Synchronous reading: learning French orthography by audiovisual training

Gérard Bailly and *Will Barbour*

GIPSA-lab, UMR 5216 CNRS/INPG/UJF/U. Stendhal

gerard.bailly@gipsa-lab.grenoble-inp.fr

Abstract

We assess here the potential benefit of a karaoke-style reading system for learning sound-to-letter mapping in irregular languages. We have developed a framework that eases the development of interactive systems exploiting the alignment of text with audio at various levels (letters, phones syllables, words, chunks, etc). Synchronous reading consists of using time-aligned text with speech at the phone level to displace a cursor – here a virtual finger – on the text in synchrony with its verbalization. We demonstrate here that this bimodal reading implicitly facilitates the learning of the correspondence between sounds and letters in French for native and foreign subjects. Native subjects are shown to benefit more strongly from synchronous reading.

Index Terms: synchronous reading, phonetic alignment

1 Introduction

Impressive performance of speech technology has raised the expectations placed on the computer as a potential learning tool and computer-assisted language learning (CALL) has emerged as a way to supplement or replace traditional student-teacher interaction. Automatic Speech Recognition (ASR) is notably used in applications for pronunciation training [1-2] whereby the speech of a learner is compared to a model utterance, in order to give a feedback score. One of the main advantages of this technology is that it encourages the learner to practice speaking in the second language. However, one problem is that the technology is not perfect and errors made by the software can be highly frustrating and off-putting. More importantly, the process of feedback generation is still a research topic; computers are still unable to reliably give feedback to the users which will help them correct errors in their pronunciation.

Most of our procedural knowledge comes both from explicit and implicit training [3-5]. There is notably considerable evidence that, even without formal instruction, young children produce spellings that correspond to orthographic regularities and phonetic features of words [6]. Much of incidental vocabulary learning comes from context during reading: people who read more know more vocabulary – this relationship between print exposure and vocabulary holds even when intelligence is controlled [7] – and reading affords acquisition of language structure – from patterns of grapheme-phoneme to syntactic constructions.

We question here the ability of synchronous reading – computerized finger-point reading [8] – to drive cognitive attention and ease implicit learning of grapheme-phoneme mappings.

2 Synchronous reading

There are a number of language teaching applications [9-10] aimed at children and adults alike, which offer a karaoke style of reading, whereby the text is highlighted as a human voice reads out the passage. This multimodal activity, whereby the same information is presented both orally and visually, is claimed to improve learning. It is argued that the redundancy of information in this scenario helps the user to overcome the problem of word decoding in a text only environment or utterance deciphering in a pure audio system.

The Talking Books Project [11] was an in depth study conducted in 10 infant classrooms with the aim of assessing the benefit of using electronic books in reading education. The software displayed text and images of a story book and introduced `read-aloud` features, whereby a child could select to hear a sentence be vocalized and watch the words being highlighted as they are read (text and audio are synchronized). Testing the children's comprehension, via graded story re-telling, and word decoding ability, through word accuracy and error quantification, were conducted before and after use of the software and were contrasted with the results obtained under traditional teaching methods. The results concluded that there was a significant improvement in both comprehension and decoding when electronic books were used. Our study, although being far smaller in scale and targeted at adults, was aimed at answering a similar question but in more specific terms – i.e. whether the multimodality of synchronous reading systems is beneficial in language learning.

This study in fact investigates whether karaoke style reading, where the mapping between orthography and pronunciation is made explicit (with a cursor), increases the learning ability of users when compared to an unsynchronized environment, where no cues are given to the user about which words are being vocalized. The target language is French which has like English a strong irregular grapheme-to-phoneme correspondence.

2.1 Method

Quantifying the amount of information that has been acquired by a user after exposure to a particular computer application is notoriously difficult. In this study we decided to model user performance on one particular feature of his knowledge, word orthography. The principle was that there were two reading environments for participants to use, one application was text-audio synchronized and the other simply displayed the text while its vocal recording is played. Initially, participants would complete an orthography quiz to ascertain their existing knowledge. Improvements to this base level would be evaluated with the same quiz, after each use of the systems under test. We could thus judge one system to be more effective than the other if it improved the user's performance more than the other.

