

Using LARA to create image-based and phonetically annotated multimodal texts for endangered languages*

Branislav Bédi

The Árni Magnússon Institute
for Icelandic Studies, Iceland
branislav.bedi@arnastofnun.is

Hakeem Beedar

The University of Adelaide, Australia
hbeedar@hotmail.com.au

Belinda Chiera

The University of South Australia
Adelaide, Australia
Belinda.Chiera@unisa.edu.au

Nedelina Ivanova

The Communication Centre
for the Deaf and Hard of Hearing, Iceland
nedelina@shh.is

Christèle Maizonniaux

Flinders University
Adelaide, Australia
christele.maizonniaux@flinders.edu.au

Neasa Ní Chiaráin

Trinity College, Dublin, Ireland
Neasa.NiChiarain@tcd.ie

Manny Rayner

FTI/TIM
University of Geneva, Switzerland
Emmanuel.Rayner@unige.ch

John Sloan

FTI/TIM
University of Geneva, Switzerland
sloanjo@tcd.ie

Ghil'ad Zuckermann

The University of Adelaide, Australia
Ghilad.Zuckermann@adelaide.edu

Abstract

We describe recent extensions to the open source Learning And Reading Assistant (LARA) supporting image-based and phonetically annotated texts. We motivate the utility of these extensions both in general and specifically in relation to endangered and archaic languages, and illustrate with examples from the revived Australian language Barngarla, Icelandic Sign Language, Irish Gaelic, Old Norse manuscripts and Egyptian hieroglyphics.

1 Introduction

When people are reading documents written in a language less than completely familiar to them, it can often be useful to present the text in multimedia form. This can give the reader access to annotations — typically audio recordings and translations — with a single click, conferring immediate and obvious advantages compared with reading a printed text and looking words up. Many such frameworks now exist; prominent examples in-

clude LingQ¹, Learning With Texts², the Perseus Digital Library's Scaife viewer³ and Clilstore⁴. In our paper from the 2021 edition of this conference, (Zuckerman et al., 2021), we described the Learning and Reading Assistant (LARA; <https://www.unige.ch/callector/lara/>), another platform of this general nature. What primarily distinguishes LARA from the other frameworks is its strongly open source nature, where new features are added in a bottom-up process driven by the demands of a diverse community involved in many different kinds of language-related projects. We argued that this makes it a good fit to endangered languages, which often pose special requirements, and illustrated with three case studies, for Irish Gaelic, Icelandic Sign Language and the revived Australian Aboriginal language Barngarla (Zuckerman et al., 2021).

The version of LARA from last year's paper represented the document as a text string and al-

* Authors in alphabetical order.

¹<https://www.lingq.com/>

²<https://sourceforge.net/projects/lwt/>

³<https://scaife.perseus.org/>

⁴<http://multidict.net/clilstore/>

lowed annotations to be attached to units at two levels, words and segments (typically a segment is a sentence). Experience since then has revealed two important ways in which the above needs to be further generalised. First, thinking of a written document abstractly as a text string obscures the important fact that it is also a visual object. For many texts (picture-books, posters, handwritten manuscripts), the visual dimension is as significant as the words. Second, it is often necessary to go below the word level and think about the relationship between sounds and letters or other primitive written signs. If the student is uncertain about the writing system, the sound system, or the relationship between them, annotations at the character level can be helpful. These observations are particularly relevant to endangered languages, and indeed it is largely because of our close interaction with the endangered language community that we have become so aware of them. We will have more to say about this later, when we discuss specific languages.

In the rest of this paper, we will describe recent work where we have extended LARA to allow image-based and phonetic annotations to be added to texts, and we again illustrate with concrete case studies. Section 2 presents the new functionality, after which Sections 3 to 5 present examples of how it has been used for Barnagarla, Icelandic Sign Language and Irish Gaelic. Section 6 briefly describes how the same features are also useful for annotating historical texts available in manuscript or related form. The final section concludes and suggests further directions.

2 Supporting image-based and phonetic annotations

In this section, we briefly present the overall architecture of LARA and then describe the new functionality which forms the subject of this paper. Full details are available in the online documentation (Rayner et al., 2020).

2.1 Overview of LARA

For a conventional text-based document, the process of converting it into LARA form goes through three stages. The first step is to add annotations dividing the text into pages and segments, tagging inflected words by lemma, and possibly adding HTML markup including links to images defined by instances of the HTML `` tag.

For well-resourced languages, the labour-intensive tasks of segmentation and lemma tagging can be performed automatically by tools already integrated into LARA, followed by some post-editing (Akhlaghi et al., 2020). For smaller languages, where the necessary resources often do not exist, all this work may need to be done manually.

