGE MARMALADE', but to her great disappointment it was empty: 🔊

← in fact she was now more than nine feet high, and she at once took up the little golden key and hurried off to the garden door. 🔊

← Alice took up the fan and gloves, 🔊

Frequency index

Alphabetical index

Figure 1: Example (online here) constructed using current LARA prototype showing a page from the personalised reading progress. The learner has read *Peter Rabbit* followed by the first three chapters of *Alice in Wonderland*. The left side shows the marked-up text, where the learner has just clicked on the word "took". The right 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 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.

constructing LARA resources, and the LARA community.

2.1. Platform

The platform consists of two layers, the *core LARA engine* and the *LARA portal*. The core engine, implemented in Python 3, constitutes the code which performs compile-time and runtime backend processing. At compile-time, it converts marked-up text and associated resources (audio files, translations, images etc) into sets of multimedia web pages, and also produces intermediate data which aids the user in performing the conversion process, for example compiling scripts to do audio recording. At runtime, the core engine updates the learner's personalised set of LARA web pages as they add new material to their reading history. The portal, implemented in PHP, wraps the core engine's functionality as a user-friendly mouse-and-menu web interface to support both construction and accessing of LARA resources. The functionality of the core engine and the portal are presented at greater length elsewhere (Akhlaghi et al., 2019; Habibi, 2019); full details are available in the online documentation (Rayner et al., 2019). Both levels make integral use of external software components, in ways we describe immediately below.

2.2. Building LARA resources

Converting a piece of text into a LARA resource involves three main steps:

Annotation: The text is marked up by adding suitable annotations (Figure 2). As can be seen, the greater part of the work consists of adding lemma tags to the words. To do this efficiently, we incorporate morphology analysis resources (taggers and lemmatizers) into the pipeline so that an initial version of the annotated form is produced automatically and then post-edited manually.

Adding translations: LARA allows translations to be attached to both words and segments in LARA texts. In the version of LARA used for these experiments, word-level translations were attached to the lemma, so all different inflected forms of a word received the same translation: thus for example in an English text annotated for French readers, "go", "gone", "going" and "went" will all get the single translation annotation *aller*. We discuss this further in the final section.

In practice, the most common way to create translations is to take the spreadsheet produced by LARA, run the source language column through Google Translate or a similar MT system, and clean up the result by hand.

Audio recording: In contrast to many online reading platforms, which use TTS to add audio, LARA has, for both research and educational reasons, consistently prioritised recorded human audio. Recording has been performed using the LiteDevTools online platform, described in §3.6..

```
MR. JONES, of the Manor Farm, had#have# locked#lock# the hen-|houses#house#
for the night, but was#be# too drunk to remember to shut the popholes#pophole#.||
With the ring of light from his#he# lantern dancing#dance# @from side to side@,
he lurched#lurch# across the yard, kicked#kick# off his#he# boots#boot# at the
back door, drew#draw# himself a last glass of beer from the barrel in the
scullery, and made#make# his#he# way up to bed, where Mrs. Jones was#be#
already snoring#snore#.||
```

Figure 2: Example of annotated LARA text showing tags for lemmas, (`#have#`, `#lock#`), multiword expressions (`@from side to side@`), compound words (`hen-|houses`) and segment boundaries (`||`). It is also possible to use standard HTML annotation to mark italics, boldface, headings, images, tabular layout, etc.

2.3. The LARA community

LARA is a free open source tool, primarily intended to be used by an online community of people interested in building and sharing resources for language learning. There is a close connection to enetCollect[4], a European COST network which links together several hundred people interested in the intersection of CALL and crowdsourcing. enetCollect's sponsorship permitted an initial hands-on LARA workshop in November 2019, which attracted 45 attendees[5]. The project has from the start had a strong focus on ethical issues, with an emphasis on decentralisation and planning for long-term maintainability (Chua et al., 2019; Chua and Rayner, 2019).

There are two main kinds of potential users: content providers who create LARA content, and content users who want to use the content to learn through reading. Content providers have already created multimodal LARA texts for a wide range of different levels of language learners. The texts can be used as a supplement to a traditional language course or made freely available on the Internet.

Although practical language teaching and learning has been the primary motivation for the development of content to date ((Akhlaghi et al., 2019; Bédi et al., 2019)), content has also been developed 'for fun' (in a world of declining reading interest, this is welcome) and even to investigate the possibilities of LARA for making linguistics papers interactive. So far, commercial gain has not been a factor.