2.2 Target orthography

In a number of studies it has been shown that children and adults alike have difficulty in the correct usage of double consonants in French words [4, 12]. Knowing where

consonants should and should not be doubled can be challenging, especially for irregular words such as “colonel” vs. the miss-written “colonnel”. A list of 120 word pairs was created on this basis. Each pair would give two alternative spellings, one correct and the other false, of a word, for example (quitter, quiter) and (afrique, affrique). The spelling alternatives we based on the principle of doubling the consonants *b,c,d,f,l,m,n,p,r,t,* sometimes it is appropriate (*quitter*) and sometimes not (*affrique*).
To make this more compact for more rapid quiz testing, a subset of the 40 most difficult pairs to spell was extracted. To do this, a web-based[1] orthographic quiz with multiple choice input was conducted on all 167 word pairs on a population of 57 people, with a range of abilities in French. The following table gives a breakdown:

Ability in French *(Common European Framework)*	**% participants**
A1 & A2	14.0%
B1 & B2	15.8%
C1 & C2	17.6%
Native	52.6%

The words that were consistently misspelled most often, across all language ability groups, were selected for the final list of 40 words (see Figure 1 and section 5).

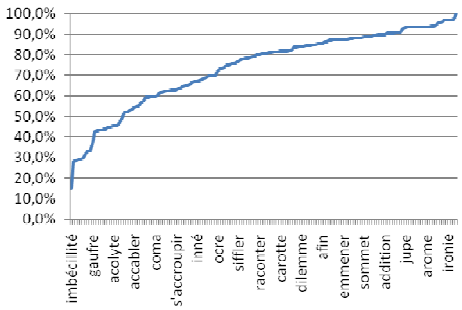

Figure 1: Words ordered by misspelling rate with some being enlightened. The worst word is “imbécillité” often spelt even by French subjects with only one “l” as in the English “imbecility”.

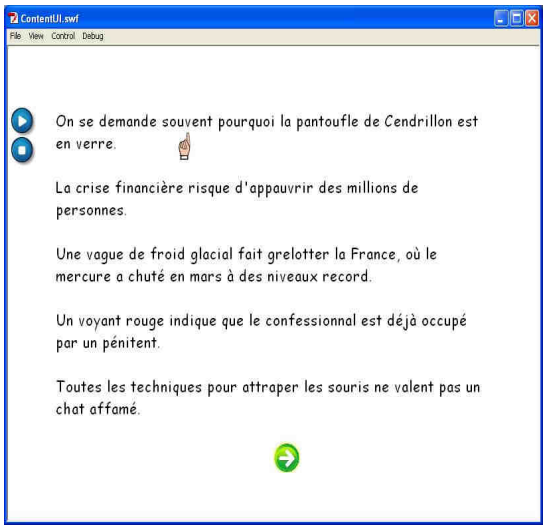

Figure 2. Sample karaoke reading of texts. Widgets enable users to stop and go anywhere. When in karaoke mode, a virtual hand moves smoothly under the current read grapheme.

[1] www.gipsa-lab.fr/~william-seamus.barbour/questions.php

2.3 Karaoke reading

The reading system displays sentences over a set of pages, with buttons to navigate between pages. On each page the user is able to play the audio recording of the text on display. The audio and text are synchronized at the phoneme level. The grapheme-to-phoneme alignment is performed by a data-driven phonetizer trained on an aligned lexicon of 200000 French entries [13]. This alignment allows a cursor to move in real-time with the graphemes that correspond to the currently vocalized phonemes. A smoothing procedure is used to interpolate the motion between synchronization points placed at phone boundaries.
For the non-synchronized version of the system the cursor graphic is not shown.

2.4 Implicit learning of orthography

Since there are two reading systems (synchronous vs. non-synchronous reading) under test, two sets of ten sentences were prepared that embedded twenty words out of these 40 difficult words. These twenty sentences were uttered by a native French male speaker at a normal speech rate (13.9 phones/s, 6 syllables/s). The emotional content of sentences is neutral and the speaker was instructed to read aloud the sentences with no break. None of the target words were enlightened in the text either by the font style or by their positions in the sentences: none of the target words were actually focused by prosody.
We check the spelling performance of all subjects before and after the reading of each set of sentences. The performance was simply monitored by the same orthographic quiz as previously detailed but limited to the 40 most difficult words presented in random order.
The aim of these quizzes was to be able to assess several points:

1. We verify that the spelling performance of the remaining 20 words that are not embedded into the two sets remains stable.
2. We want to observe an orthographic learning effect of exposure to one text, independently from exposure to the other.

