LARA: A Learning and Reading Assistant

Elham Akhlaghi, Branislav Bedi, Cathy Chua
Hanieh Habibi, Manny Rayner*

February 25, 2019

Abstract

This working paper presents an overview of LARA (Learning and Reading Assistant), a set of tools we are in the early stages of developing. The idea of LARA is to support reading as a learning strategy for improving the learner's command of a new language. LARA offers a range of options for semi-automatically marking up text in ways that can help the learner. These currently include construction of a personalised concordance based on the learner's reading history, addition of recorded audio files, and insertion of links to translations and online linguistic resources. We illustrate with initial examples in English, Farsi and Icelandic and sketch ideas for further work. This will help to pave the road to including additional languages. Our medium-term goal is to include LARA as a component platform in CALLector, a social network currently under construction which will link together creators and users of online CALL content.

1 Introduction and motivation

In this working paper, we present an overview of LARA (Learning and Reading Assistant), a tool we are in the early stages of developing. The idea of LARA is to support reading as a way to improve the learner's command of a new language. Generally, this method is referred to as the reading strategy, or reading with understanding when learning a foreign or second (L2) language (Oxford, 1990). It is one of the four language competences (speaking, writing, reading, listening) that learners can self-develop in an interactive way. The basic intuition is simple and uncontroversial: other things being equal, reading helps you improve your command of a language.

Every learner employs a different language-learning strategy and style. The choice is influenced by a multitude of factors, including aptitude, motivation, I.Q., personality and age, and all of these can affect linguistic and nonlinguistic learning outcomes (Skehan, 1991). Other important variables are the type of classroom instruction and tools used. All of this makes it

*Authors in alphabetical order.

extremely difficult to construct systematic studies designed to compare the effect of different ways of learning by reading, and few have been reported (Tarone et al., 2013).

Whichever version of the reading strategy the learner is using, however, three immediate problems stand out. First, one of the most effective strategies for learning a new word is for the learner to compare several occurrences in the text which show how it is used in different contexts; but unless the word occurs twice in close succession, it may be difficult to remember previous occasions when it has turned up. The second problem is simply that reading, by its nature, neglects the sound of words. A learner who prioritises reading can easily end up internalising incorrect guessed pronunciations. For languages with complex morphology, a third problem arises: it can be hard to recognise variant inflected forms of a word. For example, the beginner learner of French may not even suspect that *lu* could be the past participle of *lire*, or that *aura* is the third person singular future of *avoir*.

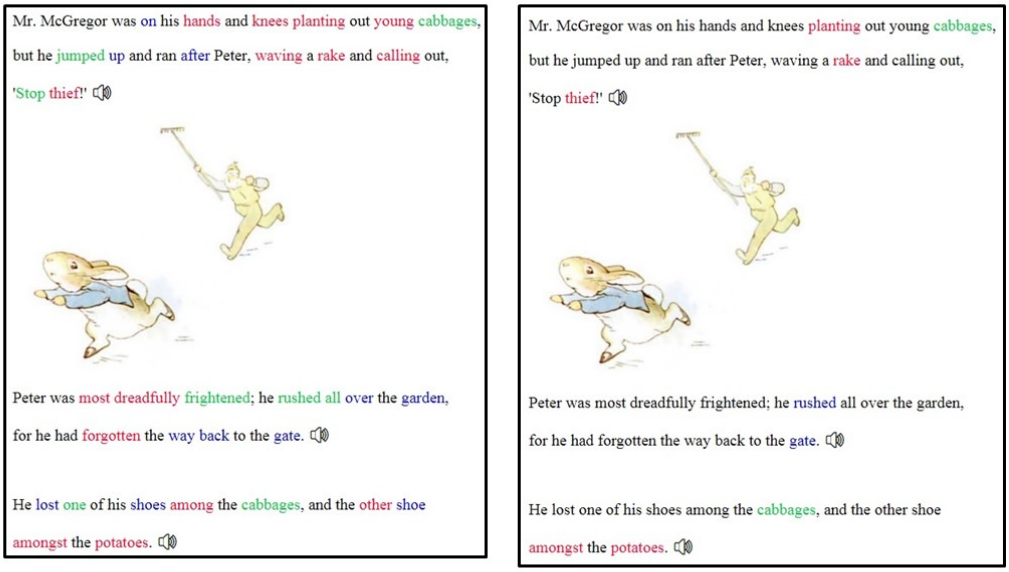

Figure 1: Example of text shown by LARA: two paragraphs from *Peter Rabbit* marked up (left) at a point where the learner has just completed *Peter Rabbit* but yet read anything else, and (right) at a point where the learner has read both *Peter Rabbit* and *Alice in Wonderland*. Colours are used to indicate how often each word has occurred in the learner's reading to date.

