Reconciling historical data and modern computational models in corpus creation

Anonymous SCiL submission

1 Overview
We live in a time of unprecedented access to linguistic data, from audio recordings to corpora of billions of words. Linguists have used these resources to advance their research and understanding of language. Historical linguistics has lagged behind in this regard. However, this is due to several unique challenges that face the subfield. Historical linguistic data is plagued by two problems: a lack of overall data due to the ravages of time and a lack of model-ready data that have gone through standard NLP processing. Barring the discovery of more texts, the former issue cannot be solved; the latter can be, though it is time-consuming and resource-intensive. These problems have only begun to be addressed.

Modern models are not designed to work with historical data: they depend on large volumes of data and pre-tagged training sets that are not available for the majority of historical languages. This paper addresses this issue by treating historical data in the same way as a low-resource language (Fang and Cohn, 2017; Buys and Botha, 2016; Mishra et al., 2018) and integrating data from modern descendant languages. Through these approaches, we are able to POS tag a number of new texts in Old Slavic languages. With these problems overcome, we can create new corpora of historical language and thus dramatically increase both the number and type of diachronic linguistic investigations.

2 Modern approaches to historical data
Historical Data as low-resource language. This challenge is not unique to historical data. Thousands of languages across the world also lack the necessary resources for standard computational analyses and models. These low-resource languages have not been sufficiently documented and thus do not have adequate datasets for model-training. Many different approaches have been proposed on how to deal with this issue for low-resource languages, including the use of parallel corpora (Buys and Botha, 2016) and feature projection (Mishra et al., 2018). In this paper, we exploit the approach called Model Transfer (Fang and Cohn, 2017).

Extending modern language data. Historical data does not exist within a vacuum. One avenue that we could exploit is its relationship to descendant and related languages, i.e. how Modern English a descendant of Middle English. We might leverage the large amount of pre-processed data available for the modern languages to help create the models for their older stages. We call this Model Extension, where a model is created to tag one language using training data from a related language. In this paper, we train models on modern data and use them to tag the older texts. Thus the model is extending to a new linguistic domain.

3 Data
For this paper, we experiment on Old Slavic languages, focusing on Old Church Slavonic (OCS; 46 texts: 10 tagged, 36 untagged), Old East Slavic (OES; 35 texts: 32 tagged, 3 untagged), and Old Polish (OP; 20 untagged texts). These are good candidates because there are (1) resources for some of the languages (OCS and OES) and (2) well-documented modern descendant languages, i.e. Bulgarian for OCS, Russian for OES, and Polish for OP. Some pre-tagged texts for OCS and
OES were taken from the TOROT treebank (Eckhoff and Berdiceviskis) to be used as training and test data. Untagged texts in all three languages were taken from sites like Thesaurus Indogermanischer Text- und Sprachmaterialien. OCS was the only language for which an extensive dictionary could be found, thus it is the only language to use Model Transfer. Word-embeddings were trained for the languages using the gathered texts. Models for the modern language were trained using data taken from Universal Dependencies.

### 4 Models and results

In order to tag the new texts, we use a BiLSTM-CRF neural network (Huang et al., 2015; Reimers and Gurevych, 2017). Word-embeddings were trained using Word2Vec (Mikolov et al., 2013). Based on this architecture, we trained three types of models: (1) Normal Models using the pre-tagged data for OCS and OES, (2) Model Transfer for OCS using an OCS-English dictionary and the British National Corpus, and (3) Modern Model Extensions using Universal Dependency models for Bugarian, Russian, and Polish.

All models were subject to the same test set in each of the languages. The test set for Old Polish was hand-tagged for this project. This determined their POS tagging accuracies, which are compiled in Table 1. In general we can see that the use of Model Transfer and Model Extension does not negatively impact the POS tagging accuracy. The Extended model for OCS is lower than for the others, but this is likely due to dramatic morphological differences between OCS and Bugarian. While the overall accuracies are not as high as most modern language models, they are not so low as to be discouraging. They do show that, in the instance of a language like Old Polish, Model Transfer and Extension are serviceable methods for tagging new texts. Even at a 70% POS-tagging accuracy, these methods provide a great first-pass run in the pipeline of corpus creation for a language with no resources. Moreover, this maintenance of a comparably high accuracy shows that we can leverage different stages of a language to fill in gaps in our models. This is still likely dependent on other diachronic factors, e.g. we might expect a lower accuracy for an older morphologically-complex language when its descendant form is much more morphologically-simple.

| Language | Normal | Transfer | Extensions |
|----------|--------|----------|------------|
| OCS      | 75.63  | 76.54    | 65.23      |
| OES      | 69.60  | N/A      | 70.95      |
| Old Polish | N/A   | N/A      | 69.82      |

Table 1: Accuracies for test set tagging in each language across different models

### 5 Conclusions

This paper attempts to fill in a gap that continues to plague historical linguistics. The results are still lacking, but they show signs of improvement. With more time and resources, other methods could be explored, particularly those that depend on extensive pre-tagged data. Nevertheless, through efforts like these, we can improve the quality of data within historical linguistics, making it more approachable to all linguists and matching the standards already established in the rest of the field.

### References

Jan Buys and Jan A. Botha. 2016. Cross-lingual morphological tagging for low-resource languages. *CoRR*, abs/1606.04279.

Hanne Martin Eckhoff and Aleksanders Berdiceviskis. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. In *Scripta e-Scripta 14-15*, pages 9–25.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593, Vancouver, Canada. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Pruthvik Mishra, Vandan Mujadid, and Dipti Sharma. 2018. Pos tagging for resource poor indian languages through feature projection.

Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark.