Understanding Cross-Lingual Syntactic Transfer in Multilingual Recurrent Neural Networks

Prajit Dhar **Arianna Bisazza**
Center for Language and Cognition
University of Groningen
{p.dhar, a.bisazza}@rug.nl

Abstract

It is now established that modern neural language models can be successfully trained on multiple languages simultaneously without changes to the underlying architecture. But what kind of knowledge is really shared among languages within these models? Does multilingual training mostly lead to an alignment of the lexical representation spaces or does it also enable the sharing of purely grammatical knowledge? In this paper we dissect different forms of cross-lingual transfer and look for its most determining factors, using a variety of models and probing tasks. We find that exposing our LMs to a related language does not always increase grammatical knowledge in the target language, and that optimal conditions for *lexical-semantic* transfer may not be optimal for *syntactic* transfer.

1 Introduction

One of the most important NLP discoveries of the past few years has been that a *single* neural network can be successfully trained to perform a given NLP task in *multiple* languages without architectural changes compared to monolingual models (Östling and Tiedemann, 2017; Johnson et al., 2017). Besides important practical advantages (fewer parameters and models to maintain), such multilingual Neural Networks (mNNs) provide an easy but powerful way to transfer task-specific knowledge from high- to low-resource languages (Devlin et al., 2019; Conneau and Lample, 2019; Aharoni et al., 2019; Neubig and Hu, 2018; Arivazhagan et al., 2019; Artetxe and Schwenk, 2019; Chi et al., 2020). These success stories have led to a need for understanding *how* exactly cross-lingual transfer works within these models. Figure 1 illustrates different possible characterizations of a trained mNN: While the no-transfer scenario is rather easy to rule out, understanding which linguistic categories are shared, and to what extent, is more challenging.

In this work, we focus on the transfer of *syntactic* knowledge among languages and look for evidence that mNNs induce a shared syntactic representation space *while not receiving any direct cross-lingual supervision*. To be clear, if we measure transfer among languages X and Y, every training sentence for language modeling will be either in language X or Y, while for machine translation every sentence pair will be either in language pair (X, Z) or (Y, Z). Thus, the only pressure to share linguistic representations is given by the sharing of the hidden layer parameters (as well as, possibly, some of the word embeddings).

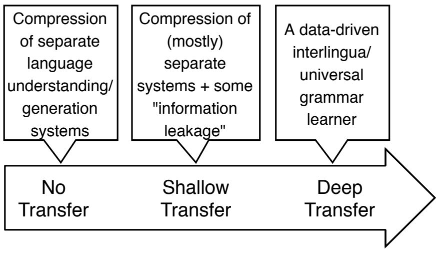
Compression of separate language understanding/ generation systems
Compression of (mostly) separate systems + some "information leakage"
A data-driven interlingua/ universal grammar learner
No Transfer
Shallow Transfer
Deep Transfer

Figure 1: Possible characterizations of a trained mNN in terms of cross-lingual transfer levels.

Neural language models have been shown to implicitly capture non-trivial structure-sensitive phenomena like long-range number agreement (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018). However most of these studies have been confined to monolingual models. We then investigate the following questions:

1. Does mNNs' implicit syntactic knowledge of L2 increase by exposure to a related L1?

2. Do mNNs induce a common representation space with shared syntactic categories?

Our research questions are reminiscent of well-known questions in the fields of psycholinguistic and second language acquisition, where work has shown that both lexical and syntactic representations are shared in the mind of bilinguals (Hartsuiker et al., 2004a; Vasilyeva et al., 2010). Taking inspiration from this body of work, we investigate what factors are needed for mNNs to successfully transfer linguistic knowledge, including vocabulary overlap, language relatedness, number of training languages, training regime (joint *vs* sequential) and training objective (next word prediction *vs* translation to a third language).

In contrast to the current mainstream focus on BERT-like models (Rogers et al., 2020), we evaluate more classical LSTM-based models trained for next word prediction or translation over a moderate number of languages (2 or 9). We choose this setup because (i) it allows for more controlled and easy-to-replicate experiments in terms of both training data and model configuration and (ii) LSTMs trained on a standard sequence prediction objective are more cognitively plausible and directly applicable to our main probing task, namely agreement prediction. In this setup, we find limited and rather inconsistent evidence for the transfer of implicit grammatical knowledge when semantic cues are removed (Gulordava et al., 2018). While moderate PoS category transfer occurs, truly language-agnostic syntactic categories (such as *noun* or *subject*) do not seem to emerge in our mNN representations. Finally, we find that optimal conditions for lexical-semantic transfer may not be optimal for syntactic transfer.

