Spelling-Aware Construction of Macaronic Texts for Teaching Foreign-Language Vocabulary

Adithya Renduchintala and **Philipp Koehn** and **Jason Eisner**
Center for Language and Speech Processing
Johns Hopkins University
{adi.r,phi}@jhu.edu jason@cs.jhu.edu

Abstract

We present a machine foreign-language teacher that modifies text in a student's native language (L1) by replacing some word tokens with glosses in a foreign language (L2), in such a way that the student can acquire L2 vocabulary simply by reading the resulting *macaronic* text. The machine teacher uses no supervised data from human students. Instead, to guide the machine teacher's choice of which words to replace, we equip a cloze language model with a training procedure that can incrementally learn representations for novel words, and use this model as a proxy for the word guessing and learning ability of real human students. We use Mechanical Turk to evaluate two variants of the student model: (i) one that generates a representation for a novel word using only surrounding context and (ii) an extension that also uses the spelling of the novel word.

1 Introduction

Reading plays an important role in building our native language (L1) vocabulary (Nation, 2001). While some novel words might require the assistance of a dictionary, a large portion are acquired through *incidental learning* – where a reader, exposed to a novel word, tries to infer its meaning using clues from the surrounding context and spelling (Krashen, 1989). An initial "rough" understanding of a novel word might suffice for the reader to continue reading, with subsequent exposures refining their understanding of the novel word.

Our goal is to design a machine teacher that uses a human reader's incidental learning ability to teach foreign language (L2) vocabulary. The machine teacher's modus operandi is to replace L1 words with their L2 glosses, which results in a *macaronic* document that mixes two languages in an effort to ease the human reader into understanding the L2. While some of our prior work (Renduchintala et al., 2016b,a) considered incorporating other features of the L2 such as word order and fixed phrases, in this paper we only consider simple lexical substitutions.

Our hope is that such a system can augment traditional foreign language instruction. As an example, consider a native speaker of English (learning German) presented with the following sentence: **Der** Nile is a **Fluss** in **Afrika**. With a little effort, one would hope the student could infer the meaning of the German words because there is sufficient contextual information and spelling information for the cognate **Afrika**.

In our previous papers on foreign language teaching (Renduchintala et al., 2016b; Knowles et al., 2016; Renduchintala et al., 2017), we focused on fitting detailed models of students' learning when the instructional stimuli (macaronic or otherwise) were chosen by a simple random or heuristic teaching policy. In the present paper, we flip the emphasis to choosing *good* instructional stimuli—machine teaching. This still requires a model of student learning. We employ a reasonable model that is not trained on any human students at all, but only on text that a generic student is presumed to have read. Thus, our model is not personalized, although it may be specialized to the domain of L1 text that it was initially trained on.

That said, our model is reasonably sophisticated and includes new elements. It uses a neural cloze language model (in contrast to the weaker pairwise CRF model of Renduchintala et al. (2016b)) to intelligently guess the meaning of L2 words in full macaronic sentential context. Guessing actually takes the form of a learning rule that jointly improves the embeddings of all L2 words in the sentence. This is our simulation of incidental learning which accumulates over repeated exposures to the same L2 words in different contexts.

Our machine teacher tries to construct macaronic sentences that the human student ought to understand, given all the learning that our generic model predicts would have happened from the previous

Sentence	The	Nile	is	a	river	in	Africa
Gloss	**Der**	**Nil**	**ist**	**ein**	**Fluss**	**in**	**Afrika**
Macaronic Configurations	**Der**	Nile	**ist**	a	river	in	Africa
	The	Nile	is	a	**Fluss**	in	Africa
	Der	**Nil**	**ist**	**ein**	river	in	Africa

Table 1: An example English (L1) sentence with German (L2) glosses. Using the glosses, many possible macaronic configurations are possible. Note that the gloss sequence is not a fluent L2 sentence.

macaronic sentences shown to the student. Our teacher does not yet attempt to monitor the human student's *actual* learning. Still, we show that it is useful to a beginner student and far less frustrating than a random (or heuristic based) alternative.

A "pilot" version of the present paper appeared at a recent workshop (Renduchintala et al., 2019): it experimented with three variants of the generic student model, using an artificial L2 language. In this paper, we extend the best of those models to consider an L2 word's spelling (along with its context) when guessing its embeddings. We therefore conduct our experiments on real L2 languages (Spanish and German).

2 Related Work

Our motivation is similar to that of commercially available prior systems such as Swych (2015) and OneThirdStories (2018) that also incorporate incidental learning within foreign language instruction. Other prior work (Labutov and Lipson, 2014; Renduchintala et al., 2016b) relied on building a model of the student's incidental learning capabilities, using supervised data that was painfully collected by asking students to react to the actions of an initially untrained machine teacher. Our method, by contrast, constructs a generic student model from unannotated L1 text alone. This makes it possible for us to quickly create macaronic documents in any domain covered by that text corpus.

3 Method

Our machine teacher can be viewed as a search algorithm that tries to find the (approximately) best macaronic configuration for the next sentence in a given L1 document. We assume the availability of a "gold" L2 gloss for each L1 word: in our experiments, we obtained these from bilingual speakers using Mechanical Turk. Table 1 shows an example English sentence with German glosses and three possible macaronic configurations (there are exponentially many configurations). The machine teacher must assess, for example, how accurately a student would understand the meanings of **Der**, **ist**, **ein**, and **Fluss** when presented with the following candidate macaronic configuration: **Der** Nile **ist ein Fluss** in Africa.[1] Understanding may arise from inference on this sentence as well as whatever the student has learned about these words from previous sentences. The teacher makes this assessment by presenting this sentence to a generic student model (§§3.1–3.3). It uses a L2 embedding scoring scheme (§3.4) to guide a greedy search for the best macaronic configuration (§3.5).