Although learning an L2 by reading with understanding has some attractive properties, in particular that it leads rapidly towards an ability to think in the L2, these core problems are severe enough that it is easy to see why most learners are reluctant to make it a large component

of their overall language learning strategy. When we began this project, we wondered what we could do to help learners who are intuitively well-disposed towards learning by reading, but are put off by the practical difficulties. Following the early example of Johns's Data-Driven Learning (Johns, 2002), it seemed to us that, by suitably marking up the text with concordance-oriented information, we could provide significant assistance.

LARA is designed to address the key problems of the reading strategy. In brief, it answers two questions that will frequently occur to all readers, and particularly to false beginners and advanced L2 learners:[1] "Where have I seen that word before?" and "What does that sentence sound like?". In slightly more detail, LARA is a script which processes text into a hyperlinked marked-up form with the following properties. First, each sentence is linked to a recorded audio file. Second, there is a personalised hyperlinked concordance which shows the learner where each word has previously occurred *in their own reading experience*. Other capabilities include addition of links to translations and existing online resources.

Figure 1 illustrates LARA's core functionality using a short passage from *The Tale of Peter Rabbit* (Potter, 1904), a classic English children's story. The marked-up version of *Peter* is shown at two points in the learner's reading progress: on the left, where the learner has read the whole text and nothing else, and on the right, where they have read both *Peter Rabbit* and *Alice in Wonderland*. Colours are used to indicate how many times each lemma occurs in the text; words in black have occurred more than five times, words in red only once, while blue and green show intermediate values. As the picture shows, the colours effectively track the learner's increased exposure to vocabulary between the two snapshots. In the first snapshot, only function words and a few content words central to the story appear in black; in the second, many of the words marked in red have turned black or blue, indicating that they have been read several times during the intervening period.

The text has been manually divided into segments and recorded in audio form. A loudspeaker icon marks the end of each segment, and the learner can listen to the segment in question by clicking on the icon. The learner can click on any word and get a personalised concordance page containing up to ten segments where the word appears. Figure 2 shows the page for the word "shoe".

In the next section, we go on look at other LARA functionalities, this time in the more challenging environments of Icelandic and Farsi.

2 More LARA functionality: examples in Icelandic and Farsi

Our second example uses *Tína fer í frí* (Skriver, 1989), the Icelandic edition of a Danish children's book commonly used as a reading text in Icelandic primary schools. *Tína* contains about 2700 words and about 480 unique lemmas, and has a reading level between A1 and A2. The left-hand side of Figure 3 shows the beginning of the marked-up version of *Tína*, produced at a point where the reader has read the whole text. As before, colours are used to indicate how many times each lemma occurs in the text, and every word in the text is linked to the information page for that word; the right-hand side of Figure 3 shows the page for the word *tveir* ("two"),

[1] A large proportion of the Icelandic class about to start testing LARA falls into these categories. We will have more to say about this elsewhere.

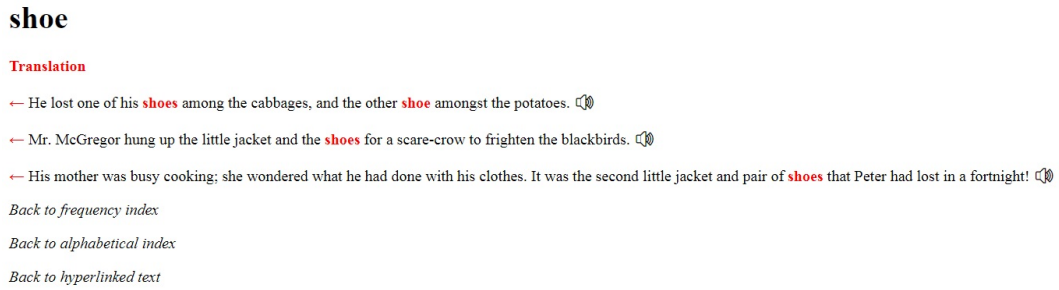

Figure 2: Concordance page for the word "shoe" in the *Peter Rabbit* corpus.