2 Previous Work

Multilingual Machine Translation Early work on multilingual NMT focused on building dedicated architectures (Dong et al., 2015; Firat et al., 2016; Johnson et al., 2017). Starting from (Johnson et al., 2017), m-NMT models are mostly built with the same architecture as their monolingual counterparts, by simply adding language identifying tags to the training sentences. Using a small set of English sentences and their Japanese and Korean translations, Johnson et al. (2017) showed that semantically equivalent sentences form well-defined clusters in the high-dimensional space induced by a NMT encoder trained on large-scale proprietary datasets. Kudugunta et al. (2019) analyze the similarity of encoder representations of different languages within a massively m-NMT model. They find that representation similarity correlates strongly with linguistic similarity and that encoder representations diverge based on the target language. However they do not disentangle the syntactic aspect from other types of transfer.

Multilingual Sentence Encoders A related line of work focuses on mapping sentences from different languages into a common representation space to be used as features in downstream tasks where training data is only available in a different language than the test language. Artetxe and Schwenk (2019) use the encoder representations produced by a massively multilingual NMT system similar to (Johnson et al., 2017) to perform cross-lingual textual entailment (XNLI) and document classification. m-BERT (Devlin et al., 2019; Devlin, 2018) and XLM (Conneau and Lample, 2019) are large-scale mNNs trained on a masked LM (MLM) objective using mixed-language corpora. This results in general-purpose contextualized word representations that are multilingual in nature, *without* requiring any parallel data. m-BERT representations have been proved particularly successful for transferring dependency parsers to low- (or zero-)resource languages (Wu and Dredze, 2019; Kondratyuk and Straka, 2019; Tran and Bisazza, 2019). On the task of cross-lingual textual entailment (Conneau et al., 2018b), XLM-based classifiers come surprisingly close to systems that use fully-supervised MT as part of their pipeline (to translate the training or test data).

Implicit Learning of Linguistic Structure NNs trained for downstream tasks such as language modeling, translation or textual entailment, have been shown to implicitly encode a great deal of linguistic structure such as morphological features (Belinkov et al., 2017; Bisazza and Tump, 2018; Bjerva and Augenstein, 2018), number agreement (Linzen et al., 2016; Gulordava et al., 2018) and other structure-sensitive phenomena (Marvin and Linzen, 2018). Studies such as (Tenney et al., 2019b,a; Jawahar et al., 2019) have extended these findings to BERT representations showing positive results on a variety of syntactic probing tasks. Extensive overviews of this body of work are presented in (Belinkov and Glass, 2019) and (Rogers et al., 2020).

Cross-lingual Transfer in Multilingual NNs Recent studies (Wu and Dredze, 2019; Pires et al., 2019; Chi et al., 2020) have found evidence of *syntactic* transfer in m-BERT using POS tagging and dependency parsing experiments. On the other hand, Libovický et al. (2019) find that m-BERT representations capture cross-lingual *semantic* equivalence sufficiently well to allow for word-alignment and sentence retrieval, but fail at the more difficult task of MT quality estimation. While this massive Transformer-based (Vaswani et al., 2017) architecture has received overwhelming attention in the past year, we believe that smaller, better understood, and easier to replicate model configurations can still play an important role in the pursuit of NLP model explainability. Moreover, the large number of m-BERT training languages (ca. 100) added to the uneven language data distribution and the highly shared subword vocabulary, make it difficult to isolate transfer effects in any given language pair. Mueller et al. (2020) recently tested a LSTM trained on five languages on a multilingual extension of the subject-verb agreement set of Marvin and Linzen (2018). They found signs of harmful interference rather than positive transfer across languages. In Section 4 we corroborate this rather surprising finding by using a more favourable setup for transfer, that is: (i) only two, related, training languages, (ii) a simulated low-resource setup for the target language, and (iii) eliminating vocabulary overlap during training with language IDs.

Cross-lingual Transfer in the Bilingual Mind Measuring the extent to which dual-language representations are shared in the mind of bilingual subjects is a long-standing problem in the field of second language acquisition (Kellerman and Sharwood Smith, 1986; Odlin, 1989; Jarvis and Pavlenko, 2008; Kootstra et al., 2012). Among others, Hartsuiker et al. (2004b) present evidence of cross-lingual *syntactic priming* in bilingual English-Spanish speakers, which are more inclined to produce English passive sentences after having heard a Spanish passive sentence. Using neuroimaging techniques in a reading comprehension experiment with in German-English bilinguals, Tooley and Traxler (2010) report that the processing of L1 and L2 sentences activates the same brain areas, pointing to the shared nature of syntactic processing in the bilingual mind. Taking inspiration from this body of work, we investigate what factors trigger cross-lingual transfer of syntactic knowledge within mNNs.

