Evaluating morphological typology in zero-shot cross-lingual transfer

Antonio Martínez-García  
Universitat de Barcelona  
ao.martinez.garcia@gmail.com

Toni Badia  
Universitat Pompeu Fabra  
tbadia@upf.edu

Jeremy Barnes  
University of Oslo  
jeremycb@ifi.uio.no

Abstract

Cross-lingual transfer has improved greatly through multi-lingual language model pretraining, reducing the need for parallel data and increasing absolute performance. However, this progress has also brought to light the differences in performance across languages. Specifically, certain language families and typologies seem to consistently perform worse in these models. In this paper, we address what effects morphological typology has on zero-shot cross-lingual transfer for two tasks: Part-of-speech tagging and sentiment analysis. We perform experiments on 19 languages from four language typologies (fusional, isolating, agglutinative, and introflexive) and find that transfer to another morphological type generally implies a higher loss than transfer to another language with the same morphological typology. Furthermore, POS tagging is more sensitive to morphological typology than sentiment analysis and, on this task, models perform much better on fusional languages than on the other typologies.

1 Introduction

Cross-lingual transfer uses available annotated resources in a source language to learn a model that will transfer to a target language. Earlier work used machine translation (Mihalcea et al., 2007), parallel data (Padó and Lapata, 2009), or delexicalized models (Zeman and Resnik, 2008; McDonald et al., 2011; Søgaard, 2011) to bridge the gap between languages. However, recent improvements (Devlin et al., 2019) have reduced the need for parallel data, instead relying on multi-lingual language models, trained on the concatenation of monolingual corpora. Fine-tuning these multilingual language models on a task in a source language can lead to strong performance when applied directly to the target-language task (zero-shot transfer).

This progress has uncovered gaps in performance, as transfer is generally easier between similar languages, and some language families consistently perform worse (Artetxe et al., 2020; Conneau et al., 2020a). So far, however, the analysis of these differences has only been anecdotal, rather than centered as a research question of its own merit. For these cases, linguistic typology has important implications, as it gives us ways to quantify the similarity of languages along certain variables, such as shared morphological or syntactic features (Bender, 2013). While previous work has studied the effects of morphological typology on language modeling (Gerz et al., 2018; Cotterell et al., 2018; Mielke et al., 2019), this effect on cross-lingual transfer has not been looked at in detail.

In this paper we attempt to answer (RQ1) to what degree morphological typology affects the performance of state-of-the-art cross-lingual models, (RQ2) whether morphological typology has a stronger effect than other variables, e.g., the amount of data for pretraining the LM or domain mismatches between source and target, (RQ3) whether there is a different effect on a low-level structural task (POS tagging) vs. a semantic task (sentiment analysis).

To answer these questions we experiment with two state-of-the-art cross-lingual models: multilingual BERT and XLM RoBERTa. We fine-tune the models for part-of-speech tagging and sentiment analysis on 19 languages from four morphologically diverse typologies. Our results show that POS tagging is more sensitive to morphological typology than sentiment analysis and that the models perform much better on fusional languages, such as German, than on the other typologies. We release the code and data1 in order to reproduce the experiments and facilitate future work in this area.

1Code and data available at https://github.com/jerbarnes/typology_of_crosslingual.
2 Related Work

Cross-lingual transfer has become ubiquitous in recent years, including cross-lingual POS tagging (Täckström et al., 2013; Huck et al., 2019) and cross-lingual sentiment analysis (Mihalcea et al., 2007; Balahur and Turchi, 2014; Barnes and Klinger, 2019). While earlier research focused on annotation projection (Yarowsky et al., 2001; Banea et al., 2008) or cross-lingual embeddings (Kim et al., 2017; Artetxe et al., 2017; Barnes et al., 2018b), multi-lingual pretraining currently leads to state-of-the-art results (Devlin et al., 2019; Lample and Conneau, 2019). These approaches rely on training transformer-based language models (Vaswani et al., 2017) on unlabeled data from multiple languages, while using careful data selection methods to avoid the over-representation of larger languages.

Although these approaches have led to large improvements on many cross-lingual tasks, it is clear that the success of zero-shot cross-lingual transfer depends on the typological similarity of the source and target language (Conneau et al., 2020b; Libovický et al., 2020). Pires et al. (2019) find POS performance correlates with word order features taken from the World Atlas of Language Structures (WALS) database (Dryer and Haspelmath, 2013). Similarly, morphologically complex languages tend to achieve poorer performance. In agglutinative languages, morphemes tend to be neatly segmentable and carry a single feature, whereas in fusional languages, a single morpheme often carries multiple grammatical, syntactic, and semantic features. Finally, in introflexive languages root words are based on consonant stems, where vowels introduced around and between them lead to syntactic and semantic changes (see Plank (1999); Bickel and Nichols (2005); Gerz et al. (2018) for a more in-depth discussion).