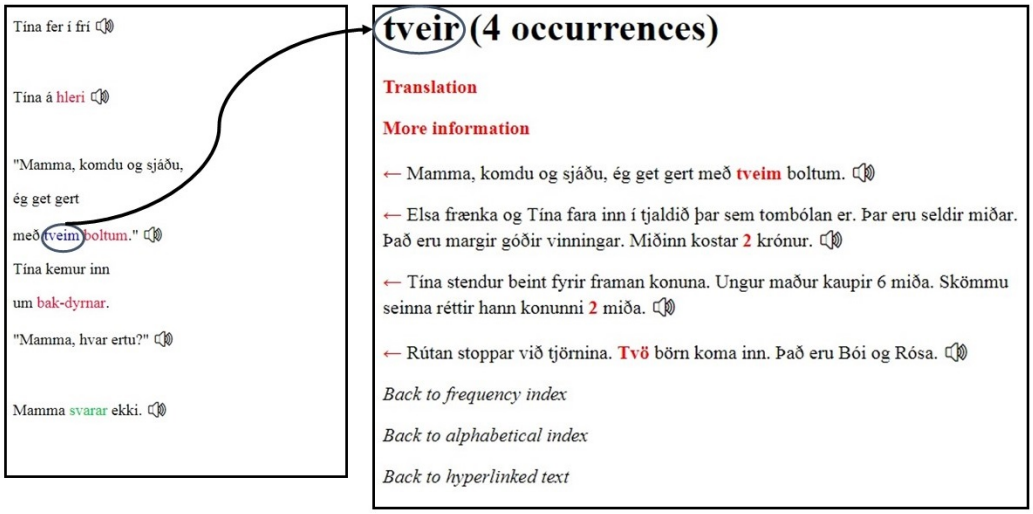

Figure 3: Example of Icelandic text marked up by LARA: first two paragraphs of *Tína fer í frí* plus the concordance page for the word *tveir* ("two").

accessed by clicking on the word *tveim* in the third segment from the marked-up text on the left. We examine the format of the word information page in a little more detail.

The main body of the page collects together segments where different inflected forms of *tveir* appear; note here, for example, that *tveim*, the word in the main text on the left, is distinct from *tveir*, the head word for the word information page on the right. As before, the loudspeaker icons allow any segment to be played in audio form, and each word is clickable; the back arrow at the beginning of the line links to the place in the main text where the segment occurs.

The learner can get an English translation by hovering the mouse over the red word **Translation** just under the title. Clicking on **More information** brings up a popup with the relevant page from an online Icelandic morphology resource (Icelandic is a morphologically rich language). The page for *tveir* is shown in Figure 4. The script also produces two vocabulary lists, one ordered by frequency and one ordered alphabetically. Parts of these lists for *Tína* are shown

in Figure 5. On the left, we see the first 14 entries by frequency. The right side of the figure shows the end of the alphabetically ordered table.