Cross-Lingual Dependency Parsing Finally, our work is also related to the productive field of cross-lingual and multilingual dependency parsing (Naseem et al., 2012; Zhang and Barzilay, 2015; Täckström et al., 2012; Ammar et al., 2016, *inter alia*), with the important difference that we are interested in models that are *not* explicitly trained to recognize syntactic structure but acquire it indirectly while optimizing next word prediction or translation objectives. Among others, Ahmad et al. (2019) have shown that the difficulty of transferring a dependency parser cross-lingually depends on typological differences between the source and target languages, with word order differences playing an important role. In this paper, we mainly consider source-target languages that are related, like French or Spanish (source) and Italian (target), where we expect implicit syntactic knowledge to be more easily transferable.

3 Probing Tasks

To answer our RQ1 (are mNNs capable of implicitly transferring syntactic knowledge between languages?) we choose the task of Number Agreement. For our RQ2 (are mNNs able to induce a common representation with shared syntactic categories?) we look at less complex syntactic tasks such as PoS tag classification and Dependency relation classification, and contrast them with a lexical-semantic task (word translation retrieval). We choose these tasks because they can be framed as simple classification (or ranking) problems and have a direct linguistic interpretation. We do not consider parsing because it is a complex task with a highly structured prediction space requiring dedicated model components. The probed models are LSTM-based language models and translation models, trained at the word-level. More details are provided below.

3.1 Number Agreement

Number agreement describes the instance where a phrase and its arguments or modifiers must agree in their number feature. Number agreement can occur between a subject-predicate pair (*the son$_{sg}$ of my neighbors goes$_{sg}$)*, noun-quantifier pair *(many$_{pl}$ huge trees$_{pl}$)*, etc. Linzen et al. (2016) first proposed the subject-verb agreement task to assess the ability of a LSTM-based LM to capture non-trivial language structure, by checking if the correct verb form was assigned a higher probability than the wrong one, e.g. if prob(*were*|context) > prob(*was*|context) in the sentence *The boys, who were lost in the forest* ***were/was*** *found.* LM performance was shown to be mostly affected by the number of agreement attractors.

Probing Dataset We adopt the benchmark by Gulordava et al. (2018), henceforth called G18, which extends the evaluation of Linzen et al. (2016) to more languages and more agreement constructions, automatically harvested from corpora using POS patterns. G18 also introduced two conditions to test whether a model relies on semantic cues or purely grammatical knowledge to predict agreement:

1. Original : Sentences automatically extracted from corpora;

2. Nonce : Nonsensical but grammatical sentences created by randomly replacing all content words in the original sentence with random words with same morphological class.

Thus, this is one of few existing tasks that allow us to measure the transfer of grammatical knowledge in isolation. Using the G18 benchmark, we compare mNNs with monolingually trained models, in order to compare if the addition of a related language improves the long-range agreement accuracy of the monolingual model. We expect this to happen for languages that have the same number agreement patterns, like French and Italian.

Probed Models Similar to G18, we train 2-layer LSTMs with embedding and hidden layer size of 650, for 40 epochs, using a dataset of crawled Wikipedia articles. These language models are trained on next word prediction and do not receive any specific supervision for the syntactic task. $L1$ is our helper language and $L2$ is the target language where we measure agreement accuracy. Fig. 2 shows our different training setups. To simulate a low-resource setup and possibly increase the chances of transfer, we train our bilingual LMs on a shuffled mix of a larger L1 corpus (L1$_{large}$, 80M tokens) and a smaller L2 corpus ($L2_{small}$, 10M tokens). L2 is oversampled to approximately match the amount of L1 sentences. This bilingual model (LM$_{L1+L2_{small}}$) is compared to a baseline monolingual LM trained on a small L2 corpus (LM$_{L2_{small}}$). As upper bound, we also show the results of a model trained on more L2 data (80M). This model performs closely to the results reported by G18 with a similar setup.

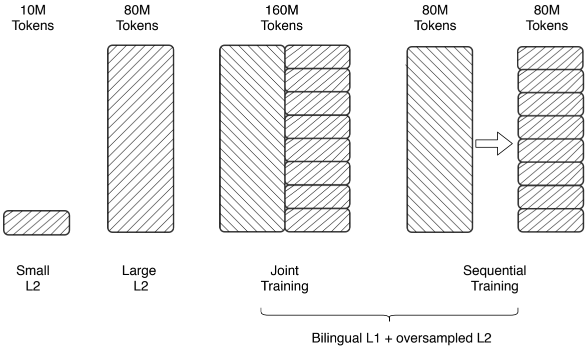

Figure 2: Monolingual and bilingual LM training schemes used in our agreement experiments.