In the second step, the annotated LARA text is passed through a script which internalises it and organises data to support creation of annotations, most obviously translations and audio. Thus for example a script is created which can either be uploaded to an integrated voice recording tool or used to invoke a suitable TTS engine, if available. The annotator fills in this data. In the third step, another script combines the internalised text and the annotations created in the second step and adds metadata to create the final multimodal document. In particular, this metadata includes automatically generated concordances and indexes.

The above steps can either be performed using command-line tools, or carried out through the LARA Portal (<https://lara-portal.unige.ch/>), a free online service which provides a wizard-style interface. Links to LARA documents in many languages can be found on the LARA examples page, <https://www.unige.ch/collector/lara-content>.

2.2 Image-based text

We now describe how the above processing flow has been extended to support image-based text. We first define more exactly what we mean by this term. Intuitively, a piece of LARA image-based text is a portion of a LARA document where the text content and annotations are as they would be in a normal LARA document, but all the visual formatting is determined by an image in JPEG or PNG form. For this to be possible, there needs to be exactly one image for each piece of image-based text, and extra information needs to be supplied to define the image locations with which words in the text are associated. In the compiled LARA document, annotations are accessed by clicking or hovering over the defined locations.

The nature of the visual content at the location associated with a given word is arbitrary. The simplest possibility is that it is a written representation of the word; thus the image could be a page containing a manuscript version of the text, with each text word mapping to the correspond-

(a)

```
<annotated_image>  
  
chair man glass ||  
table ||  
glass woman chair ||  
</annotated_image>
```

(b)