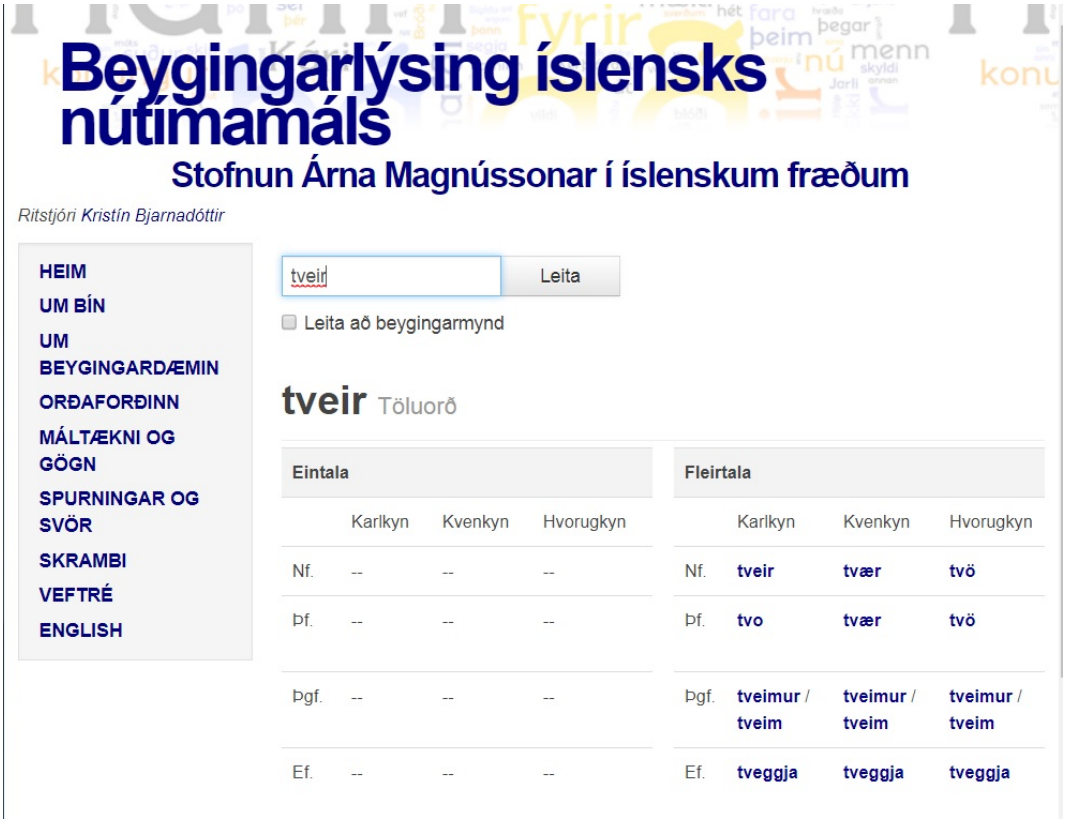

Figure 4: Icelandic morphology page from `http://bin.arnastofnun.is`, obtained by clicking **More information** on the word page for *tveir* ("two").

LARA uses unicode throughout and is consequently able to handle non-European scripts. Figure 6 presents an example for Farsi, a right-to-left language using an extended Arabic script.

3 The LARA content-creation process

The LARA content-creation scripts are designed so that they should already produce a useful result on plain text corpora. The result is marked-up text with links to pages for the words occurring in the text. The text is automatically segmented, using punctuation, into units which are approximately sentences. If the content-creator is prepared to invest more effort, they can improve the result in several ways. The list of currently available LARA texts in the next section gives links to examples.

Manual segmentation: The content-creator can explicitly mark segment boundaries in the text, using the convention that `||` marks a boundary. In practice, we find that a good

Rank	Word	Freq	Cumul
1	að vera	121	4.41%
2	að	121	8.82%
3	hún	117	13.09%
4	Tína	101	16.77%
5	og	86	19.91%
6	ég	84	22.97%
7	að segja	74	25.67%
8	í	70	28.22%
9	það	47	29.93%
10	Rósa	47	31.64%
11	að fara	46	33.32%
12	á	44	34.93%
13	Bói	42	36.46%
14	ekki	32	37.62%

útihátíð	2
ýmislegur	1
þangað	2
þar	10
það	47
þaðan	1
þegar	13
þessi	10
þig	1
þinn	1
þreyttur	1
þriðji	1
þá	20
þó	1
þú	28

Figure 5: Example of frequency-ordered and alphabetically-ordered vocabulary tables produced by LARA for *Tína fer í frí*.

segment is typically about 10-20 words; this is equivalent to 2–3 short sentences, one long sentence, or part of a very long sentence.

Tagging: The content-creator can tag an individual word, or parts of an individual word, by adding a tag enclosed in hash-signs (#) after the tagged element. The effect is to link the word to the page indicated by the tag.

Adding images: Images can be included using the standard HTML <img> tag.

Figure 7 shows how segment boundaries, tags and images have been used to mark up the example passage from the *Peter Rabbit* corpus.

Recording audio for segments and words: The LARA processing script automatically produces separate lists of segments and words in a format that can be uploaded to the Lite-DevTools recording tool (Ahmed et al., 2016). This provides a simple and ergonomically efficient web interface which allows recording tasks to be assigned to registered users. At any point, the currently available recorded results can be downloaded and automatically incorporated into the generated LARA resources.

مردم دِه، صدای پسرك چوپان را شنیدند. آنها برای کمك به پسرك چوپان و گوسفندهایش به طرف تپه دویدند.

ولی وقتی با نگرانی و دلهره به بالای تپه رسیدند، پسرك را خندان دیدند، او می‌خندید و می‌گفت : من سر به سر شما گذاشتم.

مردم از این کار او ناراحت شدند و با عصبانیت به دِه برگشتند.

از آن ماجرا مدت‌ها گذشت.

یك روز پسرك نشسته‌بود و به گذشته فکر می‌کرد.

به یاد آن خاطره خنده‌دار خود افتاد و تصمیم‌گرفت دوباره سر به سر مردم بگذارد.