Most experiments in this paper are performed by *joint training*, i.e. the model is trained on the mixed language data since initialization. However in Sect. 4.2 we also evaluate *pre-training*: i.e. the LM is first trained on L1 data, then after convergence, it continues training on L2 data (see Fig. 2). A language tag is introduced at the beginning of each sentence. The vocabulary for each language consists of the 50k most frequent tokens, with the remaining tokens replaced by the unknown tag. The bilingual vocabulary is the union of the language-specific vocabularies, resulting in a total of 88k words in our main language pair (French-Italian). In Sect. 4.2 we compare this setup (called *natural overlap*) to a *no-overlap* setup where all words are prepended with a language tag, resulting in a bilingual vocabulary of 100k words.

3.2 Cross-lingual Syntactic Category Classification

To verify whether basic syntactic categories are shared among different language representations in mNNs, we inspect the activations of our trained LMs when processing a held-out corpus. Specifically we build linear classifiers to predict either the PoS tag or the Dependency label (type of relation to the head) of a word from its hidden layer representation. This setup is similar to previous work (Blevins et al., 2018; Tenney et al., 2019b), however our diagnostic classifiers are trained on L1 and tested on L2.[1] If syntactic categories are shared, we expect to see minor drops in classification accuracy compared to a classifier trained and

[1] Another difference regards the dependency classification: Blevins et al. (2018) uses constituency parsing and Tenney et al. (2019b) predicts dependency arcs given word *pairs*.

tested on L2. In other words, we ask whether, e.g., French and Italian adjectives or subjects are recognizable by the same NN activations.

Several studies such as (Bisazza and Tump, 2018; Hewitt and Liang, 2019; Pimentel et al., 2020) have criticised diagnostic classifiers for overestimating the ability of neural networks to capture linguistic information. We partly address these pitfalls by comparing classification accuracy on top of our trained mNNs with that of their corresponding randomly initialized counterparts.

Probing Dataset We probe our models on manually annotated coarse-grained PoS and Dependency labels taken from Universal Dependency Treebanks (Nivre et al., 2019). Specifically, we use French-GSD (389k tokens), Italian-ISDT (278k), Spanish-AnCora (548k), and German-GSD (288k). UD sentences are fed to a trained model's encoder and the resulting last-layer activations are used to build the probing classifiers.

Probed Models We first apply the PoS and Dependency probing tasks to the Wikipedia-based LMs described in Sect. 3.1. To study the effect of training objective (next word prediction *vs* translation to a third language), in Sect. 5.2 we perform another set of controlled experiments using the Europarl[2] parallel corpus. Our dataset consists of $L1 \rightarrow$ English parallel sentences, where $L1$ is one of nine languages chosen from three different families: French, Italian, Portuguese, Spanish (Romance); German, Dutch, Swedish and Danish (Germanic) and Finnish (Uralic), with about 45.9M tokens for each language pair. The NMT models implement a standard attentional sequence-to-sequence architecture based on 4-layer bidirectional LSTMs (Bahdanau et al., 2015) with embedding and hidden layer size of 1024. To maximize comparability between translation and language modeling objectives, the LMs in these experiments are also 4-layer bidirectional (BiLMs, à la Peters et al. (2018)) with the same hidden layer size, trained on the source-side portion of our Europarl dataset.

3.3 Word Translation Retrieval

To put syntactic transfer in contrast with other types of transfer effects, we also experiment with word translation retrieval (henceforth abbreviated as WTR). This was used as a probing task for cross-lingual word embeddings in (Lample et al., 2018; Conneau et al., 2018a) and involves calculating the distance (measured by cosine similarity) between the embedding of a source language word (e.g., *bonjour*) and that of its translation (e.g., *buongiorno*). Since the task is context independent, only the word-type embeddings are probed. We interpret precision in this task as a measure of the alignment of two word embedding spaces, that is *lexical-semantic* transfer.

Lexicon The bilingual lexicon from MUSE (Lample et al., 2018) is used as gold standard for this task. MUSE is available for several language pairs and includes polysemous words (many-to-many pairs). For each language pair, we use 1.5k source and 200k target words.

4 Does Exposure to L1 Improve Implicit Syntactic Knowledge on a Related L2?

To answer RQ1 we use the number agreement task, which is explained in detail in Sect. 3.1. We choose Italian (IT) and Russian (RU) from the G18 dataset as our target languages $L2$. As helper languages, $L1$, we choose French (FR) and Spanish (ES) for $L2$ IT, and FR and Ukrainian (UK) for $L2$ RU, which allows us to study the impact of language relatedness. Accuracy is calculated as follows: for each sentence in the $L2$ benchmark, if the probability of the correct verb form is higher than the incorrect form, the agreement is said to be correct, and incorrect otherwise.

4.1 Main Results

Figure 3 shows the results. In this set of experiments, the bilingual models are trained by *joint training* using the union of the vocabularies in the two languages (*natural overlap*). See also Sect. 3.1. As in (Gulordava et al., 2018), the frequency baseline selects the most frequent word form (singular or plural) for each sentence.