3.1 Generic Student Model

Our model of a "generic student" (GSM) is equipped with a cloze language model that uses a bidirectional LSTM to predict L1 words in L1 context (Mousa and Schuller, 2017; Hochreiter and Schmidhuber, 1997). Given a sentence $\mathbf{x} = [x_1, \ldots, x_t, \ldots, x_T]$, the cloze model defines $p(x_t \mid \mathbf{h^f}_t, \mathbf{h^b}_t) \forall t \in \{1, \ldots, T\}$, where:

$$\mathbf{h^f}_t = \text{LSTM}^{\mathbf{f}}([\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}]; \boldsymbol{\theta}^{\mathbf{f}}) \in \mathbb{R}^D \quad (1)$$

$$\mathbf{h^b}_t = \text{LSTM}^{\mathbf{b}}([\mathbf{x}_T, \ldots, \mathbf{x}_{t+1}]; \boldsymbol{\theta}^{\mathbf{b}}) \in \mathbb{R}^D \quad (2)$$

are hidden states of forward and backward LSTM encoders parameterized by $\boldsymbol{\theta}^{\mathbf{f}}$ and $\boldsymbol{\theta}^{\mathbf{b}}$ respectively. The model assumes a fixed L1 vocabulary of size V, and the vectors $\mathbf{x}_t$ above are embeddings of these word types, which correspond to the rows of an embedding matrix $\mathbf{E} \in \mathbb{R}^{V \times D}$. The cloze distribution at each position t in the sentence is obtained using

$$p(\cdot \mid \mathbf{h^f}, \mathbf{h^b}) = \text{softmax}(\mathbf{E}\, h([\mathbf{h^f}; \mathbf{h^b}]; \boldsymbol{\theta}^{\mathbf{h}})) \quad (3)$$

where $h(\cdot; \boldsymbol{\theta}^{\mathbf{h}})$ is a projection function that reduces the dimension of the concatenated hidden states from $2D$ to D. We "tie" the input embeddings and output embeddings as in Press and Wolf (2017).

We train the parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}^{\mathbf{f}}; \boldsymbol{\theta}^{\mathbf{b}}; \boldsymbol{\theta}^{\mathbf{h}}; \mathbf{E}]$ using Adam (Kingma and Ba, 2014) to maximize $\sum_{\mathbf{x}} \mathcal{L}(\mathbf{x})$, where the summation is over sentences $\mathbf{x}$ in a large L1 training corpus, and

$$\mathcal{L}(\mathbf{x}) = \sum_t \log p(x_t \mid \mathbf{h^f}_t, \mathbf{h^b}_t) \quad (4)$$

We set the dimensionality of word embeddings and LSTM hidden units to 300. We use the WikiText-103 corpus (Merity et al., 2016) as the L1 training corpus. We apply dropout ($p = 0.2$) between the word embeddings and LSTM layers, and between the LSTM and projection layers (Srivastava et al., 2014). We assume that the resulting model represents the entirety of the student's L1 knowledge.

[1] By "meaning" we mean the L1 token that was originally in the sentence before it was replaced by an L2 gloss.

3.2 Incremental L2 Vocabulary Learning

Our generic student model (GSM) supposes that to learn new vocabulary, the student continues to try to improve $\mathcal{L}(\mathbf{x})$ on additional sentences. Thus, if x_i is a new word, the student will try to adjust its embedding to increase all summands of (4), both the $t=i$ summand (making x_i more predictable) and the $t\neq i$ summands (making x_i more predictive of x_t).

For our purposes, we do not update $\boldsymbol{\theta}$ (which includes L1 embeddings), as we assume that the student's L1 knowledge has already converged. For the L2 words, we use another word-embedding matrix, $\mathbf{F}$, initialized to $\mathbf{0}^{V'\times D}$, and modify (3) to consider both the L1 and L2 embeddings:

$$p(\cdot \mid [\mathbf{h^f};\mathbf{h^b}]) = \text{softmax}([\mathbf{E};\mathbf{F}]\cdot h([\mathbf{h^f};\mathbf{h^b}];\boldsymbol{\theta^h}))$$

We also restrict the softmax function here to produce a distribution not over the full bilingual vocabulary of size $|V| + |V'|$, but only over the bilingual vocabulary consisting of the L1 types V together with only the L2 types $v' \subset V'$ that actually appear in the macaronic sentence. (In the above example macaronic sentence, $|v'|=4$.) This restriction prevents the model from updating the embeddings of L2 types that are not visible in the macaronic sentence, on the grounds that students are only going to update the meanings of what they are currently reading (and are not even aware of the entire L2 vocabulary).

We assume that when a student reads a macaronic sentence $\mathbf{x}$, they update (only) $\mathbf{F}$ so as to maximize

$$\mathcal{L}(\mathbf{x}) - \lambda \|\mathbf{F} - \mathbf{F^{prev}}\|^2 \tag{5}$$

As mentioned above, increasing the $\mathcal{L}$ term adjusts $\mathbf{F}$ so that the surrounding context can easily predict each L2 word, and each L2 word can, in turn, easily predict the surrounding context (both L1 and L2). However, the penalty term with coefficient $\lambda > 0$ prevents $\mathbf{F}$ from straying too far from $\mathbf{F^{prev}}$, which represents the value of $\mathbf{F}$ before this sentence was read. This limits the degree to which a *single* sentence influences the update to $\mathbf{F}$. As a result, an L2 word's embedding reflects *all* the past sentences that contained that word, not just the most recent such sentence, although with a bias toward the most recent ones, which is realistic. Given a new sentence $\mathbf{x}$, we (approximately) maximize the objective above using 10 steps of gradient ascent (with step-size of 0.1), which gave good convergence in practice. In principle, λ should be set based on human-subject experiments. In practice, in this paper, we simply took $\lambda = 1$.

3.3 Spelling-Aware Extension

So far, our generic student model ignores the fact that a novel word like **Afrika** is guessable simply by its spelling similarity to Africa. Thus, we augment the generic student model to use character n-grams. In addition to an embedding per word type, we learn embeddings for character n-gram types that appear in our L1 corpus. The row in $\mathbf{E}$ for a word w is now parameterized as:

$$\tilde{\mathbf{E}}\cdot\tilde{\mathbf{w}} + \sum_n \tilde{\mathbf{E}}^{\mathbf{n}}\cdot\tilde{\mathbf{w}}^n \frac{1}{\mathbf{1}\cdot\tilde{\mathbf{w}}^n} \tag{6}$$

where $\tilde{\mathbf{E}}$ is the full-word embedding matrix and $\tilde{\mathbf{w}}$ is a one-hot vector associated with the word type w, $\tilde{\mathbf{E}}^{\mathbf{n}}$ is a character n-gram embedding matrix and $\tilde{\mathbf{w}}^n$ is a *multi*-hot vector associated with all the character n-grams for the word type w. For each n, the summand gives the average embedding of all n-grams in w (where $\mathbf{1}\cdot\tilde{\mathbf{w}}^n$ counts these n-grams). We set n to range from 3 to 4 (see Appendix B). This formulation is similar to previous sub-word based embedding models (Wieting et al., 2016; Bojanowski et al., 2017).

