Proceedings of the International Workshop
on Modern Science and Technology
Kitami, Japan/September 2010

Co-Mix Project: Towards Artificial Tutors Using Code Mixing as Foreign Language Teaching Method

Michal Mazur, Rafal Rzepka**, Kenji Araki***

** Department of Information and Management Science, Otaru University of Commerce, Otaru, Japan; mazzi@mazzi.pl*

*** Graduate School of Information Science and Technology
Hokkaido University, Sapporo, Japan; {kabura, araki}@media.eng.hokudai.ac.jp*

Abstract: It has been proven that code-mixing phenomenon may be a useful method to teach foreign language vocabulary. However, this effective method has faced some potential drawbacks caused by limiting its application to aural comprehension. In this paper we propose an innovative idea of an artificial companion that would assist in acquiring the foreign language and progress with the research on code-mixing in the field of natural language processing. The main goal of Co-Mix project is to create a novel chatterbot that would be capable of holding a simple conversation, utilizing the core set of phrases and allowing the user to elaborate on them. I present the current state of research on artificial tutors, the result of preliminary experiment and the idea of an algorithm that would estimate how many, and what kind of words should be replaced in accordance to code-mixing fundamental assumptions and individual language skills of a given student.

Keywords: Artificial Tutors, Code-Mixing, Human-Computer Interaction, Naturally Talking Machines, Second Language Acquisition

1. Introduction

Nowadays getting proper education has been essential in one's effort to find a good employment opportunity. As the world is becoming a "global village" the knowledge of foreign languages is a necessary skill to face the expectations of a modern world. Lessons with native speakers greatly enhance the acquisition of a foreign language, but it is very difficult to have such opportunity on a daily basis. The artificial tutor might be the answer to this problem by taking into account that learning process is continuous and provides the guidance on a daily basis. When access to real teachers is often limited, a tutoring system might be a sufficient solution in a situation where there are no other alternatives and make it possible to learn both at home or at any remote location. Free talking conversational agents offer a chance for the users to interact and learn in a practical and effective way, virtually anytime and anywhere. Such tutors could be placed in natural human environment, i.e. among children, that can enhance their linguistic knowledge.

2. Code-mixing

Code-mixing is often mentioned as one of the stages of bilingual children development, as they naturally mix elements of different languages and move from one system to another without noticeable discrimination [1]. This phenomenon reflects the transition between linguistic units (words) of one language into another, within one sentence, when original grammar of native language is usually preserved unchanged. Code-mixing occurs when a speaker takes some components of one language to use it while speaking another. Sridhar explains it as a kind of transition of units of one language (L1; mother language) to using those of another (L2; foreign language) within a single sentence [2], e.g. "Sorry, I can't talk right now. I'm going to start 発表 (happyou) soon." (English-Japanese code-mixed phrase; a real life example) for "Sorry, I can't talk right now. I'm going to start presentation soon".

During English conversational classes teaching vocabulary is a significant part of the course. Numerous methods exist to improve students' knowledge of foreign words. Celic believes that one's L2 vocabulary and other linguistic components are organized in the same way the synonymic expressions are organized in L1 [4]. To be precise, there might be a link between two languages by associations and semantics, therefore it is possible to consciously switch elements between them.

Our interest centers on using a code-mixing effect to introduce new linguistic items. The units such as words, phrases or clauses can be interchanged and by understanding the context of the whole phrase a student would be able to understand a foreign word without knowing the direct translation. By inserting a single L2 words into the L1 sentence it is possible to understand its meaning naturally. Some of the existing works on teaching vocabulary through code-mixing, especially the one by Celic show that this phenomenon can be effectively used during EFL (English as a foreign language) classes [4]. Usually the new vocabulary is presented by various pictures, language-to-language translations or mimicry. Code-mixing brings a chance for students to think deeply about new linguistic units and make connections between meanings in a natural way.

3. State of the Art

Code-mixing has been successfully used to introduce new vocabulary during English classes. Celic used oral method giving the students a kind of a listening task. He introduced the story and elaborated on it by inserting the new vocabulary in L1. He concluded his research with an assertion that the careful use of code-mixing can lead to successful teaching and learning of new vocabulary in foreign language class. An absence of visual support of the input was mentioned as one of the potential drawbacks. Students could not see the phrases in correctly spelt form and often misspelled newly-acquired words. It is possible that combining the oral method with visual support provided by a machine may help to avoid this problem.