Looking at the Original sentences, we see that the bilingual models outperform the respective small monolingual models in the closely related pairs ES→IT (86.8 *vs* 79.8) and UK→RU (90.4 *vs* 88.2). However the addition of FR data results in lower accuracies on both $L2$s. While this was expected in the unrelated pair FR→RU, the large drop in FR→IT is harder to explain.

When semantic cues are removed (Nonce sentences), ES→IT is the only bilingual model to outperform its monolingual counterpart (80.7 *vs*

[2] http://www.statmt.org/europarl/v7/

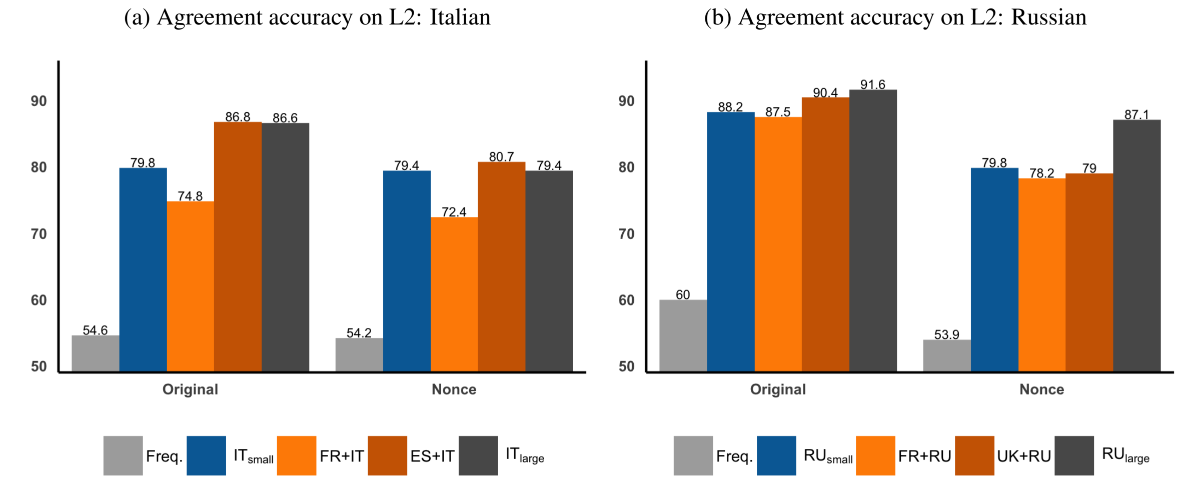

Figure 3: Probing Wikipedia-based monolingual and bilingual LMs on the agreement benchmark of Gulordava et al. (2018). Freq. is the Frequency baseline. Blue and black bars represent small and large L2 models, respectively. Orange bars represent bilingual models.

79.4), while the accuracy drop in FR→IT gets even larger (72.4 *vs* 79.4). This shows that exposing the model to a related language L1 is not guaranteed to improve implicit syntactic knowledge of L2, even when the rules of number agreement are largely shared between L1 and L2. On the contrary, our experiments suggest that in some cases L1 negatively interferes with the task in L2.

4.2 Effect of Training Regime and Vocabulary Overlap on Agreement

Could transfer in FR→IT be hampered by some of our experimental choices? To consolidate our findings, we experiment with a different training regime (*pre-training*) and a different vocabulary construction method (*no-overlap*). As shown in Table 1, both training regime and vocabulary overlap have a visible effect on the transfer of syntactic knowledge between FR and IT. Pre-training considerably reduces the negative interference effect observed in joint training, and even leads to a higher accuracy on Original sentences in the no-overlap setup (83.2 *vs* 79.8). Eliminating vocabulary overlap (None) also leads to better agreement scores in most cases. The best gain overall is obtained by the jointly trained model with no overlap (85.7 *vs* 79.8) in the Original sentences, whereas no gain is observed in the Nonce sentences.

In summary, we find limited and inconsistent evidence of transfer of purely grammatical knowledge in our bilingual models. Also contrary to our expectations, sharing more parameters (natural overlap) and mixing languages since the beginning of training leads to more negative interference than positive transfer in the FR-IT pair.

		Bilingual (FR+IT$_{small}$)				
	IT$_{small}$	Joint Training		Pre-Training		IT$_{large}$
		Natural	None	Natural	None	
Original	79.8	74.8	**85.7**	79.8	83.2	86.6
Nonce	79.4	72.4	77.6	**77.7**	76.8	79.4

Table 1: Impact of training regime and vocabulary overlap on agreement accuracy (FR→IT).

5 Do mNNs Induce Shared Syntactic Categories?

Predicting long-range agreement is a rather complex task: in principle, besides learning agreement rules, the model has to discern several syntactic categories such as number, PoS and dependencies (e.g. distinguishing subject from other noun phrases). In practice, previous work (Ravfogel et al., 2018) showed that LSTMs sometimes resort to shallow heuristics when predicting agreement.