3. External software resources

As already noted, the LARA platform requires various external software resources for morphological processing and audio recording. We describe them here.

3.1. TreeTagger and Punkt/NLTK

The default tagging/lemmatisation tool used by LARA is TreeTagger (Schmid, 1999), a popular freeware system that has now been under development for over twenty years. Parameter files for many languages can be downloaded from the TreeTagger home page.[6] In the experiments described here, we used TreeTagger for Dutch, English, French, German and Swedish. The parameter files for the first four languages are well-trained and stable. The Swedish parameter file was only added recently (Q3 2019), and appears less mature than the others.

The interface between LARA and TreeTagger is simple and straightforward. A plain version of the source LARA text is passed to TreeTagger, which returns a list of ⟨surface-word, POS-tag, lemma⟩ triples. LARA uses these to add lemma tags to the words in the text. The default is that the lemma tag consists just of the lemma; the user can optionally specify that it should include POS information as well. The same generic interface has been used, with minor adaptations, to connect LARA to the other morphology resources described in the rest of this section.

The version of the Punkt sentence tokeniser bundled with the Python NLTK package (Kiss and Strunk, 2006; Bird et al., 2009) is used to perform segmentation into sentences, in order to add default `||` annotations (cf. Figure 2).

3.2. Turkish NLP pipeline

The complex morphology of Turkish makes the use of morphological processing tools useful for lemmatization purposes. An automatic morphological analyzer would produce all possible lemmas for a word surface form and hopefully a morphological disambiguator would choose the most probable one in the given context. While creating Turkish resources in LARA, we first make an automatic morphological processing of the texts and provide automatically selected lemmas for each word and then let the annotators to update/correct them if needed.

For the preprocessing stage (namely; sentence splitting, tokenization and lemmatization), we use ITU Turkish NLP Pipeline (Eryiğit, 2014) which provides Turkish specific language processing tools as a web service[7]. The service includes several layers ranging between text normalization, morphological and syntactic analysis as well as their pipelined versions. For the integration with LARA, a new pipelined service (sentence splitter || tokenizer || morphological analyzer || morphological disambiguator) has been offered by the Turkish team.

3.3. Tagging and lemmatizing Icelandic

Traditionally a morphosyntactic tagset containing around 670 tags has been employed for tagging Icelandic. For lemmatization, such a fine-grained tagset is necessary, as detailed analyzis of grammatical function is often the only

[4] https://enetcollect.net

[5] unige.ch/callector/files/3715/7244/0129/LatestVersionOfProgram.pdf

[6] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[7] Turkish NLP web service is available from http://tools.nlp.itu.edu.tr/.

way to disambiguate when a word form from different paradigms can look the same. For lemmatizing Icelandic in LARA we need to run three modules. For sentence splitting and tokenization we use a rule-based tokenizer from the Greynir package [8]. For tagging we use the state-of-the-art ABLtagger (Steingrímsson et al., 2019), which has been shown to surpass the accuracy of other taggers by a substantial margin when working with Icelandic. And finally for lemmatizing Icelandic, Nefnir (Ingólfsdóttir et al., 2019), employing substitution rules and a large morphological database, gives considerably better results than other tools. When creating Icelandic resources in LARA, we use an open web service that provides the above mentioned tools[9]. The web service takes whole untagged texts, tokenizes and splits them into sentences, tags and lemmatizes the texts and returns them to the LARA platform in JSON format.

3.4. Hazm

Morphological processing for Farsi is performed using the hazm package[10]. So far, the LARA wrapper is very basic and only uses the word tokenizer and lemmatizer.

3.5. Irish lemmatizer

The Irish lemmatizer used is Irishfst (Uí Dhonnchadha, 2010). It is available online[11] and in source code[12].

3.6. LiteDevTools

Audio is recorded using Geneva University's LiteDevTools platform[13]. During processing of a text, the LARA core engine creates lists of words and segments to be recorded, and the portal automatically uploads them to LiteDevTools. After the voice talent has recorded the audio, the portal automatically downloads a zipfile of results, and the core engine links the audio files and metadata into the final LARA document. Files can be rerecorded at any time, with the portal doing the necessary bookkeeping.

3.7. Speech Synthesis for Irish

In a process similar to using LiteDevTools to record audio, an external Irish language DNN synthesiser was used to produce the audio for the Irish text. In case of pronunciation error, transcriptions were corrected before synthesis. The voice used was a female speaker of the Munster dialect. The synthesiser is freely available online at abair.ie. For more on the ABAIR initiative see (Ní Chasaide et al., 2017) and on the potential of TTS in CALL see (Ní Chiaráin and Ní Chasaide, 2020).