Twenty sentences were thus created, each including one word from the word list, recorded and forced-aligned with phone HMM. Manual correction is performed using Praat [14]. During the experiment, participants would be exposed to the first ten sentences in system A and the last ten sentences in system B. System A would be the karaoke version in half the experiments and the non-karaoke version in the rest (see the synopsis of the experiment in Figure 3).

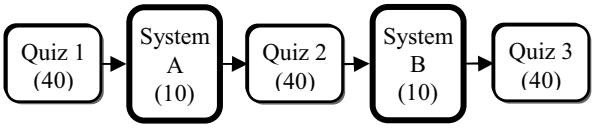

Figure 3: synopsis of the experiment. Three orthographic quizzes are performed to assess the impact of a passive listening of read sentences.

2.5 Experiments

Due to time constraints, not all of the experiments could be undertaken in the laboratory and so a web version of the

application[2] was created to allow remote experiments to take place. To date, 89 people completed the entire experiment. Figure 4 gives a breakdown of the population according to age and proficiency for French. Most non native subjects originate from England and Italy but we have scores from 12 different countries.

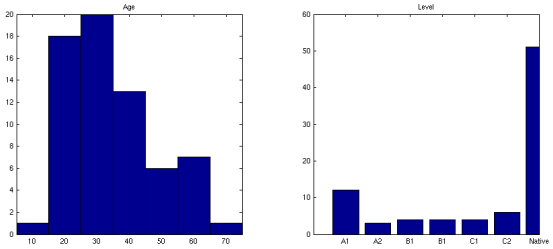

Figure 4: Breakdown of the subjects according to age and proficiency for French.

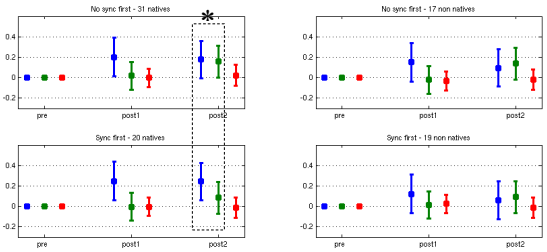

Figure 5. Relative improvement in orthographic quiz score after using each system (bottom synchronous reading first; top: second) in comparison with the pretest for native (left) vs. non native (right) readers. For each of the three quiz, the performance is given for three sets of words from left to right: the 10 words read by system A, the 10 words read by system B and the remaining 20 words of each quiz never read. Significant contrasts ($p<0.1$).are cued by stars.

For each participant, the results from orthography quizzes were normalized into a percentage of the quiz 1, to give an improvement value. The mean 'improvement' score for all participants pooled was calculated for both the words seen in the reading systems and for the unseen words. These mean values are shown in Figure 5. The error bars indicate the standard deviation.

It shows an improvement in orthographic performance for the words seen in the reading systems, which can be contrasted with little to no improvement in the spelling of the unseen words (relative improvements close to 0 for the third set of words in Figure 5). Looking more closely, system B, when in karaoke mode, leads to a greater mean performance increase than when in standard mode for native readers, whatever the order of presentation. Despite overlapping error bars, the mean improvements across system types are statistically significant ($F=2.5$; $p<0.1$). It should be noted that subjects have no training of the system and experienced only one unique right spelling of the target words in context.

The greater benefit of synchronous reading for native subjects can be explained by the fact that the correct spelling of words may refresh prior exposure and knowledge of native speakers. On the contrary, a unique reading experience by non native subjects trigger multimodal exemplars or episodes [15] that are perhaps difficult to combine as such with long-term memory and lexical access (see the tendency of the performance of the first set of words to decline for non-native readers in Figure 5 whatever the order of systems' presentation).

