Are All Languages Equally Hard to Language-Model?

Ryan Cotterell[1] and **Sebastian J. Mielke**[1] and **Jason Eisner**[1] and **Brian Roark**[2]
[1] Department of Computer Science, Johns Hopkins University [2] Google
{ryan.cotterell@,sjmielke@,jason@cs.}jhu.edu roark@google.com

Abstract

For general modeling methods applied to diverse languages, a natural question is: how well should we expect our models to work on languages with differing typological profiles? In this work, we develop an evaluation framework for fair cross-linguistic comparison of language models, using translated text so that all models are asked to predict approximately the same information. We then conduct a study on 21 languages, demonstrating that in some languages, the textual expression of the information is harder to predict with both n-gram and LSTM language models. We show complex inflectional morphology to be a cause of performance differences among languages.

1 Introduction

Modern natural language processing practitioners strive to create modeling techniques that work well on all of the world's languages. Indeed, most methods are portable in the following sense: Given appropriately annotated data, they should, in principle, be trainable on any language. However, despite this crude cross-linguistic compatibility, it is unlikely that all languages are equally easy, or that our methods are equally good at all languages.

In this work, we probe the issue, focusing on *language modeling*. A fair comparison is tricky. Training corpora in different languages have different sizes, and reflect the disparate topics of discussion in different linguistic communities, some of which may be harder to predict than others. Moreover, bits per character, a standard metric for language modeling, depends on the vagaries of a given orthographic system. We argue for a fairer metric based on the bits per utterance using utterance-aligned multi-text. That is, we train and test on "the same" set of utterances in each language, modulo translation. To avoid discrepancies in out-of-vocabulary handling, we evaluate open-vocabulary models.

We find that under standard approaches, text tends to be harder to predict in languages with fine-grained inflectional morphology. Specifically, language models perform worse on these languages, in our controlled comparison. Furthermore, this performance difference essentially vanishes when we remove the inflectional markings.[1]

Thus, in highly inflected languages, either the utterances have more content or the models are worse. (1) Text in highly inflected languages may be *inherently harder to predict* (higher entropy per utterance) if its extra morphemes carry additional, unpredictable information. (2) Alternatively, perhaps the extra morphemes are *predictable in principle*—for example, redundant marking of grammatical number on both subjects and verbs, or marking of object case even when it is predictable from semantics or word order—and yet our current language modeling technology fails to predict them. This might happen because (2a) the technology is biased toward modeling words or characters and fails to discover intermediate morphemes, or because (2b) it fails to capture the syntactic and semantic predictors that govern the appearance of the extra morphemes. We leave it to future work to tease apart these hypotheses.

2 Language Modeling

A traditional closed-vocabulary, word-level language model operates as follows: Given a fixed set of words $\mathcal{V}$, the model provides a probability distribution over sequences of words with parameters to be estimated from data. Most fixed-vocabulary language models employ a distinguished symbol UNK that represents all words not present in $\mathcal{V}$; these words are termed out-of-vocabulary (OOV).

Choosing the set $\mathcal{V}$ is something of a black art: Some practitioners choose the k most com-

[1] One might have expected *a priori* that some difference would remain, because most highly inflected languages can also vary word order to mark a topic-focus distinction, and this (occasional) marking is preserved in our experiment.

Proceedings of NAACL-HLT 2018, pages 536–541
New Orleans, Louisiana, June 1 - 6, 2018.

mon words (e.g., Mikolov et al. (2010) choose $k = 10000$) and others use all those words that appear at least twice in the training corpus. In general, replacing more words with UNK artificially improves the perplexity measure but produces a less useful model. OOVs present something of a challenge for the cross-linguistic comparison of language models, especially in morphologically rich languages, which simply have more word forms.

2.1 The Role of Inflectional Morphology

Inflectional morphology can explode the base vocabulary of a language. Compare, for instance, English and Turkish. The nominal inflectional system of English distinguishes two forms: a singular and plural. The English lexeme BOOK has the singular form *book* and the plural form *books*. In contrast, Turkish distinguishes at least 12: *kitap*, *kitablar*, *kitabı*, *kitabın*, etc.

To compare the degree of morphological inflection in our evalation languages, we use **counting complexity** (Sagot, 2013). This crude metric counts the number of inflectional categories distinguished by a language (e.g., English includes a category of 3rd-person singular present-tense verbs). We count the categories annotated in the language's UniMorph (Kirov et al., 2018) lexicon. See Table 1 for the counting complexity of evaluated languages.

2.2 Open-Vocabulary Language Models

To ensure comparability across languages, we require our language models to predict every character in an utterance, rather than skipping some characters because they appear in words that were (arbitrarily) designated as OOV in that language. Such models are known as "open-vocabulary" LMs.

Notation. Let $\cup$ denote disjoint union, i.e., $A \cup B = C$ iff $A \cup B = C$ and $A \cap B = \emptyset$. Let Σ be a discrete alphabet of characters, including a distinguished unknown-character symbol $\star$.[2] A character LM then defines $p(\boldsymbol{c}) = \prod_{i=1}^{|\boldsymbol{c}|+1} p(c_i \mid \boldsymbol{c}_{<i})$, where we take $c_{|\boldsymbol{c}|+1}$ to be a distinguished end-of-string symbol EOS. In this work, we consider two open-vocabulary LMs, as follows.

Baseline n-gram LM. We train "flat" hybrid word/character open-vocabulary n-gram models (Bisani and Ney, 2005), defined over strings Σ^+ from a vocabulary Σ with mutually disjoint subsets: $\Sigma = W \cup C \cup S$, where single characters $c \in C$ are distinguished in the model from single character full words $w \in W$, e.g., a versus the word a. Special symbols $S = \{\underline{\text{EOW}}, \text{EOS}\}$ are end-of-word and end-of-string, respectively. N-gram histories in H are either word-boundary or word-internal (corresponding to a whitespace tokenization), i.e., $H = H_b \cup H_i$. String-internal word boundaries are always separated by a single whitespace character.[3] For example, if foo, baz $\in W$ but bar $\notin W$, then the string foo bar baz would be generated as: foo b a r EOW baz EOS. Possible 3-gram histories in this string would be, e.g., [foo b] $\in H_i$, [r EOW] $\in H_b$, and [EOW baz] $\in H_b$.

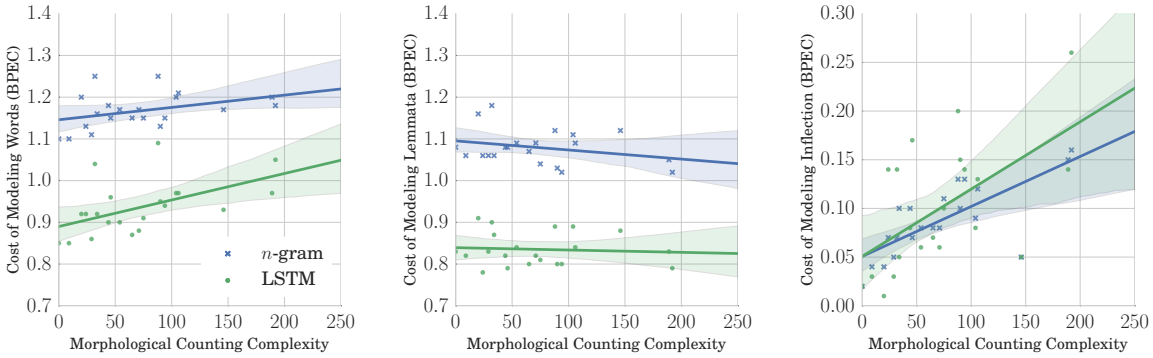
Symbols are generated from a multinomial given the history h, leading to a new history h' that now includes the symbol and is truncated to the Markov order. Histories $h \in H_b$ can generate symbols $s \in W \cup C \cup \{\text{EOS}\}$. If $s = \text{EOS}$, the string is ended. If $s \in W$, it has an implicit EOW and the model transitions to history $h' \in H_b$. If $s \in C$, it translitions to $h' \in H_i$. Histories $h \in H_i$ can generate symbols $s \in C \cup \{\underline{\text{EOW}}\}$ and transition to $h' \in H_b$ if $s = \underline{\text{EOW}}$, otherwise to $h' \in H_i$.

We use standard Kneser and Ney (1995) model training, with distributions at word-internal histories $h \in H_i$ constrained so as to only provide probability mass for symbols $s \in C \cup \{\underline{\text{EOW}}\}$. We train 7-gram models, but prune n-grams hs where the history $h \in W^k$, for $k > 4$, i.e., 6- and 7-gram histories must include at least one $s \notin W$. To establish the vocabularies W and C, we replace exactly one instance of each word type with its spelled out version. Singleton words are thus excluded from W, and character sequence observations from all types are included in training. Note any word $w \in W$ can also be generated as a character sequence. For perplexity calculation, we sum the probabilities for each way of generating the word.