4. Texts

The participants in the experiment each started by selecting one or more texts that they would convert into LARA form. The choice of text was left to the discretion of the people concerned, except that we agreed on the way in which the texts would be marked up and only to use texts published after 1900. In practice, people chose texts that they liked as works of literature and which were of a length suitable to the amount of time they had available. The texts used for each language were the following:

English George Orwell's *Animal Farm* (1945), one of the most widely read English novels of the 20th century, is in English-speaking countries often set as a middle school text. It contains about 30K words.

French Georges Simenon's *Le chien jaune* (1931) is an early book in the popular Maigret series. It contains about 38K words. The language in the Maigret books is generally considered fairly simple by the standards of literary French.

Swedish *Kallocain* (1940) is a dystopian science-fiction novel by Karin Boye, containing about 55K words. Compared to the Orwell and Simenon books, Boye's language is rich and complex, using a large vocabulary which includes many inventive coinings.

Icelandic *Litli prinsinn* is the Icelandic edition of Antoine de Saint-Exupéry's classic children's story *Le petit prince* (1943). It contains about 16K words. As in the French original, the language is simple and direct.

German *Der Flüchtling: Episode am Genfer See* (1927) is a short story by Stefan Zweig containing about 2.5K words. In contrast to the other texts, *Der Flüchtling* was marked up including POS information in the lemma tags (cf. 3.1.).

Turkish *Nasreddin Hodja Stories* is a collection of short stories about Nasreddin Hodja, a Turkish folk hero who lived in the 1200s. This collection contains about 2.5K words, and the language used is simple and direct.

Farsi *Farsi reader* is a collection of short texts, totalling 1.7K words, which together make up the "reader" section of one volume of the standard textbook used by the Ferdowsi University of Mashhad (FUM) in their intermediate level Farsi as a foreign language course. The six month course is an obligatory requirement for foreign students who wish to study at FUM. The university plans to introduce the LARA version of the textbook as course material during 2020.

Irish *Mar a Baisteadh Fionn* recounts a story about Fionn MacCumhaill, the most important person in the Fenian cycle of Irish mythology. This short piece (1774 surface word tokens) was adapted from an oral collection done in South West Kerry in the 1920s. It retains many of the characteristics of the oral tradition, making it particularly suitable as a text that should be presented in an oral format. Hence the use of Munster TTS.

Dutch For Dutch we LARAfied two short stories. *Aladdin and the Wonderful Lamp*, which also is called *Aladdin and the Magic Lamp* (443 words), is a well-known fairy tale from the *1001 Nights*. This fairy tale probably does not have an Arabic origin, but was added later to the European translations by Antoine Galland in the 18th century. *De jongen en de spreeuw* (410 words) is one of the stories from the 1905 book 'Weet je nog wel van toen?' by Henriette van Noorden with drawings by Albert Hahn.

[8] https://github.com/mideind/Tokenizer

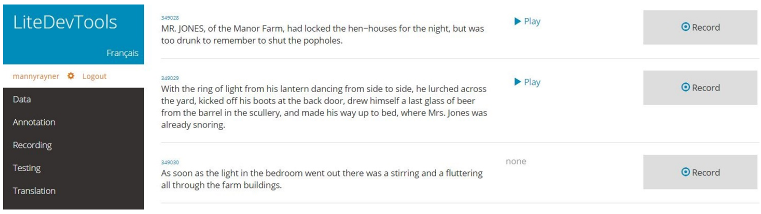
[9] https://malvinnsla.arnastofnun.is

[10] http://www.sobhe.ir/hazm/docs

[11] https://www.scss.tcd.ie/~uidhonne/irish.utf8.htm

[12] https://github.com/uidhonne/irishfst

[13] https://regulus.unige.ch/litedevtools/client



349028
MR. JONES, of the Manor Farm, had locked the hen−houses for the night, but was too drunk to remember to shut the popholes.

349029
With the ring of light from his lantern dancing from side to side, he lurched across the yard, kicked off his boots at the back door, drew himself a last glass of beer from the barrel in the scullery, and made his way up to bed, where Mrs. Jones was already snoring.

