Speakers Fill Lexical Semantic Gaps with Context

Tiago Pimentel[ð] **Rowan Hall Maudslay**[ð] **Damián Blasi**[ə,ʃ,ʁ] **Ryan Cotterell**[ð,ʕ]
[ð]University of Cambridge [ə]Harvard University [ʃ]MPI SHH [ʁ]HSE University [ʕ]ETH Zürich
tp472@cam.ac.uk, rh635@cam.ac.uk,
dblasi@fas.harvard.edu, ryan.cotterell@inf.ethz.ch

Abstract

Lexical ambiguity is widespread in language, allowing for the reuse of economical word forms and therefore making language more efficient. If ambiguous words cannot be disambiguated from context, however, this gain in efficiency might make language less clear—resulting in frequent miscommunication. For a language to be clear *and* efficiently encoded, we posit that the lexical ambiguity of a word type should correlate with how much information context provides about it, on average. To investigate whether this is the case, we operationalise the lexical ambiguity of a word as the entropy of meanings it can take, and provide two ways to estimate this—one which requires human annotation (using *WordNet*), and one which does not (using BERT), making it readily applicable to a large number of languages. We validate these measures by showing that, on six high-resource languages, there are significant Pearson correlations between our BERT-based estimate of ambiguity and the number of synonyms a word has in *WordNet* (e.g. $\rho = 0.40$ in English). We then test our main hypothesis—that a word's lexical ambiguity should negatively correlate with its contextual uncertainty—and find significant correlations on all 18 typologically diverse languages we analyse. This suggests that, in the presence of ambiguity, speakers compensate by making contexts more informative.

1 Introduction

Linguistic structure and meaning are often underdetermined in the linguistic signal. In an extreme case this can lead to **ambiguity**: sentences might allow more than one valid syntactic structure, and pronouns could corefer to various antecedents. Complementarily, linguistic signals can also overdetermine some aspect of the intended message—for instance, agreement patterns may require redundant marking, and word forms might occupy sparsely

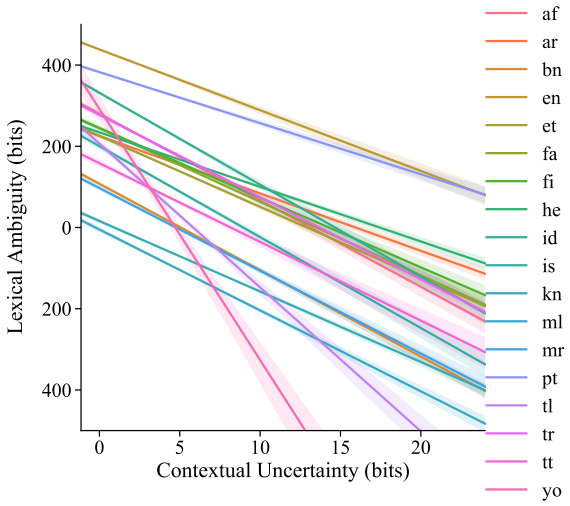

Figure 1: The relationship between contextual uncertainty—how uncertain a word is given its context—and lexical ambiguity, across a diverse set of languages.

populated parts of the phonological space (Harley and Bown, 1998).

In a tradition that goes back at least to Zipf, it has been hypothesised that individuals maintain an efficient balance between over- and under-specifying an intended message. Such balance is mediated by conflicting pressures for both **clarity** (the quality that allows the reconstruction of the intended message), and **economy of expression** (which allows for inexpensive and rapid encoding of the message in a linguistic signal).

A recent instantiation of this idea is that in an efficient language, one expects economical words (which are short or phonotactically simple) to be associated with multiple unrelated meanings, so they can be more widely used (Piantadosi et al., 2012). At first blush, this may appear to sacrifice clarity, increasing ambiguity and making it more difficult for a listener to resolve the linguistic signal. The emerging picture from psycholinguistics and

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4004–4015,
November 16–20, 2020.

pragmatics, however, is that individuals *can* fill in these ambiguous gaps, by tapping on additional linguistic or extra-linguistic cues (Tanenhaus et al., 1995; Federmeier and Kutas, 1999; Dautriche et al., 2018). An obvious example is given by the role of contextual information in reducing the ambiguity associated with the meaning of a word form. For instance, the contexts which surround the word *ruler* in the sentences 'Alice borrowed a *ruler* from her friends at school' and 'Bob rose to power and became a ruthless *ruler*' each play a crucial role in disambiguating its intended underlying meaning.

To remain robust in the presence of noise, we may expect the linguistic signal to be on average somewhat overdetermined by the speaker, leading to redundancy in how words and their contexts determine the intended meaning.[1] By analysing this redundant information-theoretically under the assumption that languages strike a balance between economy of expression and clarity, we derive that the 'amount' of lexical ambiguity in a given word type should negatively correlate with how uncertain on average the word is given its context (see §4). As communication unfolds, the efficiency of a particular word can only be modestly modified (e.g. by choosing clipped forms when available; Mahowald et al., 2013). However, contexts can be enriched or demoted dynamically, so as to complement a word with the evidence needed for disambiguation.

To investigate whether it is the case that the contexts in which a word appears are systematically adapted to enable disambiguation, we first provide an operationalisation of lexical ambiguity, grounded in information theory. We then provide two methods for estimating it, one using *WordNet* (Miller, 1995), and the other using multilingual BERT's contextualised embeddings (Devlin et al., 2019), which allows us to explore a large set of languages. We validate our lexical ambiguity measurements by comparing one to the other in six high-resource languages from four language families (Afro-Asiatic: Arabic; Austronesian: Indonesian; Indo-European: English, Persian and Portuguese; Uralic: Finnish), and find significant correlations between the number of synsets in *WordNet* and our BERT estimate (e.g. $\rho = 0.40$ in English), indicating that our annotation-free method for measuring lexical ambiguity is useful.

