Cognitive Science Comps:

Emergence of Hierarchical Syntax in Neural Machine Translation

Tyler A. Chang

Carleton College

Abstract

We explored the syntactic information encoded implicitly by neural machine translation (NMT) models with different target languages, by training NMT models from English to six target languages: English (an autoencoder), Spanish, French, Arabic, Russian, and Chinese. NMT models consist of a neural network encoder and decoder; we considered the NMT encoder state (a real-valued vector of activation values) as a connectionist model of the machine's internal state after reading a given English word. We used these NMT encoder states to predict ancestor constituent labels of the current word in a syntactic parse tree. Regardless of the NMT encoder's target language, all constituent label prediction models performed well above a per-word most-frequent-label baseline accuracy, suggesting that NMT models implicitly encode hierarchical syntax, despite being trained only on raw sentences. Furthermore, the constituent label prediction models exhibited similar behavior regardless of the NMT target language. Because the NMT models for different target languages exhibited dramatically different translation qualities but still encoded similar syntactic information, our results suggest that NMT models rely heavily on non-syntactic information when producing translations. Finally, we found that NMT models rely on explicit morphosyntactic cues (e.g. infinitives and complementizers) when extracting syntactic features such as embedded clauses and negation. Our results open up new areas of research in linguistic universals, first/second-language acquisition, unsupervised syntactic parsing, and hierarchical structures in connectionist models of cognition.

Keywords: machine translation, neural networks, generative syntax, connectionism, computational linguistics, natural language processing, machine learning

Emergence of Hierarchical Syntax in Neural Machine Translation

Modern machine translation systems use deep neural networks, a class of machine learning algorithms that "learn" to translate phrases and sentences based on large corpora of pre-translated sentences. These neural machine translation (NMT) models have produced state-of-the-art results in both academia and industry; they are currently deployed in large-scale products such as Google Translate (Turovsky, 2016; Wu et al., 2016). NMT systems are built upon artificial neural networks, computer algorithms that are often viewed as implementations of connectionist models of human cognition (McClelland, 2000). While effective, NMT systems are difficult to interpret due to large numbers (often thousands) of interconnected nodes, with learning algorithms automatically adjusting weights between nodes. Thus, previous work has investigated the types of information encoded implicitly within NMT systems, finding that NMT systems encode morphological, syntactic, and semantic information about source language words and sentences (Belinkov, Durrani, Dalvi, Sajjad, & Glass, 2017a; Poliak, Belinkov, Glass, & Van Durme, 2018; Shi, Padhi, & Knight, 2016).

However, there is relatively little work studying information encoded within NMT systems cross-linguistically. Existing cross-linguistic studies have focused either on low level morphological features such as part-of-speech (POS) or on general information content that only identifies broad distinctions between languages (e.g. clustering NMT sentence representations using correlation analysis; Belinkov, Màrquez, Durrani, Dalvi, & Glass, 2017b; Kudugunta, Bapna, Caswell, & Firat, 2019). In this project, we assessed the ability of NMT models with different target languages to predict ancestor constituent labels of words, allowing us to focus specifically on syntactic information but at various levels within each syntax tree. In this way,

we were able to evaluate whether NMT systems implicitly encode syntactic information regardless of target language and whether the types of encoded syntactic information differ across target languages. Notably, our results allow us to evaluate whether connectionist models of language (e.g. neural network models) can consistently encode hierarchical linguistic structures despite their seemingly non-hierarchical network architectures, providing evidence for the capabilities and limitations of connectionist approaches to human cognition.

Connectionism and Artificial Neural Networks

For background, we first provide an overview of connectionist models of cognition, viewing artificial neural networks as connectionist models. Connectionist approaches to human cognition claim that the human mind operates as a system of interconnected units, where each unit can be activated; when activated, units activate neighboring units in the network. In connectionist theories, concepts are represented either by individual units (local representations) or by patterns of activation over sets of units (distributed representations). For example, in a local representation, a unit representing "dog" would be activated upon seeing a dog. Connections between units are developed through some predefined "learning" algorithm (McClelland, 2000). Traditional Hebbian learning proposes that when two units are activated simultaneously, the connection between the two units is strengthened.

