# Augmenting a BiLSTM Tagger with a Morphological Lexicon and a Lexical Category Identification Step

**Steinþór Steingrímsson**[1,2] **Örvar Kárason**[1] **Hrafn Loftsson**[1]
[1]Department of Computer Science, Reykjavik University, Iceland
[2]The Árni Magnússon Institute for Icelandic Studies, Reykjavik, Iceland
{steinthor18,orvark13,hrafn}@ru.is

## Abstract

Previous work on using BiLSTM models for PoS tagging has primarily focused on small tagsets. We evaluate BiLSTM models for tagging Icelandic, a morphologically rich language, using a relatively large tagset. Our baseline BiLSTM model achieves higher accuracy than any previously published tagger not taking advantage of a morphological lexicon. When we extend the model by incorporating such data, we outperform previous state-of-the-art results by a significant margin. We also report on work in progress that attempts to address the problem of data sparsity inherent in morphologically detailed, fine-grained tagsets. We experiment with training a separate model on only the lexical category and using the coarse-grained output tag as an input for the main model. This method further increases the accuracy and reduces the tagging errors by 21.3% compared to previous state-of-the-art results. Finally, we train and test our tagger on a new gold standard for Icelandic.

## 1 Introduction

Bidirectional long short-term memory (BiLSTM) models have in recent years been shown to be effective for various sequential labelling tasks, including Part-of-Speech (PoS) tagging (Ling et al., 2015; Plank et al., 2016).

BiLSTMs are an extension of general LSTMs (Hochreiter and Schmidhuber, 1997) that perform better on sequences where the complete input sequence is available. Two LSTMs are trained on the input sequence, one on its natural reading order and the other on its reverse (Graves and Schmidhuber, 2005). In addition to word embeddings (WE), character embeddings were first used for tagging with BiLSTMs by Dos Santos and Zadrozny (2014). This entails not only examining the sequence of words in a sentence during training but also the sequences of characters within those words.

In this paper we use BiLSTM models, with both word and character embeddings, to train a PoS tagger for a morphologically rich language, Icelandic, with a fine-grained tagset of 565 morphosyntactic tags. Only a small portion of previous work using neural networks for PoS tagging has focused on languages with rich morphology and large tagsets, e.g. Sagot and Martínez Alonso (2017).

Various taggers have been developed for Icelandic: data-driven taggers (Helgadóttir, 2005), a rule-based tagger (IceTagger) (Loftsson, 2008), and a hybrid tagger (Loftsson et al., 2009). Prior to the work presented here, an averaged perceptron tagger, IceStagger (Loftsson and Östling, 2013), was the current state-of-the-art tagger, achieving an accuracy of 93.84% by employing a morphological lexicon and external word embeddings.

This paper presents the first deep neural network tagger for Icelandic. We evaluate three models. First, we confirm the effectiveness of a BiLSTM model for PoS tagging using a fine-grained tagset. Second, we supplement the base model with an external morphological lexicon, thereby obtaining state-of-the-art results. Third, we propose an approach to further increase the accuracy by creating a coarse-grained tagset from the fine-grained one and using the resulting tagset to devise a two-step process. This approach is to our best knowledge novel in the context of neural network tagging. Specifically, we train a separate model on only the lexical category and use the coarse-grained output tag as an input into the main model. Combined, this results in an overall tagging accuracy of 95.15%, which is equivalent to


*Proceedings of Recent Advances in Natural Language Processing*, pages 1161–1168,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_133

an error reduction of 21.3% compared to the previous state-of-the-art. Finally, we train and test our model on a new gold standard for Icelandic, MIM-GOLD. The new standard is larger than the older one, IFD (see Section 2), and contains more diverse texts. We achieve an accuracy of 94.17% on MIM-GOLD.

## 2 Data

In this section, we describe the data and the tagset used in our work.

