A Hyperparameters and Experimental Details

Here, we list all the hyperparameters and other experimental details necessary for the reproduction of the numbers presented in Tab. 3. The final experiments were produced with the follow setting. We performed a modest grid search over various configurations in the search of the best option on development for each component.

LSTM Morphological Tag Language Model. The morphological tag language model is a 2-layer vanilla LSTM trained with hidden size of 200. It is trained to for 40 epochs using SGD with a cross entropy loss objective, and an initial learning rate of 20 where the learning rate is quartered during any epoch where the loss on the validation set reaches a new minimum. We regularize using dropout of 0.2 and clip gradients to 0.25. The morphological tags are embedded (both for input and output) with a multi-hot encoding into $\mathbb{R}^{200}$, where any given tag has an embedding that is the sum of the embedding for its constituent POS tag and each of its constituent slots.

Lemmata Generator. The lemma generator is a single-layer vanilla LSTM, trained for 10000 epochs using SGD with a learning rate of 4, using a batch size of 20000. The LSTM has 50 hidden units, embeds the POS tags into $\mathbb{R}^5$ and each token (i.e., character) into $\mathbb{R}^5$. We regularize using weight decay (1e-6), no dropout, and clip gradients to 1. When sampling lemmata from the model, we cool the distribution using a temperature of 0.75 to generate more "conservative" values. The hyperparameters were manually tuned on Latin data to produce sensible output and fit development data and then reused for all languages of this paper.

Morphological Inflector. The reinflection model is a single-layer GRU-cell seq2seq model with a bidirectional encoder and multiplicative attention in the style of Luong et al. (2015), which we train for 250 iterations of AdaDelta (Zeiler, 2012). Our search over the remaining hyperparameters was as follows (optimal values in bold): input embedding size of [50, 100, **200**, 300], hidden size of [50, **100**, 150, 200], and a dropout rate of [0.0, 0.1, 0.2, 0.3, 0.4, **0.5**].

Lemmatizer and Morphological Tagger. The joint lemmatizer and tagger is LEMMING as described in §5.5. It is trained with default parameters, the pretrained word vectors from Bojanowski et al. (2016) as type embeddings, and beam size 3.

Wake-Sleep We run two iterations ($I = 2$) of wake-sleep. Note that each of the subparts of wake-sleep: estimating $p_{\boldsymbol{\theta}}$ and estimating $q_{\boldsymbol{\phi}}$ are trained to convergence and use the hyperparameters described in the previous paragraphs. We set γ_{wake} and γ_{sleep} to 0.25, so we observe roughly $1/4$ as many dreamt samples as true samples. The samples from the generative model often act as a regularizer, helping the variational approximation (as measured on morphological tagging and lemmatization accuracy) on the UD development set, but sometimes the noise lowers performance a mite. Due to a lack of space in the initial paper, we did not deeply examine the performance of the tagger-lemmatizer outside the context of improving inflection prediction accuracy. Future work will investigate question of how much tagging and lemmatization can be improved through the incorporation of samples from our generative model. In short, our efforts will evaluate the inference network in its own right, rather than just as a variational approximation to the posterior.

B Fake Data from the Sleep Phase

An example sentence $\tilde{\boldsymbol{f}}$ sampled via $\langle \tilde{\boldsymbol{f}}, \tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{m}} \rangle \sim p_{\boldsymbol{\theta}}(\cdot, \cdot, \cdot)$ in Portuguese:

```
dentremeticamente » isso Procusas
Da Fase » pos a acordítica
Máisringeringe Ditudis A ana ,
Urevirao Da De O linsith.muital ,
E que chegou interalionalmente Da
anundica De mêpinsuriormentais .
```

and in Latin:

```
inpremcret ita sacrum super annum
pronditi avocere quo det tuam
nunsidebus quod puela ?
```