Figure 6: LARA content in a non-European script: a passage from the marked-up version of *The Boy Who Cried Wolf* in Farsi.

```
Mr. McGregor#McGregor# was#be# on his hands#hand# and knees#knee# planting#plant out# out young cabbages#cabbage#,
but he jumped#jump# up and ran#run# after Peter#Peter#, waving#wave# a rake and calling#call# out,
'Stop thief!'||
<img src="10WavingARake.jpg" width="294" height="253"/>
Peter#Peter# was#be# most dreadfully frightened#frighten#; he rushed#rush# all over the garden,
for he had#have# forgotten#forget# the way back to the gate.||

He lost#lose# one of his shoes#shoe# among the cabbages#cabbage#, and the other shoe
amongst the potatoes#potato#.||
```

Figure 7: Marked-up source for the example from *Peter Rabbit* in Figure 1. | | marks a segment boundary and #...# marks a tag.

Adding translations: The LARA script produces a spreadsheet of headwords, which can be filled in to provide translations included in the word pages.

Linking to online resources: If online morphological or lexicon resources are available for the language, a hook is available to insert suitable links in the **Other information** fields of the word pages.

4 Currently available LARA texts

Table 1 lists the marked-up LARA texts currently available online. The texts are identified as follows:

Tina: *Tína fer í frí* (Skriver, 1989). Text marked up and audio recorded by Branislav Bedi.

Peter: *The Tale of Peter Rabbit* (Potter, 1904). Text marked up by Manny Rayner and Hanieh Habibi, audio recorded by Cathy Chua.

Aesop/FA: Farsi versions of two classic stories from the traditional Aesop's Fables. Text marked up and audio recorded by Hanieh Habibi.

Alice: *Alice in Wonderland* (Carroll, 1865). Text semi-automatically marked up by Manny Rayner using tools derived from Python NLTK (Bird et al., 2009). No audio.

LP/FR: *Le petit prince* (de Saint-Exupéry, 1945), original French edition. Text segmented by Manny Rayner but not otherwise marked up. No audio.

LP/IS: *Le petit prince*, Icelandic translation. No marking up, no audio.

Table 1: Currently available online LARA texts. "Tina" = *Tína fer í frí*; "Peter" = *The Tale of Peter Rabbit*; "Aesop/FA" = Farsi versions of Aesop; "Alice" = *Alice in Wonderland* "LP/FR" = *Le petit prince*; "LP/IS" = *Le petit prince* in Icelandic; "Virtual1" = reading history consisting of *The Tale of Peter Rabbit* plus the beginning of *Alice in Wonderland* (see §5); "#Seg" = number of segments; "#Tok" = number of surface word tokens; "#Typ" = number of word pages generated; "Seg" = text manually segmented; "Tag" = text manually tagged; "Img" = images included; "Rc" = segments recorded in audio form; "WRc" = words recorded in audio form; "Tr" = translations provided; "Inf" = references to online information provided; "Link" = link to online LARA text.

Text	Lng	#Seg	#Tok	#Typ	Seg	Tag	Img	Rc	WRc	Tr	Inf	Link
Tina	IS	207	2743	466	✓	✓	✓	✓	✓	✓	✓	☞
Peter	EN	40	942	398	✓	✓	✓	✓		✓		☞
Aesop/FA	FA	47	577	219	✓	✓	✓	✓	✓	✓		☞
Alice	EN	1473	26984	2030	✓	✓						☞
LP/FR	FR	956	14237	2657	✓							☞
LP/IS	IS	1668	13536	3042								☞
(Virtual1)	EN	65	2643	472	✓	✓	✓	✓		✓		☞

5 "Reading histories" and personalised LARA pages

The LARA texts listed in the last section have all been compiled as standalone resources. The key strength of LARA is however that it also allows construction of personalised pages, based on a reader-specific "reading history". An initial version of this functionality already exists and is being tested internally, though it will only become properly useful when it is integrated into the planned CALLector social network (cf. §6).

The idea of the reading history is straightforward. First, instead of keeping all the LARA resources for different texts on one server, we instead allow the possibility of distributing them over multiple servers anywhere on the web. Each corpus contains associated metadata. There is a master file which lists the various LARA resources. A "reading history" defines a virtual corpus as a sequence of extracts from the real corpora: thus it could for example consist of all of corpus 1, the first 120 segments from corpus 2, and segments 20 to 35 of corpus 3. The LARA compiler can create a set of personalised pages from the virtual corpus by downloading the associated metadata for each corpus and using it to build pages that link to the real corpora. The bulky multimedia files (audio and images) are not copied, but stay in place and are referenced by links. Building the pages for the reading history is thus a reasonably quick operation.

