Generating Code-switched Text for Lexical Learning

Igor Labutov
Cornell University
iil4@cornell.edu

Hod Lipson
Cornell University
hod.lipson@cornell.edu

Abstract

A vast majority of L1 vocabulary acquisition occurs through incidental learning during reading (Nation, 2001; Schmitt et al., 2001). We propose a probabilistic approach to generating code-mixed text as an L2 technique for increasing retention in adult lexical learning through reading. Our model that takes as input a bilingual dictionary and an English text, and generates a code-switched text that optimizes a defined "learnability" metric by constructing a factor graph over lexical mentions. Using an artificial language vocabulary, we evaluate a set of algorithms for generating code-switched text automatically by presenting it to *Mechanical Turk* subjects and measuring recall in a sentence completion task.

1 Introduction

Today, an adult trying to learn a new language is likely to embrace an age-old and widely accepted practice of learning vocabulary through curated word lists and rote memorization. Yet, it is not uncommon to find yourself surrounded by speakers of a foreign language and instinctively pick up words and phrases without ever seeing the definition in your native tongue. Hearing "pass *le sale* please" at the dinner table from your in-laws visiting from abroad, is unlikely to make you think twice about passing the salt. Humans are extraordinarily good at inferring meaning from context, whether this context is your physical surrounding, or the surrounding text in the paragraph of the word that you don't yet understand.

Recently, a novel method of L2 language teaching had been shown effective in improving adult lexical acquisition rate and retention [1]. This technique relies on a phenomenon that elicits a natural simulation of L1-like vocabulary learning in adults — significantly closer to L1 learning for L2 learners than any model studied previously. By infusing foreign words into text in the learner's native tongue into low-surprisal contexts, the lexical acquisition process is facilitated naturally and non-obtrusively. Incidentally, this phenomenon occurs "in the wild" and is termed *code-switching* or *code-mixing*, and refers to the linguistic pattern of bilingual speakers swapping words and phrases between two languages during speech. While this phenomenon had received significant attention from both a socio-linguistic (Milroy and Muysken, 1995) and theoretical linguistic perspectives (Belazi et al., 1994; Bhatt, 1997) (including some computational studies), only recently has it been hypothesizes that "code-switching" is a marking of bilingual proficiency, rather than deficiency (Genesee, 2001).

Until recently it was widely believed that incidental lexical acquisition through reading can only occur for words that occur at sufficient density in a single text, so as to elicit the "noticing" effect needed for lexical acquisition to occur (Cobb, 2007). Recent neurophysiological findings, however, indicate that even a single incidental exposure to a novel word in a sufficiently constrained context is sufficient to trigger an early integration of the word in the brain's semantic network (Borovsky et al., 2012).

An approach explored in this paper, and motivated by the above findings, exploits "constraining" contexts in text to introduce novel words. A state-of-the-art approach for generating such text is based on an expert annotator whose job is to decide which words to "switch out" with novel foreign words (from hereon we will refer to the "switched out" word as the *source* word and to the "switched in" word as the *target* word). Consequently the process is labor-intensive and leads to

[1] authors' unpublished work

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 562–571,
Baltimore, Maryland, USA, June 23-25 2014.

a "one size fits all solution" that is insensitive to the learner's skill level or vocabulary proficiency. This limitation is also cited in literature as a significant roadblock to the widespread adaptation of graded reading series (Hill, 2008). A reading-based tool that follows the same principle, i.e. by systematic exposure of a learner to an incrementally more challenging text, will result in more effective learning (Lantolf and Appel, 1994).

To address the above limitation, we develop an approach for automatically generating such "code-switched" text with an explicit goal of maximizing the lexical acquisition rate in adults. Our method is based on a global optimization approach that incorporates a "knowledge model" of a user with the content of the text, to generate a sequence of lexical "switches". To facilitate the selection of "switch points", we learn a discriminative model for predicting switch point locations on a corpus that we collect for this purpose (and release to the community). Below is a high-level outline of this paper.

- We formalize our approach within a probabilistic graphical model framework, inference in which yields "code-switched" text that maximizes a surrogate to the acquisition rate objective.
- We compare this global method to several baseline techniques, including the strong "high-frequency" baseline.
- We analyze the operating range in which our model is effective and motivate the near-future extension of this approach with the proposed improvements.

2 Related Work

Our proposed approach to the computational generation of code-switched text, for the purpose of L2 pedagogy, is influenced by a number of fields that studied aspects of this phenomenon from distinct perspectives. In this section, we briefly describe a motivation from the areas of socio- and psycho- linguistics and language pedagogy research that indicate the promise of this approach.