Similarly, the embedding of an L2 word w is parameterized as

$$\tilde{\mathbf{F}}\cdot\tilde{\mathbf{w}} + \sum_n \tilde{\mathbf{F}}^{\mathbf{n}}\cdot\tilde{\mathbf{w}}^n \frac{1}{\mathbf{1}\cdot\tilde{\mathbf{w}}^n} \tag{7}$$

Crucially, we initialize $\tilde{\mathbf{F}}^{\mathbf{n}}$ to $\mu\tilde{\mathbf{E}}^{\mathbf{n}}$ (where $\mu > 0$) so that L2 words can inherit part of their initial embedding from similarly spelled L1 words: $\tilde{\mathbf{F}}^{\mathbf{4}}{}_{\textbf{Afri}} := \mu\tilde{\mathbf{E}}^{\mathbf{4}}{}_{\texttt{Afri}}$.[2] But we allow $\tilde{\mathbf{F}}^{\mathbf{n}}$ to diverge over time in case an n-gram functions differently in the two languages. In the same way, we initialize each row of $\tilde{\mathbf{F}}$ to the corresponding row of $\mu\cdot\tilde{\mathbf{E}}$, if any, and otherwise to $\mathbf{0}$. Our experiments set $\mu = 0.2$ (see Appendix B). We refer to this spelling-aware extension to GSM as sGSM.

3.4 Scoring L2 embeddings

Did the simulated student learn correctly and usefully? Let $\mathcal{P}$ be the "reference set" of all (L1 word, L2 gloss) pairs from *all tokens in the entire document*. We assess the machine teacher's success by how many of these pairs the simulated student has learned. (The student may even succeed on some pairs that it has never been shown, thanks to n-gram clues.) Specifically, we measure the "goodness" of

[2] We set $\mu = 0.2$ based on findings from our hyperparameter search (see Appendix B).

the updated L2 word embedding matrix $\mathbf{F}$. For each pair $p = (e, f) \in \mathcal{P}$, sort all the words in the entire L1 vocabulary according to their cosine similarity to the L2 word f, and let r_p denote the rank of e. For example, if the student had managed to learn a matrix $\mathbf{F}$ whose embedding of f exactly equalled $\mathbf{E}$'s embedding of e, then r_p would be 1. We then compute a mean reciprocal rank (MRR) score of $\mathbf{F}$:

$$\text{MRR}(\mathbf{F}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left(\frac{1}{r_p} \textbf{ if } r_p \leq r_{\text{max}} \textbf{ else } 0 \right) \quad (8)$$

We set $r_{\text{max}} = 4$ based on our pilot study. This threshold has the effect of only giving credit to an embedding of f such that the correct e is in the simulated student's top 4 guesses. As a result, §3.5's machine teacher focuses on introducing L2 tokens whose meaning can be deduced *rather accurately* from their single context (together with any prior exposure to that L2 type). This makes the macaronic text comprehensible for a human student, rather than frustrating to read. In our pilot study we found that r_{max} substantially improved human learning.

3.5 Macaronic Configuration Search

Our current machine teacher produces the macaronic document greedily, one sentence at a time. Actual documents produced are shown in Appendix D.

Let $\mathbf{F}^{\text{prev}}$ be the student model's embedding matrix after the reading the first $n-1$ macaronic sentences. We evaluate a candidate next sentence $\mathbf{x}$ by the score $\text{MRR}(\mathbf{F})$ where $\mathbf{F}$ maximizes (5) and is thus the embedding matrix that the student would arrive at after reading $\mathbf{x}$ as the n^{th} macaronic sentence.

We use best-first search to seek a high-scoring $\mathbf{x}$. A search state is a pair $(i, \mathbf{x})$ where $\mathbf{x}$ is a macaronic configuration (Table 1) whose first i tokens may be either L1 or L2, but whose remaining tokens are still L1. The state's score is obtained by evaluating $\mathbf{x}$ as described above. In the initial state, $i = 0$ and $\mathbf{x}$ is the n^{th} sentence of the original L1 document. The state $(i, \mathbf{x})$ is a final state if $i = |\mathbf{x}|$. Otherwise its two successors are $(i+1, \mathbf{x})$ and $(i+1, \mathbf{x}')$, where $\mathbf{x}'$ is identical to $\mathbf{x}$ except that the $(i+1)^{\text{th}}$ token has been replaced by its L2 gloss. The search algorithm maintains a priority queue of states sorted by score. Initially, this contains only the initial state. A step of the algorithm consists of popping the highest-scoring state and, if it is not final, replacing it by its two successors. The queue is then pruned back to the top 8 states. When the queue becomes empty, the algorithm returns the configuration $\mathbf{x}$ from the highest-scoring final state that was ever popped.

L2	Model	Closed-class	Open-class
Es	random	0.74±0.0126(54)	0.61±0.0134(17)
	GSM	0.72±0.0061(54)	0.70±0.0084(17)
	sGSM	**0.82±0.0038**(41)	**0.80±0.0044**(21)
De	random	0.59±0.0054(34)	0.38±0.0065(13)
	GSM	0.80±0.0033(34)	0.78±0.0056(13)
	sGSM	**0.82±0.0063**(33)	**0.79±0.0062**(14)

Table 2: Average token guess quality ($\tau = 0.6$) in the comprehension experiments. The ± denotes a 95% confidence interval computed via bootstrap resampling of the set of human subjects. The % of L1 tokens replaced with L2 glosses is in parentheses. Appendix C evaluates with other choices of τ.