There have been some considerable attempts in creating an artificial tutor. Two "Robovie" robots created by scientists from Kioto and Osaka were tested in 18-day field trial held at a Japanese elementary school [6]. California Institute of Telecommunications and Information Technology's Perception Lab created a robot named RUBI that taught elementary school children some basic foreign language words [7]. To our knowledge there is no existing research concerning the use of code-mixing in creating an artificial tutor able to teach foreign language vocabulary. Therefore, Co-Mix project can be seen as a novel contribution to this field.

Fryer et al. mentioned in their paper that learning with a help of chatterbot is a best challenge for more advanced students and do not always meet beginner students' need [3]. Therefore, it is important to make this technology more accessible also for the inexperienced English students. 211 students participating in the experiment had 20 minutes-long conversation sessions with two popular conversational agents, ALICE and Jabberwacky [3]. Chatterbots offering an engaging and non-stressful conversation, are able to repeat the same material with a student as long as necessary and, compared to their human counterparts, do not lose their patience. The real extension of the usefulness of artificial conversational agents is wide, and enriching it with synthesized speech technology, the ability of recognizing and reacting to emotions and sense of humor might offer much more than just mere engaging dialogue.

4. Co-Mix Project

The main purpose of Co-Mix project is to create a solid foundation for further research on artificial tutors by providing a functional chatterbot system able to teach user a foreign language using context-learning method. We would like to create a system that could be a valuable language learning tool, at first used as a supplement to normal learning, but ultimately able to function as an independent mean for teaching. The first step in development of such artificial tutor is to obtain vocabulary and create a lexical corpus. The next goal is to implement a context learning method based on code-mixing phenomenon into chatterbot. This method is based on code-mixed phrases used in the human-machine conversation. Firstly, the chatterbot would use a simple QA (Question Answering) strategy, but gradually we would like to improve its use with additional functionalities. During such conversation students will encounter L2 lexis and by establishing a connection between L1 and L2 meaning of a certain word they might be able to equate their meaning. That would help students to establish new connections between related lexical fields in their inner L1/L2 lexicons.

We expect Co-Mix project not only to be a useful way to introduce new vocabulary, but also to practice various language structures. Studying the dialogue logs create new possibilities for teachers to check the type of language structures and vocabulary students use and give feedback to students' efforts. It opens a possibility to extend the use of such logs to create a program that would analyze them and learn from the most common mistakes. It might be a chance of finding some statistically wrong tendencies that normally would escape human teacher's attention. Co-Mix project aims on creating a chatterbot system that would provide students with a chance to acquire foreign language in a natural way, give a necessary practice and a chance of self-improvement in a non-classroom environment.

Co-Mix Project development stages:

1. Lexical resources acquisition (obtain L1 and corresponding L2 vocabulary to create a lexical corpus).

2. System development (create a system able to generate code-mixed phrases).

2. Implementation into a chatterbot

3. Evaluation experiment

5. Co-Mix method

Our current point of interest is code-mixing between English and Japanese languages, because Japanese are quite familiar with borrowing foreign words and drawing them into their own language. The best example of this process is the popularity of katakana, one of the Japanese syllabifies used to represent foreign words and names, e.g. television is written "terebi" (テレビ) and communication is written "komunikeeshon" (コムニケーション). Katakana language has adapted many different loanwords, especially in the fields of economy and media, where the degree of usability is exceptionally high. However, there are no established norms of using those words, as they do not keep their original accents and are mostly pronounced in accordance to Japanese phonology. Therefore, they tend to be undistinguishable to non-Japanese speakers. They usually have one semantic meaning, differently than in English, where, on the contrary, words tend to have more than one meaning. It happens with words such as "naive" that in Japanese has the positive meaning of delicacy and sensitivity, when in English also presents the negative connotation. The loan words in Japanese sometimes bring new meaning, e.g. "teeburu" (テーブル) does not mean the same as English "table" and "sukuuru" (スクール) is not the synonym of the word "school". This uniqueness of Japanese language shows the challenge an English teacher has to face, because the students usually are not able to distinguish the foreign and native words if they are not presented with the visual input, such as phrase written in katakana. It is important to make students aware of differences between both languages, but due to a fact that many loanwords already exist in their minds it is possible to expedite their foreign vocabulary learning.