We then test our main hypothesis—that the contextual uncertainty about a word should negatively correlate with its degree of lexical ambiguity. First, we test this on the same set of six high-resource languages for which we have *WordNet* annotation, and find significant negative correlations on five of them. We then extend our evaluation, using our BERT-based measure, to cover a much more representative set of 18 typologically diverse languages: Afrikaans, Arabic, Bengali, English, Estonian, Finnish, Hebrew, Indonesian, Icelandic, Kannada, Malayalam, Marathi, Persian, Portuguese, Tagalog, Turkish, Tatar, and Yoruba.[2] In this set, we find significant negative correlations for all languages (see Figure 1).

2 Ambiguity in Language

While the pervasiveness of ambiguity in language encumbers the algorithmic processing of natural language (Church and Patil, 1982; Manning and Schütze, 1999), people seamlessly overcome ambiguity through both linguistic and non-linguistic means. World knowledge, pragmatic inferences, and expectations about discourse coherence all contribute to rapidly decoding the intended message out of potentially ambiguous signals (Wasow, 2015). While sometimes ambiguity might indeed result in an observed processing burden (Frazier, 1985), which could lead communication astray, individuals can in response retrace and reanalyse their inferences (as it has been famously shown in garden-path sentences like "The horse raced past the barn fell"; Bever, 1970).

This outstanding capacity to navigate ambiguous linguistic signals calls for a reexamination of the presence of ambiguity found in language. If the linguistic signal was deterministically and uniquely decodable—as, for instance, in the universal language proposed by Wilkins (Borges, 1964)—then all of the para-linguistic evidence would be redundant, and the code underlying the signal would be substantially more cumbersome. On the other hand, if linguistic signals present individuals with too many compatible inferences, communication would break down. An extreme case is represented by Louis Victor Leborgne, an aphasia patient described by Paul Broca (Mohammed et al., 2018). Louis, in spite of immaculate comprehension and mental functions, was unable to utter anything else than the syllable "tan" in his attempts to communicate.

[1] We refer to overdetermination with relation to redundancies in the signal itself, rather than a precise intended meaning.

[2] We refer to these using ISO 639-1 codes.

The most influential explanation offered for why natural languages are seemingly far from both extremes derives from the seminal work of Zipf (1949). In that work, Zipf proposed several aspects of human cognition and behaviour could be derived from the principle of least effort. Languages should aim to minimise the complexity and cost of linguistic signals as much as possible, under the sole constraint that the signal can be decoded efficiently.

2.1 Lexical Ambiguity

We are concerned exclusively with **lexical ambiguity**. A classic example is the English word *bank*, which can refer to either an establishment where money is kept, or the patch of land alongside a river. A significant source of lexical ambiguity is word types which exhibit multiple senses, which are said to be **polysemous** or **homonymous**.[3] Dautriche (2015) estimates that about 4% of word forms are homophones: "such variation is the rule rather than the exception" (Cruse, 1986).

Lexical ambiguity is, in general, a fuzzy concept. Not only can it be unclear what it means for two senses to be distinct, but different linguistic annotators will also have different opinions on what constitutes a word sense versus a productive use of metaphor. Often the 2nd or 3rd definitions of a word in a dictionary blur this line (Lakoff and Johnson, 1980)—in *WordNet* (Miller, 1995), for instance, the third sense of *attack* (intense adverse criticism, e.g. "the government has come under attack") could be viewed as a metaphorical usage of the first (a military offensive against an enemy, e.g. "the attack began at dawn"), projected from one domain to another. Indeed, this fuzziness has led some researchers to prefer unsupervised word sense induction methods, as they obviate the potentially problematic annotation altogether (e.g. Panchenko et al., 2017). Such unsupervised methods are not without problems, though, with one example being their overreliance on topical words (Amrami and Goldberg, 2019). These difficulties motivate us to opt for using two distinct representation of a word's lexical ambiguity: one hand-annotated and discrete, the other unsupervised and continuous.

2.2 Accounts of Lexical Ambiguity

When investigating the relationship between ambiguity and word frequency, Zipf argued that ambiguity results as a trade-off from opposing forces between speaker and listener, together optimising the communication channel via a principle of least effort: the listener wants to easily disambiguate, the speaker wants to choose words which required little effort to utter, and to avoid excessively searching their lexicon.

Building on Zipf's (1949) theories, Piantadosi et al. (2012) posit that, when viewed information-theoretically, ambiguity is in fact a *requirement* for a communication system to be efficient. Focusing on economy of expression, Piantadosi et al. suggest that lexical ambiguity serves a purpose when the context allows for disambiguation—it allows the re-use of simpler word forms.[4] They support their hypothesis by demonstrating a correlation between the number of senses for a word listed in *WordNet* (Miller, 1995) and a number of measures of speaker effort—phonotactic well-formedness, word length and the word's log unigram probability (based on a maximum-likelihood estimate from a large corpus).

More recently, Dautriche et al. (2018) showed that languages' homophones are more likely to appear across distinct syntactic and semantic categories, and will therefore be naturally easier to disambiguate. In this work, we show that speakers compensate for lexical ambiguity by making contexts themselves more informative in its presence.

We note an important detail in one of Piantadosi et al.'s experiments. In their work, they employ unigram surprisal (i.e. $-\log p_{unigram}(\cdot)$, where $p_{unigram}(\cdot)$ is the unigram distribution) as a proxy for ease of production, correlating this with polysemy. They justify this approximation based on the fact that more frequent words are, in general, processed more quickly (Reder et al., 1974). However, this measure has a confounder with our hypothesis: a word's frequency correlates with its contextual uncertainty. We believe our proposed measure to be more directly connected with lexical ambiguity.