LSTM LM. While neural language models can also take a hybrid approach (Hwang and Sung, 2017; Kawakami et al., 2017), recent advances indicate that full character-level modeling is now competitive with word-level modeling. A large part of this is due to the use of recurrent neural networks (Mikolov et al., 2010), which can generalize about

[2] The set of graphemes in these languages can be assumed to be closed, but external graphemes may on rare occasion appear in random text samples. These are rare enough to not materially affect the metrics.

[3] The model can be extended to handle consecutive whitespace characters or punctuation at word boundaries; for this paper, the tokenization split punctuation from words and reduced consecutive whitespaces to one, hence the simpler model.



(a) BPEC performance of n-gram (blue) and LSTM (green) LMs over word sequences. Lower is better.

(b) BPEC performance of n-gram (blue) and LSTM (green) LMs over lemma sequences. Lower is better.

(c) Difference in BPEC performance of n-gram (blue) and LSTM (green) LMs between words and lemmata.

Figure 1: The primary findings of our paper are evinced in these plots. Each point is a language. While the LSTM outperforms the hybrid n-gram model, the relative performance on the highly inflected languages compared to the more modestly inflected languages is almost constant; to see this point, note that the regression lines in Fig. 1c are almost identical. Also, comparing Fig. 1a and Fig. 1b shows that the correlation between LM performance and morphological richness disappears after lemmatization of the corpus, indicating that inflectional morphology is the origin for the lower BPEC.

how the distribution $p(c_i \mid \boldsymbol{c}_{<i})$ depends on $\boldsymbol{c}_{<i}$.

We use a long short-term memory (LSTM) LM (Sundermeyer et al., 2012), identical to that of Zaremba et al. (2014), but at the character-level. To achieve the hidden state $\mathbf{h}_i \in \mathbb{R}^d$ at time step i, one feeds the left context $\boldsymbol{c}_{i-1}$ to the LSTM: $\mathbf{h}_i = \text{LSTM}(c_1, \ldots, c_{i-1})$ where the model uses a learned vector to represent each character type. This involves a recursive procedure described in Hochreiter and Schmidhuber (1997). Then, the probability distribution over the i^{th} character is $p(c_t \mid \boldsymbol{c}_{<i}) = \text{softmax}(\mathbf{W}\,\mathbf{h}_i + \mathbf{b})$, where $\mathbf{W} \in \mathbb{R}^{|\Sigma| \times d}$ and $\mathbf{b} \in \mathbb{R}^{|\Sigma|}$ are parameters.

Parameters for all models are estimated on the training portion and model selection is performed on the development portion. The neural models are trained with SGD (Robbins and Monro, 1951) with gradient clipping, such that each component has a maximum absolute value of 5. We optimize for 100 iterations and perform early stopping (on the development portion). We employ a character embedding of size 1024 and 2 hidden layers of size 1024.[4] The implementation is in PyTorch.

3 A Fairer Evaluation: Multi-Text

Effecting a cross-linguistic study on LMs is complicated because different models could be trained and tested on incomparable corpora. To avoid this problem, we use **multi-text**: k-way translations of the same semantic content.

What's wrong with bits per character? Open-vocabulary language modeling is most commonly evaluated under **bits per character** (BPC) $= \frac{1}{|\boldsymbol{c}|+1} \sum_{i=1}^{|\boldsymbol{c}|+1} \log p(c_i \mid \boldsymbol{c}_{<i})$.[5] Even with multi-text, comparing BPC is not straightforward, as it relies on the vagaries of individual writing systems. Consider, for example, the difference in how Czech and German express the phoneme /tʃ/: Czech uses *č*, whereas German *tsch*. Now, consider the Czech word *puč* and its German equivalent *Putsch*. Even if these words are both predicted with the *same* probability in a given context, German will end up with a lower BPC.[6]

Bits per English Character. Multi-text allows us to compute a fair metric that is invariant to the orthographic (or phonological) changes discussed above: **bits per English character** (BPEC). BPEC $= \frac{1}{|\boldsymbol{c}_{English}|+1} \sum_{i=1}^{|\boldsymbol{c}|+1} \log p(c_i \mid \boldsymbol{c}_{<i})$, where $\boldsymbol{c}_{English}$ is the English character sequence in the utterance aligned to $\boldsymbol{c}$. The choice of English is arbitrary, as any other choice of language would simply scale the values by a constant factor.