2.1 Code-switching as a natural phenomenon

Code-switching (or code-mixing) is a widely studied phenomenon that received significant attention over the course of the last three decades, across the disciplines of sociolinguistics, theoretical and psycholinguistics and even literary and cultural studies (predominantly in the domain of *Spanish-English* code-switching) (Lipski, 2005).

Code-switching that occurs naturally in bilingual populations, and especially in children, has for a long time been considered a marking of incompetency in the second language. A more recent view on this phenomenon, however, suggests that due to the underlying syntactic complexity of code-switching, code-switching is actually a marking of bilingual fluency (Genesee, 2001). More recently, the idea of employing code-switching in the classroom, in a form of conversation-based exercises, has attracted the attention of multiple researchers and educators (Moodley, 2010; Macaro, 2005), yielding promising results in an elementary school study in South-Africa.

2.2 Computational Approaches to Code-switching

Additionally, there has been a limited number of studies of the computational approaches to code-switching, and in particular code-switched text generation. Solorio and Liu (2008), record and transcribe a corpus of Spanish-English code-mixed conversation to train a generative model (Naive Bayes) for the task of predicting code-switch points in conversation. Additionally they test their trained model in its ability to generate code-switched text with convincing results. Building on their work, (Adel et al., 2012) employ additional features and a recurrent network language model for modeling code-switching in conversational speech. Adel and collegues (2011) propose a statistical machine translation-based approach for generating code-switched text. We note, however, that the primary goal of these methods is in the faithful modeling of the natural phenomenon of code-switching in bilingual populations, and not as a tool for language teaching. While useful in generating coherent, syntactically constrained code-switched texts in its own right, none of these methods explicitly consider code-switching as a vehicle for teaching language, and thus do not take on an optimization-based view with an objective of improving lexical acquisition through the reading of the generated text. More recently, and concurrently with our work, Google's Language Immersion app employs the principle of

code-switching for language pedagogy, by generating code-switched web content, and allowing its users to tune it to their skill level. It does not, however, seem to model the user explicitly, nor is it clear if it performs any optimization in generating the text, as no studies have been published to date.

2.3 Computational Approaches to Sentence Simplification

Although not explicitly for teaching language, computational approaches that facilitate accessibility to texts that might otherwise be too difficult for its readers, either due to physical or learning disabilities, or language barriers, are relevant. In the recent work of (Kauchak, 2013), for example demonstrates an approach to increasing readability of texts by learning from unsimplified texts. Approaches in this area span methods for simplifying lexis (Yatskar et al., 2010; Biran et al., 2011), syntax (Siddharthan, 2006; Siddharthan et al., 2004), discourse properties (Hutchinson, 2005), and making technical terminology more accessible to non-experts (Elhadad and Sutaria, 2007). While the resulting texts are of great potential aid to language learners and may implicitly improve upon a reader's language proficiency, they do not explicitly attempt to promote learning as an objective in generating the simplified text.

2.4 Recent Neurophysiological findings

Evidence for the potential effectiveness of code-switching for language acquisition, stem from the recent findings of (Borovsky et al., 2012), who have shown that even a single exposure to a novel word in a constrained context, results in the integration of the word within your existing semantic base, as indicated by a change in the N400 electrophysiological response recorded from the subjects' scalps. N400 ERP marker has been found to correlate with the semantic "expectedness" of a word (Kutas and Hillyard, 1984), and is believed to be an early indicator of word learning. Furthermore, recent work of (Frank et al., 2013), show that word surprisal predicts N400, providing concrete motivation for artificial manipulation of text to explicitly elicit word learning through natural reading, directly motivating our approach. Prior to the above findings, it was widely believed that for evoking "incidental" word learning through reading alone, the word must appear with sufficiently high frequency within the text, such as to elicit the "noticing" effect — a prerequisite to lexical acquisition (Schmidt and Schmidt, 1995; Cobb, 2007).

3 Model

3.1 Overview

The formulation of our model is primarily motivated by two hypotheses that have been validated experimentally in the cognitive science literature. We re-state these hypotheses in the language of "surprisal":

1. Inserting a *target* word into a **low surprisal** context increases the rate of that word's integration into a learner's lexicon.

2. Multiple exposures to the word in **low surprisal** contexts increases rate of that word's integration.

Hypothesis 1 is supported by evidence from (Borovsky et al., 2012; Frank et al., 2013), and hypothesis 2 is supported by evidence from (Schmidt and Schmidt, 1995). We adopt the term "low-surprisal" context to identify contexts (e.g. n-grams) that are highly predictive of the target word (e.g. trailing word in the n-gram). The motivation stems from the recent evidence (Frank et al., 2013) that low-surprisal contexts affect the N400 response and thus correlate with word acquisition. To realize a "code-switched" mixture that adheres maximally to the above postulates, it is self-evident that a non-trivial optimization problem must be solved. For example, naively selecting a few words that appear in low-surprisal contexts may facilitate their acquisition, but at the expense of other words within the same context that may appear in a larger number of low-surprisal contexts further in the text.