3 Ambiguity and Uncertainty

We formulate both lexical ambiguity and contextual uncertainty information-theoretically. Let $\mathcal{M}$ be a space of all lexical meaning representations, $\mathcal{W}$ be the space of all words and $\mathcal{C}$ be the space of all contexts. We denote the $\mathcal{M}$-, $\mathcal{W}$-, and $\mathcal{C}$-valued random variables as M, W and C, respectively, and name elements of those sets m, w and c. We take $\mathcal{M}$ to be an either discrete or continuous mean-

[3] We make no distinction between polysemy, homonymy, and other sources of lexical ambiguity a word may exhibit.

[4] Recent work, though, has shed some doubt in the interpretation behind these results, showing they might arise solely due to a language's phonotactics distribution (Trott and Bergen, 2020; Caplan et al., 2020).

ing space, $\mathcal{W}$ to be the set of words in a language (excluding the beginning-of- and end-of-sequence symbols, BOS and EOS) and

$$\mathcal{C} = \{\langle \text{BOS} \circ \mathbf{p}, \mathbf{s} \circ \text{EOS}\rangle \mid \mathbf{p} \circ w \circ \mathbf{s} \in \mathcal{W}^*\} \quad (1)$$

where $\circ$ denotes string concatenation, and $\mathbf{p}$ and $\mathbf{s}$ are the prefix and suffix context strings respectively. This set contains every possible context that could surround a word, padded with beginning-of-sequence and end-of-sequence symbols. We additionally define $\tilde{\mathbf{p}} = \text{BOS} \circ \mathbf{p}$ and $\tilde{\mathbf{s}} = \mathbf{s} \circ \text{EOS}$.

3.1 Lexical Ambiguity

We start with a formalisation of lexical ambiguity. Specifically, we formalise the lexical ambiguity of *an entire language* as

$$\mathrm{H}(M \mid W) = - \sum_{w \in \mathcal{W}} p(w) \int p(m \mid w) \log_2 p(m \mid w) \, \mathrm{d}m \quad (2)$$

Interpreting entropy as uncertainty, this definition implies that the harder it is to predict the meaning of a word from its form alone, the more lexically ambiguous that word must be.

We will generally be interested in the half-pointwise entropy, rather than the entropy itself. In the case of lexical ambiguity, we consider the following half-pointwise entropy

$$\mathrm{H}(M \mid W = w) = - \int p(m \mid w) \log_2 p(m \mid w) \, \mathrm{d}m \quad (3)$$

This half-pointwise entropy tells us how difficult it is to predict the meaning when you know the specific word *without* considering its context. We will not generally have access to the true distribution $p(m \mid w)$, so we will need to approximate this entropy. This is discussed in §5.1. A unique feature of this operationalisation of lexical ambiguity is that it is language independent.[5] However, the quality of a possible approximation will vary from language to language, depending on the models and the data available in that language.

A final note is that mutual information between M and W *as a function of* w is equivalent, up to an additive constant, to the conditional entropy

$$\mathrm{I}(M; W = w) = \mathrm{H}(M) - \mathrm{H}(M \mid W = w) \quad (4)$$

[5]We acknowledge the abuse of this bigram in the NLP literature (Bender, 2009), and use it in the following specific sense: the operationalisation may be applied to *any* language independent of its typological profile.

where $\mathrm{H}(M)$ is constant with respect to w. This equation asserts something rather trivial: that lexical ambiguity is inversely correlated with how informative a word is about its meaning.

3.2 Contextual Uncertainty

The predictability of a word in context is also naturally operationalised information-theoretically. We take the contextual uncertainty, once again defined for *an entire language*, as

$$\mathrm{H}(W \mid C) = - \sum_{w \in \mathcal{W}} p(w) \sum_{\mathbf{c} \in \mathcal{C}} p(\mathbf{c} \mid w) \log_2 p(w \mid \mathbf{c}) \quad (5)$$

Again, we are mostly interested in the half-pointwise entropy, which tells us how predictable a given word is, averaged over all contexts:

$$\mathrm{H}(W = w \mid C) = - \sum_{\mathbf{c} \in \mathcal{C}} p(\mathbf{c} \mid w) \log_2 p(w \mid \mathbf{c}) \quad (6)$$

We take this as our operationalisation of contextual uncertainty. We note that this definition is different to typical uses of surprisal in computational psycholinguistics (Hale, 2001; Levy, 2008; Seyfarth, 2014; Piantadosi et al., 2011; Pimentel et al., 2020). Most work in this vein attempts to maintain cognitive plausibility, usually calculating surprisal based on only the unidirectional left piece of the context, as $-\log p(w \mid \mathbf{c}_{\leftarrow})$.

Although surprisal is the operationalisation we are interested in here, we note that a word may have low surprisal if it is frequent *across many* contexts and not just in a specific one under consideration. Sticking with our notion of half-pointwiseness, we define contextual informativeness as

$$\mathrm{I}(W = w; C) = \mathrm{H}(W = w) - \mathrm{H}(W = w \mid C) \quad (7)$$

where we define a word's pointwise entropy (also known as surprisal) as

$$\mathrm{H}(W = w) = -\log_2 p(w) \quad (8)$$

The mutual information between a word and its context was studied before by Bicknell and Levy (2011), Futrell and Levy (2017) and Futrell et al. (2020)—although only using the unidirectional left piece of the context. Eq. (7) again asserts something trivial: low contextual uncertainty implies in an informative context. This informativeness itself is upper-bounded by the word's absolute negative log-probabiliy (i.e. the unigram surprisal).