349030
As soon as the light in the bedroom went out there was a stirring and a fluttering all through the farm buildings.

none

Figure 3: Using LiteDevTools to record the first few segments of *Animal Farm*. The voice talent hits "Record" to start recording, speaks, and hits it again to finish. The "Play" button is use to review completed recordings. Items can be completed in any order and in multiple sessions.

Table 1: Texts used for experiments. "Lng" = language;"#Seg" = number of segments; "#Tok" = number of surface word tokens; "#Typ" = number of lemma types; "Links" = links to online material; "Src" = tagged and postedited source; "LARA" = compiled LARA pages; "Effort" = person-hours required to perform tasks; "Annot" = tagging cleanup and other post-editing of annotation; "Audio" = recording audio; "Trns" = translation; "Edits" = number of changes made during tagging cleanup; "Word" = surface words changed; "Lemma" = lemma tags changed.

Text	Lng	#Seg	#Tok	#Typ	Links		Effort (person-hours)			Edits	
					Src	LARA	Annot	Audio	Trns.	Word	Lemma
Animal Farm	EN	1168	30201	2983	☞	☞	5.9	9.75	—	301	496
Le chien jaune	FR	2861	37599	3489	☞	☞	7.8	—	—	100	361
Kallocain	SE	3461	56820	4284	☞	☞	22.5	—	8.0	1784	4294
Litli prinsinn	IS	1394	13546	1968	☞	☞	2.3	—	5.7	117	196
Der Flüchtling	DE	151	2455	857	☞	☞	1.6	—	—	0	46
Nasreddin Hodja	TR	359	2558	648	☞	☞	1.6	—	—	107	189
Mar a baisteadh Fionn	GA	119	1774	356	☞	☞	2	<1	3	10	35
Farsi Reader	FA	147	1654	512	☞	☞	5	1.5	4	341	431
Aladdin	NL	17	422	199	☞	☞	1.0	0.3	2.5	19	38
De jongen	NL	30	410	173	☞	☞	0.3	0.3	1.5	8	23

5. Experiments

Each participant completed the "annotation" phase of LARA processing and then optionally performed the "audio recording" and "translation" phrases. Due to time constraints, we only carried out the recording and translation phases for a few texts; previous experience with small texts had led us to believe that they were conceptually simple and similar in nature across different texts and languages. The annotation task, on the other hand, is nontrivial and varies widely in difficulty cross-linguistically, both due to the nature of the languages and the quality of the available morphology resources. In more detail, the three tasks were performed as follows.

5.1. Annotation

LARA provides a number of tools that help support efficient annotation. Here, we used the following pipeline.

1. Process the text through LARA to produce a first cut.

2. Perform a first round of cleaning by manually examining the list of lemmas produced by LARA and searching for items which appear to be inflected words. These words have most likely been incorrectly processed due to not appearing in the lemmatizer's lexicon. LARA's vocabulary list links each word to a list of examples where it appears in the text, and in most cases this immediately suggests what the correct lemma should be. Edit the text to fill in this lemma.

3. Perform a second round of cleaning by examining words which are tagged in more than one way in the text. LARA produces a list of such words, associating each one with the contexts in which it appears. Since these are the words where the lemmatizer is known to have made a choice, they are the ones where it has most likely made a mistake.

4. Perform a third round of cleaning by reading through the entire text checking for multiword expressions (MWEs) and compound words, and marking them where appropriate. A caveat: as discussed in the last two sections, we adopted an overly conservative strategy with respect to MWEs.

5.2. Translation

For some texts, annotators also added translations at the word level. The procedure used was to take the translation spreadsheet produced by the LARA engine, which gives the words in alphabetical order, and fill in the L1 translations for a selected L1. The compiled set of LARA pages contains an alphabetical list of lemmas, with each lemma linked to examples of its occurrences in the text. In general, we found it was often useful to have this list open at the same time.

5.3. Audio recording

Thinking that this would be enough for the current experiment, we carried out careful audio recording of only one large text, *Animal Farm*. In fact the process turned out to involve more choices than we had expected, since we had underestimated the importance of the initial step where we determined the lengths of the segments. The text had originally been segmented by the Punkt sentence tokeniser (cf. §3.1.), but this produced some sentences that were very short, which the person acting as the voice talent thought might result in unnatural prosody. We consequently performed a second pass, where the initial segments were consolidated so that they had a minimum length of 15 words. The average length was 25.9 words, which turned out to be a bit too long for comfort, and most likely slowed down recording significantly. The figure of 9.75 hours to record the whole book (3.1K words/hour) may thus be higher than necessary. The voice talent considered they were fussier and therefore slower than others would find necessary.