3 Data

We select five languages from each category except introflexive (four), shown in Table 1. A short example sentence in a fusional (Norwegian), isolating (Indonesian), agglutinative (Basque), and introflexive (Maltese) language with glosses and translation in English is shown in Example 1.

(1) 

| Language | Meaning | POS | Morphemes |
|----------|---------|-----|-----------|
| no       | autobus | ART | com:PERF |
| in       | ting   | ART | datang, tardlambat |
| eu       | bus    | ART | berandu, etorri, zen |
| mt       | bus    | ART | waslet, PRT.3S |

‘The bus came late.’

We use the following color combinations to denote isolating, agglutinative, introflexive, and fusional languages.

---

2 We use the following color combinations to denote isolating, agglutinative, introflexive, and fusional languages.
### Table 1: Number of examples for each task, language and dataset

| Type          | Language | Part-of-Speech | Sentiment Analysis |
|---------------|----------|----------------|--------------------|
|               |          | train | dev | test | train | dev | test |
| ● Fusional    | German   | 38,102 | 18,434 | 18,459 | 6,444 | 772 | 1,490 |
| ● Fusional    | Spanish  | 14,305 | 1,654 | 1,721 | 1,029 | 147 | 296 |
| ● Fusional    | Slovak   | 8,483 | 1,060 | 1,061 | 3,560 | 522 | 1,042 |
| ● Fusional    | Norwegian | 15,696 | 2,409 | 1,939 | 2,675 | 516 | 417 |
| ● Fusional    | Greek    | 1,662 | 403 | 456 | 5,936 | 383 | 767 |
| ✧ Isolating   | Mandarin | 3,997 | 500 | 500 | 12,348 | 2,591 | 4,896 |
| ✧ Isolating   | Vietnamese | 1,400 | 800 | 800 | 2,384 | 331 | 685 |
| ✧ Isolating   | Thai      | 0 | 0 | 1,000 | 8,103 | 1,153 | 2,344 |
| ✧ Isolating   | Cantonese | 0 | 0 | 1,004 | 28,204 | 4,459 | 8,915 |
| ✧ Isolating   | Indonesian | 4,477 | 559 | 557 | 7,926 | 1,132 | 2,266 |
| ✧ Agglutinative | Finnish  | 12,217 | 1,364 | 1,555 | 4,432 | 633 | 1,267 |
| ✧ Agglutinative | Basque   | 5,396 | 1,798 | 1,799 | 789 | 113 | 227 |
| ✧ Agglutinative | Korean   | 23,010 | 2,066 | 2,287 | 36,000 | 1,333 | 2,667 |
| ✧ Agglutinative | Japanese | 7,027 | 501 | 543 | 9,831 | 1,677 | 2,552 |
| ✧ Introspective | Arabic   | 6,075 | 909 | 680 | 2,468 | 353 | 706 |
| ✧ Introspective | Hebrew   | 5,241 | 484 | 491 | 6,621 | 1,184 | 2,305 |
| ✧ Introspective | Algerian | 997 | 136 | 143 | 564 | 75 | 92 |
| ✧ Introspective | Maltese  | 1,123 | 433 | 518 | 595 | 85 | 171 |

### 3.1 Part-of-speech

We obtain the data for the part-of-speech tagging task from the Universal Dependencies project (Ze- man et al., 2020), which currently gathers data annotated with universal POS tags for more than 90 languages, although there are differences in size and domain. For Algerian we use the annotations from Seddah et al. (2020). We found no training sets available for Thai and Cantonese, hence we use them for testing only. For more details on these datasets, see Table 5 in the Appendix.

### 3.2 Sentiment Analysis

For sentiment analysis, however, there is no centralized repository of similar data. Therefore, we collect data from a number of sources and process them to create binary (positive, negative) sentence-level sentiment datasets. For convenience, we list the origin of each dataset in Table 2 and their full characteristics in Table 6 in the Appendix.

### 4 Methods

We fine-tune both multilingual BERT (mBERT) (Xu et al., 2019) and XLM RoBERTa (XLM-R) (Conneau et al., 2020a) models on the available training data in each language, using a shared set of hyperparameters selected from recommended values according to the characteristics of our data. We set the learning rate to 2e-5, maximum sequence length of 256, batch size of 8 or 16, and perform early stopping once the validation score has not improved in the last epochs, saving the model that performs best on the dev set. We then test each model on all languages, giving us a matrix of test scores, where the diagonal is in-language, and all others are cross-lingual. We use accuracy as our metric for POS and macro F₁ for sentiment, as the latter often contains unbalanced classes, and define

---

1Including https://github.com/dimitrakatseli/review_sentiment_analysis
2https://github.com/1jw9609/SentimentAnalysis
3https://github.com/e9t/nsmc
4https://github.com/Darkmap/japanese_sentiment
5Including https://github.com/ozturkaslii/analyze-turkish-sentiment
6Depending on the size of the training set, model architecture and available GPU memory.
Table 2: Origin of the data for sentiment analysis.