4 Hypothesis: Why Should Ambiguity Correlate with Uncertainty?

As discussed in §1, we expect the linguistic signal to be on average somewhat overdetermined or redundant—such redundancy leads to **robustness** in noisy situations, when part of the signal may be lost during its implementation. A natural measure of robustness is the three-way mutual information between the context of a word, the word itself, and meaning—$\mathrm{I}(M;C;W)$—which represents how much information about the meaning is redundantly encoded in both the context and the word. The half-pointwise tripartite mutual information can be decomposed as

$$\begin{aligned}
&\mathrm{I}(M;C;W=w)\\
&\quad=\mathrm{I}(M;W=w)-\mathrm{I}(M;W=w\mid C)\\
&\quad=\mathrm{I}(M;W=w)-\mathrm{H}(W=w\mid C)\\
&\qquad+\cancel{\mathrm{H}(W=w\mid M,C)}\\
&\quad\approx \underbrace{\mathrm{I}(M;W=w)}_{(1)}-\underbrace{\mathrm{H}(W=w\mid C)}_{(2)} \qquad (9)
\end{aligned}$$

In this equation, we assume there are no true synonyms under a specific context—i.e. given a meaning and a context there is no uncertainty about the word choice: $\mathrm{H}(W=w\mid M,C)\approx 0$. Term 1 is the information a word shares with its meaning (which is inversely correlated with lexical ambiguity; see eq. (4)) and term 2 is the predictability of a word in context or the contextual uncertainty (which is itself inversely correlated with contextual informativeness; see eq. (7)).

For a language to be efficient, it may reuse its optimal word forms (as defined by their utterance effort), increasing lexical ambiguity (Piantadosi et al., 2012) and reducing the amount of information a word contains about its meaning (term 1). This reduces redundancy though, increasing the chance of miscommunication in the presence of noise. Speakers can compensate for this by making contexts more informative for these words (term 2 smaller). A negative correlation between contextual uncertainty and lexical ambiguity then arises from the trade-off between clarity and economy.

5 Computation and Approximation

Our information-theoretic operationalisation requires approximation. First, we do not know the true distributions over words, their meanings and their contexts. Second, even if we did, eq. (3) and eq. (6) would likely be hard to compute.

5.1 Lexical Ambiguity

In this section, we provide two approximations for lexical ambiguity. One assumes discrete word senses and requires data annotation (*WordNet*), while the other considers continuous meaning spaces (BERT) and allows us to extend our analysis to languages with fewer of these resources.

Discrete senses *WordNet* (Miller, 1995) is a valuable resource available in high-resource languages, which provides a list of synsets for word types. By taking these synsets to be the possible meanings of a word, and assuming a uniform distribution over them, we approximate the entropy as

$$\mathrm{H}(M\mid W=w)\approx \log_2(\#\textit{senses}[w]) \qquad (10)$$

Continuous meaning space We now describe how to approximate ambiguity using BERT (Devlin et al., 2019).[6] Let $w\in\mathcal{W}$ be a word and let $\mathbf{c}=\langle\tilde{\mathbf{p}},\tilde{\mathbf{s}}\rangle\in\mathcal{C}$ be a padded context. We assume that a word's contextual embedding in BERT (i.e. its final hidden state) is a good approximation for its meaning in a given sentence.[7] We define the hidden state of a word w in a context $\mathbf{c}$ as

$$\mathbf{h}_{\langle w,\mathbf{c}\rangle}=\mathrm{BERT}(\tilde{\mathbf{p}}\circ w\circ\tilde{\mathbf{s}}) \qquad (11)$$

and we approximate the true distribution over words, meanings and contexts by

$$p(w,m,\mathbf{c})\approx\delta(m\mid w,\mathbf{c})\,p(w,\mathbf{c}) \qquad (12)$$

where we define $\delta(m\mid w,\mathbf{c})$ to place probability 1 on the point $m=\mathbf{h}_{\langle w,\mathbf{c}\rangle}$ and 0 on every other point. In other words, we assume the meaning is a deterministic function of a word–context pair, and that it is approximated by BERT's hidden state.

This alone is not enough to estimate eq. (3), though, since we still do not have access to the true distribution $p(w,\mathbf{c})$. Furthermore, estimating the marginal distribution $p(m|w)$ directly is infeasible, given the sparsity of the meaning space. Instead, we approximate an upper bound of the entropy directly—exploiting the fact that a Gaussian distribution $\mathcal{N}(\mu,\Sigma)$ will have an entropy that is

[6] We used the implementation of Multilingual BERT made available by Wolf et al. (2019).

[7] Since BERT returns embeddings for WordPiece units (Wu et al., 2016) rather than words, we average them per word to get embeddings at the word-level. We acknowledge that this is a naïve method of compositionality; improving the method would likely strengthen our results.

greater than or equal to any other distribution with the same finite and known (co)variance (Cover and Thomas, 2012, Chapter 8):[8]

$$\begin{aligned} &\mathrm{H}(M \mid W = w) \qquad (13)\\ &\quad \leq \mathrm{H}(\mathcal{N}(\boldsymbol{\mu}_w, \Sigma_w)) = \frac{1}{2}\log_2 \det\left(2\pi e \Sigma_w\right) \end{aligned}$$

We estimate this covariance based on a corpus of N word–context pairs $\{\langle w, \mathbf{c}\rangle_i\}_{i=1}^{N}$, which we assume to be sampled according to the true distribution p (our corpora comes from Wikipedia dumps and is described in §6).[9]

The tightness of this upper bound on the entropy depends on both the accuracy of the covariance matrix estimation and the nature of the true distribution $p(m \mid w)$. If $p(m \mid w)$ is concentrated in a small region of the meaning space (corresponding to a word with nuanced implementations of the same sense), the bound in eq. (13) could be relatively tight. In contrast, a word with several unrelated homophones would correspond to a highly structured $p(m \mid w)$ (e.g. with multiple modes in far distant regions of the space) for which this normal approximation would result in a very loose upper bound.