**The IFD Corpus:** The taggers developed for Icelandic so far have all been trained and tested on the Icelandic Frequency Dictionary (IFD) corpus (Pind et al., 1991), a balanced corpus containing about 590 thousand tokens. The IFD corpus was collected in the early 1990s and contains texts from published books, primarily fiction (60%) but also biographies (20%) and scholarly work (20%). As with the other taggers referenced in this paper, we use the so-called *corrected version* of the corpus, with the reduced tagset (565 tags) and ten-fold split from Loftsson et al. (2009).[1] The morphosyntactic tags in this tagset are mnemonic encodings, i.e. character strings where each character has a particular function. The first character denotes the *lexical category*. For each category there is a predefined number of additional characters (at most six), which describe morphological features, like *gender*, *number* and *case* for nouns; *degree* and *declension* for adjectives; *voice*, *mood* and *tense* for verbs, etc. To illustrate, consider the word form *maður* "man". The corresponding tag is *nken*, denoting noun (*n*), masculine (*k*), singular (*e*), and nominative (*n*) case.

**The MIM-GOLD Corpus:** MIM-GOLD[2] (Loftsson et al., 2010), a subset of the MIM corpus (Helgadóttir et al., 2012), contains a greater diversity of texts than the IFD corpus. In addition to texts from published books, it contains texts from news media, blogs, parliamentary speeches and more. Furthermore, MIM-GOLD contains approximately 1 million running words, about twice as many as IFD. The tagset used in MIM-GOLD consists of the same reduced tagset of 565 tags, mentioned above.

**Morphological Lexicon:** The Database of Modern Icelandic Inflections (DMII) is a lexicon of about 280 thousand paradigms and close to six million inflectional forms (Bjarnadóttir, 2012). The output from the database used in this project contains word form and morphological features. By incorporating DMII, the average unknown word rate in testing, using the IFD ten-fold split, drops from 6.8% to 1.1% (Loftsson et al., 2011).

[1] IFD can be downloaded from http://malfong.is/?pg=ordtidnibok.

[2] MIM-GOLD can be downloaded from http://malfong.is/?pg=gull.

## 3 The Three Models

### 3.1 Word and Character Embeddings

Word embeddings are vector representations of words based on their context in training data. Adding recurrent character embeddings has been shown to significantly improve performance for handling of unknown words (e.g. Plank et al. 2016; Dos Santos and Zadrozny 2014). For each word, both forward and backward expressions are generated, containing the sequence of characters in the word, as well as word initial and word final markers. This helps the model grasp morphological details.

In our baseline model, which is similar to Plank et al. (2016), both word embeddings and recurrent character embeddings are used as input. The character embeddings for a given word are input into a BiLSTM. The output from the BiLSTM is concatenated to the word embedding and the combined vector input into another BiLSTM, whose output is input into a hidden layer. The hidden layer feeds the output layer, which selects a PoS tag.

### 3.2 Using Data from an External Morphological Lexicon

Horsmann and Zesch (2017) replicated Plank et al. (2016) using a collection of corpora annotated with fine-grained tagsets of varying sizes, in contrast to the coarse-grained Universal Dependencies (UD) tagset in the previous study (17 tags). The replication confirmed the superior performance of the BiLSTM tagger, also on fine-grained tagsets. Furthermore, they found that the advantages of the BiLSTM tagger over other taggers grow proportionally with the tagset size of the corpus. However, they also claim that for large tagsets of morphologically rich languages, hand-crafted morphological lexicons are still necessary to reach state-of-the-art performance.

Using a morphological lexicon has become common practice for enriching training data for PoS taggers. Hajič (2000) marked the importance

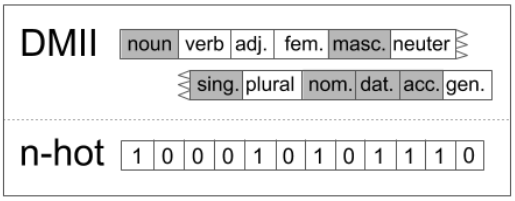


Figure 1: A partial n-hot vector and the corresponding features from DMII. The example shows 12 features, including the active features for the word form *strætó* "bus". The word, a noun, has the same form for nominative, dative and accusative and therefore all corresponding labels are activated. An actual vector in our model has 61 labels, which are either active, `1`, or inactive, `0`.

of this for morphologically rich languages. It was first done for Icelandic in Loftsson et al. (2011).