| Language | Data Origin |
|----------|-------------|
| German   | Wojatzki et al. (2017) |
| Spanish  | Agerri et al. (2013) |
| Slovak   | Pecar et al. (2019) |
| Norwegian| Øvrelid et al. (2020) |
| Greek    | Kalamatianos et al. (2015) |
| Mandarin | Github repository |
| Vietnamese| Cuong et al. (2016) |
| Thai     | Xiang (2019) |
| Cantonese | Purwarianti and Crisdayanti (2019) |
| Indonesian| Github repository |
| Finnish  | Lindén et al. (2020) |
| Basque   | Barnes et al. (2018a) |
| Korean   | Github repository |
| Japanese | Github repository |
| Turkish  | Pontiki et al. (2016) |
| Arabic   | Abdulla et al. (2013) |
| Hebrew   | Amram et al. (2018) |
| Algerian | Touileb and Barnes (2021) |
| Maltese  | Dingli and Sant (2016) |

Table 2: Origin of the data for sentiment analysis.

5 Results

Once our scores matrix is built, we average\(^9\) the score of each fine-tuned model, which we refer to as language-to-language cross-lingual scores, over the other languages in each morphological group, thus obtaining each model’s average cross-lingual performance per target group (language-to-group cross-lingual scores). Next, we average again for each source language group. This yields the average cross-lingual performance values per training and testing language groups (group-to-group cross-lingual scores), which we report in Table 3.

In the part-of-speech task, the best group-to-group cross-lingual performance always corresponds to models fine-tuned in a language of the same morphological group, regardless of the model’s architecture. Fusional models, in particular, obtain a remarkably higher score when tested on other fusional languages (over 80%). On the other hand, the group-to-group cross-lingual scores where the target language is introflexive are considerably lower than the rest (always below 50%).

In contrast, both model architectures show different patterns in the sentiment analysis task. For the XLM-R models, the best group-to-group cross-lingual scores are all achieved by those trained on a fusional language, while for the mBERT it is mainly models trained on an isolating language that achieve the best scores. In any case, all scores are within a similar range of values. In fact, the main difference in this task seems to be due to XLM-R’s considerably higher scores.

In order to capture the cross-lingual phenomenon more accurately, we introduce transfer loss, a relative metric defined in Equation 1:

\[
TL_{x \rightarrow y} = S_{x \rightarrow x} - S_{x \rightarrow y} \tag{1}
\]

where \(TL_{x \rightarrow y}\) is the transfer loss experienced by a model fine-tuned in language \(x\) when transferring to language \(y\) (language-to-language transfer loss) and \(S_{x \rightarrow y}\) is the score achieved when testing a model fine-tuned in language \(x\) on language \(y\). Thus, it is a measure of the performance lost in the zero-shot transfer process: the better the transfer between both languages, the lower it will be.

We also define its averaged variants:

\[
\overline{TL}_{x \rightarrow A} = S_{x \rightarrow x} - \frac{1}{N_A} \sum_{i \in A, i \neq x} S_{x \rightarrow i} \tag{2}
\]

\[
\overline{TL}_{A \rightarrow B} = \frac{1}{N_A} \sum_{i \in A} \overline{TL}_{i \rightarrow B} \tag{3}
\]

where \(\overline{TL}_{x \rightarrow A}\) denotes the average transfer loss from language \(x\) to languages belonging to morphological type \(A\) (language-to-group transfer loss), \(\overline{TL}_{A \rightarrow B}\) refers to the average transfer loss experienced by languages from morphological group \(A\) to languages from group \(B\) (group-to-group transfer loss) and \(N_A\) is the number of languages (other than \(x\)) included in the experiment that belong to group \(A\). Table 4 shows the resulting group-to-group transfer loss values for each task.

\(^9\)Note that, throughout this paper, when we average across morphological groups, we do so with a weighted average so that all groups are equally represented regardless of how many languages they include.

\(^{10}\)The score metric will depend on the task: accuracy in POS and macro F\(_1\) in sentiment analysis.
Models fine-tuned in all groups except agglutinative experience the lowest performance drop when transferring to fusional languages in the part-of-speech task, whereas in the sentiment analysis task there is no clear pattern. It is also worth noting that the XLM-R models tend to transfer better compared to mBERT, only slightly in part-of-speech tagging but more drastically in sentiment analysis. Additionally, the cases of worst transfer happen when the target language is introflexive (especially for XLM-R).

Next, to address RQ1 more directly, we compare two different types of transfer: *intra-group* transfer, where both the fine-tuning and target languages belong to the same morphological group, and *inter-group* transfer, where the two differ in morphological type. We calculate an average for both types of transfer and for each training group, model architecture and task. We present the resulting values in Figure 1.

Generally, transfer to another morphological type implies a higher cost in terms of performance, except for the introflexive models. This difference in transfer loss appears to be similar for all groups in the sentiment task, yet it varies considerably in the part-of-speech task. More specifically, there are two extremes in this latter case: fusional models suffer large performance drops when switching morphological groups, whereas isolating models experience similar transfer losses in both conditions.