5.2 Contextual Uncertainty

How uncertain the context is about a specific word is formalised in the half-pointwise entropy presented in eq. (6). We may get an upper bound on this entropy from its cross-entropy:

$$\begin{aligned} \mathrm{H}(W = w \mid C) &\leq \mathrm{H}_{q_{\boldsymbol{\theta}}}(W = w \mid C) \qquad (14)\\ &= -\sum_{\mathbf{c}\in\mathcal{C}} p(\mathbf{c} \mid w) \log q_{\boldsymbol{\theta}}(w \mid \mathbf{c}) \end{aligned}$$

where $q_{\boldsymbol{\theta}}$ is a cloze language model that we train to approximate p (as we explain later in this section). This equation, though, still requires an infinite sum over $\mathcal{C}$. We avoid that by using an empirical estimate of the cross-entropy:

$$\mathrm{H}_{q_{\boldsymbol{\theta}}}(W = w \mid C) \approx -\sum_{i=1}^{N_w} \log q_{\boldsymbol{\theta}}(w_i \mid \mathbf{c}_i) \qquad (15)$$

where N_w is the number of samples we have for a specific word type w.

To choose an appropriate distribution $q_{\boldsymbol{\theta}}(w \mid \mathbf{c})$, we train a model on a masked language modelling task. Defining MASK as a special type in vocabulary V, we take a masked hidden state as

$$\mathbf{h}_{\mathbf{c}} = \mathrm{BERT}(\tilde{\mathbf{p}} \circ \textsc{mask} \circ \tilde{\mathbf{s}}) \qquad (16)$$

We then use this masked hidden state to estimate the distribution

$$q_{\boldsymbol{\theta}}(w \mid \mathbf{c}) = \mathrm{softmax}(W^{(2)}\sigma(W^{(1)}\mathbf{h}_{\mathbf{c}})) \qquad (17)$$

where $W^{(\cdot)}$ are linear transformations, and bias terms are omitted for brevity. We fix BERT's parameters and train this model with Adam (Kingma and Ba, 2015), using its default learning rate in PyTorch (Paszke et al., 2019). We use a ReLU as our non-linear function σ and 200 as our hidden size, training for only one epoch. By minimising cross-entropy loss we achieve an estimate for p.

We do not use BERT directly as our model $q_{\boldsymbol{\theta}}$ because its multilingual version was trained on multiple languages, and, thus, was not optimised on each individually. We found this resulted in poor approximations on the lowest-resource languages. Furthermore, we note that BERT gives probability estimates for word pieces (as opposed to the words themselves), and combining these piece-level probabilities to word-level ones is non-trivial. Indeed, doing so would require running BERT several times per word, increasing the already high computational requirements of this study. To compute the probability of a word composed of two word pieces, for example, we would need to run the model with two masks, i.e. $\mathrm{BERT}(\tilde{\mathbf{p}} \circ \textsc{mask} \circ \textsc{mask} \circ \tilde{\mathbf{s}})$, and combine the pieces' probabilities. To correctly estimate the probability distribution over the entire vocabulary (i.e. $q_{\boldsymbol{\theta}}(w \mid \mathbf{c})$), we would need to replace each position with an arbitrary number of MASKs and normalise these probability values.

6 Data

We used Wikipedia as the main data source for all our experiments. Multilingual BERT[10] was trained on the 104 languages with the largest Wikipedias[11]—of these, we subsampled a diverse set of 18 for our experiments: Afrikaans, Arabic, Bengali, English, Estonian, Finnish, Hebrew, Indonesian, Icelandic, Kannada, Malayalam, Marathi, Persian, Portuguese, Tagalog, Turkish, Tatar, and Yoruba.

[8] We note that, unlike its discrete counterpart, differential entropy values can be negative.

[9] We explain how to approximate the covariance matrix Σ_w per word type in App. A.

[10] Information about multilingual BERT can be found in: https://github.com/google-research/bert/blob/master/multilingual.md

[11] List of Wikipedias can be found in https://meta.wikimedia.org/wiki/List_of_Wikipedias

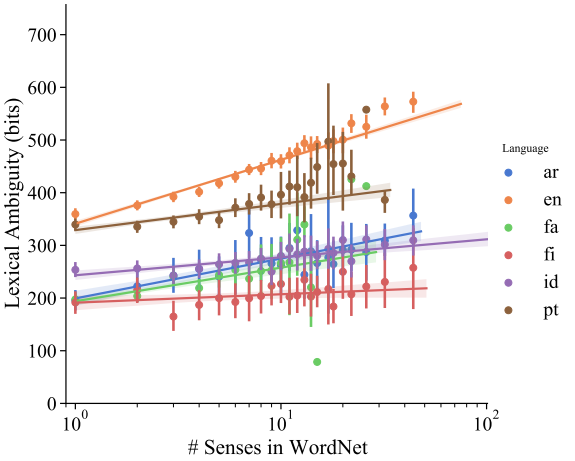

Figure 2: Correlating our BERT-based estimate of lexical ambiguity with the number of senses in *WordNet*

Language	# Types	Pearson	Spearman
Arabic	836	0.25**	0.30**
English	6995	0.40**	0.40**
Finnish	1247	0.06*	0.07*
Indonesian	3308	0.12**	0.13**
Persian	2648	0.14**	0.13**
Portuguese	3285	0.13**	0.13**

** $p < 0.01$ * $p < 0.1$

Table 1: Correlations between a word's lexical ambiguity as estimated with BERT or *WordNet*.

For each of these languages, we first downloaded their entire Wikipedia, which we sentencized and tokenized using language specific models in spaCy (Honnibal and Montani, 2017)—our definition of a word here is, thus, a token as given by the spaCy tokenizer. We then subsampled 1 million random sentences per language for our analysis and another 100,000 random sentences to train the model $q_{\boldsymbol{\theta}}$. We run multilingual BERT on the 1 million analysis sentences to acquire both $\mathbf{h}_{\langle w,\mathbf{c}\rangle}$ and $\mathbf{h}_{\mathbf{c}}$ (eq. (11) and eq. (16)) for each word in these corpora—discarding any word for which we do not have at least 100 contexts in which the word occurs. For the purpose of our analysis, we also discarded any word containing characters not in the individual scripts of the analysed language. The final number of word types used in our analysis can be found in Tables 1 and 3.

