Appendices

A Morphological Lexicons and Training Corpora

Here we provide additional information about the training corpora for the skip-gram model and the morphological lexicons used by our Gaussian graphical model.

	Tokens	Types	TTR
Czech	83,048,682	1,663,361	0.02
English	1,778,938,441	7,403,109	0.004
German	609,406,708	8,582,032	0.0141
Spanish	396,443,513	6,719,014	0.0169
Turkish	58,746,962	2,272,946	0.0387

Table 5: The number of types and tokens and their ratio (TTR) in each Wikipedia corpus.

	Lexicon	Types	OOV
Czech	MorfFlex	25,226,946	97.9
English	CELEX	79,208	2.3
German	CELEX	365,530	43.9
Spanish	Wiktionary	668,681	66.1
Turkish	Wiktionary	118,786	32.2

Table 6: Sizes of the various morphological lexicons and their origin. We note that our method is compatible with any morphologically annotated lexicon as well as finite-state analyzers that have the capacity to analyze an unbounded number of words. We also report OOV, the percentage of the types in the morphological lexicon that are *not* attested in the Wikipedia corpus.

B Additional Results

We include additional results for the experiment in section 8.2. In addition to showing results for all our test languages in Figure 4, we also show a different breakdown of the results in Figure 5.

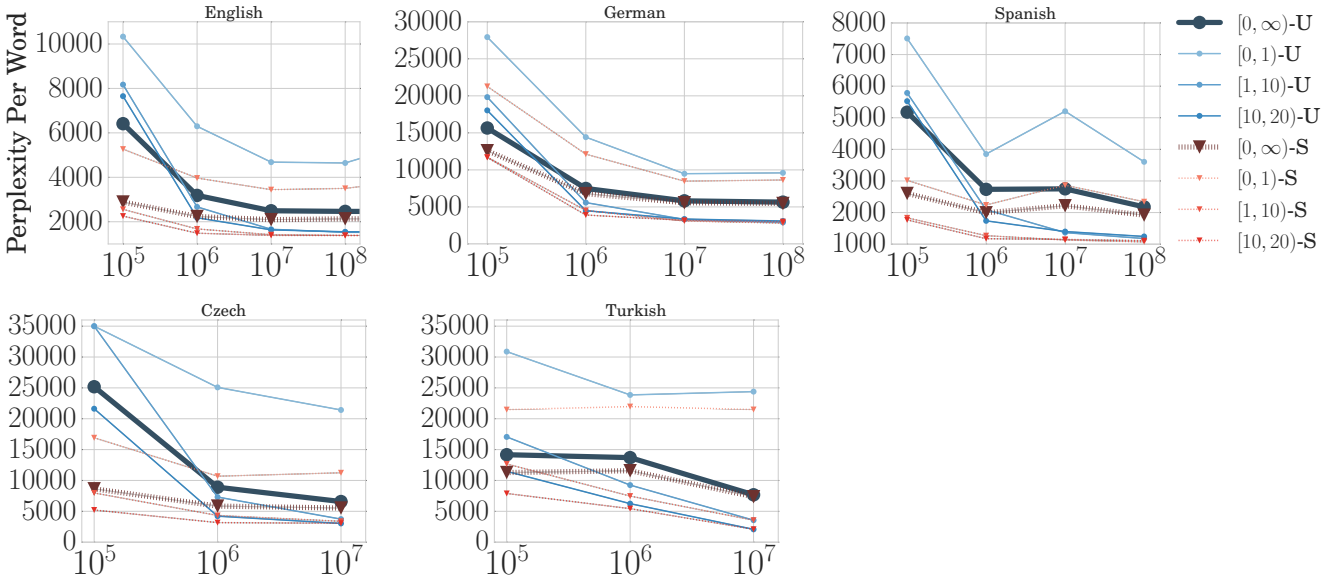

Figure 4: A full version of Figure 3, with all 5 languages.