4 Evaluation

Does our machine teacher generate useful macaronic text? To answer this, we measure whether *human* students (i) comprehend the L2 words in context, and (ii) retain knowledge of those L2 words when they are later seen without context.

We assess (i) by displaying each successive sentence of a macaronic document to a human student and asking them to guess the L1 meaning for each L2 token f in the sentence. For a given machine teacher, all human subjects saw the same macaronic document, and each subject's comprehension score is the average quality of their guesses on all the L2 tokens presented by that teacher. A guess's quality $q \in [0,1]$ is a thresholded cosine similarity between the embeddings[3] of the guessed word $\hat{e}$ and the original L1 word e: $q = \text{cs}(e, \hat{e})$ **if** $\text{cs}(e, \hat{e}) \geq \tau$ **else** 0. Thus, $\hat{e} = e$ obtains $q = 1$ (full credit), while $q = 0$ if the guess is "too far" from the truth (as determined by τ).

To assess (ii), we administer an L2 vocabulary quiz after having human subjects *simply* read a macaronic passage (without any guessing as they are reading). They are then asked to guess the L1 translation of each L2 word type that appeared at least once in the passage. We used the same guess quality metric as in (i).[4] This tests if human subjects naturally learn the meanings of L2 words, in informative contexts, well enough to later translate them out of context. The test requires only short-term retention, since we give the vocabulary quiz immediately after a passage is read.

We compared results on macaronic documents constructed with the generic student model (GSM), its spelling-aware variant (sGSM), and a random

[3] Here we used pretrained word embeddings from Mikolov et al. (2018), in order to measure actual semantic similarity.

[4] If multiple L1 types e were glossed in the document with this L2 type, we generously use the e that maximizes $\text{cs}(e, \hat{e})$.

baseline. In the baseline, tokens to replace are randomly chosen while ensuring that each sentence replaces the same number of tokens as in the GSM document. This ignores context, spelling, and prior exposures as reasons to replace a token.

Our evaluation was aimed at native English (L1) speakers learning Spanish or German (L2). We recruited L2 "students" on Amazon Mechanical Turk (MTurk). They were absolute beginners, selected using a placement test and self-reported L2 ability.

4.1 Comprehension Experiments

We used the first chapter of Jane Austen's "Sense and Sensibility" for Spanish, and the first 60 sentences of Franz Kafka's "Metamorphosis" for German. Bilingual speakers provided the L2 glosses (see Appendix A).

For English-Spanish, 11, 8, and 7 subjects were assigned macaronic documents generated with sGSM, GSM, and the random baseline, respectively. The corresponding numbers for English-German were 12, 7 and 7. A total of 39 subjects were used in these experiments (some subjects did both languages). They were given 3 hours to complete the entire document (average completion time was ≈ 1.5 hours) and were compensated \$10.

Table 2 reports the mean comprehension score over all subjects, broken down into comprehension of function words (closed-class POS) and content words (open-class POS).[5] For Spanish, the sGSM-based teacher replaces *more* content words (but fewer function words), and furthermore the replaced words in both cases are *better understood* on average, which we hope leads to more engagement and more learning. For German, by contrast, the number of words replaced does not increase under sGSM, and comprehension only improves marginally. Both GSM and sGSM do strongly outperform the random baseline. But the sGSM-based teacher only replaces a few additional cognates (**hundert** but not **Mutter**), apparently because English-German cognates do not exhibit large *exact* character n-gram overlap. We hypothesize that character skip n-grams might be more appropriate for English-German.

4.2 Retention Experiments

For retention experiments we used the first 25 sentences of our English-Spanish dataset. New participants were recruited and compensated \$5. Each participant was assigned a macaronic document generated with the sGSM, GSM or random model (20, 18, and 22 participants respectively). As Table 3 shows, sGSM's advantage over GSM on comprehension holds up on retention. On the vocabulary quiz, students correctly translated > 30 of the 71 word types they had seen (Table 8), and more than half when near-synonyms earned partial credit (Table 3).

L2	Model	Closed-class	Open-class
Es	random	$0.47 \pm 0.0058 (60)$	$0.40 \pm 0.0041 (46)$
	GSM	$0.48 \pm 0.0084 (60)$	$0.42 \pm 0.0105 (15)$
	sGSM	$\mathbf{0.52 \pm 0.0054} (47)$	$\mathbf{0.50 \pm 0.0037} (24)$

Table 3: Average type guess quality ($\tau = 0.6$) in the retention experiment. The % of L2 gloss types that were shown in the macaronic document is in parentheses. Appendix C evaluates with other choices of τ.

5 Future Work

We would like to explore different character-based compositions such as Kim et al. (2016) that can potentially generalize better across languages. We would further like to extend our work beyond simple lexical learning to allow learning phrasal translations, word reordering, and morphology.

Beyond that, we envision machine teaching interfaces in which the student reader *interacts* with the macaronic text—advancing through the document, clicking on words for hints, and facing occasional quizzes (Renduchintala et al., 2016b)—and with other educational stimuli. As we began to explore in Renduchintala et al. (2016a, 2017), interactions provide feedback that the machine teacher could use to adjust its model of the student's lexicons (here $\mathbf{E}, \mathbf{F}$), inference (here $\boldsymbol{\theta}^{\mathbf{f}}, \boldsymbol{\theta}^{\mathbf{b}}, \boldsymbol{\theta}^{\mathbf{h}}, \mu$), and learning (here λ). In this context, we are interested in using models that are *student-specific* (to reflect individual learning styles), *stochastic* (since the student's observed behavior may be inconsistent owing to distraction or fatigue), and able to model *forgetting* as well as learning (e.g., Settles and Meeder, 2016).

6 Conclusions

We presented a method to generate macaronic (mixed-language) documents to aid foreign language learners with vocabulary acquisition. Our key idea is to derive a model of student learning from only a cloze language model, which uses both context and spelling features. We find that our model-based teacher generates comprehensible macaronic text that promotes vocabulary learning.

[5] https://universaldependencies.org/u/pos/

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Rebecca Knowles, Adithya Renduchintala, Philipp Koehn, and Jason Eisner. 2016. Analyzing learner understanding of novel L2 vocabulary. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 126–135, Berlin, Germany.