Studies conducted by Fryer and al. [3] on using chatterbots during English classes concluded with a result stating that most students enjoyed talking to chatterbot and felt more comfortable having a conversation with bots than with a teacher or other student. Over 85% respondents enjoyed the conversation with the artificial agent. This research also underlines the fact that the classical classroom setting proves to be non-interactive enough due to limited chance of practice with a real teacher. The other problems include a general shyness and the lack of confidence of students, as well as lack of time to check their mistakes in grammar or pronunciation. In the Co-Mix method we would like to emphasize hybridization of a language and the necessity of constant repetition to memorize the new vocabulary. Repeating the same phrases with different changes would increase child's performance. Using the conversational agent system gives a chance to reach beyond the normal task-oriented learning, and also create an opportunity for dialogue in English. In our research we would like to use Modalin, a non-task oriented keyword-based agent developed in Language Media Laboratory at Hokkaido University [5]. Using it as a foundation implementing new solutions and adapting to perform new tasks might be highly beneficial contribution to the subject of chatterbots as learning tools.

6. Preliminary experiment

In the experiment we asked two groups of 10 participants to memorize a set of cards representing different code-mixed phrases. The average age of participants was 31,4. The purpose of this task was to evaluate whether the code-mixing method might be useful for acquiring the new L2 vocabulary. We assumed that by understanding the context of the whole phrase, participants would guess the foreign word without knowing the direct translation. To prove this hypothesis we decided to use three core phrases in English, and then modified them with appropriate Japanese vocabulary for the purpose of the experiment. Our main interest centered on how efficient the participant could acquire new vocabulary. All the cards were presented twice – firstly we made a trial and secondly the real test. We selected participants in accordance to their language skills. In the first experiment all participants were native L1 speakers (Japanese) of different age and gender, and represented beginner level of English. In the second experiment participant's L1 varied, but they all could use upperintermediate/advanced/native level of English and got first contact with Polish vocabulary.

Core phrases:

1. This is a red apple.

2. I watch television every day.

3. My friend lives in a big apartment.

During experiment the first phrase appeared three times in the following form:

a) これは赤い APPLE だ。(Kore wa akai APPLE da.)

b) これは RED りんごだ。(Kore wa RED ringo da.)

c) This is a 赤いりんご。(This is a akai ringo.)

The second phrase was modified in the following way:

a) 毎日、television をみる。(Mainichi, television wo miru.)

b)Everyday, テレビを見る。(Everyday,terebi wo miru.).

c) 毎日、テレビを watch。(Mainichi,terebi wo watch.).

The last phrase was presented in the following form:

a) My 友だち lives in a big apartment. (My tomodachi lives in a big apartment.).

b) My friend lives in a 大きいなマンション。(My friend lives in a Nokii na manshon.).

c) My friend 住んでいる a big apartment. (My friend sundeiru a big apartment.).

The whole experiment took approximately 8 minutes for each person. The participants were presented with cards containing modified phrases and their task was to memorize them and, finally, reconstruct them in both L1 and L2.

One of the drawbacks that we took into consideration was the fact that English and Japanese grammar system are very distant and some problems may occur while memorizing the vocabulary. In addition, we performed the same experiment with Polish and English language, whose grammar systems share some similarities.

Example 1:

a) My przyjaciel lives in a big apartment.

b) My friend mieszka w big apartment.

c) My friend lives in a duzym mieszkaniu.

Example 2:

a) Everyday watch telewizje.

b) Codziennie watch television.

c) Everyday ogladam television.

Example 3:

a)This is a czerwone apple.

b)To jest czerwone apple.

c)This is a red jablko.

7. Results

In the first experiment 5 out of 10 participants successfully memorized the vocabulary and recreated the core phrases both in L1 and L2 (50%). 3 out of 10 participants partially recreated the phrase (30%), and 2 out of 10 participant did not manage to perform the task in the sufficient way (20%).

Table 1 Experiment results for English-Japanese code-mixed phrases.