7 Discussion: *WordNet* vs. BERT-based approximations

The novel continuous (BERT-based) approximation of lexical ambiguity has two important virtues over the alternative *WordNet*-based measure. On the practical side, it can be readily computed for many languages. Since we are using multilingual BERT for our continuous approximation, as discussed in §5, this quantity is easily obtainable for the 104 languages on which it was trained. Second, on more theoretical grounds, the continuous representation of the space of meanings might better capture the gradient that goes from subtle but distinct senses of the same word to completely unrelated homophones (Cruse, 1986, p. 51). Alternatively, the *WordNet*-based measure of lexical ambiguity is supported by expert human annotation and extensive research on its linguistic and psycholinguistic correlates, e.g. Sigman and Cecchi (2002) and Budanitsky and Hirst (2006).

These differences notwithstanding, we expect both measures to correlate to a certain degree. To evaluate this, we run an experiment comparing both estimates in six languages from four different families for which *WordNet* is available: Arabic, English, Finnish, Indonesian, Persian, and Portuguese.

Figure 2 and Table 1 show that indeed both measures are positively correlated, although the association may be modest in some languages. The Pearson correlation between our estimates is $\rho = 0.40$ for English, but only $\rho = 0.06$ for Finnish—other languages lie in the range between the two.[12] This correlation seems to increase with the quality of the BERT model for the language under consideration—English has the largest Wikipedia, so multilingual BERT should naturally be better modelling it, while Finnish has the smallest Wikipedia among these six languages. A complementary explanation is that *WordNet* itself might be better for English than other languages—while English's *WordNet* contains synsets for 147,306 words, Persian only has them for 17,560. This suggests that the modest associations found should be taken as pessimistic lower bounds.

A potential underlying problem in the above study is that the number of senses a word has in *WordNet* might rely on word frequency (this beyond a true underlying relationship with it)—e.g. annotating senses for frequent words may be easier than for infrequent ones. Furthermore, the number

[12] For all tests of significance in this paper, we apply Benjamini and Hochberg's correction (1995).

Language	# Types	*WordNet*	Frequency
Arabic	836	0.28**	0.30**
English	6995	0.38**	0.21**
Finnish	1247	0.07*	0.35**
Indonesian	3308	0.09**	0.37**
Persian	2648	0.13**	0.14**
Portuguese	3285	0.13**	0.29**

** $p < 0.01$ * $p < 0.1$

Table 2: Parameters (and their significance) of a multivariate linear regression predicting our BERT-based measure of ambiguity from both our *WordNet* estimate and the word's frequency. All analysed variables were normalised to have zero mean and unit variance.

of samples a word has in our corpus will affect its sample density in the embedding space and thus its estimated BERT entropy. As a second evaluation, we therefore train a multivariate linear regressor predicting our BERT-based measure not only from the log of the number of senses a word has in *WordNet*, but also the word's frequency (i.e. its number of occurrences in the corpus). This analysis is presented in Table 2, where we can see that both our estimates of lexical ambiguity still correlate when controlling for frequency. This table also shows that our BERT-based estimate still correlates with the word's frequency when controlling for the number of senses the word has in *WordNet*. Future work could delve further into what this correlation implies, with the potential to improve our proposed annotation-free estimate of lexical ambiguity.

8 Lexical Ambiguity Correlates With Contextual Uncertainty

We now test whether lexical ambiguity negatively correlates with contextual uncertainty, the main hypothesis of our paper. We first evaluate this on a set of six high-resource languages, using our *WordNet* estimate for the lexical ambiguity of a word. The top half of Table 3 shows the results: for five of the six languages, there is a negative correlation between the number of senses of a word and contextual uncertainty ($p < 0.01$). The top half of Figure 3 further presents these results. In these Figures we see that, especially for highly ambiguous words, contextual uncertainty tends to be very small. This supports our hypothesis, but only on a restricted set of languages for which *WordNet* is available.

With that in mind, we now consider a larger and more diverse set of 18 languages, analysed using

Language	# Types	Pearson	Spearman
Lexical ambiguity as WordNet			
Arabic (ar)	836	-0.14**	-0.15**
English (en)	6995	-0.07**	-0.11**
Finnish (fi)	1247	0.01	-0.00
Indonesian (id)	3308	-0.09**	-0.14**
Persian (fa)	2648	-0.11**	-0.12**
Portuguese (pt)	3285	-0.10**	-0.11**
Lexical ambiguity as BERT			
Afrikaans (af)	4505	-0.41**	-0.52**
Arabic (ar)	10181	-0.33**	-0.41**
Bengali (bn)	8128	-0.43**	-0.44**
English (en)	7097	-0.33**	-0.35**
Estonian (et)	4482	-0.40**	-0.44**
Finnish (fi)	3928	-0.38**	-0.45**
Hebrew (he)	13819	-0.34**	-0.37**
Indonesian (id)	4524	-0.45**	-0.57**
Icelandic (is)	3578	-0.44**	-0.46**
Kannada (kn)	9695	-0.42**	-0.41**
Malayalam (ml)	6203	-0.47**	-0.46**
Marathi (mr)	5821	-0.39**	-0.40**
Persian (fa)	6788	-0.39**	-0.49**
Portuguese (pt)	5685	-0.31**	-0.45**
Tagalog (tl)	3332	-0.45**	-0.50**
Turkish (tr)	4386	-0.40**	-0.46**
Tatar (tt)	2997	-0.34**	-0.39**
Yoruba (yo)	417	-0.55**	-0.64**

** $p < 0.01$

Table 3: Correlation between lexical ambiguity and contextual uncertainty.

our BERT-based estimator of lexical ambiguity. Figures 1 and 3 show the relationship between contextual uncertainty and lexical ambiguity—in all 18 analysed languages, we find negative correlations, further supporting our hypothesis. These correlations are presented in the bottom half of Table 3, and range from Pearson $\rho = -0.31$ in Portuguese to $\rho = -0.55$ in Yoruba ($p < 0.01$).