The last line of Table 1 gives a link to an example of a set of personalised pages for a reading history which consists of the whole of *Peter Rabbit* plus the first 25 segments of *Alice in Wonderland*. As can be seen, the concordance pages combine examples from both texts; for example, the page for "rabbit" refers both to Peter Rabbit and the White Rabbit.

6 Next steps

Our own experiences of using the resources so far developed leave us feeling optimistic that LARA is a viable idea that, suitably developed, could be of real help to language learners. We organise our tentative plans for further work under the following headings.

Evaluation The initial content needs to be tested with real learners, and user feedback collected. During the first half of 2019, we plan to do this with Icelandic content at the University of Iceland and Farsi content at the Ferdowsi University of Mashhad.

Technical issues Several technical improvements are clearly desirable: a) Most urgently, we need to provide better tools to support the manual annotation process and make it less laborious and error-prone. We have constructed an initial tool for English using the Python Natural Language Toolkit (Bird et al., 2009); on the *Alice* text, this made tagging about ten times faster. We are currently investigating options for constructing similar tools in the other languages we are using. b) The compilation mechanism should be made incremental, so that the marked-up resources can be minimally regenerated as new text is added. This would make it possible for readers to perform continuous updating of personalised online resources in response to their reading progress.

Copyright The copyright issues are in general complex and unclear. Here, we are leaning towards a model which considers the LARA tools as having a status similar to that of an

open source compiler; users are free to download the tools and use them as they wish to produce marked-up resources, but must take responsibility for deploying these resources appropriately. We will be cautious in releasing our own LARA content, since we primarily wish to focus on technical and scientific questions.

Ethical issues Several ethical issues may arise. Since the system is designed to perform detailed tracking of learners' reading progress, we must be mindful of the fact that many users will regard this as sensitive information that needs to be protected. The crowd-sourced nature of the content-creation process introduces further issues; some content-creators will require their contributions to be suitably acknowledged, while others may conversely wish to remain anonymous and will not appreciate being identified. We need to accommodate both types of content-creator.

Integration At some point, probably beginning during the second half of 2019, the intention is to incorporate LARA as a component platform in CALLector[2], a social network we are currently developing to support construction and use of online CALL resources. The network will support interfaces for both resource constructors and readers, in particular so that individual learners have personalised sets of LARA pages as outlined in the previous section. Not only do learners have different reading histories; in general, they want their pages marked up in different ways. The social network must thus at a minimum be able to give each user an account plus functionality for choosing mark-up options, posting reading updates, and accessing their personalised pages.

Acknowledgements

The original idea for LARA was suggested by Cathy Chua. We would very much like to thank Lionel Nicolas and Verena Lyding for their support in getting this project started and fostering collaboration between different groups within the enetCollect network.

References

Farhia Ahmed, Pierrette Bouillon, Chelle Destefano, Johanna Gerlach, Angela Hooper, Manny Rayner, Irene Strasly, Nikos Tsourakis, and Catherine Weiss. Rapid construction of a web-enabled medical speech to sign language translator using recorded video. In *Proceedings of FETLT 2016*, Seville, Spain, 2016.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

Lewis Carroll. *Alice's Adventures in Wonderland*. Macmillan, 1865.

Antoine de Saint-Exupéry. *Le petit prince: avec des aquarelles de l'auteur*. Gallimard, 1945.

[2] `https://groups.google.com/forum/#!forum/callector`

Tim Johns. Data-driven learning: The perpetual challenge. In *Teaching and learning by doing corpus analysis*, pages 105–117. Brill Rodopi, 2002.

Rebecca L. Oxford. Language learning strategies. In Anne Burns and Jack C. Richards, editors, *The Cambridge Guide to Learning English as a Second Language*, chapter 9, pages 81–89. Cambridge University Press, Cambridge, 1990.

Beatrix Potter. *The tale of Peter Rabbit*. Frederick Warne & Co., 1904.

Peter Skehan. Individual differences in second language learning. *Studies in second language acquisition*, 13(2):275–298, 1991.

Esther Skriver. *Tína fer í frí*. Námsgagnstofnun, 1989.

E Tarone, K Hansen, and M Bigelow. Alphabetic literacy and second language acquisition by older learners. In Julia Rogers Herschensohn and Martha Young-Scholten, editors, *The Cambridge handbook of second language acquisition*, pages 180–203. Cambridge University Press, 2013.