Participant #	Pass / Tested	Percentage
1	3/3	100%
2	2/3	66%
3	2/3	66%
4	3/3	100%
5	3/3	100%
6	3/3	100%
7	1/3	33%
8	3/3	100%
9	1/3	33%
10	2/3	66%

Table 2 The percentage of correct word associations by participants (English-Japanese).

100%	66%	33%
5	3	2

In the second experiment 6 out of 10 participants managed to recreate sentences in both L1 an L2 languages (60%). 3 out of 10 participants passed 2 of 3 tests (30%) and 1 participant did not manage to perform the task in the sufficient way (10%).

Table 3 Experiment results for English-Polish code – mixed phrases.

Participant #	Pass / Tested	Percentage
1	3/3	100%
2	3/3	100%
3	2/3	66%
4	2/3	66%
5	3/3	100%
6	3/3	100%
7	1/3	66%
8	3/3	100%
9	1/3	33%
10	3/3	100%

Table 4 The percentage of correct word associations by participants (Polish-English).

100%	66%	33%
6	3	1

Both performed experiments proved the Co-Mix method to be effective way of acquiring a new vocabulary. Most of the participants succeeded in remembering and recreating phrases completely or partially, with a small percentage of participants who did not manage to perform the task in a sufficient way. A slightly better result for Polish-English code-mixing experiment may indicate that coping with distant grammar systems within one sentence may, as expected, cause some possible drawbacks. Every language has its own characteristics so it might be necessary to adjust the method to the specificity of each system.

8. Co-Mix Algorithm

In the future work we would like to use Co-Mix algorithm (Fig.1.) to create a system able to generate code-mixed sentences and then implement it into chatterbot. The algorithm contains the following elements:

a) Level assessment
b) Vocabulary sets
c) Extended database
d) Progress estimation

We divided the sets of words into four age groups (0-5; 6-10; 11-15: 16-20 years) according to their intricacy and one optional set with more specialized vocabulary for advanced learners. We would like to acquire words straight from Internet from various websites by calculating commonness degree of extracted words. The sets will be evaluated by human judges who would comment the selection and give reference if gathered lexis is appropriate for specific age group and understood.

The algorithm will also analyze the content of websites to provide a credible list of most frequent words. The additional content for some specific age groups (children) could be acquired from free internet services such as Project Gutenberg that offers a great selection of children books and online dictionaries[8]. The other possible sources are websites like Children's Storybook Online that contain children's stories for kids of all ages [9].

The algorithm is designed to estimate the students' progress and if they succeed in acquiring the vocabulary of certain level it automatically proceeds to the next level (next set of words). If the progress is not satisfying the students' will continue learning on the certain level as long as it is necessary. One of the problems we would like to solve in the future work is the issue of generating code-mixed phrases using two languages from distant grammars. The research on this subject is ongoing and results are planned to be presented soon.

9. Conclusion

In this paper we introduce the idea of an artificial tutor able to teach English vocabulary using Co-Mix method. We presented results of our preliminary experiment and have shown that our method might be an effective way of expanding one's L2 vocabulary. The next step will be to create the set of words to be used by chatterbot and check how this method will function in the casual conversation. Technical possibilities given by such tutor could lead to enrichment of the code mixing teaching method with additional layers, for instance, visual support (by seeing the phrase in its correctly spelt form, or by illustrating sentences with images). Impression survey carried out by Fryer at al. shown the positive attitude of students towards the artificial tutors, but a more precise empirical research has to be done to answer the question whether artificial tutors can be a valuable learning tools. With naturally speaking artificial tutors assisting people on daily basis a process of learning can become an enjoyable experience.

References:

[1] King K. A. (2006), "Child language acquisition," in Fasold R. and Connor-Linton J. *An Introduction to Language and Linguistics*, Cambridge University Press, pp. 205-224.

[2] Sridhar S. N. and Kamal K. (1980), The syntax and psycholinguistics of bilingual code-mixing, *Canadian Journal of Psychology 34(4)*: 407-416.

[3] Fryer L. and Carpenter R. (2006), "Emerging technologies: Bots as language learning tools", *Language Learning & Technology, 10*(3), pp. 8-14.

[4] Celic M. (2003), Teaching vocabulary through code-mixing, *ELT Journal 57(4):* 361-369.