Comparing the top and bottom half of Table 3, we see that the correlations are larger when using our BERT estimate rather than the *WordNet* one. We believe this may result from one or all of the following: (i) there is a confounding effect caused by the use of the same model (BERT) to estimate both ambiguity and surprisal, (ii) the assumption that the senses in *WordNet* are uniformly distributed may be simplistic, and (iii) our BERT-based ambiguity estimate may capture a more subtle sense of ambiguity than *WordNet*, which may result in a

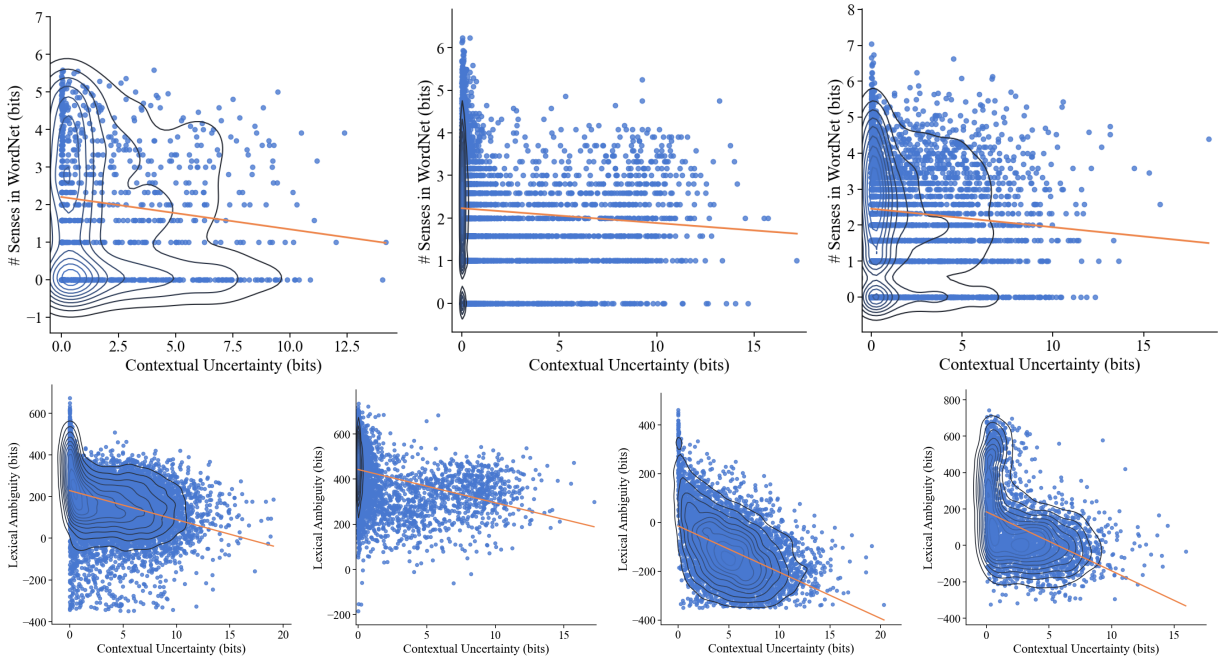

Figure 3: Contextual uncertainty versus lexical ambiguity in a selection of languages. Each plot contains the scatter points (representing each word type), a robust linear regression and kernel density estimate regions. (From left to right; Top) *WordNet*: Arabic, English, Indonesian; (Bottom) BERT: Arabic, English, Malayalam, Tagalog.

stronger correlation with contextual uncertainty.[13] Nonetheless, even if there is a confounding effect in this second batch of experiments (using BERT to estimate lexical ambiguity), the first batch (with *WordNet*) has no such confounding factor—providing strong support for our main hypothesis.

A quick visual inspection of Figure 3 indicates this data might be heteroscedastic—it might have unequal variance across distinct ambiguity levels. To investigate this, we run White's (1980) test on the uncertainty–ambiguity pairs. This verifies the intuition that this distribution is heteroscedastic for both our *WordNet* and BERT measures ($p < 0.01$). Future work should investigate the impact of this heteroscedasticity in lexical ambiguity.

Limitations This work focuses on proposing new information-theoretic approximations for both lexical ambiguity and bidirectional contextual uncertainty and on positing that these two measures should negatively correlate. In this experiment section, we tested the hypothesis on a set of typologically diverse languages. Nonetheless, our experiments are restricted to Wikipedia corpora. This data is naturally limited. For instance, while dialog utterances may rely on extra-linguistic clues, sentences in Wikipedia cannot. Furthermore, due to its ample audience target, the text in Wikipedia may be over descriptive. Future work should investigate if similar results apply to other corpora.

[13] Cruse (1986, p. 51) argues there are two ways in which context affects a word's semantics—selection between units of distinct senses, or contextual modification of a single sense.

9 Conclusion

In this paper we hypothesised that, were a language economical in its expressions *and* clear, then the contextual uncertainty of a word should negatively correlate with its lexical ambiguity—suggesting speakers compensate for lexical ambiguity by making contexts more informative. To investigate this, we proposed an information-theoretic operationalisation of lexical ambiguity, together with two methods of approximating it, one using *WordNet* and one using BERT. We discuss the relative advantages of each, and provide experiments using both. With our *WordNet* approximation, we found significant negative correlations between lexical ambiguity and contextual uncertainty in five out of six high-resource languages analysed, supporting our hypothesis in this restricted setting. With our BERT approximation, we then expanded our analysis to a larger set of 18 typologically diverse languages and found significant negative correlations between lexical ambiguity and contextual uncertainty in all of them, further supporting our hypothesis that contextual uncertainty negatively correlates with lexical ambiguity.