Stephen Krashen. 1989. We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, 73(4):440–464.

Igor Labutov and Hod Lipson. 2014. Generating code-switched text for lexical learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–571, Baltimore, Maryland. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Amr Mousa and Björn Schuller. 2017. Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1023–1032, Valencia, Spain.

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

OneThirdStories. 2018. OneThirdStories. `https://onethirdstories.com/`. Accessed: 2019-02-20.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain.

Adithya Renduchintala, Rebecca Knowles, Philipp Koehn, and Jason Eisner. 2016a. Creating interactive macaronic interfaces for language learning. In *Proceedings of ACL-2016 System Demonstrations*, pages 133–138, Berlin, Germany.

Adithya Renduchintala, Rebecca Knowles, Philipp Koehn, and Jason Eisner. 2016b. User modeling in language learning with macaronic texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1859–1869, Berlin, Germany.

Adithya Renduchintala, Philipp Koehn, and Jason Eisner. 2017. Knowledge tracing in sequential learning of inflected vocabulary. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 238–247, Vancouver, Canada.

Adithya Renduchintala, Philipp Koehn, and Jason Eisner. 2019. Simple construction of mixed-language texts for vocabulary learning. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, Florence.

Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1848–1858, Berlin, Germany.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Swych. 2015. Swych. `http://swych.it/`. Accessed: 2019-02-20.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515, Austin, Texas.

A Obtaining L2 Glosses

Using Amazon Mechanical Turk (MTurk), we asked bilingual speakers (English-Spanish or English-German) to produce a gloss for each token in English documents. Each Turker was given a sentence in English (L1) and produced a L2 gloss for each L1 token. Each sentence was annotated by three Turkers. For each L1 token, we then selected the L2 gloss that was agreed on by a majority of the Turkers. If there was no such gloss, we did not select any L2 gloss, meaning that the machine teacher was required to leave that token as L1.

B Hyperparameter Search

We tuned the model hyperparameters by hand on separate English-Spanish data, namely the second chapter of "Sense and Sensibility," equipped with glosses. Hyperparameter tuning results are reported in this appendix. All other English-Spanish results in the paper are on the first chapter of "Sense and Sensibility," which was held out for testing. We might have improved the results on English-German by tuning separate hyperparameters for that setting.

The tables below show the effect of different hyperparameter choices on the quality MRR($\mathbf{F}$) of the embeddings learned by the simulated student. Recall from §3.4 that the MRR score evaluates $\mathbf{F}$ using all glosses, not just those used in a particular macaronic document. Thus, it is comparable across the different macaronic documents produced by different machine teachers.

QueueSize (§3.5) affects only how hard the machine teacher searches for macaronic sentences that will help the simulated student. We find that larger QueueSize is in fact valuable.

The other choices (Model, n-grams, μ) affect how the simulated student actually learns. The machine teacher then searches for a document that will help that particular simulated student learn as many of the words in the reference set as possible. Thus, the MRR score is high to the extent that the simulated student "can be successfully taught." By choosing hyperparameters that achieve a high MRR score, we are assuming that human students are adapted (or can adapt online) to be teachable.

The scale factor μ (used only for sGSM) noticeably affects the macaronic document generated by the machine teacher.Setting it high ($\mu = 1.0$) has a adverse effect on the MRR score. Table 4 shows how the MRR score of the simulated student (§3.4) varies according to the student model's μ value. Tables 5 and 6 show the result of the same hyperparameter sweep on the number of L1 word tokens and types replaced with L2 glosses.

Note that μ only affects initialization of the $\mathcal{F}$ parameters. Thus, with $\mu = 0$, the L2 word and subword embeddings are initialized to $\mathbf{0}$, but the simulated sGSM student still has the ability to learn subword embeddings for both L1 and L2. This allows it to beat the simulated GSM student.

We see that for sGSM, $\mu = 0.2$ results in replacing the most words (both types and tokens), and also has very nearly the highest MRR score. Thus, for sGSM, we decided to use $\mu = 0.2$ and allow both 3-gram and 4-gram embeddings.

Model	n-grams	QueueSize	Scale Factor μ					
			1.0	0.4	0.2	0.1	0.05	0.0
sGSM	2,3,4	1	0.108	0.207	0.264	0.263	0.238	0.175
sGSM	3,4	1	0.113	0.199	0.258	0.274	0.277	0.189
sGSM	3,4	4	-	-	0.267	0.286	-	-
sGSM	3,4	8	-	-	0.288	0.292	-	-
GSM	∅	1						0.159
GSM	∅	4						0.171
GSM	∅	8						0.172

Table 4: MRR scores obtained with different hyperparameter settings.

Model	n-grams	QueueSize	Scale Factor μ					
			1.0	0.4	0.2	0.1	0.05	0.0
sGSM	2,3,4	1	149	301	327	275	201	247
sGSM	3,4	1	190	340	439	399	341	341
sGSM	3,4	4	-	-	462	440	-	-
sGSM	3,4	8	-	-	478	450	-	-
GSM	∅	1						549
GSM	∅	4						557
GSM	∅	8						530

Table 5: Number of L1 tokens replaced by L2 glosses under different hyperparameter settings.

Model	n-grams	QueueSize	Scale Factor μ					
			1.0	0.4	0.2	0.1	0.05	0.0
sGSM	2,3,4	1	39	97	121	106	75	88
sGSM	3,4	1	44	97	125	124	112	99
sGSM	3,4	4	-	-	124	127	-	-
sGSM	3,4	8	-	-	145	129	-	-
GSM	∅	1						106
GSM	∅	4						111
GSM	∅	8						114

Table 6: Number of distinct L2 word types present in the macaronic document under different hyperparameter settings.