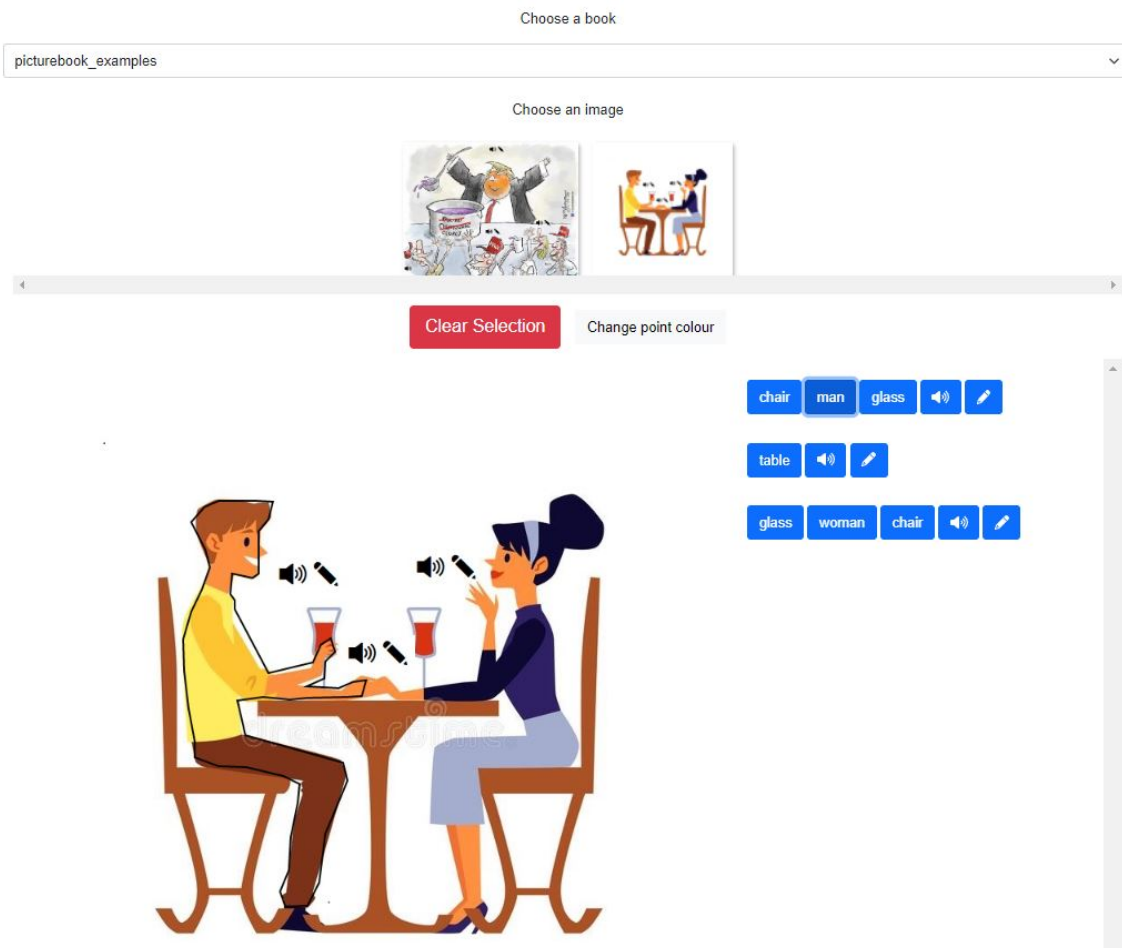


Figure 1: Toy example of a piece of image-based text based on a simple cartoon. The LARA source (a) is given above. The screenshot below (b) shows the tool used to create the word locations file. The top control allows the annotator to choose the text to annotate, after which the slider with the series of thumbnails allows them to choose a page by its image. The bottom left pane presents the selected image, and the bottom right pane the associated words. The annotator can draw a polygon on the left and save it to a word, or select a word on the right to show the current polygon. Here, the annotator has just selected the word “man” on the right, showing the polygon for the picture of the man on the left. The speaker and pencil icons optionally associate audio or text with a whole line. The LARA document is online [here](#).

ing manuscript word. But the visual content can equally well be an image representing the word. Thus for example, in an alphabet book, the text word “apple” could either map to the visual word “apple”, or it could map to a part of the image that contains a picture of an apple.

In the concrete LARA implementation, a piece

of image-based text is delimited by the tag `<annotated_image>`. Links between words and locations are defined by a “word locations file”, a JSON file with a hierarchical structure whose levels are pages, segments and words. A word is optionally associated with a list of three or more coordinate pairs that specify a polygon. Fol-

lowing the usual LARA processing flow outlined in §2.1, the first processing step creates an uninstantiated or partially instantiated version of the word locations file. This can be efficiently filled in using an online graphical tool. which presents the information and allows the user to draw polygons and associate them with words by pointing and clicking. In the compiled LARA document, hovering over a polygon area outlines it as well as performing the usual LARA functions based on the annotations attached to the area, such as playing audio or displaying translations. Figure 1 presents a toy example with a piece of image-based text and a screenshot showing use of the graphical tool.

2.3 Phonetic annotations

We now move on to describe how we have also extended LARA to support texts annotated at the phonetic level. As outlined in §2.1, a normal LARA text is hierarchically divided into pages, segments and words, where the words are associated with lemmas. In contrast, a *phonetic* LARA text is hierarchically divided into pages, words and letter-groups, where each letter-group is associated with a phonetic value. The same notation is used for both types of text, and nearly all of the processing associated with normal (word-oriented) LARA texts carries over to phonetic texts. In particular, a compiled phonetic text contains a phonetic concordance, giving examples of contexts where each phonetic value occurs.

It would be extremely laborious to construct phonetic LARA texts by hand, and there is a script that converts a normal text into the corresponding phonetic version. This post-processes the internalised text to convert each word into a corresponding phonetic version, while keeping formatting unchanged. For languages which are written completely phonetically (common for endangered languages which only recently have acquired a written form), this only requires the annotator to supply the list of phonetically meaningful letter groups defining the orthography of the language. We present an example for the revived Australian language Barnjarla in §3 below.

For languages where online phonetic lexica exist, phonetic versions of most words can be read off the lexicon; free phonetic lexica for many languages are for example available from the IPA-dict project (<https://github.com/open-dict-data/ipa-dict>). The challenge is to

align the letters with the phonetic symbols. At the moment, the strategy used is for the conversion script to help the annotator compile an aligned phonetic lexicon, where typical entries are as illustrated in Figure 2. The script creates new entries automatically using a simple dynamic programming method which maximises the number of alignments already seen in the lexicon (this idea is partly inspired by the one from (Jiampojarn and Kondrak, 2010)), after which a human annotator cleans up the result. Once a few hundred examples of aligned words have been collected, error rates become low and the cleaning-up process is quick. This work will be described in more detail elsewhere.

"admirateur"
"a d m i r a t eu r"
"a d m i ɛ a t œ ɛ"
"ainsi"
"ain s i"
"ɛ̃ s i "
"alors"
"a l o r s"
"a l ɔ ɛ "

Figure 2: Examples of entries from French aligned pronunciation lexicon. Several letters can map into one (beginning of "ainsi"), and letters can map into the empty string (end of "alors").

2.4 Combining LARA documents

LARA includes functionality that allows multiple LARA documents to be linked together. One possibility is sequential linking: the texts are concatenated in a way that combines their metadata, in particular creating a concordance which includes entries from all the component documents. The practical import is that someone reading a later document will easily be able to see when words also occurred in earlier documents, strengthening memory links across their reading history.

Here, we will be more concerned with a new capability, *parallel* linking. For this to make sense, the linked documents must all be different variants of the same text, organised so that page divisions are consistent. In the compiled versions, links are inserted so that each page in one compiled document is connected to the corresponding pages in the other documents.