Finally, we average again to obtain a single transfer loss value for each task and model, and use it to establish a comparison in Figure 2. Here we observe that: (1) the difference in transfer loss between an intra-group and inter-group transfer is higher on the part-of-speech task, (2) transfer is also generally worse on this task\(^\text{11}\), (3) XLM-R models perform better cross-lingual transfers in general (especially on the sentiment analysis task), and (4) the difference between intra-group and inter-group transfer is similar on both model architectures.

### Analysis

In this section, we run several statistical tests to verify our conclusion to RQ1 and detail several points of analysis that relate to RQ2 and RQ3. Namely, to what degree do other variables contribute to effects on cross-lingual transfer.

#### 6.1 Testing the effect of transfer type

We run a set of statistical tests to validate the observations made from Figure 2 in Section 5. In the part-of-speech tagging task, an analysis of variance (ANOVA) reveals there is a statistically significant, although weak, difference in transfer loss between the intra- and inter-group conditions, for both model architectures (\(\eta^2 \approx 0.06, p < 0.01\) in both cases). In contrast, a Kruskal-Wallis analysis of variance\(^\text{12}\) finds no significant difference

\(^{11}\)Strictly speaking, we use different metrics for both tasks, which are not necessarily comparable.

\(^{12}\)The normality condition for ANOVA is not met.
Table 4: Group-to-group transfer loss (in percentage points) in the part-of-speech tagging (top) and sentiment analysis (bottom) tasks for each fine-tuning (column) and testing (row) language’s morphological group, as well as each model architecture. Minimum values in each fine-tuning group and architecture are highlighted. Lower is better.

Figure 1: Average transfer loss (in percentage points) to other languages of the same group (intra-group) and to languages that belong to the other groups (inter-group) in the part-of-speech tagging (top) and sentiment analysis (bottom) tasks. Lower is better.

between the two types of transfer in the sentiment analysis task, in neither mBERT or XLM-R models ($p > 0.01$ in both cases). We also test for differences in transfer loss between model architectures and find a significant difference in the sentiment analysis task (Kruskal-Wallis, $p < 0.01$), but not in the part-of-speech tagging task (ANOVA, $p > 0.01$). This is all consistent with our previous observations.
6.2 Linear regression model for transfer loss

Additionally, we model language-to-language transfer loss with a linear regression model, using transfer type, as well as other variables, as possible predictors. This allows us to (a) test whether the intra-/inter-group difference retains its statistical significance in the presence of other variables and (b) evaluate its effect in comparison to other predictors.

First, we select a set of variables that might be relevant in cross-lingual transfer, and remove those that are highly correlated with the rest to avoid multicollinearity in the model (see Table 7 in the Appendix for the final list of selected variables). We standardize all of the remaining features so that their units are comparable and, consequently, so are their regression coefficients.

Again, we find transfer type (intra-/inter-group) to be a significant predictor in both regression models for part-of-speech tagging ($p < 0.01$), but not in sentiment analysis. In the former case, it has the second strongest effect with a standardized coefficient of 8.6, the first being presence of the target language in pretraining with a coefficient of -25.9. In other words, transferring to a language on which the model has not been pretrained implies an additional performance drop of 25.9 percentage points, while transferring to another morphological group incurs an additional 8.6.

The remaining predictors for this task are average test example length (measured in tokens, coefficient of 4.0) and in-language score (3.3). The first is a complex variable because differences in text length can be due to their domain or to the languages themselves but, in either case, its coefficient confirms our intuition that longer sequences generally make the task more difficult. The second could indicate some overfitting to the fine-tuning language, as higher in-language score entails slightly poorer transfer.

XLM-R adds another predictor: the proportion of words that have been split into subword tokens in the test data (2.1). This variable is related to the size of the pretraining corpus for each language: a richer pretraining vocabulary will ensure more words are considered frequent during Byte Pair Encoding and, therefore, assigned a single token, instead of being broken down into subword tokens by the tokenizer. This means that high-resource languages will have a lower word split probability and, hence, it will be slightly easier to transfer to them. However, it is worth pointing out that this bias has little effect and is only statistically significant in XLM-R.

In the case of sentiment analysis, relevant predictors are: presence of the fine-tuning (coefficient of -11.8 for mBERT and -18.7 for XLM-R) and target (-10.3 and -16.3) languages in pretraining, in-language score (6.8 and 6.5), proportion of words split into subword tokens in the training data (3.3 and 2.7) and proportion of examples labeled as positive in the test set (-2.8, XLM-R only).

Curiously, sentiment analysis is more sensitive to variables related to the training data compared to part-of-speech tagging, whereas sequence length only affects the latter. On the other hand, language inclusion in pretraining and in-language score are useful predictors in both tasks, yet the former is far stronger in POS and the latter is more relevant in sentiment analysis. In summary, we verify that transferring to a different morphological type has a relevant effect in part-of-speech tagging but not in sentiment analysis, regardless of the model architecture.