L2	τ	Model	Closed-class	Open-class
Es	0.0	rand	$0.81 \pm 0.0084(54)$	$0.72 \pm 0.0088(17)$
		GSM	$0.80 \pm 0.0045(54)$	$0.79 \pm 0.0057(17)$
		sGSM	$0.86 \pm 0.0027(41)$	$0.84 \pm 0.0032(21)$
	0.2	rand	$0.81 \pm 0.0085(54)$	$0.72 \pm 0.0089(17)$
		GSM	$0.80 \pm 0.0045(54)$	$0.79 \pm 0.0057(17)$
		sGSM	$0.86 \pm 0.0027(41)$	$0.84 \pm 0.0033(21)$
	0.4	rand	$0.79 \pm 0.0101(54)$	$0.66 \pm 0.0117(17)$
		GSM	$0.76 \pm 0.0057(54)$	$0.75 \pm 0.0071(17)$
		sGSM	$0.84 \pm 0.0033(41)$	$0.82 \pm 0.0039(21)$
	0.6	random	$0.74 \pm 0.0126(54)$	$0.61 \pm 0.0134(17)$
		GSM	$0.72 \pm 0.0061(54)$	$0.70 \pm 0.0084(17)$
		sGSM	$0.82 \pm 0.0038(41)$	$0.80 \pm 0.0044(21)$
	0.8	rand	$0.62 \pm 0.0143(54)$	$0.46 \pm 0.0124(17)$
		GSM	$0.59 \pm 0.0081(54)$	$0.58 \pm 0.0106(17)$
		sGSM	$0.71 \pm 0.0052(41)$	$0.67 \pm 0.0062(21)$
	1.0	rand	$0.62 \pm 0.0143(54)$	$0.45 \pm 0.0124(17)$
		GSM	$0.59 \pm 0.0081(54)$	$0.55 \pm 0.0097(17)$
		sGSM	$0.70 \pm 0.0052(41)$	$0.64 \pm 0.0063(21)$
De	0.0	random	$0.70 \pm 0.0039(34)$	$0.56 \pm 0.0046(13)$
		GSM	$0.85 \pm 0.0023(34)$	$0.84 \pm 0.0039(13)$
		sGSM	$0.87 \pm 0.0045(33)$	$0.84 \pm 0.0044(14)$
	0.2	random	$0.69 \pm 0.0042(34)$	$0.56 \pm 0.0047(13)$
		GSM	$0.85 \pm 0.0024(34)$	$0.84 \pm 0.0039(13)$
		sGSM	$0.87 \pm 0.0046(33)$	$0.84 \pm 0.0044(14)$
	0.4	random	$0.64 \pm 0.0052(34)$	$0.45 \pm 0.0064(13)$
		GSM	$0.83 \pm 0.0029(34)$	$0.81 \pm 0.0045(13)$
		sGSM	$0.84 \pm 0.0055(33)$	$0.81 \pm 0.0054(14)$
	0.6	random	$0.59 \pm 0.0054(34)$	$0.38 \pm 0.0065(13)$
		GSM	$0.80 \pm 0.0033(34)$	$0.78 \pm 0.0056(13)$
		sGSM	$0.82 \pm 0.0063(33)$	$0.79 \pm 0.0062(14)$
	0.8	random	$0.45 \pm 0.0058(34)$	$0.25 \pm 0.0061(13)$
		GSM	$0.72 \pm 0.0037(34)$	$0.66 \pm 0.0081(13)$
		sGSM	$0.75 \pm 0.0079(33)$	$0.65 \pm 0.0077(14)$
	1.0	random	$0.45 \pm 0.0058(34)$	$0.24 \pm 0.0061(13)$
		GSM	$0.71 \pm 0.0040(34)$	$0.63 \pm 0.0082(13)$
		sGSM	$0.75 \pm 0.0079(33)$	$0.63 \pm 0.0081(14)$

Table 7: An expanded version of Table 2 (human comprehension experiments), reporting results with various values of τ.

C Results Varying τ

A more comprehensive variant of Table 2 is given in Table 7. This table reports the same human-subjects experiments as before; it only varies the measure used to assess the quality of the humans' guesses, by varying the threshold τ. Note that $\tau = 1$ assesses exact-match accuracy, $\tau = 0.6$ as in Table 2 corresponds roughly to synonymy (at least for content words), and $\tau = 0$ assesses average *unthresholded* cosine similarity. We find that sGSM consistently outperforms both GSM and the random baseline over the entire range of τ. As we get closer to exact match, the random baseline suffers the largest drop in performance.

L2	τ	Model	Closed-class	Open-class
Es	0.0	random	$0.67 \pm 0.0037(60)$	$0.60 \pm 0.0027(46)$
		GSM	$0.67 \pm 0.0060(60)$	$0.62 \pm 0.0076(15)$
		sGSM	$0.71 \pm 0.0035(47)$	$0.68 \pm 0.0028(24)$
	0.2	random	$0.67 \pm 0.0037(60)$	$0.60 \pm 0.0027(46)$
		GSM	$0.67 \pm 0.0061(60)$	$0.61 \pm 0.0080(15)$
		sGSM	$0.71 \pm 0.0036(47)$	$0.67 \pm 0.0029(24)$
	0.4	random	$0.60 \pm 0.0051(60)$	$0.50 \pm 0.0037(46)$
		GSM	$0.60 \pm 0.0086(60)$	$0.51 \pm 0.0106(15)$
		sGSM	$0.66 \pm 0.0044(47)$	$0.61 \pm 0.0037(24)$
	0.6	random	$0.47 \pm 0.0058(60)$	$0.40 \pm 0.0041(46)$
		GSM	$0.48 \pm 0.0084(60)$	$0.42 \pm 0.0105(15)$
		sGSM	$0.52 \pm 0.0054(47)$	$0.50 \pm 0.0037(24)$
	0.8	random	$0.40 \pm 0.0053(60)$	$0.30 \pm 0.0032(46)$
		GSM	$0.41 \pm 0.0078(60)$	$0.37 \pm 0.0097(15)$
		sGSM	$0.46 \pm 0.0055(47)$	$0.41 \pm 0.0041(24)$
	1.0	random	$0.40 \pm 0.0053(60)$	$0.29 \pm 0.0031(46)$
		GSM	$0.40 \pm 0.0077(60)$	$0.36 \pm 0.0092(15)$
		sGSM	$0.45 \pm 0.0053(47)$	$0.39 \pm 0.0042(24)$

Table 8: An expanded version of Table 3 (human retention experiments), reporting results with various values of τ.