Sagot and Martínez Alonso (2017) first used morphological lexicons as supplemental input for PoS tagging with BiLSTM taggers and showed that it yields consistent improvement. Following their work, we extend the baseline model by adding an input layer that contains token-wise features obtained from the DMII lexicon (see Section 2). The input vector for a given word is an n-hot vector where each active value corresponds to one of 61 possible labels in the lexicon. An example of an n-hot vector is given in Figure 1.

The vector is concatenated to the two vectors described in the previous section, i.e. the word embedding and the character embedding, and the result is then fed into the BiLSTM layer. Previous taggers using DMII have had to map the information to the IFD tagset. As the tagsets of IFD and DMII are not completely compatible some information has been lost in the mapping process. Our method allows the model to use and learn from all the information encoded in the morphological lexicon, even though it uses a tagset slightly different from our training data.

### 3.3 Stepwise Tagging Model

When employing a fine-grained tagset with mnemonic encoding, the model does not place different significance on two tags when they differ in lexical category, on one hand, or share a lexical category but differ in morphological features, on the other. A human, however, would consider the former a more significant error than the latter. A PoS tagger is especially prone to such errors when the tagset is large and the amount of training data is insufficient to detect all the subtle differences between labels, as sometimes is the case for under-resourced or domain-specific languages.

To place a higher emphasis on assigning the correct lexical category, we devise a two-step process. First, we simplify the tagset from 565 to 10 tags by using only the first letter of the fine-grained tag mnemonic, i.e. the letter denoting the lexical category. We then train our model on this new coarse-grained tagset, using word and character embeddings as well as the morphological lexicon. This results in a lexical category tagger with very high accuracy, 98.97% in our case. In the second step, the output of that tagger is embedded as a one-hot vector and concatenated to the vectors input into the BiLSTM layer of the main model. This guides the tagger to the correct lexical category and eliminates some of the errors caused by insufficient training data.

This is a work in progress and other morphological features in the tags are promising for evolving this stepwise approach and further increasing overall accuracy. Thus, separate models for detecting gender, number and case agreement, for example, might be considered at each step.

We are not aware of other implementations of stepwise PoS tagging using BiLSTMs, but Horsmann and Zesch (2016) employ such a method in a slightly different setting. They use a Support Vector Machine for training, assume the coarse-grained tags are correct and then have their tagger assign the fine-grained tags based on them. In their Bidir tagger, Dredze and Wallenberg (2008) tag case separately in a second-pass, after running a general first-pass that uses the whole tagset. The second-pass tagger has access to the output of the first-pass, and is permitted to change its case and gender selections.

## 4 Experiments and Results

### 4.1 Experimental Setup

Our models were built using DyNet[3] (Neubig et al., 2017). We use the same hyperparameters for all models, SGD training with the initial learning rate of 0.13, which decays 5% in each epoch and runs for 30 epochs. The network has 128-dimensional embeddings for words and 20 for characters. The supplemental embeddings have 61

[3]The Dynamic Neural Network Toolkit, see `http://dynet.io`.