Note that this metric is essentially capturing the overall *bits per utterance*, and that normalizing using English characters only makes numbers independent of the overall utterance length; it is not critical to the analysis we perform in this paper.

[4] As Zaremba et al. (2014) indicate, increasing the number of parameters may allow us to achieve better performance.

[5] To aggregate this over an entire test corpus, we replace the denominator and also the numerator by summations over all utterances $\boldsymbol{c}$.

[6] Why not work with *phonological* characters, rather than orthographic ones, obtaining */putʃ/* for both Czech and German? Sadly this option is also fraught with problems as many languages have perfectly predictable phonological elements that will artificially lower the score.

			BPEC / ΔBPC (·e-2)			
	data (M)		hybrid n-gram		LSTM	
lang	wds / ch	MCC	form	lemma	form	lemma
bg	0.71/4.3	96	1.13/ 4	1.03/ 1	0.95/ 3	0.80/ 1
cs	0.65/3.9	195	1.20/ -8	1.05/-12	0.97/ -6	0.83/ -9
da	0.70/4.1	15	1.10/ -1	1.06/ -4	0.85/ -1	0.82/ -3
de	0.74/4.8	38	1.25/ 17	1.18/ 13	1.04/ 14	0.90/ 10
el	0.75/4.6	50	1.18/ 13	1.08/ 5	0.90/ 10	0.82/ 4
en	0.75/4.1	6	1.10/ 0	1.08/ -3	0.85/ 0	0.83/ -3
es	0.81/4.6	71	1.15/ 12	1.07/ 7	0.87/ 9	0.80/ 5
et*	0.55/3.9	110	1.20/ -8	1.11/-15	0.97/ -6	0.89/-12
fi*	0.52/4.2	198	1.18/ 2	1.02/-11	1.05/ 1	0.79/ -9
fr	0.88/4.9	30	1.13/ 17	1.06/ 13	0.92/ 14	0.78/ 10
hu*	0.63/4.3	94	1.25/ 5	1.12/ -9	1.09/ 5	0.89/ -7
it	0.85/4.8	52	1.15/ 16	1.08/ 14	0.96/ 14	0.79/ 10
lt	0.59/3.9	152	1.17/ -6	1.12/ -7	0.93/ -5	0.88/ -6
lv	0.61/3.9	81	1.15/ -6	1.04/ -9	0.91/ -5	0.81/ -7
nl	0.75/4.5	26	1.20/ 11	1.16/ 4	0.92/ 8	0.91/ 4
pl	0.65/4.3	112	1.21/ 6	1.09/ -1	0.97/ 5	0.84/ -1
pt	0.89/4.8	77	1.17/ 16	1.09/ 9	0.88/ 12	0.82/ 7
ro	0.74/4.4	60	1.17/ 8	1.09/ 0	0.90/ 6	0.84/ 0
sk	0.64/3.9	40	1.16/ -6	1.06/-11	0.92/ -5	0.87/ -9
sl	0.64/3.8	100	1.15/-10	1.02/-10	0.90/ -8	0.80/ -7
sv	0.66/4.1	35	1.11/ -2	1.06/ -8	0.86/ -2	0.83/ -7

Table 1: Results for all configurations and the typological profile of the 21 Europarl languages. All languages are Indo-European, except for those marked with * which are Uralic. Morpholical counting complexity (MCC) is given for each language, along with bits per English character (BPEC) and the ΔBPC, which is BPEC minus bits per character (BPC). This is blue if BPEC > BPC and red if BPEC < BPC.

A Potential Confound: Translationese. Working with multi-text, however, does introduce a new bias: all of the utterances in the corpus have a source language and 20 translations of that source utterance into target languages. The characteristics of translated language has been widely studied and exploited, with one prominent characteristic of translations being simplification (Baker, 1993).

Note that a significant fraction of the original utterances in the corpus are English. Our analysis may then have underestimated the BPEC for other languages, to the extent that their sentences consist of simplified "translationese." Even so, English had the lowest BPEC from among the set of languages.