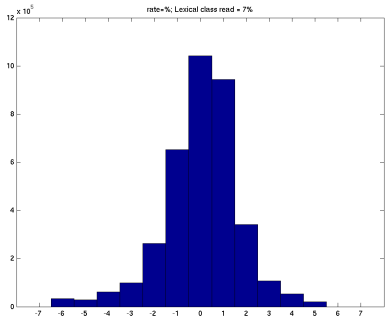

Figure 6: Size of the optimal number of contextual letters (left and right to the current letter) necessary to trigger the appropriate letter-to-sound rule in French.

3 Conclusions and perspectives

The concept of investigation oriented language learning applications [16], where authentic content is the key ingredient, was the motivating factor behind the research. The hope was that by creating a framework, within which applications could be easily and naturally defined, a new breed of CALL software could be introduced to the language student. It is the author's opinion that the increasing prevalence of mobile computing platforms, with tablet devices being a prime example, is providing developers with an exciting opportunity to create a new breed of highly interactive and engaging multimedia CALL applications.

Motivated by the increasing number of reading systems that include this feature, being aimed at children, it was logical to wonder if the same practice could be beneficial to adult learners of a foreign language. Therefore, the main contribution of the experiment, acting as the first of its kind, is to investigate the benefits of karaoke style reading systems in adult learning. Despite the limited size of the study, results show that passive but synchronous reading of audio books provides an implicit learning - or refreshing! - of letter-to-sound mapping in the target language.

Further experimentation is needed to evaluate the learning impact of using karaoke style reading systems for adult language learning. Following on from the work undertaken in our study, the principal requirement is to acquire more experimental data, i.e. finding more volunteers to take part. We should also have added a control experiment using only monomodal/textual presentation. Secondly the use of eye gaze data could be combined with the pre/post orthography tests to form a more concrete analysis of the subjects' behavior. This data was captured in our study but in an insufficient quantity.

We have in project to test the impact of longer exposure to synchronous reading especially with young learners. This is the aim of a large project on the importance and role of the "the visual attention span" [17] in orthographic processing and learning. With this respect, it would be interesting to test if visual information on the size of the optimal orthographic context necessary to trigger the appropriate letter-to-sound rule may contribute to learning performance. Figure 6 shows that up to 6 letters on the left and 5 letters on the right are necessary to trigger the correct spelling of letters in French. This data was obtained using a decision tree applied on aligned corpora [18]. An idea would be to modulate the size of the highlighted text with the size of the context window.

[2] www.gipsa-lab.fr/~william-seamus.barbour/experiment.php

Another key aspect of sound-to-letter mapping is reading proficiency: an improved mapping should enhance reading fluency. Audio-assisted reading has been shown to increase fluency [19]: we expect synchronous reading to greatly impact chunking, decoding, phrasing and thus reading fluency.

As mentioned earlier, one of the premises behind investigation based learning in the language learning context is on providing as many opportunities to the learner to notice features and patterns. With text this can be accomplished easily, with applications highlighting particular aspects of words or sentences, such as in [9]. Doing the same thing with recorded speech by selectively amplifying or otherwise enhancing the signal to exaggerate a specific feature is more challenging, but could be very desirable in CALL. Two relevant studies have been conducted by Hazan et al. [20-21].

4 Acknowledgments

We thank Edouard Gentaz, Eric Petris, Michel Zorman and Marie-Line Bosse for their help. This work has been partially financed by the ANR AMORCES and supported by COST 2102.

5 Annex: most difficult words

Word	Mean. Correct	Word	Mean. Correct
Imbécillité	15,1%	Alaise	45,5%
(Le) Soufre	28,2%	Camouflet	47,2%
Bonhomie	28,6%	Mamelle	49,6%
Apaiser	28,9%	Trafic	52,0%
Oculaire	29,2%	Sabbat	52,3%
Rafler	29,6%	Chuchoter	52,7%
Peccadille	31,6%	Botter	53,3%
Rabbin	33,1%	Accabler	54,2%
Mufle	33,1%	Girafe	54,8%
Gaufre	36,4%	Echafaud	55,3%
Pantoufle	42,6%	Rate	56,8%
Appauvrir	43,1%	Colonel	57,4%
Grelotter	43,3%	Saccade	59,4%
Confessionnal	43,4%	Comité	59,5%
Attraper	43,8%	Illusionniste	59,6%
Sacoche	43,9%	Bagarre	59,8%
Taffetas	44,8%	Coma	59,8%
Acompte	45,0%	Gouffre	60,2%
Acolyte	45,2%	Accroc	61,5%
Agrafe	45,5%	Atroce	61,6%