To address this problem, we approach it with a formulation of a factor graph that takes global structure of the text into account. Factor graph formalism allows us to capture local features of individual contexts, such as lexical and syntactic surprisal, while inducing dependencies between consequent "switching decisions" in the text. Maximizing likelihood of the joint probability under the factorization of this graph yields an optimal sequence of these "switching decisions" in the entirety of the text. Maximizing joint likelihood, as we will show in the next section, is a surrogate to maximizing the probability of the learner acquiring novel words through the process of reading the generated text.

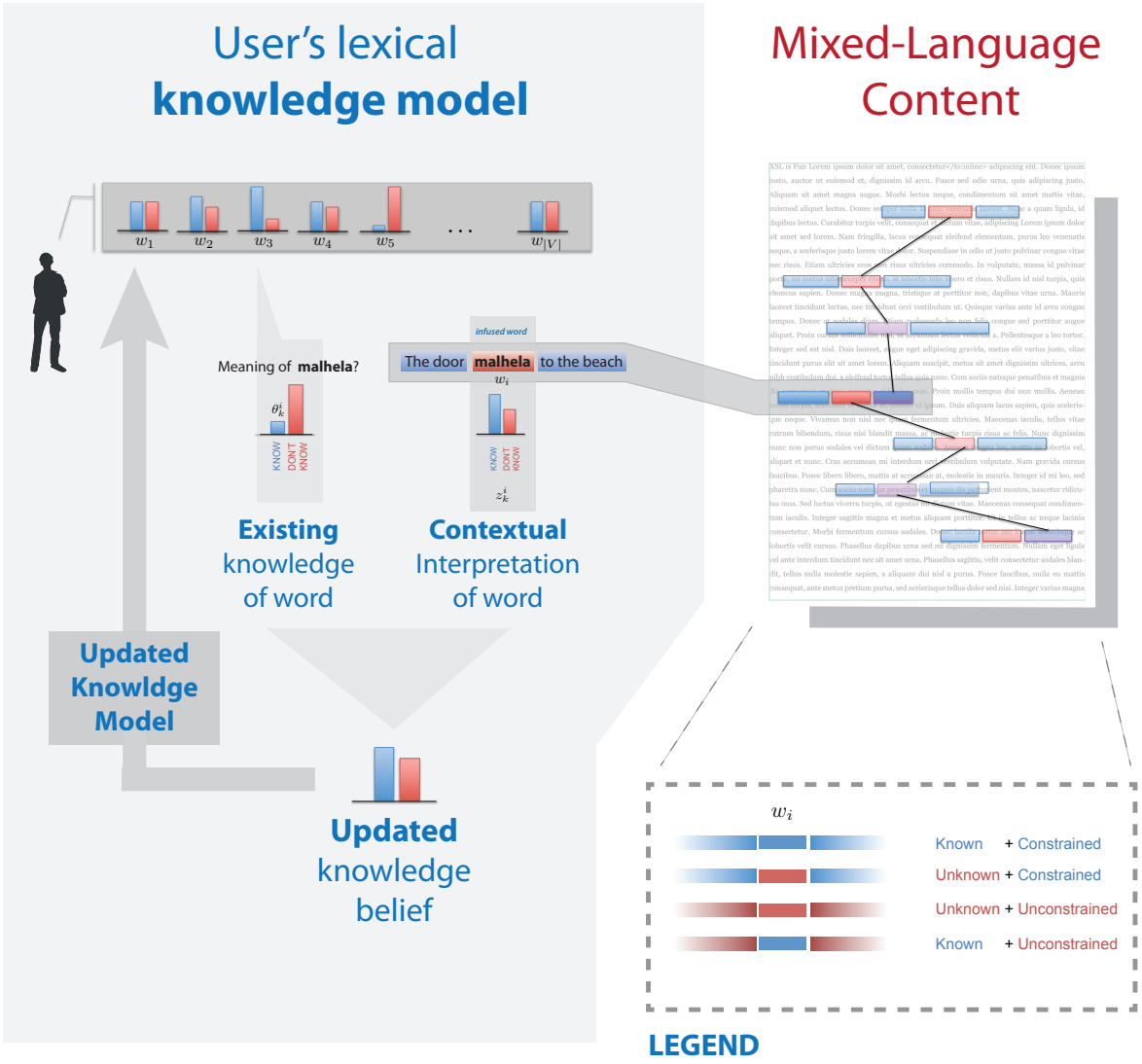

Figure 1: Overview of the approach. Probabilistic learner model (PLM) provides the current value of the belief in the learner's knowledge of any given word. Local contextual model provides the value of the belief in learning the word from the context alone. Upon exposure of the learner to the word in the given context, PLM is updated with the posterior belief in the user's knowledge of the word.