4 Experiments and Results

Our experiments are conducted on the 21 languages of the Europarl corpus (Koehn, 2005). The corpus consists of utterances made in the European parliament and are aligned cross-linguistically by a unique utterance id. With the exceptions (noted in Table 1) of Finnish, Hungarian and Estonian, which are Uralic, the languages are Indo-European.

While Europarl does not contain quite our desired breadth of typological diversity, it serves our purpose by providing large collections of aligned data across many languages. To create our experimental data, we extract all utterances and randomly sort them into train-development-test splits such that roughly 80% of the data are in train and 10% in development and test, respectively.[7] We also perform experiments on *lemmatized* text, where we replace every word with its lemma using the UDPipe toolkit (Straka et al., 2016), stripping away its inflectional morphology. We report two evaluation metrics: BPC and BPEC (see §3). Our BPEC measure always normalizes by the length of the original, not lemmatized, English.

Experimentally, we want to show: (i) When evaluating models in a controlled environment (multi-text under BPEC), the models achieve lower performance on certain languages and (ii) inflectional morphology is the primary culprit for the performance differences. However, we repeat that we do not in this paper tease apart whether the models are at fault, or that certain languages inherently encode more information.

5 Discussion and Analysis

We display the performance of the n-gram LM and the LSTM LM under BPC and BPEC for each of the 21 languages in Fig. 1 with full numbers listed in Table 1. There are several main take-aways.

The Effect of BPEC. The first major take-away is that BPEC offers a cleaner cross-linguistic comparison than BPC. Were we to rank the languages by BPC (lowest to highest), we would find that English was in the middle of the pack, which is surprising as new language models are often only tuned on English itself. For example, BPC surprisingly suggests that French is easier to model than English. However, ranking under BPEC shows that the LSTM has the easiest time modeling English itself. Scandinavian languages Danish and Swedish have BPEC closest to English; these languages are typologically and genetically similar to English.

n-gram versus LSTM. As expected, the LSTM outperforms the baseline n-gram models across the board. In addition, however, n-gram modeling yields relatively poor performance on some languages, such as Dutch, with only modestly more complex inflectional morphology than English.

[7] Characters appearing < 100 times in train are ⋆.

Other phenomena—e.g., perhaps, compounding—may also be poorly modeled by n-grams.

The Impact of Inflectional Morphology. Another major take-away is that rich inflectional morphology is a difficulty for both n-gram and LSTM LMs. In this section we give numbers for the LSTMs. Studying Fig. 1a, we find that Spearman's rank correlation between a language's BPEC and its counting complexity (§2.1) is quite high ($\rho = 0.59$, significant at $p < 0.005$). This clear correlation between the level of inflectional morphology and the LSTM performance indicates that character-level models do not automatically fix the problem of morphological richness. If we lemmatize the words, however (Fig. 1b), the correlation becomes insignificant and in fact slightly negative ($\rho = -0.13$, $p \approx 0.56$). The difference of the two previous graphs (Fig. 1c) shows more clearly that the LM penalty for modeling inflectional endings is greater for languages with higher counting complexity. Indeed, this penalty is arguably a more appropriate measure of the complexity of the inflectional system. See also Fig. 2.

The differences in BPEC among languages are reduced when we lemmatize, with standard deviation dropping from 0.065 bits to 0.039 bits. Zooming in on Finnish (see Table 1), we see that Finnish forms are harder to model than English forms, but Finnish lemmata are *easier* to model than English ones. This is strong evidence that it was primarily the inflectional morphology, which lemmatization strips, that caused the differences in the model's performance on these two languages.

6 Related Work

Recurrent neural language models can effectively learn complex dependencies, even in open-vocabulary settings (Hwang and Sung, 2017; Kawakami et al., 2017). Whether the models are able to learn particular syntactic interactions is an intriguing question, and some methodologies have been presented to tease apart under what circumstances variously-trained models encode attested interactions (Linzen et al., 2016; Enguehard et al., 2017). While the sort of detailed, construction-specific analyses in these papers is surely informative, our evaluation is language-wide.

MT researchers have investigated whether an English sentence contains enough information to predict the fine-grained inflections used in its foreign-language translations (see Kirov et al., 2017).