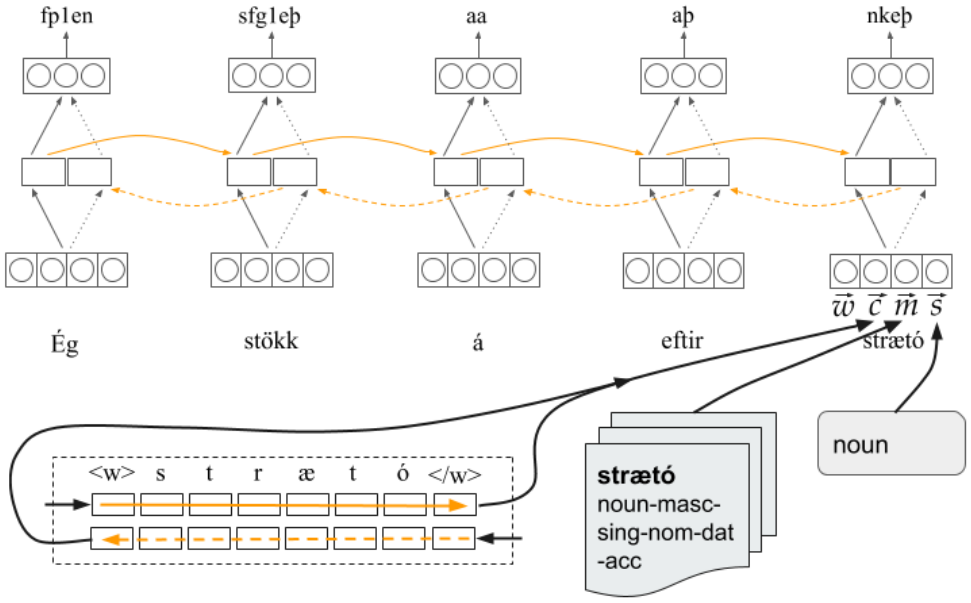


Figure 2: Our full model, employing word embeddings, character embeddings, a morphological lexicon, and the output of the first-pass of the stepwise model. The hidden layer is omitted for simplicity. Figure adapted from (Plank et al., 2016).

dimensions for the lexicon and 10 for the lexical categories. The hidden layer has 32 hidden states.[4]

Our experiments consist of three models:

**Baseline:** The first model uses word and character embeddings only. This corresponds to the model described in Section 3.1.

**DMII:** The second model adds external morphological data from DMII to the baseline model by encoding the information in n-hot vectors as described in Section 3.2.

**LC:** The third and full model then adds the lexical category embeddings created by a coarse-grained tagging step described in Section 3.3. This full model is shown in Figure 2.

### 4.2 Part-of-Speech Tagging Results

The test results for all three models are shown in Table 1, with the full model reaching 95.15% accuracy after 30 epochs. The baseline model (93.25%) already gets close to state-of-the-art results and surpasses existing taggers when not using an external morphological lexicon (cf. Loftsson and Östling 2013).

The substantial gain achieved by using DMII confirms the advantages of using an external morphological lexicon as discussed in Section 3.2. The accuracy gain is considerably higher than the corresponding gain in IceStagger (1.59 vs. 0.88 percentage points).

| | Acc. | Known | Unknown |
|---|---|---|---|
| Baseline | 93.25% | 95.19% | **66.84%** |
| + DMII | 94.84% | 95.17% | 54.61% |
| + LC | **95.15%** | **95.48%** | 54.06% |

Table 1: Accuracy of the three models trained and tested on IFD. Note that when DMII is employed the number of unknown words falls almost 90%, from 4,036 to 476 out of an average total of 58,977 words in the splits.

By employing the stepwise model discussed in Section 3.3 we try to guide the tagger to the highly accurate lexical category given by the coarse-grained tagger. This helps in assigning rare or ambiguous tags in the fine-grained tagset by raising the accuracy of the lexical category, resulting in a further 0.31 percentage point gain.

Note that the baseline model achieves the highest accuracy for unknown words because when adding data from DMII the unknown word ratio drops considerably (see Table 1), from 6.8%

[4]The source code for our implementation is available from https://github.com/steinst/ABLTagger

| | Acc. | Known | Unknown |
|---|---|---|---|
| TnT | 90.45% | 91.82% | 71.82% |
| IceTagger | 92.73% | 93.84% | **77.47%** |
| + DMII | 93.48% | 93.85% | 60.50% |
| IceStagger | 92.82% | 93.97% | 77.03% |
| + DMII | 93.70% | 94.02% | 61.45% |
| + DMII,WE | 93.84% | 94.15% | 61.99% |
| **Our model** | **95.15%** | **95.48%** | 54.06% |

Table 2: Comparison to other taggers for Icelandic.

to 0.8%. This is in line with results of previous taggers (see Section 2), further reduction in unknown words is due to us using the latest version of DMII, while previous results were published in 2011. When DMII is employed the remaining unknown words are more likely to be foreign words, typos or to be irregular in some other way and therefore more difficult to tag. This explains the drop in accuracy for unknown words.