2.5 An illustrative example

In order to show how the different functionalities introduced in this section can be usefully combined, we present an example in a familiar language, a multimodal French alphabet book based on *Le petit prince* where each page occurs in three different parallel-linked versions. Figure 3 illustrates. Note that in the second, “phonetic”, version, the “picture-book” and “phonetic” functionalities have been combined.

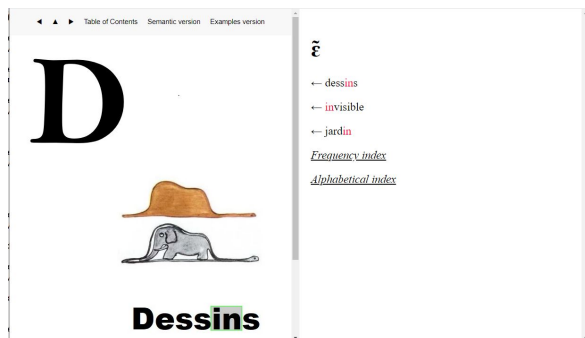


Figure 3: “Phonetic” version of an example page from a “Little Prince” themed LARA alphabet book for French; each page is in three versions, “phonetic”, “semantic” and “example”. Hovering over the word **Dessins** (“Drawings”) outlines phonetically meaningful letter groups; clicking plays audio for the phonetic content selected and shows a phonetic concordance. The student has selected the letter group **in**, showing on the right other words containing the nasalised vowel / \tilde{e} / that this letter group usually represents in French. In the “semantic” version, hovering over the picture on the left outlines it and clicking on it plays audio for the word. The “example” version shows an annotated example sentence. The document can be found [here](#).

3 Barngarla

Barngarla is an Australian Aboriginal language belonging to the Thura-Yura language group, a subgroup of the large Pama-Nyungan language family. Typically for a Pama-Nyungan language, Barngarla has a phonemic inventory featuring three vowels ([a], [i], [u]) and retroflex consonants, an ergative grammar with many cases, and a complex pronominal system.

During the twentieth century, Barngarla was intentionally eradicated under Australian ‘stolen generation’ policies, the last original native speaker dying in 1960. Language reclamation efforts were launched in 2011 (Zuckermann, 2020). Since then, a series of language reclamation workshops have been held in which about 120 Barn-

garla people have participated. The primary resource used has been a dictionary, including a brief grammar, written by the German Lutheran missionary Clamor Wilhelm Schürmann (Schürmann, 1844; Clendon, 2015). A number of educational texts have now been constructed using Schürmann material as the base; as described in last year’s paper, several of them have been converted into LARA form. This has highlighted two issues, both of which materially contributed to motivating the new functionality we describe here.

First, the original texts are always created as a collaboration between ethnic Barngarla people and non-Barngarla expert linguists: usually, design aspects are the responsibility of the Barngarla members of the team. When converting the texts into LARA form, it is thus important to maintain a format that is as close as possible to the original text layout. Second, even though revised Barngarla is written phonetically, the orthography is not transparent to people whose linguistic heritage is primarily anglophone. A particularly important example is retroflex consonants, which are written using an ‘r’ before the corresponding non-retroflex version: thus the voiced retroflex plosive [d̪] (similar to the final sound of Swedish *nord*, “north”) is written ‘rd’ as for example in Barngarla *yarda*, “country”. It is however all too easy for the anglophone reader to interpret this as representing a lengthened preceding vowel followed by [d], as for example in the usual Australian pronunciation of “card” or “herd”. Another important problem is ambiguous phonetic segmentation. Barngarla orthography contains both the unigraph ‘w’, representing the velar approximant [w] and the digraph ‘aw’, representing the diphthong [aʊ]. When Barngarla digraph ‘aw’ is followed by letters representing a vowel, as for example in the common words *bawoo* (“hello”), *gawoo* (“water”), the anglophone reader most naturally segments the words as b|a|w|oo, g|a|w|oo; in fact, they should be b|aw|oo, g|aw|oo.

These issues came to a head during the creation of the latest Barngarla text, *Mangiri Yarda* (Zuckermann and Richards, 2021). The main Barngarla contributor, Emma Richards, invested a substantial amount of effort in the design of the book, and it was clear that the approach used for previous Barngarla LARA texts, trying to reproduce the layout using HTML formatting, would not yield a good result. The issues with pronunciation also

became apparent when recording the audio.

The new functionality developed here however made it possible to address both the layout and phonetic issues in a logical way. The draft book is available online [here](#). It is organised as a LARA picture-book exactly reproducing the text layout, in which all the Barngarla words are annotated with audio information, coupled with a parallel track organised as a “phonetic” LARA book, where the reader can spell through each word a letter-group at a time and listen to the associated phonetic value. By the time of the conference, we expect that the book will have been tested with enough Barngarla readers to be able to present initial feedback.