6. Results

The overall results are summarised in Table 1. We present the texts used and the amount of time required for the various tasks concerned. For correction of markup, the main focus of this paper, we give the number of edits at the levels of surface words and lemmas. For reasons we examine in the next two sections, the figures should not at this stage be considered as more than initial rough data; nonetheless, we can draw some tentative conclusions. We divide up the material by language.

English and French The LARA annotation process already seemed to work well for English and French. Only 1–2% of the tags assigned by the automatic tagger/lemmatizers were edited, 1.0% for French and 1.6% for English. As discussed in §7., many of these edits were arguably not even necessary, and the work was easy enough that texts as long as short novels, 30–40K words, could be quickly converted into LARA form, with a post-editing rate of about 5K words/hour. We ascribe this success to two factors. First, we have mature and reliable morphology resources available; second, English and French are inherently fairly easy languages to deal with.

Icelandic The results for Icelandic were similar to those for the first two languages, with 1.5% of tags edited and a post editing rate of about 6K words/hour on a substantial text of 13.5K words. The numbers suggest that the Icelandic morphology resources are also very good, though we noted inconsistency in tagging of certain words due to the morphological complexity of the language. For instance, the inflected variants of loan words from English "baóbabbur-tré" (baobab tree) and numerals "tveir" (two), "þrír" (three) and "hundrað" (hundred) were among them. Some multiwords had to be tagged manually, the most common being "eins og" ("such as") with 53 occurrences. Like all the annotators except the German one, the Icelandic annotator chose not to include POS information in the tags. This created problems when specifying translations, since some words which would have been disambiguated by POS information had the same tag. A typical example was "þá", which needs to be translated as "then" when it is a temporal adverbial and "they" when it is a personal pronoun.

German The picture is not quite as clear for German. It is not possible to make a direct comparison with English, French and Icelandic. The annotator chose a shorter text (2.5K words); they also decided to use the option of including POS information in the tags, which made the editing process more burdensome. (The upside is that the resulting LARA document is significantly more user-friendly). The post-editing speed here was lower, around 1.5K words/hour. However, the proportion of tags edited, 1.8%, is only slightly higher. We tentatively guess that the lower post-editing speed is mostly due to the more ambitious strategy of including the POS information in the tags.

Swedish The Swedish text gave results clearly worse than those for the preceding languages. 7.6% of the tags were edited, and post-editing speed was 2.5K words/hour. There are two obvious contributory factors. First, the Swedish TreeTagger package is less mature than the English, French and German ones; second, the Swedish text was definitely the most challenging of the ones we attempted.

Turkish The Turkish text gave similar results to the Swedish one in terms of tag editing. 7.4% of the tags were edited, and post-editing speed was 1.6K words/hour. When we investigate the edited lemmas and surface forms, we notice that the used automatic lemmatizer is very mature and performed quite well on the given text. The excessive presence of multi-word expressions in Turkish seems to make the editing percentage relatively high and the tagging process slower. Most of the edit time was due to the need for merging MWE components (104 out of 189 changes). We observed that the Turkish automatic lemmatizer works quite well in order to be used without a need for manual checking. No more than 10 lemmas underwent substantial changes. Most of the lemma changes seem to be due to the annotator choices on lowercasing the produced lemma for proper nouns. For example the lemma "Hoca" (*Hodja*) produced by the automatic lemmatizer was constantly lowercased by the annotator to make it consistent with its lowercase occurences (51 out of 189 changes was due to this specific example and many more like that). It looks like the integration of MWE preprocessors would be extremely useful for quick LARA resource creation in case of Turkish.

Irish The Irish lemmatizer performed very well, with only about 2% of the tags requiring correction. This divided roughly equally between missing words, some of which were archaic ("crostua", an old type of axe); ambiguity, e.g. "léim" can be either the first person singular present tense of 'I read' or the verb 'jump'; compounds, e.g. 'rédhuine' ('second person') was tagged simply as 'duine' ('person'); and multiwords, e.g. "go dtí" ("to/towards").

Irish was the only language where we used TTS. This involved a small amount of effort (< 1 hour) of prooflistening and making minor corrections to the automatically generated phonetic transcription to improve the quality of the speech output.