### 4.3 Comparison to Other Taggers

A comparison of our model to other previously published taggers for Icelandic is shown in Table 2. The results for TnT, IceTagger and IceStagger are presented in (Loftsson et al., 2009; Loftsson, 2008; Loftsson and Östling, 2013), respectively. All the reported results are fully comparable as they are based on exactly the same cross-validation split of the IFD corpus, with the exception that the TnT tagger does not employ data from DMII, and has therefore a higher ratio of unknown words.

Our model outperforms all previous taggers by a substantial margin, equaling a 21.3% reduction in errors compared to the highest accuracy (93.84%) obtained by IceStagger. It also has the highest accuracy for known words, i.e. those seen in the training data, including DMII. It should be noted though, that the numbers for accuracy of known and unknown words are not very well comparable between the different models, as using DMII eliminates a substantial part of unknown words, but the ones that remain tend to be more irregular, and can thus be harder to tag correctly.

### 4.4 Error Analysis

When comparing the most frequent kinds of errors our tagger makes to the errors of IceStagger, two differences stand out. The frequency of

| | Our model | | IceStagger |
|---|---|---|---|
| No. | Proposed tag > gold tag | Error rate | Proposed tag > gold tag |
| 1. | aþ>ao | 3.28% | aþ>ao |
| 2. | ao>aþ | 2.99% | ao>aþ |
| 3. | nveo>nveþ | 1.80% | nveo>nveþ |
| 4. | nveþ>nveo | 1.72% | nveþ>nveo |
| 5. | ao>aa | 1.18% | **sng>sfg3fn** |
| 6. | aa>ao | 1.09% | ao>aa |
| 7. | nkeo>nkeþ | 0.98% | **sfg3eþ>sfg1eþ** |
| 8. | nheo>nhfo | 0.92% | aa>ao |
| 9. | nkeþ>nkeo | 0.82% | nheo>nhen |
| 10. | ct>c | 0.81% | nhen>nheo |

Table 3: Ten most frequent kinds of errors.

sng>sfg3fn and sfg3eþ>sfg1eþ are drastically reduced and are no longer among the ten most frequent kinds of errors (see Table 3). These are verbs that are assigned infinitive mood instead of indicative mood (sn...>sf...) and third person instead of first (sfg3...>sfg1...), respectively. These kinds of errors occur when the subject is far away from the verb itself and the more frequent tag for the word form is selected instead of the correct one. This corroborates that LSTMs are better at handling long-distance dependencies (Linzen et al., 2016) than other methods that have a limited context window during training.

The remaining kinds of errors in the top ten list are for the most part mistakes in case assignment. For example, prepositions are often wrongly marked as governing accusative instead of dative and vice versa (1 and 2) and there is often a confusion between prepositions and adverbs (5 and 6). The same goes for nouns (7 to 9) and, in addition, they are often assigned the wrong number, i.e. singular instead of plural (8). The last kind of error (10) is caused by a lack of syntactic and contextual information: a conjunction is marked as a relativizer, i.e. conjunction introducing a relative clause.

The nearly even distributions (1+2, 3+4, 5+6, 7+9) at which these kinds of errors occur indicate that there is nothing in the training data to discern which tag to select in these instances. One way forward to try to tackle these errors is to supplement the model further, e.g. with verb subcategorization frames.

| | Acc. | Known | Unknown |
|---|---|---|---|
| MIM-GOLD | 94.04% | 95.13% | 68.34% |
| + IFD | 94.17% | 95.62% | 68.18% |

Table 4: Accuracy when training and testing on MIM-GOLD.