6.3 Testing pretrained languages only

Given the considerable effect pretraining seems to have on transfer loss (discussed in Section 6.2), we re-evaluate our results after removing the languages that were not present during the pretraining of either of the two model architectures (Cantonese, Algerian and Maltese) and check whether there are relevant differences with our previous results.

\[^{14}\text{In fact, we do not include pretraining data size as a predictor because of its correlation with the variable in question.}\]

\[^{13}\text{Since the regression models for mBERT and XLM-R are quite similar, we report the averaged coefficients here.}\]

Figure 2: Comparison across tasks of the average transfer loss (in percentage points) to other languages of the same group (intra-group) and to languages that belong to the other groups (inter-group). Lower is better.
Of course, we observe an improvement in cross-lingual scores involving either an isolating or an introflexive language, because these are the groups the excluded languages belong to. Overall, however, re-running the statistical tests does not modify our previous conclusions (see Figure 3).

### 6.4 Balanced in-language scores

Since in-language score is relevant in all regression models considered in 6.2 (and the value of transfer loss is relative to it), we decide to re-train all models, this time preventing them from increasing said score above a fixed threshold value (we choose the minimum in-language score achieved previously in each task and model architecture) and re-evaluate our previous conclusions.

The intra-/inter-group difference in transfer loss is still statistically significant in part-of-speech tagging and not in sentiment analysis. Similarly, there is still a statistically significant difference in transfer loss between both models only in the sentiment analysis task. All of this can be seen in Figure 3. The only remarkable difference is in the part-of-speech task, where the average inter-group transfer loss values for all morphological groups seem to converge to the same value (see Figure 5 in the Appendix). For more information, see Figures 5 and 6, as well as Tables 8 and 9, all of which can be found in the Appendix.

### 6.5 Effect of training data size

We also test the effect that training with considerably more data has on cross-lingual transfer. We select two languages, each with around 150,000 examples available: German for the part-of-speech tagging task and Korean for sentiment analysis. We train four models with increasingly more data and then test them on all languages.

In German, we notice an important decline in cross-lingual scores when increasing data size from 80,000 to 150,000 examples (see Figure 4). More specifically, in mBERT models there is an average decrease of 15.6 and 9.0 points when the cross-lingual transfer is intra- and inter-group, respectively. In XLM-R, the corresponding values are 25.4 and 19.5. Hence, it appears that a phenomenon of language specialization takes place, one to which XLM-R is more susceptible and that has more important consequences in intra-group transfer. To ensure this is a language and not a domain/dataset specialization, we test these models on another
German dataset (PUD) and find no decrease in performance.

In contrast, average Korean cross-lingual scores remain relatively constant (see Figure 4). Therefore, the language specialization phenomenon could be more characteristic of part-of-speech tagging than sentiment analysis.

6.6 Domain effects

Conneau et al. (2020b) find that domain mismatch in pretraining of multilingual LMs is more problematic than domain mismatch in fine-tuning. Yet given the variety of domains present in the sentiment data, we decided to test its effect. Proxy A-distance (Glorot et al., 2011) measures the generalization error of a linear SVM trained to discriminate between two domains. We translate 1000 sentences from each dataset to English using GoogleTranslate and then compute the proxy A-distance. For POS tagging, there are small but insignificant negative effects of proxy A-distance on results for both models (a Pearson coefficient of -0.07, \( p > 0.01 \) and -0.07, \( p > 0.01 \) for mBERT and XLM-R, respectively). On the sentiment task, there is no significant domain effect for mBERT (-0.06, \( p > 0.01 \)), while there is a small negative effect for XLM-R (-0.27, \( p < 0.01 \)). This suggests that most of the transfer loss is not due to domain mismatch.

7 Discussion and Future Work

In this paper, we have conducted an extensive analysis of the effects of morphological typology on cross-lingual transfer and attempted to isolate these factors from other variables. We have compared performance of two state-of-the-art zero-shot cross-lingual models on two tasks (part-of-speech tagging and sentiment analysis) for 19 languages across four morphological typologies. We have found that transfer to another morphological type generally implies a higher performance loss than transfer to another language with the same morphological typology. Additionally, part-of-speech tagging is more sensitive to morphological differences than sentiment analysis, while sentiment analysis is more sensitive to variables related to the fine-tuning data and is less predictable in general.

We have tested this sensitivity to morphology after balancing other influential factors, such as in-language score, and, still, the intra-/inter-group difference remains. However, the effect of morphological typology, while significant, is not strong, given that most of the variability in transfer loss is due to other factors.

We have also confirmed that XLM-R generally transfers better than mBERT, especially on sentiment analysis. In part-of-speech tagging, we have reported considerably better transfer within fusional languages, as well as easier transfer from the other groups towards the fusional type. Moreover, we have found a case that suggests that fine-tuning on large training sets might lead to language specialization and, consequently, be detrimental to cross-lingual transfer.