4 Icelandic Sign Language

Icelandic Sign Language (*Íslenskt táknmál*; ÍTM) is a natural language and the first language of about 250–300 people in Iceland, almost exclusively Deaf people and their children. A peculiarity of ÍTM, compared to other sign languages, is that hereditary deafness hardly exists in Iceland. This means that Deaf children are much less likely to have Deaf parents than in other countries, rendering more difficult the intergenerational transmission of the language and contributing to its endangered status.

Zuckerman et al., 2021 gave further background and outlined some initial experiments in which LARA was used to create annotated texts for Deaf readers, with audio replaced by signed video. Here, we describe two sample image-based texts of this kind. Both are direct multimodal transpositions of existing paper texts designed for the ÍTM community, whose general purpose is to introduce ÍTM signs, and in particular the handshape inventory, to beginner signers. The signed video content has been taken from YouTube videos linked from Icelandic SignWiki (<https://is.signwiki.org/>).

4.1 Background: handshape inventory

There is a long tradition of using the fingerspelling alphabet in signed conversations. The fingerspelling alphabet is a visual representation of the spoken language’s alphabet, and it is used to spell out proper names and other words when a sign is lacking or not known. A sign language’s fingerspelling alphabet in no way corresponds to the phonemic inventory of a spoken language. This

role is filled by the handshape inventory.

Research on ÍTM’s phonemic handshapes has been carried out by Deaf signers and researchers at the Communication Centre for the Deaf and Hard of Hearing. Because there is no corpus for ÍTM, analysis of the frozen lexicon of ÍTM has been slow. In 2019, 33 handshapes were identified as phonemic. Work is still continuing and there may be a slight change in the number of the handshapes. The handshape inventory for ÍTM was developed on the basis of HamNoSys (Schmalting and Hanke, 2001; Smith, 2013). It has two forms, one designed for sign language linguists and one for learners. Further details are available in (Ivanova et al., in press).

4.2 Handshape poster

A poster with the 33 ÍTM handshapes was published in December 2019 in connection with celebrations of the Center’s 30 year jubilee. The poster was intended to spread awareness among children, both Deaf and hearing, about the phonemes of ÍTM, and serve as a basic teaching resource. The design was chosen to be colourful and eye-catching, and includes 33 handshapes. For each handshape, there is a drawing representing a sign that exemplifies the handshape in question, together with a disambiguating gloss in Icelandic.

As an initial exercise, we created a LARA version of the poster, linking the 33 shape/picture combinations to Icelandic SignWiki videos so that clicking on a picture plays the video. The result is posted [here](#). Despite the document’s very simple construction, we were surprised by the enthusiastic reception it received from the Deaf members of the Center. One memorable comment was “It makes the poster as alive as sign language”.

4.3 Pocket dictionary

In 2020 and 2021, the Icelandic Student Innovation Fund, in cooperation with the Center, financed the work of two students for three months each year to develop bilingual ÍTM-Icelandic pocket dictionaries for families of signing children. The model used was the *I am Deaf: Let’s talk* series of booklets produced by Deaf Aotearoa⁵, in which every sign has an equivalent in written English, a morphological description, a drawing of the sign, and a photo representing the sign’s meaning. Six

⁵<https://www.deaf.org.nz/resources/lets-talk-booklets/>

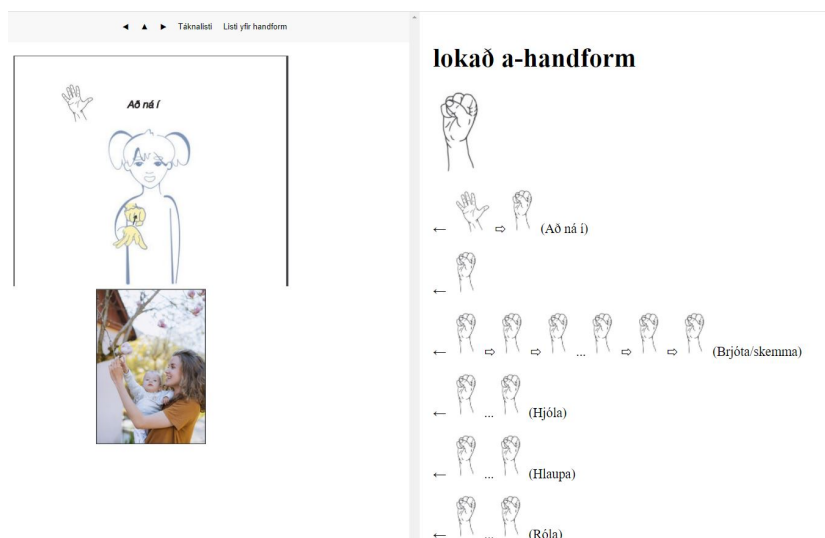


Figure 4: Screenshot of a page from the ÍTM “pocket dictionary”; the user has just clicked on the yellow “closed a-handform” on the upper left, showing examples on the right. The document can be found [here](#).

ÍTM-language booklets were developed, each containing 25 signs grouped by a common theme: “Baby’s first signs”, “The family’s first signs”, “Food”, “Actions”, “Adjectives” and “More signs for the family”. The New Zealand model was developed further by adding an image of each sign’s handshape and a QR code for SignWiki video. In order to stress that ÍTM is the source language and Icelandic the target language, the signs are not ordered alphabetically but rather by handshape, using the canonical ordering of the handshape inventory and subordered by the movement in the sign.

We converted one of the booklets, “Actions”, to LARA form, using a method which tried to respect the core ideas in the project and extend them. Following the principle that sign language is primary, we eliminated all Icelandic text except short phrases naming the actions. Each page (cf. Figure 4) is divided into two halves. The lower half contains the picture illustrating the action; clicking on this picture plays the SignWiki video. The upper half contains the diagram illustrating production of the sign. Here, the reader can click on any hand. This shows the relevant handshape on the right-hand side of the screen, together with a list of other examples where the same handshape is used; the handshapes are shown graphically.

5 Irish

Irish, from the Celtic branch of the Indo-European family, is the first official and national language of the Republic of Ireland and is now a full work-

ing language of the EU. English is the second official language in Ireland. Despite the official status of Irish, however, an erosion of first-language speaker communities is clear and according to the UNESCO Atlas of the World’s Languages in Danger, the language is considered “definitely endangered” (Moseley, 2012).

Irish is spoken as a community language in pockets in the rural West of Ireland called ‘Gaeltacht’ areas. Speakers in urban areas tend to be mostly in individual homes and Irish is relatively rarely overheard on the street. Irish is a compulsory subject until school leaving age. There are approximately 700,000 learners of Irish in the education system in the Republic of Ireland (Ní Chiaráin, 2014). There are also large numbers in the education system in Northern Ireland and many learning Irish abroad, although these numbers are more difficult to quantify.

Irish shares distinctive features with other Celtic languages such as a verb-subject-object (VSO) word-order and rich morphology (Stenson, 1981). As in other Celtic languages, initial consonants undergo mutations in specific grammatical contexts, e.g., the lenition of stops to fricatives/approximants; of voiceless stops to voiced stops; of voiced stops to voiced nasals. Verbs are inflected for tense, number and person, while nouns are inflected for number and case. Prepositions can inflect for person and number. Nouns are either masculine or feminine in grammatical gender.

	Labial	Dental	Alveolar	Alveolo-palatal	Palatal	Velar	Glottal
Plosive	p ^v b ^v p ^j b ^j	t ^v d ^v		t ^j d ^j	c ɟ	k g	
Fricative/ Approx.	f ^v w f ^j v ^j		s ^v	ç	ç j	x ɣ	h
Nasal	m ^v m ^j	ŋ ^v	n	ɲ ^j	ɲ	ŋ	
Tap			r ^v r ^j				
Lateral Approx.		l ^v	l	l ^j			

Figure 5: Consonantal system of Irish (Ní Chasaide, 1999), where there is a fundamental contrast between velarised [C^v] and palatalised [C^j] phonemes

Irish has three main dialects and a number of sub dialects. These dialects differ at many levels, including their structural features, vocabulary and particularly in their pronunciation. A written standard “An Caighdeán Oifigiúil” was first introduced in 1958 and the most recent update to this was published online in 2017. However, as with many minority languages, there is no single spoken standard and all dialect variants hold equally. The fact that the writing system does not match in a simple way to any one of the spoken dialects presents challenges to learners.

A major feature of the Irish sound system is the contrast between palatalised and velarised pairs of consonants as illustrated in Figure 5. The contrast of palatalised and velarised segments not only differentiates words, e.g., /L^jO:N^s/ leon ‘lion’ vs. /L^vO:N^s/ lón ‘lunch’, but serves for grammatical differentiation of the same lexical item, as in /L^vO:N^s/ (nominative) vs. /L^vO:N^j/ (genitive).

Latin script is used for the language’s writing system, with an alphabet which is superficially similar to English, excluding j, k, q, v, w, x, y, z, (except in loan words). However, the consonants are not marked for the fundamental contrast of palatalisation and velarisation of Irish; rather, the palatalisation-velarisation difference is shown by the adjacent vowel letter used (‘i’, ‘e’, mark palatalisation and ‘a’, ‘o’, ‘u’ mark velarisation). All this makes it very complex for learners to acquire the link between the orthography and the sounds of the language. There is also a contrast between long and short vowels, which differentiates words, e.g. /m^ji:n^j/ min ‘(oat)meal’ and /m^ji:m^j/

mín ‘smooth’. Long vowels are orthographically marked with an acute accent, as in: á é í ó ú.