3.2 Language Learner Model

A simplified model of the learner, that we shall term a *Probabilistic Learner Model* (PLM) serves as a basis for our approach. PLM is a model of a learner's lexical knowledge at any given time. PLM models the learner as a vector of independent Bernoulli distributions, where each component represents a probability of the learner knowing the corresponding word. We motivate a probabilistic approach by taking the perspective of measuring *our* belief in the learner's knowledge of any given word, rather than the learner's uncertainty in own knowledge. Formally, we can fully specify this model for learner i as follows:

$$U_i = (\pi_0^i, \pi_1^i, \ldots, \pi_{|V|}^i) \quad (1)$$

where V is the vocabulary set — identical across all users, and π_j^i is our degree of belief in the learner i's knowledge of a target word $w_j \in V$. Statistical estimation techniques exist for estimating an individual's vocabulary size, such as (Bhat and Sproat, 2009; Beglar, 2010), and can be di-

rectly employed for estimating the parameters of this model as our prior belief about user i's knowledge.

The primary motivation behind a probabilistic user model, is to provide a mechanism for updating these probabilities as the user progresses through her reading. Maximizing the parameters of the PLM under a given finite span of code-switched text, thus, provides a handle for generating optimal code-switched content. Additionally, a probabilistic approach allows for a natural integration of the user model with the uncertainty in other components of the system, such as uncertainty in determining the degree of constraint imposed by the context, and in bitext alignment.

3.3 Model overview

At the high level, as illustrated in Figure 1, our approach integrates the model of the learner (PLM) with the local contextual features to update the PLM parameters incrementally as the learner progresses through the text. The fundamental assumption behind our approach is that the learner's knowledge of a given word *after* observing it in a sentence is a function of 1) the learner's previous knowledge of the word, *prior* to observing it in a given sentence and 2) a degree of constraint that a given context imposes on the meaning of the novel word, and is directly related to the surprisal of novel word in that context. Broadly, as the learner progresses from one sentence to the next, exposing herself to more novel words, the updated parameters of the language model in turn guide the selection of new "switch-points" for replacing source words with the target foreign words. In practice, however, this process is carried out implicitly and off-line by optimizing the estimated progress of the learner's PLM, without dynamic feedback. Next, we describe the model in detail.

3.4 Switching Factor Graph Model

To aid in the specification of the factor graph structure, we introduce new terminology. Because the PLM is updated progressively, we will refer to the parameters of the PLM for a given word w_i after observing its k^{th} appearance (instance) in the text, as the learner's *state* of knowledge of that word, and denote it as a binary random variable z_k^i.

$$P(z_k^i = 1) = \begin{cases} \text{Probability that} \\ \text{word } w_i \in V \\ \text{is understood on } k^{th}\text{exposure} \end{cases}$$

Without explicit testing of the user, this variable is hidden. We can view the prior learning model as the parameters of the vector of random variables $(z_0^0, z_0^1, \ldots z_0^{|V|})$.

The key to our approach is in how the parameters of these hidden variables are updated from repeated exposures to words in various contexts. Intuitively, an update to the parameter of z_k^i from z_{k-1}^i occurs after the learner observes word w_i in a context (this may be an n-gram, an entire sentence or paragraph containing w_i, but we will restrict our attention to fixed-length n-grams). Intuitively an update to the parameter of z_{k-1}^i will depend on how "constrained" the meaning of w_i is in the given context. We will refer to it as the "learnability", denoted by L_i^k, of word w_i on its k^{th} appearance, given its context. Formally, we will define "learnability" as follows:

$$P(L_k^i = 1 | w^i, \mathbf{w}^{\setminus i}, \mathbf{z}_k^{\setminus i}) = P(\text{constrained}(w_i) = 1 | \mathbf{w}) \prod_{i \neq j} P(z_k^j = 1) \quad (2)$$

where $\mathbf{w}^{\setminus i}$ represents the set of words that comprise the context window of w_i, not including w_i, and $\mathbf{z}_k^{\setminus i}$ are the states corresponding to each of the words in $\mathbf{w}^{\setminus i}$. $P(\text{constrained}(w_i) = 1 | \mathbf{w})$ is a real value (scaled between 0 and 1) that represents the degree of constraint imposed on the meaning of word w_i by its context. This value comes from a binary prediction model trained to predict the "predictability" of a word in its context, and is based on the dataset that we collected (described later in the paper). Generally, this value may come directly from the surprisal quantity given by a language model, or may incorporate additional features that are found informative in predicting the constraint on the word. Finally, the quantity is weighted by the parameters of the state variables corresponding to the words other than w_i contained in the context. This encodes an intuition that a degree of predictability of a given word given its context is related to the learner's knowledge of the other words in that context. If, for example, in the sentence "pass me the salt and pepper, please", both "salt" and "pepper" are substituted with their foreign translations that the learner is unlikely to know, it's equally unlikely that she will learn them after being exposed to this context, as the context itself will not offer sufficient

information for both words to be inferred simultaneously. On the other hand, substituting "salt" and "pepper" individually, is likely to make it much easier to infer the meaning of the other.

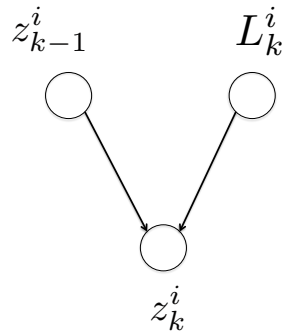

Figure 2: A noisy-OR combination of the learner's previous state of knowledge of the word z^i_{k-1} and the word's "learnability" in the observed context L^i_k

The updated parameter of z^i_k is obtained from a noisy-OR combination of the parameters of z^i_{k-1} and L^i_k:

$$P(z^i_k = 1 | z^i_{k-1}, L^i_k) = \\ 1 - [1 - P(L^i_k = 1)][1 - P(z_{k-1} = 1)]$$

A noisy-OR-based CPD provides a convenient and tractable approximation in capturing the intended intuition: updated state of knowledge of a given word will increase if the word is observed in a "good" context, or if the learner already knows the word.

Combining Equation 2 for each word in the context using the noisy-OR, the updated state for word w_i will now be conditioned on $z^i_{k-1}, \mathbf{z}^{\setminus i}_k, \mathbf{w}_k$. Because of the dependence of each z in the context on all other hidden variables in that context, we can capture the dependence using a single factor per context, with all of the z variables taking part in a clique, whose dimension is the size of the context.

We will now introduce a dual interpretation of the z variables: as "switching" variables that decide whether a given word will be replaced with its translation in the foreign language. If, for example, all of the words have high probability of being known by a learner, than maximizing the joint likelihood of the model will result in most of the words "switched-out" — a desired result. For an arbitrary prior PLM and the input text, maximizing joint likelihood will result in the selection of "switched-out" words that have the highest final probability of being "known" by the learner.

3.5 Inference

The problem of selecting "switch-points" reduces to the problem of inference in the resulting factor graph. Unfortunately, without a fairly strong constraint on the collocation of switched words, the resulting graph will contain loops, requiring techniques of approximate inference. To find the optimal settings of the z variables, we apply the loopy max-sum algorithm. While variants of loopy belief propagation, in general, are not guaranteed to converge, we found that the convergence does indeed occur in our experiments.

3.6 Predicting "predictable" words

We carried out experiments to determine which words are likely to be inferred from their context. The collected data-set is then used to train a logistic regression classifier to predict which words are likely to be easily inferred from their context. We believe that this dataset may also be useful to researchers in studying related phenomena, and thus make it publicly available.

For this task, we focus only on the following context features for predicting the "predictability" of words: n-gram probability, vector-space similarity score, coreferring mentions. N-gram probability and vector-space similarity [2] score are all computed within a fixed-size window of the word (trigrams using Microsoft N-gram service). *Coreference* feature is a binary feature which indicates whether the word has a co-referring mention in a 3-sentence window preceding a given context (obtained using Stanford's CoreNLP package). We train L2-regularized logistic regression to predict a binary label $L \in \{\text{Constrained}, \text{Unconstrained}\}$ using a crowd-sourced corpus described below.

3.7 Corpus Construction

For collecting data about which words are likely to be "predicted" given their content, we developed an Amazon Mechanical Turk task that presented turkers with excerpts of a short story (English translation of "The Man who Repented" by

[2] we employ C&W word embeddings from `http://metaoptimize.com/projects/wordreprs/`

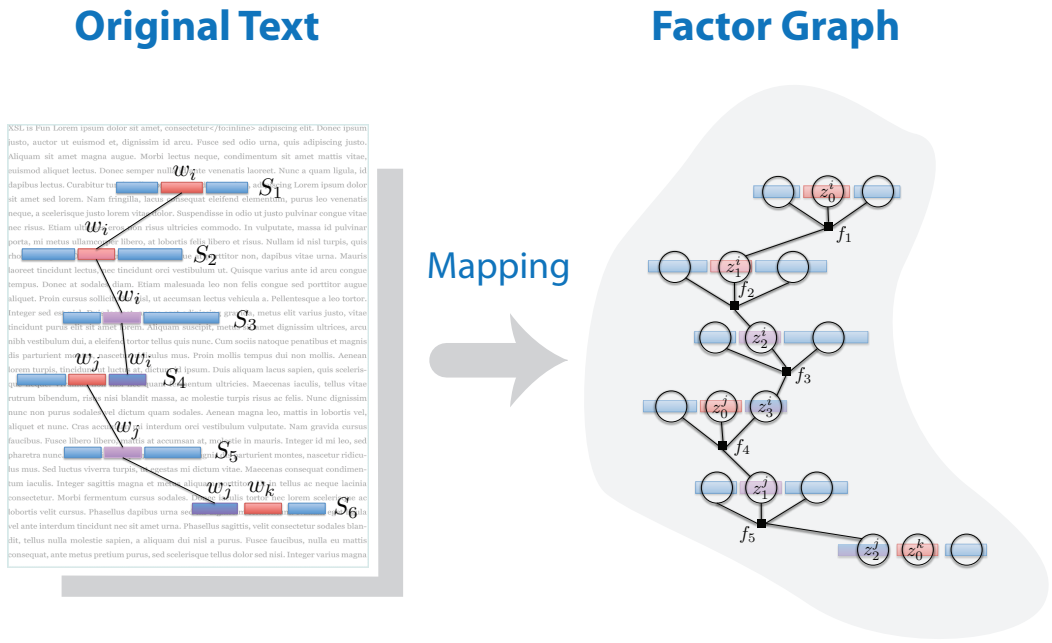

Figure 3: Sequence of sentences in the text (left) is mapped into a factor graph, whose nodes correspond to specific occurences of individual words, connected in a clique corresponding to a context in which the word occurs.

Ana Maria Matute), with some sentences containing a blank in place of a word. Only content words were considered for the task. Turkers were required to type in their best guess, and the number of semantically similar guesses were counted by an average number of 6 other turkers. A ratio of the median of semantically similar guesses to the total number of guesses was then taken as the score representing "predictability" of the word being guessed in the given context. All words corresponding to blanks whose scores were equal to and above 0.6 were than taken as a positive label (Constrained) and scores below 0.6 were taken as a negative label (Unconstrained). Turkers that judged the semantic similarity of the guesses of other turkers achieved an average Cohen's kappa agreement of 0.44, indicating fair to poor agreement.

4 Experiments

We carried out experiments on the effectiveness of our approach using the Amazon Mechanical Turk platform. Our experimental procedure was as follows: 162 turkers were partitioned into four groups, each corresponding to a treatment condition: OPT (N=34), HF (N=41), $RANDOM$ (N=43), MAN (N=44). Each condition corresponded to a model used to generate the presented code-switched text. For all experiments, the text used was a short story "Lottery" by Shirley Jackson, and a total number of replaced words was controlled (34). Target vocabulary consisted of words from an artificial language, generated statically by a mix of words from several languages. Below we describe the individual treatment conditions:

The café was dark and pokey.

The front overlooked the road and the rear the beach.

The door leading to the beach was hung with a reed curtain swaying in the breeze.

With every gust of wind it crackled a little like a slight rattle of bones.

Old Tomeu was sitting on the doorstep, leisurely stroking a well-worn tobacco pouch of black leather.

Figure 4: Visualization of the most "predictable" words in an excerpt from the "The Man who Repented" by Ana Maria Matute (English translation). Font-size correlates with the score given by judge turkers in evaluating guesses of other turkers that were presented with the same text, but the word replaced with a blank. Snippet of the dataset that we release publicly.

RANDOM (Baseline): words for switching are

selected at random from content only words.

HF (*High Frequency*) Baseline: words for switching are selected at random from a ranked list of words that occur most frequently in the presented text.

MAN (*Manual*) Baseline: words for switching are selected manually by the author, based on the intuition of which words are most likely to be guessed in context.

OPT (*Optimization-based*): factor graph-based model proposed in this paper is used for generating code-switched content. The total number of switched words generated by this method is used as a constant for all baselines.

Turkers were solicited to participate in a study that involved "reading a short story with a twist" (title of HIT). Not the title, nor the description gave away the purpose of the study, nor that it would be followed by a quiz. Time was not controlled for this study, but on average turkers took 27 minutes to complete the reading. Upon completing the reading portion of the task, turkers were presented with novel sentences that featured the words observed during reading, where only one of the sentences used the word in a semantically correct way. Turkers were asked to select the sentence that "made the most sense". An example of the sentences presented during the test:

> *Example 1*
>
> ✓ My edzino loves to go shopping every weekend.
>
> The edzino was too big to explore on our own, so went with a group.
>
> English word: **wife**
>
> *Example 2*
>
> ✓ His unpreadvers were utterly confusing and useless.
>
> The unpreadvers was so strong, that he had to go to a hospital.
>
> English word: **directions**

A "recall" metric was computed for each turker, defined as the ratio of correctly selected sentences to the total number of sentences presented. The "grand-average recall" across all turkers was then computed and reported here.

5 Results

We perform a one-way ANOVA across the four groups listed above, with the resulting $F = 11.38$ and $p = 9.7e{-}7$. Consequently, multiple pairwise comparison of the models was performed with the Bonferroni-corrected pairwise t-test, yielding the only significantly different recall means between $HF - MAN$ ($p = 0.00018$), $RANDOM - MAN$ ($p = 2.8e-6$), $RANDOM - OPT$ ($p = 0.00587$). The results indicate that, while none of the automated methods ($RANDOM$, HF, OPT) outperform manually generated code-switched text, OPT outperforms the $RANDOM$ baseline (no decisive conclusion can be drawn with respect to the $HF - RANDOM$ pair). Additionally, we note, that for words with frequency less than 4, OPT produces recall that is on average higher than the HF baseline (p=0.043, Welch's t-test), but at the expense of higher frequency words.

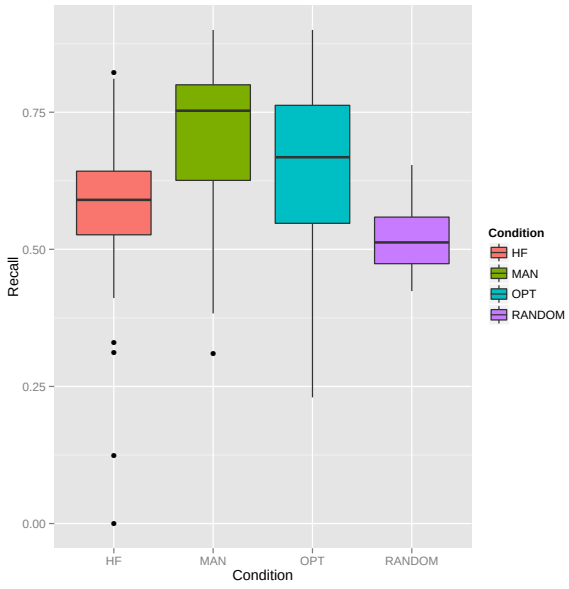

Figure 5: Results presented for 4 groups, subjected to 4 treatment conditions: $RANDOM$, HF, MAN, OPT. Recall performance for each group corresponds to the average ratio of selected sentences that correctly utilize code-switched words in novel contexts, across all turkers.

6 Discussion

We observe from our experiments that the optimization-based approach does not in general outperform the HF baseline. The strength of the

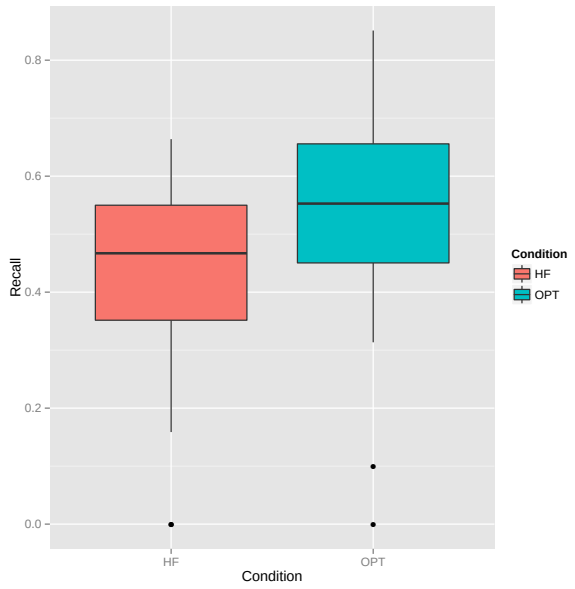

Figure 6: Subset of the results for 2 of the 4 treatment conditions: HF and OPT that correspond to recall only for words with item frequency in the presented text below 4.

frequency-based baseline is attributed to a well-known phenomenon that item frequency promotes the "noticing" effect during reading, critical for triggering incidental lexical acquisition. Generating code-switched text by replacing high frequency content words, thus, in general is a simple and viable approach for generating effective reading-based L2 curriculum aids. However, this method is fundamentally less flexible than the optimization-based method proposed in this paper, for several reasons:

- The optimization-based method explicitly models the learner and thus generates code-switched text progressively more fit for a given individual, even across a sequence of multiple texts. A frequency-based baseline alone would generate content at approximately the same level of difficulty consistently, with the pattern that words that tend to have high frequency in the natural language in general to be the ones that are "switched-out" most often.