It is worth noting that we do not explore whether the type of script used by the languages has an effect on cross-lingual transfer. This is hard to control in our experimental setup, as there are some scripts that are either unique to a language or only have one with enough data to represent it, making it impossible to make comparisons.

The recent cross-lingual suite Xtreme (Hu et al., 2020) includes a number of benchmark tasks in 40 languages. While this dataset is a useful collection of cross-lingual tasks, it is unfortunately not sufficient for our purposes. The POS data is the same as we use, while other tasks either a) do not contain a representative sample of language typologies b) use translation, introducing problems of ‘translationese’, or c) are automatically created and not manually curated Named Entity Recognition data. Our experimental setup avoids these problems by focusing on binary sentiment analysis, which is a task that has data available in many languages and does not require translation to get multilingual data.

Finally, this work ties in with the increasing interest in typological questions in NLP (Takamura et al., 2016; Ponti et al., 2019; Bjerva et al., 2019; Nooralahzadeh et al., 2020; Bjerva and Augenstein, 2021), which often try to directly predict typological features, or use these to analyze model performance.

In the future, it would be interesting to train multi-lingual language models on specific language families in order to find maximal benefits from shared morphology. Finally, as typology seems to affect tasks differently, it would be interesting to explore other tasks, e.g., dependency parsing or semantic role labeling.

\[\text{Implementation adapted from the code available at } \text{https://github.com/rpryzant/proxy-a-distance.}\]
References

Nawaf Abdulla, Nizar A. Ahmed, Mohammed Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. pages 1–6.

Rodrigo Agerri, Montse Cuadros, Seán Gaines, and German Rigau. 2013. Openener: Open polarity enhanced named entity recognition. Procesamiento del Lenguaje Natural, 51(0):215–218.

Adam Amram, Anat Ben David, and Reut Tsarfaty. 2018. Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from modern Hebrew. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2242–2252, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mikel Arteste, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Arteste, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4623–4637, Online. Association for Computational Linguistics.

bact’, Pattarawat Chormai, Charin, and ekapolc. 2019. Pythainlp/wisesight-sentiment: First release.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech Language, 28(1):56 – 75.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 127–135, Honolulu, Hawaii. Association for Computational Linguistics.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018a. MultiBooked: A corpus of basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan. European Languages Resources Association (ELRA).

Jeremy Barnes and Roman Klinger. 2019. Embedding projection for targeted cross-lingual sentiment: Model comparisons and a real-world study. Journal of Artificial Intelligence Research, 66:691–742.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018b. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2483–2493, Melbourne, Australia. Association for Computational Linguistics.

Emily M. Bender. 2013. Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax. Morgan amp; Claypool Publishers.

Balthasar Bickel and Johanna Nichols. 2005. Inflectional morphology. In Timothy Shopen, editor, Language Typology and Syntactic Description. Cambridge University Press, Cambridge. 2nd edition.

Johannes Bjerva and Isabelle Augenstein. 2021. Does typological blinding impede cross-lingual sharing?

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019. Uncovering probabilistic implications in typological knowledge bases. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3924–3930, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6022–6034, Online. Association for Computational Linguistics.

Keith Cortis and Brian Davis. 2019. A social opinion gold standard for the Malta government budget 2018. In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pages 364–369, Hong Kong, China. Association for Computational Linguistics.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Le Anh Cuong, Ng. T. Minh Huyen, and Ng. Viet Hung. 2016. Vlsp 2016 shared task: Vietnamese analysis. In VLSP 2016.
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.

Alexiei Dingli and Nicole Sant. 2016. Sentiment analysis on maltese using machine learning. In Proceedings of The Tenth International Conference on Advances in Semantic Processing (SEMAPRO 2016), pages 21–25.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korbonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML’11, page 513–520, Madison, WI, USA. Omnipress.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Matthias Huck, Diana Dutka, and Alexander Fraser. 2019. Cross-lingual annotation projection is effective for neural part-of-speech tagging. In Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 223–233, TOBEFILLED-Ann Arbor, Michigan. Association for Computational Linguistics.

Georgios Kalamatianos, Dimitrios Mallis, Symeon Symeonidis, and Avi Arampatzis. 2015. Sentiment analysis of greek tweets and hashtags using a sentiment lexicon. pages 63–68.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2832–2838, Copenhagen, Denmark. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. Advances in Neural Information Processing Systems (NeurIPS).

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4483–4499, Online. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1663–1674, Online. Association for Computational Linguistics.

Krister Lindén, Tommi Jauhiainen, and Sam Hardwick. 2020. Finnssentiment – a finnish social media corpus for sentiment polarity annotation.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Sebastian J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 976–983, Prague, Czech Republic. Association for Computational Linguistics.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. ASTD: Arabic sentiment tweets dataset. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.

Garrett Nicolai, Kyle Gorman, and Ryan Cotterell, editors. 2020. Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. Association for Computational Linguistics, Online.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4547–4562, Online. Association for Computational Linguistics.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In Proceedings of the 12th Language
Sebastian Pado and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. Journal of Artificial Intelligence Research, 36(1):307–340.