In this section we therefore investigate whether our mNNs induce at least basic syntactic categories that are shared across languages (RQ2). We assume this is a necessary condition to enable transfer of purely grammatical knowledge, like agreement in nonce sentences, and beyond.

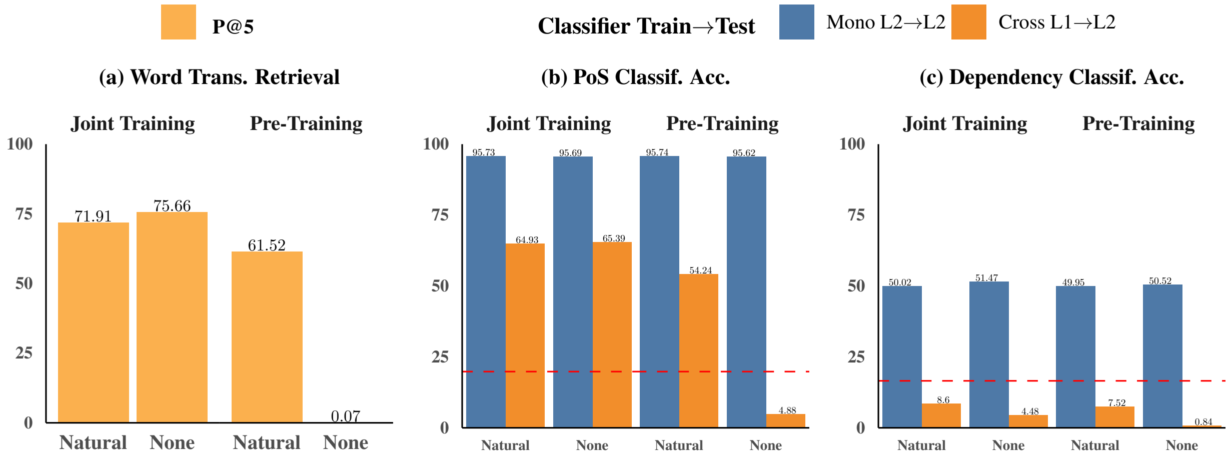

Figure 4: Semantic *vs* syntactic transfer in Wikipedia-based FR-IT bilingual LMs: (a) Word translation retrieval precision (P@5) measures lexical-semantic transfer; (b) PoS accuracy and (c) Dependency accuracy measure syntactic transfer. The classifiers are always tested on L2 (IT), and trained on either L2 or L1 (FR). If syntactic categories were perfectly shared across languages, we should observe no drop between the blue and orange bars. Dashed red lines show majority baselines for both (b) and (c).

5.1 Effect of Training Regime and Vocabulary Overlap on Syntactic Category Transfer

In this section we examine the same FR-IT Wikipedia-based LMs described in section 4.2. Figure 4(a) shows that joint training yields better alignment of the word embedding spaces compared to the pre-training setup, which confirms the findings by Ormazabal et al. (2019). Secondly, eliminating vocabulary overlap does not necessarily imply less alignment. Interestingly, work on m-BERT/XLM models has also shown that vocabulary overlap has a much smaller effect on transfer than previously believed (Wu et al., 2019). An exception to this is the combination of pre-training and disjoint vocabulary (dubbed P/D), which gives near zero alignment of both lexical and syntactic spaces. This suggests that sharing hidden layers is not a sufficient ingredient to adapt a pre-trained model on a new (even if related) language, and that specific techniques should be used when joint training is not a viable option (Wang et al., 2019; Artetxe et al., 2019).

Moving to the transfer of syntactic categories (Fig. 4(b) we find that all cross-lingually trained PoS classifiers (except P/D) perform much better than the majority baseline but notably worse than the corresponding monolingually trained classifiers. As for dependency classification (Fig. 4c), accuracies are low overall and no cross-lingual classifier outperforms the majority baseline. In summary, some form of syntactic transfer indeed occurs, but truly language-agnostic syntactic categories (such as *noun* or *subject*) have not emerged in our mNN representations.

5.2 Training Objective, Number of Input Languages, and Language Relatedness

We now study whether a different training objective, namely translation to a third language (English), leads to more syntactic transfer among input languages. We also check whether number of input languages and language relatedness play a significant role in the sharing of syntactic categories. All models in this section are jointly trained with natural vocabulary overlap on Europarl, and compared to their randomly initialized equivalents following Zhang and Bowman (2018). Dependency classification results are omitted as they were always below the majority baseline.