[5] Higuchi S., Rzepka R. and Araki K. (2008), "A casual conversation system using modality and Word associations retrieved from the Web,", Proc. EMNLP `08, Honolulu, USA, pp.382-390.

[6] Kanda T., Hirano T. and Eaton D. (2004), Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial in *Human-Computer Interaction 19*, pp. 61–84.

[7] Movellan J., Tanaka F., Fasel I. and Taylor C. (2007), The RUBI project: a progress report, *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 333-339 .

[8] http://www.gutenberg.org/

[9] http://www.magickeys.com/books/

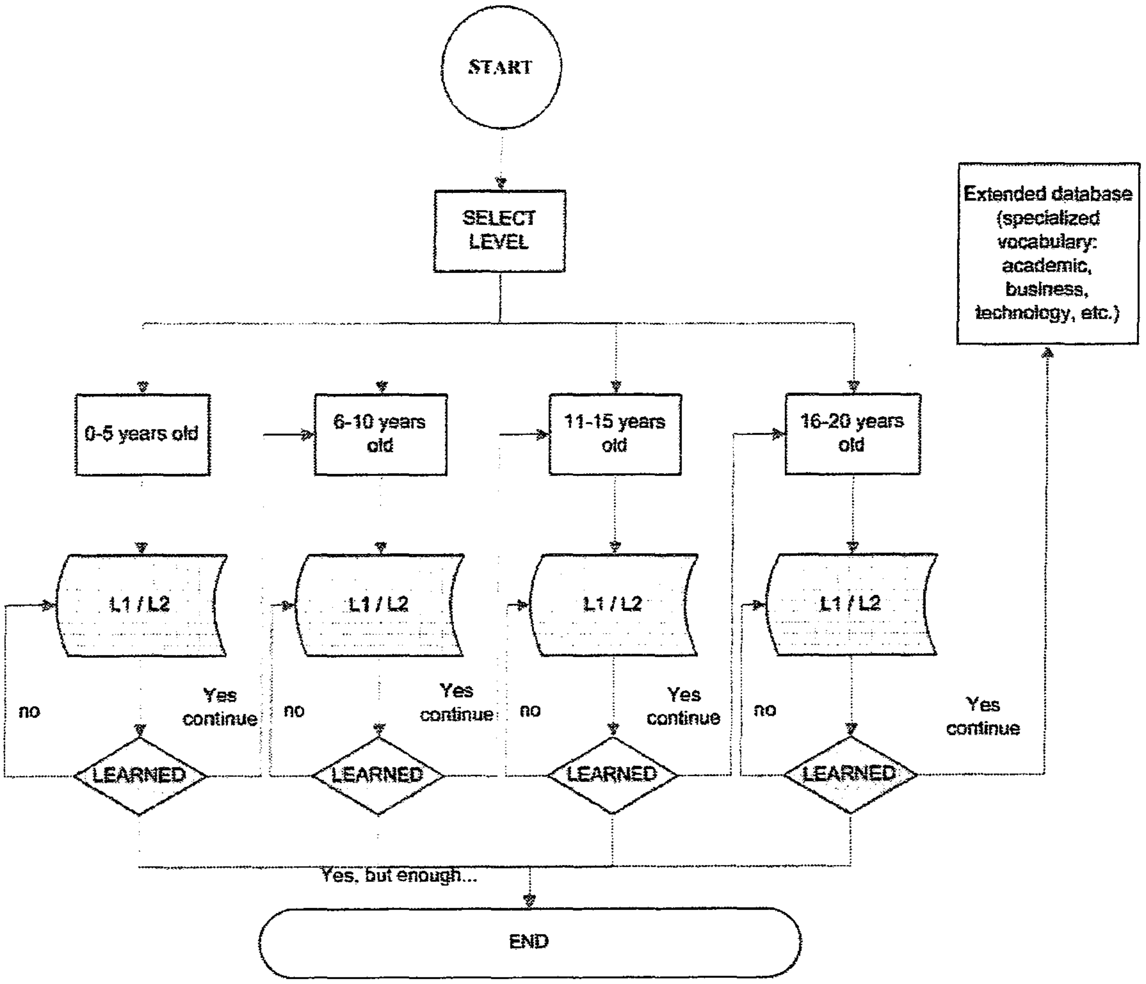

Fig.1. Co-Mix algorithm flowchart.