Similarly, Table 8 shows a expanded version of the retention results in Table 3. The gap between the models is smaller on retention than it was on comprehension. However, again sGSM $>$ GSM $>$ random across the range of τ. We find that for function words, the random baseline performs as well as GSM as τ is increased. For content words, however, the random baseline falls faster than GSM.

We warn that the numbers are not genuinely comparable across the 3 models, because each model resulted in a different document and thus a different vocabulary quiz. Our human subjects were asked to translate just the L2 words in the document they read. In particular, sGSM taught *fewer* total types (71) than GSM (75) or the random baseline (106). All that Table 8 shows is that it taught its chosen types better (on average) than the other methods taught their chosen types.

D Macaronic Examples

Below, we display the actual macaronic documents generated by our methods. Table 9 is the opening of Jane Austin's "Sense and Sensibility," converted into a macaronic English-Spanish document using our sGSM-based teacher. Table 10 shows the same passage converted into macaronic form with the GSM-based teacher. Similarly, Tables 11 and 12 show macaronic English-German versions of "The Metamorphosis."

Sense y Sensibility

La family de Dashwood llevaba long been settled en Sussex. Their estate era large, and their residencia was en Norland Park, in el centre de their property, where, for muchas generations, they habían lived en so respectable a manner as to engage el general good opinion of los surrounding acquaintance. El late owner de this propiedad was un single man, que lived to una very advanced age, y que durante many years of his life, had a constante companion and housekeeper in su sister. But ella death, que happened ten años antes his own, produced a great alteration in su home; for to supply her loss, he invited and received into his house la family of su sobrino señor Henry Dashwood, the legal inheritor of the Norland estate, and the person to whom he intended to bequeath it. En la society de su nephew y niece, y their children, el old Gentleman's days fueron comfortably spent. Su attachment a them all increased. The constant attention de Mr. y Mrs. Henry Dashwood to sus wishes, que proceeded no merely from interest, but from goodness of heart, dio him every degree de solid comfort which su age podía receive; and la cheerfulness of the children added a relish to his existencia.

By un former marriage, Mr. Henry Dashwood tenía one son : by su present lady, three hijas. El son, un steady respectable young man, was amply provided for por the fortuna de his mother, which había been large, y half of which devolved on him on his coming of edad. Por su own matrimonio, likewise, which happened soon después, he added a his wealth. To him therefore la succession a la Norland estate era no so really importante as to his sisters; para their fortuna, independent de what pudiera arise a ellas from su father's inheriting that propiedad, could ser but small. Su mother had nothing, y their father only seven mil pounds en his own disposición; for la remaining moiety of his first esposa's fortune was also secured to her child, and él tenía sólo a life-interés in it.

el anciano gentleman died : his will was read, and like almost todo other will, dio as tanto disappointment as pleasure. He fue neither so unjust, ni so ungrateful, as para leave his estate de his nephew; --but he left it a him en such terms as destroyed half the valor de el bequest. Mr. Dashwood había wished for it more por el sake of his esposa and hijas than for himself or su son; --but a his son, y su son's son, un child de four años old, it estaba secured, in tal a way, as a leave a himself no power de providing por those que were most dear para him, and who most necesitaban a provisión by any charge on la estate, or por any sale de its valuable woods. El whole fue tied arriba para the beneficio de this child, quien, in occasional visits with his padre and mother at Norland, had tan far gained on el affections de his uncle, by such attractions as are by no means unusual in children of two o three years old; una imperfect articulación, an earnest desire of having his own way, many cunning tricks, and a great deal of noise, as to outweigh all the value de all the attention which, for years, él había received from his niece and sus daughters. He meant no a ser unkind, however, y, como a mark de his affection for las three girls, he left ellas un mil libras a-piece.

Table 9: First few paragraphs of "Sense and Senibility" with the sGSM model using $\mu = 0.2$, 3- and 4-grams, priority queue size of 8, and $r_{\max=4}$.

Sense y Sensibility

La family de Dashwood llevaba long been settled en Sussex. Su estate era large, and su residence estaba en Norland Park, in el centre de their property, where, por many generations, they had lived in so respectable una manner as a engage el general good opinion de los surrounding acquaintance. El late owner de esta estate was un single man, que lived to una very advanced age, y who durante many years de su existencia, had una constant companion y housekeeper in his sister. But ella death, que happened ten years antes su own, produced a great alteration in su home; for para supply her loss, él invited and received into his house la family de su nephew Mr. Henry Dashwood, the legal inheritor de the Norland estate, and the person to whom se intended to bequeath it. In the society de su nephew and niece, and sus children, el old Gentleman's days fueron comfortably spent. Su attachment a them all increased. La constant attention de Mr. y Mrs. Henry Dashwood to sus wishes, which proceeded not merely from interest, but de goodness de heart, dio him every degree de solid comfort que his age could receive; y la cheerfulness of the children added un relish a su existence.

By un former marriage, Mr. Henry Dashwood tenía one son : by su present lady, three hijas. El son, un steady respectable joven man, was amply provided for por la fortune de su madre, que había been large, y half de cuya devolved on him on su coming de edad. By su own marriage, likewise, que happened soon después, he added a su wealth. Para him therefore la succession a la Norland estate was no so really importante as to his sisters; para their fortune, independent de what pudiera arise a them from su father's inheriting that property, could ser but small. Su madre had nothing, y su padre only siete thousand pounds in su own disposal; for la remaining moiety of his first wife's fortune era also secured a su child, y él had only una life-interest in ello.

el old gentleman died : su will was read, y like almost every otro will, gave as tanto disappointment as pleasure. He fue neither so unjust, nor so ungrateful, as to leave su estate from his nephew; --but he left it to him en such terms como destroyed half the valor of the bequest. Mr. Dashwood había wished for it más for el sake de su wife and daughters than para himself or su hijo; --but a su hijo, y his son's hijo, un child de four años old, it estaba secured, en tal un way, as a leave a himself no power of providing for aquellos who were most dear para him, y who most needed un provision by any charge sobre la estate, or por any sale de its valuable woods. El whole was tied arriba for el benefit de this child, quien, en ocasionales visits with his father and mother at Norland, had tan far gained on the affections of his uncle, by such attractions as are por no means unusual in children of two or three years old; an imperfect articulation, an earnest desire of having his own way, many cunning tricks, and a gran deal of noise, as to outweigh todo the value of all the attention which, for years, he had received from his niece and her daughters. He meant no a ser unkind, however, y, como una mark de su affection por las three girls, he left them un mil pounds a-pieza.