- An optimization-based approach is able to elicit higher recall in low frequency words, as the mechanism for their selection is driven by the context in which these words appear, rather than frequency alone, favoring those that are learned more readily through context.

  Moreover, the proposed method in this paper is extensible to more sophisticated learner models, with a potential to surpass the results presented here. Another worthwhile application of this method is as a nested component within a larger optimization-based tool, that in addition to generating code-switched text as demonstrated here, aids in selecting content (such as popular books) as units in the code-switched curriculum.

7 Future Work

In this work we demonstrated a pilot implementation of a model-based, optimization-based approach to content generation for assisting in the reading-based L2 language acquisition. Our approach is based on static optimization, and while it would, in theory progress in difficulty with more reading, its open-loop nature precludes it from maintaining an accurate model of the learner in the long-term. For generating effecting L2 content, it is important that the user be kept in a "zone of proximal development" — a tight region where the level of the taught content is at just the right difficulty. Maintaining an accurate internal model of the learner is the single most important requirement for achieving this functionality. Closed-loop learning, with active user feedback is, thus, going to be functionally critical component of any system of this type that is designed to function in the long-term.

Additionally, our approach is currently a proof-of-concept of an automated method for generating content for assisted L2 acquisition, and is limited to artificial language and only isolated lexical items. The next step would be to integrate bitext alignment across texts in two natural languages, inevitably introducing another stochastic component into the pipeline. Extending this method to larger units, like chunks and simple grammar is another important avenue along which we are taking this work. Early results from concurrent research indicate that "code-switched based" method proposed here is also effective in eliciting acquisition of multi-word chunks.

References

Heike Adel, Ngoc Thang Vu, Franziska Kraus, Tim Schlippe, Haizhou Li, and Tanja Schultz. 2012. Re-

current neural network language modeling for code switching conversational speech. ICASSP.

David Beglar. 2010. A rasch-based validation of the vocabulary size test. *Language Testing*, 27(1):101–118.

Hedi M Belazi, Edward J Rubin, and Almeida Jacqueline Toribio. 1994. Code switching and x-bar theory: The functional head constraint. *Linguistic inquiry*, pages 221–237.

Suma Bhat and Richard Sproat. 2009. Knowing the unseen: estimating vocabulary size over unseen samples. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 109–117. Association for Computational Linguistics.

Rakesh Mohan Bhatt. 1997. Code-switching, constraints, and optimal grammars. *Lingua*, 102(4):223–251.

Or Biran, Samuel Brody, and Noemie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification.

Fabian Blaicher. 2011. *SMT-based Text Generation for Code-Switching Language Models*. Ph.D. thesis, Nanyang Technological University, Singapore.

Arielle Borovsky, Jeffrey L Elman, and Marta Kutas. 2012. Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development*, 8(3):278–302.

Tom Cobb. 2007. Computing the vocabulary demands of l2 reading. *Language Learning & Technology*, 11(3):38–63.

Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 49–56. Association for Computational Linguistics.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2013. Word surprisal predicts n400 amplitude during reading. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*, pages 878–883.

Fred Genesee. 2001. Bilingual first language acquisition: Exploring the limits of the language faculty. *Annual Review of Applied Linguistics*, 21:153–168.

David R Hill. 2008. Graded readers in english. *ELT journal*, 62(2):184–204.

Ben Hutchinson. 2005. Modelling the substitutability of discourse connectives. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 149–156. Association for Computational Linguistics.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of ACL*.

Marta Kutas and Steven A Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*.

James P Lantolf and Gabriela Appel. 1994. *Vygotskian approaches to second language research*. Greenwood Publishing Group.

John M Lipski. 2005. Code-switching or borrowing? no sé so no puedo decir, you know. In *Selected Proceedings of the Second Workshop on Spanish Sociolinguistics*, pages 1–15.

Ernesto Macaro. 2005. Codeswitching in the l2 classroom: A communication and learning strategy. In *Non-native language teachers*, pages 63–84. Springer.

Lesley Milroy and Pieter Muysken. 1995. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge University Press.

Visvaganthie Moodley. 2010. Code-switching and communicative competence in the language classroom. *Journal for Language Teaching*, 44(1):7–22.

Ian SP Nation. 2001. *Learning vocabulary in another language*. Ernst Klett Sprachen.

Richard C Schmidt and Richard W Schmidt. 1995. *Attention and awareness in foreign language learning*, volume 9. Natl Foreign Lg Resource Ctr.

Norbert Schmitt, Diane Schmitt, and Caroline Clapham. 2001. Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language testing*, 18(1):55–88.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2004. Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 896. Association for Computational Linguistics.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 973–981. Association for Computational Linguistics.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368. Association for Computational Linguistics.