Learning Objective As shown in Fig. 5(a,b), the translation objective has a slightly negative impact on the alignment of word embedding spaces when all other factors are fixed. The translation objective also leads to lower PoS accuracy (monolingually probed), confirming previous results by Zhang and Bowman (2018). However, translating to English does result in visibly better cross-lingual transfer of PoS categories (mono/cross-lingual drop of -27.7 for translation *vs* -37.2 for

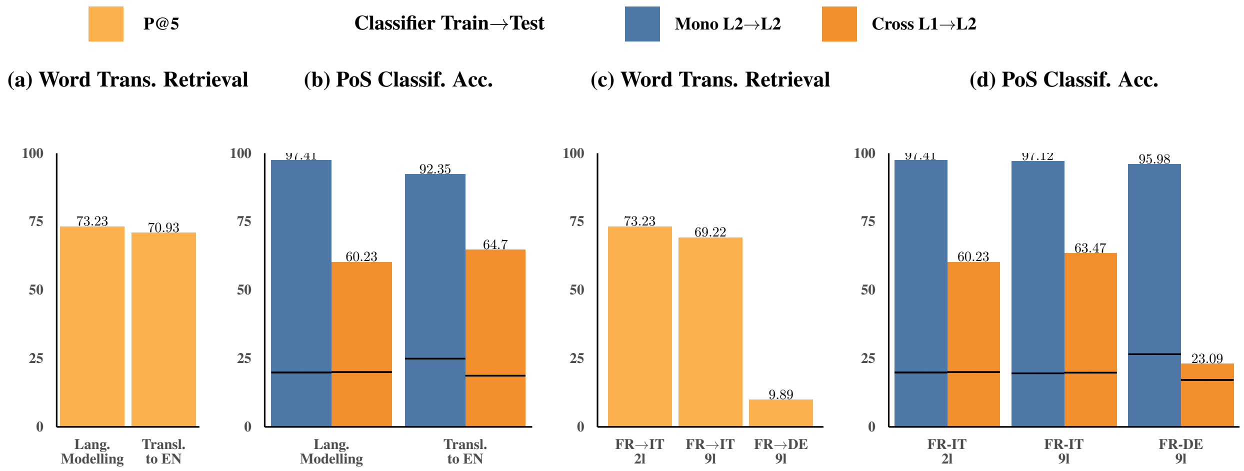

Figure 5: Semantic (word translation retrieval) *vs* syntactic (PoS classif.) transfer in Europarl-based bidirect. mNNs. (a,b) Effect of training objective: next word prediction *vs* translation to English. (c,d) Effect of number of input languages (2 *vs* 9) and language relatedness (FR-IT *vs* FR-DE) for the bidi-LM objective. Horizontal lines (b,d) refer to the corresponding randomly initialized mNNs.

language modelling), showing that what are optimal conditions for lexical-semantic may no be optimal for syntactic transfer.

Number of Source-side Languages For the remaining experiments we look at the (bidirectional) LM objective. As shown in Fig. 5(c,d), moving from 2 input languages to 9 results in lower WTR precision but higher cross-lingual PoS accuracy. This suggests that adding more languages does not cause mNN representations to lose syntactic information and actually leads to more sharing of syntactic categories across languages. The generality of this remark is however restrained by our findings on language relatedness.

Language Relatedness Fig. 5(c,d) also shows that moving from a very related pair of input languages (FR-IT) to a less related one (FR-DE) results in dramatically lower transfer of both lexical-semantics *and* syntactic categories. To substantiate this finding, we extend the analysis of our 9-language LM to more training-test pairs (we select a subset of languages for which a sizeable UD treebank exists). The results in Fig. 6 confirm that, for both lexical-semantics *and* syntax, the related languages FR, IT and ES report considerably higher values than those involving DE, while the smallest drop (-6.45) is seen between FR→FR and FR→IT. While we expected transfer to depend on relatedness, we did not expect the effect to be so large given that DE is not completely unrelated from the Romance languages.

6 Conclusions

We have presented an in-depth analysis of various factors affecting cross-lingual syntactic transfer within multilingually trained LSTM-based language (and translation) models. Our main result is a negative one: Transfer of purely grammatical knowledge (specifically long-range agreement in nonce sentences) is very limited in general – confirming recent findings by Mueller et al. (2020) – and strongly dependent on the specific choice of source-target languages. Namely, small gains were only reported on ES→IT, while a considerable drop was reported on FR→IT and almost no change was reported on UK→RU. When semantic cues were not removed (original sentences), transfer levels were overall higher with a peak of +7% absolute in ES→IT, but FR→IT still suffered a considerable loss (-5%). While ES is arguably closer to IT than FR, we cannot yet find a convincing linguistic explanation for the large differences observed. Our second set of experiments shows that POS categories are shared to a moderate extent, but dependency categories are not shared at all in our models. This suggests that syntactic knowledge transfer within our multilingual models is rather shallow, and may explain the lack of agreement transfer.