Dutch For Dutch, it was difficult to find copyright-free recent texts and we had to select older texts. A problem is that these older texts are written in old-fashioned Dutch. We modernised these texts because otherwise they would look outdated and strange to readers nowadays, and also because the tagger could not cope with these old-fashioned texts. We therefore decided to start with shorter texts. This also made it possible to study whether there would be a learning effect. This indeed turned out to be the case, as can be observed from the numbers in Table 1. There is a clear learning effect for the time spent on making the annotations and translations, but not for audio recordings. The tools for recording audio are user friendly, everything is clear from the start, and in order to record audio the words simply have to be read aloud, this is not something that can be done faster. If we compare the average number of person-hours needed for reading 1000 words, we observe that the numbers for Dutch are a bit higher than those for English and Irish. It would be interesting to study what causes these differences, e.g. differences in speaking rate or differences in the segmental structure of these languages.

Farsi The hazm-based tagger was clearly the lowest-performing morphology resource of the ones used: over 25% of the tags in the Farsi text had to be edited. The problem is that the word-tokenizer/lemmatizer pipeline we used in general turns out to be unable to identify most MWEs. These are extremely common in Farsi; to start with, most verbs are phrasal verbs. Multi-word proper names were common in this text, making the problem even worse than usual. To be able to reach an acceptable level of performance in Farsi, the first step is evidently to be able to handle MWEs properly.

7. Discussion

As already noted, we should be careful not to read too much into these preliminary results. That said, we have a decent amount of data from multiple languages, with substantial samples from at least some of them. We outline the issues.

7.1. Correcting automatic tagging

One of the central questions we are interested in here is estimating how much work needs to be done when correcting automatic tagging: we divide this up into edits made to surface words and edits made to lemmas. Surface word edits normally arise in one of two ways. The first is that a series of words can be grouped into an MWE. For example, in *Animal Farm* we find "from side to side", "human being", "at once"; in *Le chien jaune* "au fait" ("by the way"), "c'est à dire" ("that is to say"), "au fur et à mesure" ("progressively"); in *Kallocain*, "på måfå" ("randomly"), "till hands" ("available"), "i förväg" ("in advance"); in *Nasreddin Hodja*, "kabul et" ("accept"), "akşam yemeği" ("dinner"), "satın al" ("buy"). The second way a surface word edit can occur is when a compound word is split up into its components; this does not occur a great deal in English and French, but is common in Swedish, German, Icelandic and Dutch. Thus for example in *Kallocain* (Swedish), where the action takes place in an imagined totalitarian society, there are many compounds with "polis" = "police" ("polischef" = "police chief", "polissekreterare" = "police secretary"...), "tjänst" = "service" ("tjänsteplikt" = "service duty", "offertjänst" = "sacrifice service"...), etc.

Examining the various texts, we find that the question of what constitutes "correct" markup in a LARA document sometimes has a clear answer, and sometimes comes down to judgement. Some tags produced by the automatic taggers are clearly wrong and need to be corrected. For example, in the sentence "And you, Clover, where are those four foals you bore?", "bore" should not be tagged as an uninflected word, but rather is a form of the verb "bear". But there is a large grey area where words can be tagged in more than one way. A common case, which occurs in most of the languages under consideration, is participles used as adjectives. For example, should we in English tag "surprised", "broken" and "astonishing" as inflected forms of "surprise", "break" and "astonish"? Lexica often list "surprised" and "surprise" as separate lemmas, and the English TreeTagger package has been trained to make this distinction. The person doing the English tagging thought it would be more helpful to group them together, and corrected the tagging of "surprised" accordingly; but another annotator might have made the contrary judgement. In morphologically rich languages (MRLs) as well, there exist many open discussions about what the correct lemma should be. Turkish being a strong representative of MRLs allows multiple causatives and the dictionaries are not always consistent about the lemmas for different words. For example,"ölmek" means *to die*, "öldürmek" means *to kill* and "öldürtmek" means *to make one person kill another one* in Turkish. The lemma may be "öl" (*die*) for all of these whereas some Turkish dictionaries would contain different entries for them but may be not for another verb which could take similar inflections.

Similar issues arise with regard to compound nouns in German, Swedish, Icelandic and Dutch. Going back to the examples of compound nouns from *Kallocain*, TreeTagger often runs into trouble with these words. For example, consider the word "ovanjordslicenserna", "above-ground-license-PLUR-DEF". Unsurprisingly, this is not in the lexicon, and by default it is tagged as an uninflected word. This is clearly wrong, and at a minimum it needs to be marked as an inflected form of "ovanjordslicens". But the word is most likely a productive coining, and there is a good argument for splitting it up further into "ovan" + "jord" + "licens" ("above" + "ground" + "license").