## 5 Tagging a Different Gold Standard

In the previous sections, we have described the tagging process and compared the results to previous taggers using the same splits on IFD. We have demonstrated that our tagger achieves a significant gain in accuracy over previous taggers. Since the IFD corpus mainly contains literary work (see Section 2), these texts are not necessarily characteristic of texts that have to be tagged for language technology or research purposes. This is one of the reasons why a new gold standard, MIM-GOLD, was built containing more diverse texts (see Section 2). In 2015, Steingrímsson et al. (2015) trained IceStagger on MIM-GOLD, but found it had many inconsistencies and errors. Since then it has been reviewed and corrected and the final version, along with 10-fold splits, was made available in 2018.

We trained our BiLSTM tagger on these splits and measured the accuracy for our full model, employing both DMII and the two-step method. We carried out two experiments. In the first, we only trained and tested on the 10-fold splits for MIM-GOLD, but in the second we added the whole IFD corpus to the training data. As evident from Table 4, there is a substantial drop in accuracy compared to training and testing on IFD (see Table 1). The lower accuracy may, at least partly, be due to a greater variety in texts than before and a larger proportion of unknown words in the MIM-GOLD test set compared to IFD (Steingrímsson et al., 2015).

## 6 Conclusions and Future Work

We have shown that BiLSTM models with combined word and character embeddings achieve state-of-the-art accuracy in PoS tagging of Icelandic texts. We have also confirmed that BiLSTMs perform well with a fine-grained tagset, such as the one used in the Icelandic corpora, IFD and MIM-GOLD. When dealing with small corpora, as often is the case with under-resourced languages, supplementing the models with external data can be highly beneficial as shown by our experiments.

To deal with the problem of data sparsity, which is more prevalent when using fine-grained tagsets, we devised a stepwise method to guide the tagger in assigning lexical categories. This method is a work in progress – we have pinpointed morphological features that can be independently identified with very high accuracy and are therefore promising candidates for being handled in a separate step in the tagging process. Furthermore, it could be worthwhile to pre-train word embeddings on unlabeled data, such as the Icelandic Gigaword Corpus (IGC) of 1.2 billion words (Steingrímsson et al., 2018), e.g. employing the method described by Wang et al. (2015), which is specifically adapted to BiLSTMs.

The error analysis in Section 4.4 suggests that information on case governance is critical in reducing the most common errors the tagger makes. This could be external data on case governance of verbs and prepositions, or data derived from a method akin to the stepwise method that better discerns this information from the training data.

The final version of a new gold standard, MIM-GOLD, has recently been released and has not been used for training a PoS tagger for Icelandic before. IFD is heavily biased towards literary fiction but MIM-GOLD is a more balanced mix of different text genres and is thus more diverse. The lower accuracy for MIM-GOLD should thus not have been surprising, even though it has more data than IFD. Comparison of error analysis for both gold standards should reveal if there are other factors at play. We suggest that further work on developing PoS taggers for Icelandic texts focuses on this new gold standard.

## References

Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. In *LREC 2012 Proceedings: Proceedings of Language Technology for Normalization of Less-Resourced Languages, SaLTMiL 8 – AfLaT*. https://www.aflat.org/files/saltmil8-aflat2012.pdf#page=25.

Cícero Nogueira Dos Santos and Bianca Zadrozny. 2014. Learning Character-level Representations for Part-of-Speech Tagging. In *Proceedings of the* $31^{st}$ *International Conference on Machine Learning – Volume 32*. Beijing, China, ICML '14. http://dl.acm.org/citation.cfm?id=3044805.3045095.

Mark Dredze and Joel Wallenberg. 2008. Further results and analysis of Icelandic part of speech tagging. Technical report, Department of Computer and Information Science, University of Pennsylvania.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks* 18(5-6):602–610. https://doi.org/10.1016/j.neunet.2005.06.042.