Acknowledgments

Damán Blasi acknowledges funding from the framework of the HSE University Basic Research Program and is funded by the Russian Academic Excellence Project '5-100'.

References

Asaf Amrami and Yoav Goldberg. 2019. Towards better substitution-based word sense induction. *arXiv preprint arXiv:1905.12598*.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Thomas G. Bever. 1970. The cognitive basis for linguistic structures. In John R. Hayes, editor, *Cognition and the Development of Language*, pages 279–362. Wiley & Sons, Inc, New York.

Klinton Bicknell and Roger Levy. 2011. Why readers regress to previous words: A statistical analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Jorge Luis Borges. 1964. The analytical language of John Wilkins. *Other Inquisitions*, 1937–1952:101–105.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Spencer Caplan, Jordan Kodner, and Charles Yang. 2020. Miller's monkey updated: Communicative efficiency and the statistics of words in natural language. *Cognition*, 205:104466.

Kenneth Church and Ramesh Patil. 1982. Coping with syntactic ambiguity or how to put the block in the box on the table. *Computational Linguistics*, 8(3-4):139–149.

Thomas M. Cover and Joy A. Thomas. 2012. *Elements of Information Theory*. John Wiley & Sons.

David A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

Isabelle Dautriche. 2015. *Weaving an ambiguous lexicon*. Ph.D. thesis, Sorbonne Paris Cité.

Isabelle Dautriche, Laia Fibla, Anne-Caroline Fievet, and Anne Christophe. 2018. Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology*, 104:83 – 105.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kara D. Federmeier and Marta Kutas. 1999. A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4):469 – 495.

Lyn Frazier. 1985. *Syntactic Complexity*, Studies in Natural Language Processing, pages 129–189. Cambridge University Press.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3).

Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Trevor A. Harley and Helen E. Bown. 1998. What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, 89(1):151–174.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference for Learning Representations*.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Kyle Mahowald, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nasser Mohammed, Vinayak Narayan, Devi Prasad Patra, and Anil Nanda. 2018. Louis Victor Leborgne ("tan"). *World Neurosurgery*, 114:121–125.

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone Paolo Ponzetto, and Chris Biemann. 2017. Unsupervised does not mean uninterpretable: The case for word sense induction and disambiguation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 86–98, Valencia, Spain. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035. Curran Associates, Inc.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

Lynne M. Reder, John R. Anderson, and Robert A. Bjork. 1974. A semantic interpretation of encoding specificity. *Journal of Experimental Psychology*, 102:648–656.

Scott Seyfarth. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140–155.

Mariano Sigman and Guillermo A. Cecchi. 2002. Global organization of the WordNet lexicon. *Proceedings of the National Academy of Sciences*, 99(3):1742–1747.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Sean Trott and Benjamin Bergen. 2020. Why do human languages have homophones? *Cognition*, 205:104449.

Thomas Wasow. 2015. Ambiguity avoidance is overrated. In *Ambiguity: Language and Communication*, pages 29–48. De Gruyter.

Halbert White. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

John Wilkins. 1668. *An Essay Towards a Real Character, and a Philosophical Language*. The Royal Society, London.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort.* Addison-Wesley Press.

Appendices

A Gaussian Approximation for a Words' Meanings

Given our samples $\{\langle w, \mathbf{c}\rangle_i\}_{i=1}^N$ of word–context pairs (assumed to be drawn from the true distribution p), we get the subset of N_w instances of word type w. We then use an unbiased estimator of the covariance matrix:

$$\Sigma_w \approx \tag{18}$$
$$\frac{1}{N_w - 1} \sum_{i=1}^{N_w} \left(\mathbf{h}_{\langle w,\mathbf{c}\rangle_i} - \tilde{\boldsymbol{\mu}}_w\right) \left(\mathbf{h}_{\langle w,\mathbf{c}\rangle_i} - \tilde{\boldsymbol{\mu}}_w\right)^\top$$

where the sample mean is defined as

$$\tilde{\boldsymbol{\mu}}_w \approx \frac{1}{N_w} \sum_{i=1}^{N_w} \mathbf{h}_{\langle w,\mathbf{c}\rangle_i} \tag{19}$$

We note that these approximations become exact as $N_w \to \infty$ due to the law of large numbers.

Since $\mathbf{h}_{\langle w,\mathbf{c}\rangle}$ (i.e. BERT's hidden state) is a 768 dimensional vector, we might not have enough samples to fully estimate Σ_w. So we actually approximate this entropy by using only its variance $\mathrm{diag}(\Sigma_w)$. This is still an upper bound on the true entropy

$$H(\mathcal{N}(\boldsymbol{\mu}_w, \Sigma_w)) \leq H(\mathcal{N}(\boldsymbol{\mu}_w, \mathrm{diag}(\Sigma_w))) \tag{20}$$

The right side of this equation is, then, used as our actual lexical ambiguity estimate.

B ISO 639-1 Codes

In this Section, we present the set of ISO 639-1 language codes we use throughout this paper—in Table 4.

ISO Code	Language
af	Afrikaans
ar	Arabic
bn	Bengali
en	English
et	Estonian
fi	Finnish
he	Hebrew
id	Indonesian
is	Icelandic
kn	Kannada
ml	Malayalam
mr	Marathi
fa	Persian
pt	Portuguese
tl	Tagalog
tr	Turkish
tt	Tatar
yo	Yoruba

Table 4: ISO Codes and their languages