References

[1] Aist, G., *Speech recognition in computer assisted language learning*, in *Computer Assisted Language Learning (CALL): Media, Design, and Applications*, K.C. Cameron, Editor. 1999, Swets & Zeitlinger: Lisse, The Netherlands. p. 165-182.

[2] Wang, H., C.J. Waple, and T. Kawahara, *Computer Assisted Language Learning system based on dynamic question generation and error prediction for automatic speech recognition.* Speech Communication, 2009. **51**(10): p. 995-1005.

[3] Saetrevik, B., R. Reber, and P. Sannum, *The utility of implicit learning in the teaching of rules.* Learning and Instruction, 2006. **16**: p. 363-373.

[4] Deacon, S.H., N. Conrad, and S. Pacton, *A statistical learning perspective on children's learning about graphotactic and morphological regularities in spelling.* Canadian Psychology, 2008. **49**: p. 118-124.

[5] Steffler, D.J., *Implicit cognition and spelling development.* Developmental Review, 2001. **21**: p. 168-204.

[6] Treiman, R., *Beginning to spell*. 1993, New York: Oxford University Press.

[7] Stanovich, K.E. and A.E. Cunningham, *Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure.* Memory and Cognition, 1992. **20**: p. 51-68.

[8] Ehri, L. and J. Sweet, *Finger-point reading of memorized text: What enables beginners to process the print?* . Reading Research Quarterly, 1991. **26**: p. 442-462.

[9] Guez, M. *So Ouat! Edutainment apps*. 2010; Available from: http://www.so-ouat.com.

[10] Dorfman, K. *Audio-text synchronized books. Learning foreign languages through audiovisual methodology*. 2010; Available from: http://www.interactiveselfstudy.com.

[11] Medwell, J., *The talking books project: some further insights into the use of talking books to develop reading.* Reading, 1998. **32**(1): p. 3-8.

[12] Cassar, M. and R. Treiman, *The beginnings of orthographic knowledge: Children's knowledge of double letters in words.* Journal of Educational Psychology, 1997. **89** p. 631-644.

[13] Hunt, A.J. and A.W. Black. *Unit selection in a concatenative speech synthesis system using a large speech database*. in *International Conference on Acoustics, Speech and Signal Processing*. 1996. Atlanta, GA.

[14] Boersma, P. and D. Weenink, *Praat, a System for doing Phonetics by Computer, version 3.4*, in *Institute of Phonetic Sciences of the University of Amsterdam, Report 132. 182 pages*. 1996.

[15] Goldinger, S.D., *Echoes of echoes? An episodic theory of lexical access.* Psychological Review, 1998. **105**: p. 251-279

[16] Tan, M.M.L., *Conscious investigation and investigative-oriented learning (IOL) in language teaching.* CAUCE, Revista de Filologá y su Didáctica, 2001. **24**: p. 225-238.

[17] Prado, C., M. Dubois, and S. Valdois, *The eye movements of dyslexic children during reading and visual search: Impact of the visual attention span.* Vision Research, 2007. **47**: p. 2521-2530.

[18] Black, A., K. Lenzo, and V. Pagel. *Issues in building general letter-to-sound rules*. in *ESCA Workshop on Speech Synthesis*. 1998. Jenolan Caves, Australia.

[19] Rasinski, T.V., *Effects of repeated reding and listening-while-reading on reading fluency.* Journal of Educational Research, 1990. **83**(3): p. 147-150.

[20] Hazan, V. and A. Simpson, *The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects.* Language and Speech, 2000. **43**(3): p. 273-294.

[21] Iverson, P., V. Hazan, and K. Bannister, *Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults.* The Journal of the Acoustical Society of America, 2005. **118**: p. 3267-3278.