MWEs pose the same kind of problems. Some expressions are so idiosyncratic and noncompositional that it seems necessary to mark them as multiwords; e.g. "at once" in English, "au fur et à mesure" in French (literally "by price and by measure" meaning "progressively"), "ne olur ne olmaz" in Turkish (literally "what happens what not happens" meaning "just in case"). But in other cases it is again less obvious. Should "from side to side" be marked as an MWE? It is a set expression; but its meaning is easy to understand from its component words, so it does not seem wrong to leave it unmarked.

A frequent case in English, Swedish, German, Icelandic and Dutch results from verb/preposition constructions. For example, in English, "break up" is not compositional ("... the meeting broke up hurriedly"; *Animal Farm*), and it should be tagged as an MWE. But "break in" is less clear ("One of the cows broke in the door of the store-shed with her horn"). This expression inherits its central meaning from the verb "break" and is arguably compositional. It could be tagged either way. In addition, there is the problem, most frequently occurring with German and Dutch separable verbs, English particle verbs, and French reflexive verbs, that these expressions may sometimes not be contiguous. The version of LARA used here did not support explicit tagging of discontinuous constituents, on the grounds that we wished to avoid complicating the formalism. As described later, we have now added this capability. The above examples also show that there is not always a clear answer to the question of what the "correct" tagging is. The real issue is what will be most helpful to the learner who is trying to relate together the different word occurrences which occur in their reading. If the person doing the tagging has paid too much attention to fine shades of meaning, closely related usages will not be placed in the same LARA example page, and the tool will be correspondingly less helpful. If the annotator is too inclusive, occurrences will be grouped together which have radically different meanings, and the learner may be confused.

7.2. Recorded audio versus TTS

The jury is still out on whether TTS is at an acceptable level for oral language learning. (Smith et al., 2015), who reviewed old data before describing their own experiment concluded that TTS was far inferior to natural voice, but much better than it used to be. Presumably quality has improved further since then. But, as recently as 2018, one of these authors stated that '... the literature on its pedagogical applications in L2 education is still scarce' and went on to assert the viability of TTS based largely on their own research over several years (Cardoso, 2018).

The fact that TTS is so readily available and yet largely eschewed at an educational level suggests that teachers have either intuitive or demonstrated reasons for not using it. However, this is not the concern of LARA, which supports TTS simply because doing so will ensure meeting the requirement that it should be a facilitator, rather than a director. There are obvious advantages to using TTS for the audio side of content development: it is a cheap and quick solution. The Irish experience suggests that inclusion of the available TTS voices enables much wider ranges of materials to be provided by a large population of teachers who have limited access/budgets for human voice talents. This works for Irish as the newest TTS voices provide high quality native speaker speech, covering the main dialects. Correction and editing of TTS output can be utilised to filter out 'errors' but this process is not different from what would be required for human recording. The disadvantage for many languages is that the quality and range of dialect coverage, etc. may not be adequate for the purpose, making natural voice the only proper choice. The prosodic diversity of current TTS may also limit their use in certain genres.

The question is whether the extra effort involved in natural voice production of content is worth it. In the end this is a personal matter for the teacher or others who may become involved in the crowd-sourcing of content development. Of course, it will depend on the feedback of learner-users too. If they all preferred TTS, there would be no reason for the laboursome exercise of human audio.

8. Conclusions and further directions

We have described experiments where texts in nine languages were converted into LARA documents. LARA is still at an early stage of development, and work on some languages is more advanced than others; we have only recently started to use Turkish, Irish and Dutch. Although the results should so far be considered as preliminary, we are cautiously optimistic that the project is proceeding in a good direction. It was possible to use LARA for a wide variety of different languages, and for the ones where we had most experience we were able to convert novel-length texts into LARA form, admittedly with some caveats.

As a result of the experience gained here, we have since added some new capabilities, which we briefly summarise. Most simply, we have added better support for attaching translations to words (cf. §5.2.). It is now possible to attach translations not just to lemmas, but also to surface word types and surface word tokens. As we found, attaching to surface words is essential for morphologically rich languages like Turkish, and the option of attaching to surface word tokens is useful when producing high-quality LARA documents for complex literary texts.