Jan Hajič. 2000. Morphological tagging: Data vs. dictionaries. In $6^{th}$ *ANLP Conference* / $1^{st}$ *NAACL Meeting. Proceedings*. Seattle, Washington. https://dl.acm.org/citation.cfm?id=974318.

Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir, and Hrafn Loftsson. 2012. The Tagged Icelandic Corpus (MÍM). In *Proceedings of SaLTMiL-AfLaT Workshop on Language technology for normalisation of less-resourced languages*. Istanbul, Turkey, LREC 2012. https://www.aflat.org/files/saltmil8-aflat2012.pdf#page=79.

Sigrún Helgadóttir. 2005. Testing data-driven learning algorithms for PoS tagging of Icelandic. In H. Holmboe, editor, *Nordisk Sprogteknologi 2004*, Museum Tusculanums Forlag, Copenhagen.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Tobias Horsmann and Torsten Zesch. 2016. Assigning fine-grained PoS tags based on high-precision coarse-grained tagging. In *Proceedings of COLING 2016,* $26^{th}$ *International Conference on Computational Linguistics*. Osaka, Japan. http://aclweb.org/anthology/C/C16/C16-1032.pdf.

Tobias Horsmann and Torsten Zesch. 2017. Do LSTMs really work so well for PoS tagging? – a replication study. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark. https://doi.org/10.18653/v1/D17-1076.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal. https://doi.org/10.18653/v1/D15-1176.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics* 4:521–535. http://aclweb.org/anthology/Q16-1037.

Hrafn Loftsson. 2008. Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics* 31(1):47–72. https://doi.org/10.1017/S0332586508001820.

Hrafn Loftsson, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2011. Using a Morphological Database to Increase the Accuracy in POS Tagging. In *Recent Advances in Natural Language Processing*. Hissar, Bulgaria, RANLP 2011. http://www.aclweb.org/anthology/R11-1007.

Hrafn Loftsson, Ida Kramarczyk, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2009. Improving the PoS tagging accuracy of Icelandic text. In *Proceedings of the* $17^{th}$ *Nordic Conference on Computational Linguistics*. Odense, Denmark, NODALIDA 2009. https://dspace.ut.ee/handle/10062/9736.

Hrafn Loftsson and Robert Östling. 2013. Tagging a Morphologically Complex Language Using an Averaged Perceptron Tagger: The Case of Icelandic. In *Proceedings of the* $19^{th}$ *Nordic Conference on Computational Linguistics*. Oslo, Norway, NODALIDA 2013. https://aclanthology.info/papers/W13-5613/w13-5613.

Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. In Francis M. Tyers Sarasola, Kepa and Mikel L. Forcada, editors, *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*. Valetta, Malta, LREC 2010.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. DyNet: The dynamic neural network toolkit. *CoRR* abs/1701.03980. http://arxiv.org/abs/1701.03980.

Jörgen Pind, Friðrik Magnússon, and Stefán Briem. 1991. *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavik, Iceland.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of the* $54^{th}$ *Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany. https://doi.org/10.18653/v1/P16-2067.

Benoît Sagot and Héctor Martínez Alonso. 2017. Improving neural tagging with lexical information. In *Proceedings of the* $15^{th}$ *International Conference on Parsing Technologies*. Pisa, Italy. http://aclweb.org/anthology/W17-6304.

Steinþór Steingrímsson, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2015. Analysing Inconsistencies and Errors in PoS Tagging in two Icelandic Gold Standards. In *Proceedings of the* $20^{th}$ *Nordic Conference of Computational Linguistics*. Vilnius, Lithuania, NODALIDA 2015. https://www.aclweb.org/anthology/W15-1838.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. Miyazaki, Japan, LREC 2018. http://www.lrec-conf.org/proceedings/lrec2018/pdf/746.pdf.

Peilu Wang, Yao Qian, Frank K. Soong, Lei He, and Hai Zhao. 2015. Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network. *CoRR* abs/1510.06168. https://arxiv.org/pdf/1510.06168.pdf.