5.1 An Scéalai

*An Scéalai*⁶ is a purpose-built iCALL platform for Irish. It builds on the AB AIR initiative, which is concerned with the development of core speech technologies for Irish⁷ (particularly TTS to date but ASR development ongoing more recently). *An Scéalai* deploys core language technologies and presents them to learners in a pedagogically appropriate way. It is currently being used primarily as a writing tool but aims at a holistic approach to language learning, simultaneously training the four skills (for a more detailed description see (Ní Chiaráin et al., 2022)). The intention is to provide a motivational environment for learners to practise writing, and, through having TTS available at the click of a button, brings the spoken language into every aspect of the language learning, helping to compensate for the fact that native speakers are not readily to hand for most learners (one of the most common complaints from learners is the fact that they have limited opportunities to interact through the medium of Irish). As learners practise writing they are encouraged to think of spelling as a phonic-based system (see (Ní Chiaráin and Ní Chasaide, 2019) for more detail). There is an emphasis on self-correction (proofreading and prooflistening) using the available language technology tools and

⁶<https://abair.ie/scealai/>

⁷<https://www.abair.ie>

resources for Irish, such as dictionaries⁸, TTS, a grammar checker⁹, and a grammar database¹⁰, which gives inflected forms of nouns, verbs, adjectives, etc.

5.2 A LARA alphabet book for Irish

We have used the infrastructure described above to create a LARA primer for the sounds of Irish. The format is superficially that of an alphabet book: alphabet books will be regarded by most students as simple and unthreatening, while the introduction of the complex phonetic symbols, as in Figure 5, could be forbidding.

The book’s structure presents minimal pairs illustrating key phonemic differences, where words are presented in the context of short sentences combined with pictures and both TTS and human audio. The core goals are to develop phonological awareness of the velarisation-palatalisation contrast in Irish in the hope that learners make the link between the phonological contrasts and the spelling regularities of the language.

Resources of this kind are badly needed, since, remarkably, there is virtually no awareness of consonantal palatalisation/velarisation difference among learners or indeed among many teachers of Irish. It is hardly ever made explicit in teaching, and the difficulty for learners is further compounded by the fact that the L2 learners are English speaking and familiar with the English alphabet and phonics. This undermines the teaching of pronunciation, and fails to highlight the phonic basis of the orthographic system. Pronunciation training is typically not even considered in Irish language instruction.

The LARA Irish alphabet picturebook¹¹ uses visual and auditory cues to illustrate minimal pairs and help consolidate auditory memory of contrasting forms. It is designed to raise awareness of this fundamental phonological contrast of Irish. This gives a glimpse of a parallel current project *Lón don Leon*, a tablet-based app which is specifically designed to develop phonological awareness and early literacy skills in young learners. This is a multimodal app with a high level of interactivity. To consolidate memorisation and acquisition of the contrasts and of their orthographic realisations, it includes newly composed musical ditties,

⁸<https://www.teanglann.ie>

⁹<https://cadhan.com/gramadoir/>

¹⁰<https://www.teanglann.ie/en/gram/>

¹¹<https://tinyurl.com/2p8k7zfz>

stories, graphics, quizzes set on a virtual island (see description in (Ní Chasaide et al., 2019)).

We expect the resource to be useful for trainee teachers, at the very least for awareness raising, and for learners at all levels; a recent study carried out by an Irish author of this paper with advanced learners of Irish showed they do not produce the velarisation-palatalisation contrast reliably.

6 Manuscripts and other archaic texts

Although this is not the focus of the current paper, we note in passing that the functionality described here also appears to be relevant to archaic texts in manuscript or inscription form, where the visual appearance of the document is of critical importance. We illustrate with two initial examples. The first is a LARA version of an extract from the Old Norse poem *Völuspá* (Bédi et al., 2020), with each verse presented both in facsimile manuscript and plain text form. The second is an inscription in Ancient Egyptian hieroglyphics taken from (Collier and Manley, 1998), presented in parallel ‘word’ and ‘sign’ views. The two examples are posted [here](#) and [here](#).

7 Summary and further directions

We have described extensions recently added to the LARA platform to support image-based and phonetically annotated texts, and illustrated with examples from Barngarla, Icelandic Sign Language, Irish, French, Old Norse and Egyptian hieroglyphics. The work was motivated by the demands of these languages, particularly the first three. We are currently liaising with members of other endangered language communities, a leading example being the Austronesian language Iai.