Table 10: First few paragraphs of "Sense and Senibility" with the GSM model using priority queue size of 8 and $r_{\max=4}$.

Metamorphosis

One morning, **als** Gregor Samsa woke from troubled dreams, he **fand** himself transformed in **seinem** bed into **einem** horrible vermin. He lay on **seinem** armour-like back, **und** if **er** lifted **seinen** head a little he **konnte** see his brown belly, slightly domed **und** divided by arches into stiff sections. The bedding **war** hardly able **zu** cover it **und** seemed ready **zu** slide off any moment. His many legs, pitifully thin compared **mit** the size **von dem** rest **von** him, waved about helplessly as **er** looked.

``What's happened to **mir**?'' he thought. His room, **ein** proper human room although **ein** little too small, lay peacefully between its four familiar walls. **Eine** collection of textile samples lay spread out on the table - Samsa was **ein** travelling salesman - and above it there hung a picture **das** he had recently cut out **von einer** illustrated magazine **und** housed **in einem** nice, gilded frame. It showed **eine** lady fitted out **mit** a fur hat **und** fur boa who sat upright, raising **einen** heavy fur muff **der** covered the whole **von** her lower arm towards **dem** viewer.

Gregor then turned **zu** look out the window at the dull weather. Drops **von** rain could **sein** heard hitting the pane, **welche** made him **fühlen** quite sad. ``How about if **ich** sleep **ein** little bit longer and forget all **diesen** nonsense,'' he thought, **aber** that was something **er** was unable **zu** do because he **war** used **zu** sleeping **auf** his right, **und** in **seinem** present state couldn't **bringen** into that position. However hard he threw **sich** onto **seine** right, he always rolled **zurück** to where he was. He must **haben** tried it a **hundert** times, shut **seine** eyes so **dass er** wouldn't **haben zu** look at **die** floundering legs, and only stopped when **er** began **zu fühlen einen** mild, dull pain there **das** he **hatte** never felt before.

``**Ach**, God,'' he thought, ``what a strenuous career it is **das** I've chosen! Travelling day in **und** day out. Doing business like **diese** takes **viel** more effort than doing your own business at home, **und auf** top of that there's the curse **des** travelling, worries **um** making train connections, bad **und** irregular food, contact **mit** different people all the time so that **du** can **nie** get to know anyone or become friendly **mit ihnen**. It can **alles** go **zum** Hell!'' He felt a slight itch up **auf** his belly; pushed himself slowly up **auf** his back towards **dem** headboard so **dass** he **konnte** lift his head better; **fand** where **das** itch was, **und** saw that **es** was covered **mit vielen** of little **weißen** spots which he didn't know what to make of; **und als** he **versuchte** to **fühlen** the place with one **von seinen** legs he drew it quickly back because as soon as he touched it he was overcome **von** a cold shudder.

Table 11: First few paragraphs of "The Metamorphosis" with the sGSM model using $\mu = 0.2$, 3- and 4-grams, priority queue size of 8, and $r_{\text{max}=4}$.

Metamorphosis

One morning, **als** Gregor Samsa woke from troubled dreams, he **fand** himself transformed **in** his bed into **einem** horrible vermin. **Er** lay on **seinem** armour @-@ like back, **und** if **er** lifted his head a little **er** could see **seinen** brown belly, slightly domed **und** divided by arches into stiff **teile**. **das** bedding was hardly **fähig** to cover **es und** seemed ready **zu** slide off any moment. His many legs, pitifully thin compared **mit** the size **von dem** rest **von** him, waved about helplessly **als er** looked.

``What's happened to **mir**?'' **er** thought. His room, **ein** proper human room although **ein** little too **klein**, lay peacefully between **seinen** four familiar walls. **Eine** collection of textile samples lay spread out on the table - Samsa was **ein** travelling salesman - **und** above it there hung a picture that **er** had recently cut **aus** of **einer** illustrated magazine **und** housed in **einem** nice, gilded frame. **Es** showed a lady fitted out with a fur hat and fur boa who **saß** upright, raising a heavy fur muff **der** covered the whole of her lower arm towards **dem** viewer.

Gregor then turned **zu** look out the window at the dull weather. Drops **von** rain could **sein** heard hitting the pane, which **machte** him feel **ganz** sad. ``How about if **ich** sleep **ein** little bit longer and forget all **diesen** nonsense,'' he thought, but that **war** something he was unable to **tun** because **er** was used to sleeping **auf** his right, and in his present state couldn't get into that position. However hard he **warf** himself onto **seine** right, he always rolled **zurück** to **wo** he was. **Er** must **haben** tried it **ein** hundred times, shut **seine** eyes so **dass** he wouldn't **haben** to **sehen** at **die** floundering legs, **und** only stopped when he **begann** to feel **einen** mild, dull pain there that he **hatte nie** felt before.

``**Ach**, God,'' he thought, ``what a strenuous career it **ist** that I've chosen! Travelling day in **und** day **aus**. Doing business like **diese** takes much **mehr** effort than doing your own business at home, **und** on **oben** of that there's **der** curse of travelling, worries **um** making train connections, bad and irregular food, contact with different people all the time so that you **kannst nie** get to know anyone or become friendly with **ihnen**. It **kann** all go to **Teufel**!'' He felt **ein** slight itch up **auf seinem** belly; pushed himself slowly up **auf** his back towards **dem** headboard **so dass** he could lift his head better; **fand** where **das** itch was, and saw that it was **besetzt** with lots of little **weißen** spots which he didn't know what to make of; and **als** he tried to feel the place with one of his legs he drew it quickly back because as soon as he touched it he was overcome by a cold shudder.

Table 12: First few paragraphs of "The Metamorphosis" with the GSM model using priority queue size of 8 and $r_{\text{max}=4}$.