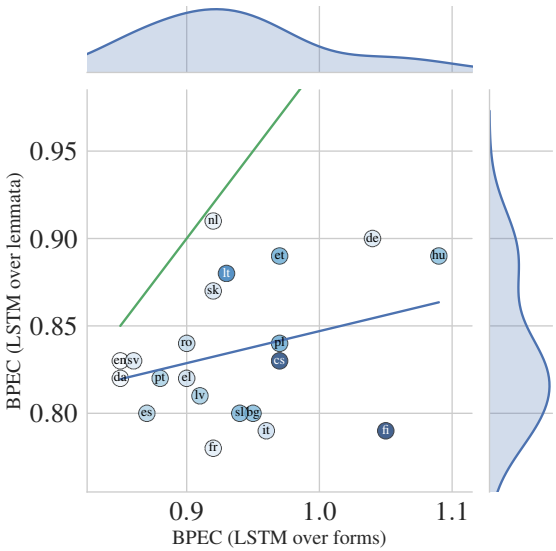


Figure 2: Each dot is a language, and its coordinates are the BPEC values for the LSTM LMs over words and lemmata. The top and right margins show kernel density estimates of these two sets of BPEC values. All dots follow the blue regression, but stay below the green line ($y = x$), and the darker dots—which represent languages with higher counting complexity—tend to fall toward the right but not toward the top, since counting complexity is correlated only with the BPEC over words.

Sproat et al. (2014) present a corpus of close translations of sentences in typologically diverse languages along with detailed morphosyntactic and morphosemantic annotations, as the means for assessing linguistic complexity for comparable messages, though they expressly do not take an information-theoretic approach to measuring complexity. In the linguistics literature, McWhorter (2001) argues that certain languages are less complex than others: he claims that Creoles are simpler. Müller et al. (2012) compare LMs on EuroParl, but do not compare performance across languages.

7 Conclusion

We have presented a clean method for the cross-linguistic comparison of language modeling: We assess whether a language modeling technique can compress a sentence and its translations equally well. We show an interesting correlation between the morphological richness of a language and the performance of the model. In an attempt to explain causation, we also run our models on lemmatized versions of the corpora, showing that, upon the removal of inflection, no such correlation between morphological richness and LM performance exists. It is still unclear, however, whether the performance difference originates from the inherent difficulty of the languages or with the models.

References

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and Technology: In Honour of John Sinclair* 233:250.

Maximilian Bisani and Hermann Ney. 2005. Open vocabulary speech recognition with flat hybrid models. In *INTERSPEECH*. pages 725–728.

Émile Enguehard, Yoav Goldberg, and Tal Linzen. 2017. Exploring the syntactic abilities of RNNs with multi-task learning. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, pages 3–14. https://doi.org/10.18653/v1/K17-1003.

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

Kyuyeon Hwang and Wonyong Sung. 2017. Character-level language modeling with hierarchical recurrent neural networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 5720–5724.

Kazuya Kawakami, Chris Dyer, and Phil Blunsom. 2017. Learning to create and reuse words in open-vocabulary neural language modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Vancouver, Canada, pages 1492–1502. http://aclweb.org/anthology/P17-1137.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Arya McCarthy, Sebastian J. Mielke, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. Unimorph 2.0: Universal morphology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Christo Kirov, John Sylak-Glassman, Rebecca Knowles, Ryan Cotterell, and Matt Post. 2017. A rich morphological tagger for English: Exploring the cross-linguistic tradeoff between morphology and syntax. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, pages 112–117. http://aclweb.org/anthology/E17-2018.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. pages 181–184.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*. volume 5, pages 79–86.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association of Computational Linguistics* 4:521–535. http://www.aclweb.org/anthology/Q16-1037.

John McWhorter. 2001. The worlds simplest grammars are creole grammars. *Linguistic Typology* 5(2):125–66.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*. pages 1045–1048. http://www.isca-speech.org/archive/interspeech_2010/i10_1045.html.

Thomas Müller, Hinrich Schütze, and Helmut Schmid. 2012. A comparative investigation of morphological language modeling for the languages of the European Union. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, pages 386–395. http://www.aclweb.org/anthology/N12-1043.

Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics* pages 400–407.

Benoît Sagot. 2013. Comparing complexity measures. In *Computational Approaches to Morphological Complexity*.

Richard Sproat, Bruno Cartoni, HyunJeong Choe, David Huynh, Linne Ha, Ravindran Rajakumar, and Evelyn Wenzel-Grondie. 2014. A database for measuring linguistic information content. In *LREC*.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *LREC*.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *CoRR* abs/1409.2329. http://arxiv.org/abs/1409.2329.