The implementation of the new functionalities is still at an early stage, and our current priority is to improve their integration into LARA and make them easier to use. In particular, we have started development of an intuitive tool which will enable simple creation of “LARA albums”, LARA picture-book/phonetic documents consisting of images paired with short captions. The intention is to lower the bar to entry for people wishing to create LARA texts, so that it can become a routine part of language teaching; this is particularly interesting in the context of the An Scéaláí Irish platform (cf. §5). We hope to be able to report on this work later in 2022.

References

- Elham Akhlaghi, Branislav Bédi, Fatih Bektaş, Harald Berthelsen, Matthias Butterweck, Cathy Chua, Catia Cucchiari, Gülşen Eryiğit, Johanna Gerlach, Hanieh Habibi, Neasa Ní Chiaráin, Manny Rayner, Steinþór Steingrímsson, and Helmer Strik. 2020. Constructing multimodal language learner texts using LARA: Experiences with nine languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 323–331.
- Branislav Bédi, Haraldur Bernharðsson, Cathy Chua, Birgitta Björg Guðmarsdóttir, Hanieh Habibi, and Manny Rayner. 2020. Constructing an interactive Old Norse text with LARA. *CALL for widening participation: short papers from EUROCALL*, pages 27–35.
- Mark Clendon. 2015. *Clamor Schürmann’s Barngarla grammar: A commentary on the first section of A vocabulary of the Parnkalla language*. University of Adelaide Press.
- Mark Collier and Bill Manley. 1998. *How to read Egyptian hieroglyphs: a step-by-step guide to teach yourself*. Univ of California Press.
- N. Ivanova, R. Sverrisdóttir, and G.T. Thorvaldsdóttir. in press. The handshake inventory for Icelandic Sign Language (ÍTM) in early intervention and teaching of ÍTM. *Croatian Review of Rehabilitation Research*, Special issue on Sign Language, Deaf Culture, and Bilingual Education.
- Sittichai Jiampojarn and Grzegorz Kondrak. 2010. Letter-phoneme alignment: An exploration. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden*, pages 780–788.
- Christopher Moseley. 2012. *The UNESCO atlas of the world’s languages in danger: Context and process*. World Oral Literature Project.
- Ailbhe Ní Chasaide. 1999. Irish. In *Handbook of the International Phonetic Association*, pages 111–116. Cambridge University Press Cambridge.
- Neasa Ní Chiaráin and Ailbhe Ní Chasaide. 2019. An iCALL approach to morphophonemic training for Irish using speech technology. *CALL and complexity*, page 314.
- Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, Andrew Murphy, Emily Barnes, and Christer Gobl. 2019. Leveraging phonetic and speech research for Irish language revitalisation and maintenance. In *Proceedings of ICPHS 2019: the 19th International Congress of Phonetic Sciences*, pages 994 – 998, Melbourne, Australia.
- Neasa Ní Chiaráin, Oisín Nolan, Madeleine Comtois, Naimhin Robinson Gunning, Harald Berthelsen, and Ailbhe Ní Chasaide. 2022. Using Speech and NLP resources to build an iCALL platform for a minority language: the story of *An Scéalaí*, the Irish experience to date. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.
- Manny Rayner, Hanieh Habibi, Cathy Chua, and Matt Butterweck. 2020. *Constructing LARA content*. <https://www.issco.unige.ch/en/research/projects/collector/LARADoc/build/html/index.html>. Online documentation.
- C. Schmaling and T. Hanke. 2001. HamNoSys 4.0. Interface definitions. ViSiCAST Deliverable D5-1. Technical report, University of Hamburg.
- Clamor Wilhelm Schürmann. 1844. *A Vocabulary of the Parnkalla Language. Spoken by the natives inhabiting the western shore of Spencer’s Gulf. To which is prefixed a collection of grammatical rules, hitherto ascertained*.
- Robert Smith. 2013. Hamnosys 4.0 user guide. Technical report, Technical Report. Institute of Technology Blanchardstown Ireland.
- Ghil’ad Zuckerman, Sigurður Vigfússon, Manny Rayner, Neasa Ní Chiaráin, Nedelina Ivanova, Hanieh Habibi, and Branislav Bédi. 2021. LARA in the service of revivalistics and documentary linguistics: Community engagement and endangered languages. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 13–23.
- Ghil’ad Zuckermann. 2020. *Revivalistics: From the Genesis of Israeli to Language Reclamation in Australia and Beyond*. New York: Oxford University Press.
- Ghil’ad Zuckermann and Emma Richards. 2021. *Mangiri Yarda (Healthy Country: Barngarla Wellbeing and Nature)*. Revivalistics Press.