Samuel Pecar, Marian Simko, and Maria Bielikova. 2019. Improving sentiment classification in Slovak language. In Proceedings of the 7th Workshop on Balto- Slavic Natural Language Processing, pages 114–119, Florence, Italy. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Frans Plank. 1999. Split morphology: how agglutination and flexion mix. Linguistic Typology, 3:279–340.

Edoardo Maria Ponti, Helen O’Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. Computational Linguistics, 45(3):559–601.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL- Smadi, Mahmoud Al- Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez- Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 19–30, San Diego, California. Association for Computational Linguistics.

A. Purwarianti and I. A. P. A. Crisdayanti. 2019. Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector. In 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA), pages 1–5.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North- African Arabic treebank: Tackling hell. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1139–1150, Online. Association for Computational Linguistics.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 682–686, Portland, Oregon, USA. Association for Computational Linguistics.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. Transactions of the Association for Computational Linguistics, 1:1–12.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 69–76, Portorož, Slovenia. European Language Resources Association (ELRA).

Samia Touileb and Jeremy Barnes. 2021. The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, Online. Association for Computational Linguistics.

Adam Tsakalidis, Symeon Papadopoulos, Rania Voskaki, Kyriaki Ioannidou, Christina Boididou, Alexandra I Cristea, Maria Liakata, and Yiannis Kompatsiaris. 2018. Building and evaluating resources for sentiment analysis in the greek language. Language resources and evaluation, 52(4):1021–1044.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30, pages 5998–6008, Curran Associates, Inc.

Michael Wojatzki, Eugen Ruppert, Sarah Holzscheiter, Torsten Zesch, and Chris Biemann. 2017. Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. In Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback, pages 1–12, Berlin, Germany.

Rong Xiang. 2019. Sentiment augmented attention network for cantonese restaurant review analysis.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis
tools via robust projection across aligned corpora. In Proceedings of the First International Conference on Human Language Technology Research.

Marcos Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann, Shervin Malmasi, and Ahmed Ali, editors. 2018. Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018). Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noémi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielle Aleksandraviciūtė, Lene Antonsen, Katya Aplonova, Angelina Aquino, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Balleseros, Esha Banerjee, Sebastian Bank, Verónica Barbu Miutul, Victoria Basmov, Colin Batchelor, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Irina Boffa, Jørg Tiedemann, Shervin Malmasi, and Ahmed Ali, Jonsdóttir, Fredrik Jørgensen, Markus Juutinen, Bartra Mititelu, Victoria Bobicev, Loïc Boizou, Emanuel Brokaitė, Aljoscha Burchardt, Marie Candito, Snezana Čobilić, Martin Karhinovček, Atalaya Castro, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Bernard Castier, Bernard Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann, Shervin Malmasi, and Ahmed Ali, editors. 2018. Proceedings of the First International Conference on Human Language Technology Research.
Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utkas, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilian Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.
A Appendix

| Language     | Dataset   | Domain                   |
|--------------|-----------|--------------------------|
| • German     | HDT (subset) | News                     |
| • Spanish    | AnCorra   | News                     |
| • Slovak     | SNK       | News, Literature         |
| • Norwegian  | Bokmaal NDT | News                    |
| • Greek      | GDT       | Parliament, Wikipedia, Web|
| ✗ Mandarin   | GSD       | Wikipedia                |
| ✗ Vietnamese | VTB       | News                     |
| ✗ Thai       | PUD       | News, Wikipedia          |
| ✗ Cantonese  | HK        | Movies, Parliament       |
| ✗ Indonesian | CSUI      | News                     |
| ★ Finnish    | TDT       | Many                     |
| ★ Basque     | BDT       | News                     |
| ★ Korean     | Kaist     | Literature, News, Academic|
| ★ Japanese   | GSD       | News, Web                |
| ★ Turkish    | IMST      | News, Literature         |
| ▲ Arabic     | PADT      | News                     |
| ▲ Hebrew     | HTB       | News                     |
| ▲ Algerian   | NArabizi  | Web, Lyrics              |
| ▲ Maltese    | MUDT      | Many                     |