Less trivially, we have introduced support for semi-automatic tagging of MWEs. The platform provides a library of MWE patterns for each language. A pattern in the library consists of a list of words, each marked by typecase as being either a surface word or a lemma. Thus for example in English the phrasal verb "catch up" is entered as `CATCH up`, indicating that "catch" can be inflected but not "up". The pattern can match discontinuous constituents, e.g. "He **caught** them both **up**". In the first phase of annotation, the platform finds all possible MWE matches in the text. These are presented to the annotator, who marks the ones they consider correct; in the second stage, the selected MWE annotations are added to the text. We have so far tested this method mostly with English, where our initial MWE library contains about 1200 entries. Initial experiments on *Animal Farm*, *Alice in Wonderland* and a few shorter texts suggest two conclusions. First, the process of tagging MWEs is reasonably efficient; manually triaging the candidate matches for a text of 25-30K words takes about an hour. Secondly, a process of this kind is necessary. Even experienced annotators frequently miss MWEs when manually annotating text, and the automatic matcher finds many examples they walk past. This work is currently under active development and will be reported elsewhere.

Returning to general issues, the point of LARA is to support learners who wish to improve their reading ability in non-L1 languages. Initial responses are promising: after a first session with the tool, most learners were very positive (Bédi et al., 2019). The next step is to gather data about how well it does when learner use is tracked over an extended period, with more serious texts. We now have all the infrastructure we need to carry out such experiments, and expect to begin during Q2 2020.

9. Bibliographical References

Akhlaghi, E., Bédi, B., Butterweck, M., Chua, C., Gerlach, J., Habibi, H., Ikeda, J., Rayner, M., Sestigiani, S., and Zuckermann, G. (2019). Overview of LARA: A learning and reading assistant. In *Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, pages 99–103.

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

Bédi, B., Chua, C., Habibi, H., Martinez-Lopez, R., and Rayner, M. (2019). Using LARA for learning Icelandic. In *Proc. EUROCALL 2019*.

Cardoso, W. (2018). Learning L2 pronunciation with a text-to-speech synthesizer. In *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018*.

Chua, C. and Rayner, M. (2019). Vegetarian vampires: why the CALL technology provider doesn't have to suck the teacher's blood. In *Proceedings of ICERI 2019*, Seville, Spain.

Chua, C., Habibi, H., Rayner, M., and Tsourakis, N. (2019). Decentralising power: how we are trying to keep CALLector ethical. In *Proceedings of the enetCollect WG3/WG5 workshop*, Leiden, Holland.

Eryiğit, G. (2014). ITU Turkish NLP web service. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden, April. Association for Computational Linguistics.

Grabe, W. and Stoller, F. L. (2012). Teaching reading. *The Encyclopedia of Applied Linguistics*.

Habibi, H. (2019). LARA portal: A tool for teachers to develop interactive text content, an environment for students to improve reading skill. In *Proceedings of ICERI 2019*, Seville, Spain.

Ingólfsdóttir, S. L., Loftsson, H., Daðason, J. F., and Bjarnadóttir, K. (2019). Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 310–315, Turku, Finland, 30 September – 2 October. Linköping University Electronic Press.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Ní Chasaide, A., Ní Chiaráin, N., Wendler, C., Berthelsen, H., Murphy, A., and Gobl, C. (2017). The ABAIR initiative: Bringing spoken irish into the digital space. In *Proceedings of Interspeech 2017*, pages 2113 – 2117, Stockholm, Sweden.

Ní Chiaráin, N. and Ní Chasaide, A. (2020). The potential of text-to-speech synthesis in computer-assisted language learning: A minority language perspective. In Alberto Andujar, editor, *Recent Tools for Computer- and Mobile-Assisted Foreign Language Learning*, chapter 7, pages 149–169. IGI Global, Hershey, PA.

Rayner, M., Habibi, H., Chua, C., and Butterweck, M., (2019). *Constructing LARA content*. https://www.issco.unige.ch/en/research/projects/callector/LARADoc/build/html/index.html. Online documentation.

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Smith, G., Cardoso, W., and Fuentes, C. G. (2015). Text-to-speech synthesizers: Are they ready for the second language classroom? In *Proceedings of the Meeting on English Language Teaching (MELT) 2015*.

Steingrímsson, S., Kárason, Ö., and Loftsson, H. (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*, Varna, Burgaria.

Uí Dhonnchadha, E. (2010). *Natural Language Processing Tools: Developing a Part-of-Speech Tagging and Partial Parsing for Irish*. LAP Lambert Academic Publishing, Köln.