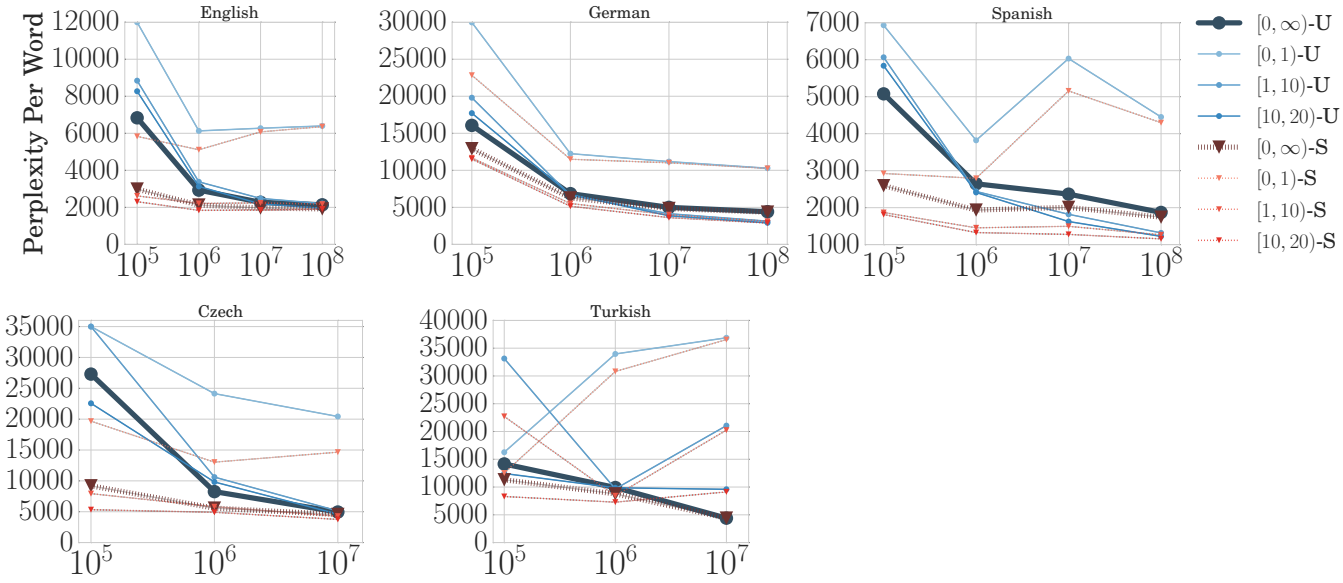

Figure 5: An alternate version of Figure 4. The aggregate curves are the same as before, but the frequency breakdown into word categories is now performed separately at each training size. These points are useful to look up or compare the performance of different word categories (novel, rare, frequent) at a given training size. However, the points along a given curve are incomparable: the $[0, 1)$ curve aggregates over many fewer types at the right than at the left. Sometimes all breakdown curves get worse at once even while their aggregate gets better, an example of Simpson's paradox.

C Coordinate Descent Algorithm for MAP Inference

We now derive the algorithm for maximizing the posterior probability $p(\boldsymbol{w}, \boldsymbol{m} \mid \boldsymbol{v})$. This is equivalent to minimizing (7), which is the negative log of the posterior probability plus a constant, repeated here:

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{m}) = \sum_k \lambda ||m_k||_2^2 + \sum_i ||w_i - \sum_{k \in M_i} m_k||_{\Sigma_i}^2 + \sum_i ||v_i - w_i||_{\Sigma_i'/n_i}^2 \tag{9}$$

Recall that $||\boldsymbol{x}||_\Sigma^2 \stackrel{\text{def}}{=} \boldsymbol{x}^T \beth \boldsymbol{x}$ where $\beth \stackrel{\text{def}}{=} \Sigma^{-1}$, the inverse covariance matrix.

We take the partial gradient with respect to a particular vector m_k (renaming the dummy variable k to j):

$$\frac{\partial \mathcal{L}}{\partial m_k} = \frac{\partial}{\partial m_k} \left[\sum_j \lambda ||m_j||_2^2 + \sum_i ||w_i - \sum_{j \in M_i} m_j||_{\Sigma_i}^2 + \sum_i ||v_i - w_i||_{\Sigma_i'/n_i}^2 \right] \tag{10}$$

$$= \frac{\partial}{\partial m_k} \left[\sum_j \lambda ||m_j||_2^2 + \sum_i ||w_i - \sum_{j \in M_i} m_j||_{\Sigma_i}^2 \right] \tag{11}$$

$$= 2\lambda m_k. + \sum_{i \in W_k} -2\beth_i (w_i - \sum_{j \in M_i} m_j) \tag{12}$$

We now set this equal to $\mathbf{0}$:

$$\mathbf{0} = \lambda m_k - \sum_{i \in W_k} \beth_i (w_i - \sum_{j \in M_i} m_j) \tag{13}$$

Rearranging terms,

$$\lambda m_k = \sum_{i \in W_k} (\beth_i w_i - \sum_{j \in M_i} \beth_i m_j) \tag{14}$$

$$\lambda m_k + \sum_{i \in W_k} \beth_i m_k = \sum_{i \in W_k} (\beth_i w_i - \sum_{j \in M_i, j \neq k} \beth_i m_j) \tag{15}$$

$$\left(\lambda I + \sum_{i \in W_k} \beth_i \right) m_k = \sum_{i \in W_k} (\beth_i w_i - \sum_{j \in M_i, j \neq k} \beth_i m_j). \tag{16}$$

Finally, we arrive at the update rule for m_k:

$$m_k \leftarrow \left(\lambda I + \sum_{i \in W_k} \beth_i \right)^{-1} \sum_{i \in W_k} \beth_i (w_i - \sum_{j \in M_i, j \neq k} m_j). \tag{17}$$

Now take the partial gradient of $\mathcal{L}$ with respect to a particular vector w_i (renaming dummy variable i to j):

$$\frac{\partial \mathcal{L}}{\partial w_i} = \frac{\partial}{\partial w_i} \left[\sum_k \lambda ||m_k||_2^2 + \sum_j ||w_j - \sum_{k \in M_j} m_k||_{\Sigma_j}^2 + \sum_j ||v_j - w_j||_{\Sigma_i'/n_i}^2 \right] \tag{18}$$

$$= \frac{\partial}{\partial w_i} \left[\sum_j ||w_j - \sum_{k \in M_j} m_k||_{\Sigma_j}^2 + \sum_j ||v_j - w_j||_{\Sigma_i'/n_i}^2 \right] \tag{19}$$

$$= 2\beth_i (w_i - \sum_{k \in M_i} m_k) - 2n_i \beth_i' (v_i - w_i) \tag{20}$$

Setting this equal to $\mathbf{0}$, we get

$$\Im_i(w_i - \sum_{k \in M_i} m_k) = n_i \Im_i'(v_i - w_i) \tag{21}$$

$$\left(n_i \Im_i' + \Im_i\right) w_i = n_i \Im_i' v_i + \Im_i \sum_{k \in M_i} m_k \tag{22}$$

This yields the update rule

$$w_i \leftarrow (n_i \Im_i' + \Im_i)^{-1} \left(n_i \Im_i' v_i + \Im_i \sum_{k \in M_i} m_k \right) \tag{23}$$