Table 5: Detailed description of the data used in part-of-speech tagging.
| Language        | Text Type     | Domain      | Annotation | Examples | Train % | Dev/Test % |
|-----------------|---------------|-------------|------------|----------|---------|------------|
| ● German        | Social media  | Trains      | Manual     | 8706     | 100     | 100        |
| ● Spanish       | Reviews       | Hotels      | Manual     | 1472     | 100     | 100        |
| ● Slovak        | Reviews       | Services    | Manual     | 5124     | 100     | 100        |
| ● Norwegian     | Reviews       | Many        | Manual     | 3608     | 100     | 100        |
| ● Greek         | Social media  | Politics    | Manual     | 661      | 3       | 39         |
|                 | Social media  | Many        | Manual     | 519      | 5       | 22         |
|                 | Reviews       | Mobile phones | User scores | 5906 | 92 | 39 |
|                   | Reviews       | Many        | User scores | 19835 | 100 | 100 |
| ❌ Mandarin      | Reviews       | Technology  | Manual     | 3400     | 100     | 100        |
| ❌ Chinese       | Reviews       | Product reviews | Manual     | 11600    | 100     | 100        |
| ❌ Cantonese     | Reviews       | Food        | User scores | 41578 | 100 | 100 |
| ❌ Vietnamese    | Reviews       | Many        | Manual     | 11324    | 100     | 100        |
| ♦ Finnish       | Social media  | Many        | Manual     | 6332     | 100     | 100        |
| ★ Basque        | Reviews       | Food/lodging | Manual     | 1129     | 100     | 100        |
| ★ Korean        | Reviews       | Movies      | User scores | 40000 | 100 | 100 |
| ★ Japanese      | Reviews       | Many        | User scores | 14060 | 100 | 100 |
| ★ Turkish       | Reviews       | Food        | Manual     | 1052     | 16      | 100        |
| ★ Arabic        | Social media  | Many        | Manual     | 1589     | 45      | 45         |
| ★ Hebrew        | Social media  | Politics    | Manual     | 10110    | 100     | 100        |
| ★ Algerian      | Social media  | Many        | Manual     | 731      | 100     | 100        |
| ★ Maltese       | Social media  | Many        | Manual     | 718      | 84      | 84         |
|                 | Social media  | Politics    | Manual     | 133      | 16      | 16         |

Table 6: Detailed description of the data used in sentiment analysis. "Train %" and "Dev/Test %" indicate what percentage of the language’s training and validation/test data, respectively, comes from the dataset in question.

Intra- and Inter-Group Transfer Losses in Part-of-Speech (Balanced In-language Scores)

Figure 5: Average transfer loss (in percentage points) to other languages of the same group (intra-group) and to languages that belong to the other groups (inter-group) in the part-of-speech tagging task after balancing in-language scores. Lower is better.
| Predictor                                      | Language | Task |
|-----------------------------------------------|----------|------|
| In-language score                             | Train    | Both |
| Average example length (tokens)               | Train    | Both |
| Average example length (tokens)               | Test     | Both |
| Included in pretraining                       | Train    | Both |
| Included in pretraining                       | Test     | Both |
| Words split into subword tokens (%)           | Train    | Both |
| Words split into subword tokens (%)           | Test     | Both |
| Proportion of positive examples               | Train    | SA   |
| Proportion of positive examples               | Test     | SA   |
| Transfer type (intra-group/inter-group)      | -        | Both |

Table 7: Variables considered in the linear regression model after eliminating multicollinearity. "Language" indicates whether the predictor was measured on the fine-tuning language (train) or the target language (test), "SA" stands for sentiment analysis.

| Test  | Train | mBERT | XLM-R | mBERT | XLM-R | mBERT | XLM-R | mBERT | XLM-R | mBERT | XLM-R |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Fusional |      |      |       |       |       |       |       |       |       |       |       |
| Isolating |     | 46.5  | 51.9  | 48.8  | 49.2  | 47.5  | 49.6  | 45.7  | 47.0  |       |       |
| Agglutinative |   | 54.4  | 55.2  | 53.3  | 50.9  | 55.2  | 54.8  | 49.7  | 46.7  |       |       |
| Introflexive |     | 39.9  | 41.9  | 37.4  | 36.7  | 36.9  | 34.8  |       |       |       |       |

| Test  | Train | mBERT | XLM-R | mBERT | XLM-R | mBERT | XLM-R | mBERT | XLM-R | mBERT | XLM-R |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Fusional |      | 48.3  | 42.8  |       |       | 45.4  | 45.4  | 41.7  | 42.4  |       |       |
| Isolating |     | 49.8  | 44.2  | 51.9  | 43.0  | 37.6  | 42.1  | 36.3  | 43.0  |       |       |
| Agglutinative |   | 46.4  | 47.1  | 48.0  | 50.7  | 40.1  | 47.3  | 41.6  | 43.5  |       |       |
| Introflexive |     | 48.0  | 42.6  | 45.5  | 41.8  | 43.4  | 45.4  | 45.0  | 45.2  |       |       |

Table 8: Group-to-group cross-lingual accuracy scores (%) for part-of-speech tagging (top) and macro F1 scores (%) in the sentiment analysis task (bottom) (after balancing in-language scores) for each fine-tuning (column) and testing (row) morphological group, and each model architecture. Maximum values in each test group and architecture are highlighted. Higher is better.
Table 9: Group-to-group transfer loss (in percentage points) in POS (top) and sentiment analysis (bottom) tasks (after balancing in-language scores) for each fine-tuning (column) and testing (row) language’s morphological group, as well as each model architecture. Minimum values in each fine-tuning group and architecture are highlighted. Lower is better.

Figure 6: Average transfer loss (in percentage points) to other languages of the same group (intra-group) and to languages that belong to the other groups (inter-group) in the sentiment analysis task after balancing in-language scores. Lower is better.