Our experiments with different training objectives and number of input languages show that

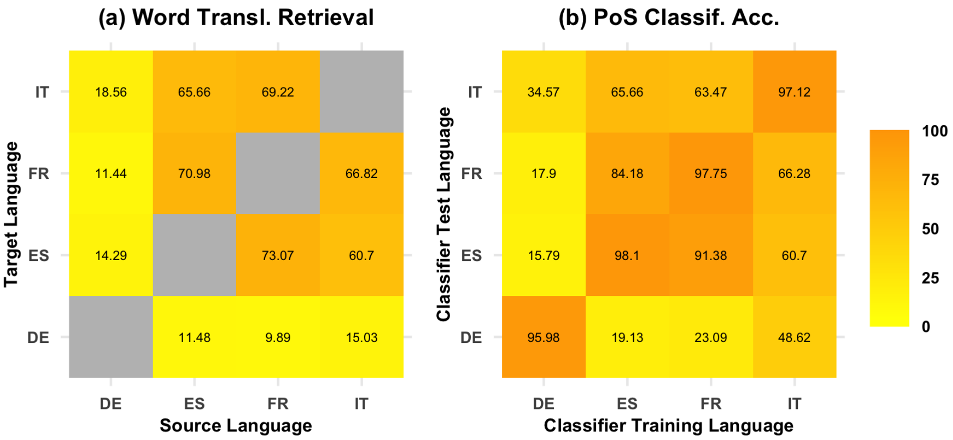

Figure 6: Pairwise semantic and syntactic transfer in the 9-language bidi-LM (a subset of languages is shown). Non-applicable (monolingual) settings in (a) are greyed out. Diagonal values in (b) are scores of monoling. L2→L2 classifiers, while remaining values are for cross-ling. L1→L2 ones.

what are optimal conditions for the alignment of word embedding spaces (lexical-semantic transfer) may not be optimal for syntactic transfer, and vice versa. Language relatedness is by far the most determining factor for both word embedding alignment and POS transfer. And finally, scaling from two languages to a mix of nine languages from three different families results in better POS transfer between related languages but considerably worse between unrelated ones. Together with the findings by Wu et al. (2019), our results suggest that scaling to highly multilingual models may improve syntactic transfer among the most related languages by decreasing the per-language capacity, but may also exacerbate the divergence among less related ones. Thus modern multilingual NNs appear still far from acquiring a true interlingua.

Acknowledgements

Arianna Bisazza was partly funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.021.646.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Joakim Nivre et al. 2019. Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations,* *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Arianna Bisazza and Clara Tump. 2018. The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2871–2876. Association for Computational Linguistics.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Melbourne, Australia. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual bert.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7057–7067. Curran Associates, Inc.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin. 2018. Multilingual BERT Readme Document. `https://github.com/google-research/bert/blob/a9ba4b8d7704c1ae18d1b28c56c0430d41407eb1/multilingual.md`.

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.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Robert J. Hartsuiker, Martin J. Pickering, and Eline Veltkamp. 2004a. Is syntax separate or shared between languages?: Cross-linguistic syntactic priming in spanish-english bilinguals. *Psychological Science*, 15(6):409–414. PMID: 15147495.

Robert J. Hartsuiker, Martin J. Pickering, and Eline Veltkamp. 2004b. Is syntax separate or shared between languages?: Cross-linguistic syntactic priming in spanish-english bilinguals. *Psychological Science*, 15(6):409–414.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Scott Jarvis and Anna Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. Routledge.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Eric Kellerman and ed. Sharwood Smith, Michael. 1986. *Crosslinguistic influence in second language acquisition*. Pergamon.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Gerrit Jan Kootstra, Janet G. Van Hell, and Ton Dijkstra. 2012. Priming of code-switches in sentences: The role of lexical repetition, cognates, and language proficiency. *Bilingualism: Language and Cognition*, 15(4):797–819.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Jindrich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *CoRR*, abs/1911.03310.

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.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Terence Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge Applied Linguistics. Cambridge University Press.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 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.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Kristen M. Tooley and Matthew J. Traxler. 2010. Syntactic priming effects in comprehension: A critical review. *Language and Linguistics Compass*, 4(10):925–937.

Ke Tran and Arianna Bisazza. 2019. Zero-shot dependency parsing with pre-trained multilingual sentence representations. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 281–288, Hong Kong, China. Association for Computational Linguistics.

Marina Vasilyeva, Heidi Waterfall, Perla B. Gámez, Ligia E. Gómez, Edmond Bowers, and Priya Shimpi. 2010. Cross-linguistic syntactic priming in bilingual children. *Journal of Child Language*, 37(5):1047–1064.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5720–5726, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *CoRR*, abs/1911.01464.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1857–1867.