The process by which one concept activates nearby concepts (concepts sharing many connections) is called spreading activation. Spreading activation has support from both behavioral and computational studies. Participants are faster to respond to lexical decision tasks (identifying whether a given string is a word) when the presented word is semantically related to the previously-presented word (Neely, 1991). For instance, a participant would respond more

quickly to the word “prince” after seeing the word “boy” than after seeing the word “hat.” This priming effect can be interpreted as one concept (that of the first word) activating related concepts in the participant’s semantic memory, leading to faster reaction times for related words. Similarly, in semantic fluency tasks (listing examples of a given category, such as “animals”), the time between participant-produced words was found to correlate with the mean number of steps between the same words as produced by a computer randomly traversing a semantic network of words (Abbott, Austerweil, & Griffiths, 2012). These results support the theory that human semantic memory functions as a network of interconnected units with activations spreading along the connections.

Artificial neural networks provide additional evidence for the plausibility of connectionist models of cognition. Similar to theoretical connectionist networks, artificial neural networks pass inputs through multiple layers of interconnected nodes to produce outputs. Each node has an activation function, a function that outputs a real number based on the activations of connected nodes in the previous layer; the final output of the neural network is a function of the activations of the nodes in the final layer (see Figure 1 for a visualization of a simple neural network architecture). A neural network that passes a single input directly through a set of layers to generate an output is called a feedforward neural network, and the state of a neural network at any given layer is the vector of real-valued activations of all nodes in the layer. Neural networks have been used successfully in tasks ranging from image classification to machine translation (Schmidhuber, 2015).

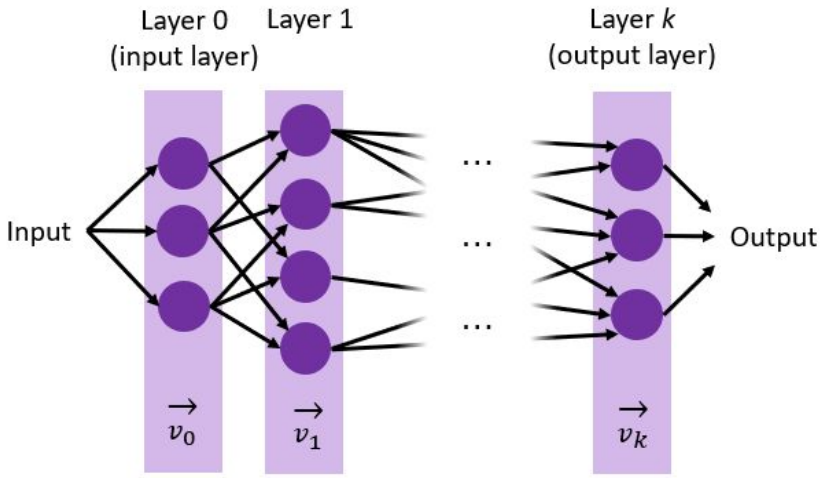


Figure 1. A visualization of a feedforward artificial neural network. Vectors v_i represent the neural network state at each layer.

When training a neural network, connection weights in the network are adjusted using a learning algorithm called backpropagation, which seeks to minimize errors between the network's predicted output and some provided "correct" output (Daumé III, 2012). The neural network's predicted output is computed by passing values forward through the network, and backpropagation propagates any errors backwards through the network layers to adjust individual connection weights between nodes. Unfortunately, at an algorithmic level, backpropagation is not neurally plausible. Backpropagation requires each neuron (node) to emit two signals: an output (passing activation values forward through the network) and an error (propagating error values backwards through the network); biological neurons have only one known mechanism for transmitting information, by sending an electrical signal called an action potential to connected neurons (Balduzzi, Vanchinathan, & Buhmann, 2015). That said, several biologically plausible error-propagation algorithms have been proposed, computing weight adjustments directly from the global error rather than recursively propagating local errors

backwards through the neural network (Balduzzi et al., 2015; Bengio, Lee, Bornschein, Mesnard, & Lin, 2015). At a cognitive level, error-propagation models are common in predictive processing models of cognition, which propose that people's senses are used only to verify or contradict our predictions about the world (Clark, 2015). According to predictive processing models, feedback from our senses allows us to constantly update our internal model of the world, analogous to updating connection weights when using backpropagation in neural networks.

Neural Machine Translation (NMT)

Then, NMT systems can be viewed as connectionist models of language. NMT systems generally use an encoder-decoder framework consisting of two neural network models: the encoder, which forms a representation of the input sentence, and the decoder, which translates that representation into a sentence in the target language (Sutskever, Vinyals, & Le, 2014). Both the encoder and decoder often use a type of neural network called a recurrent neural network. Recurrent neural networks (RNNs) allow each word in a sentence to be inputted consecutively, using the current input word and the previous output state to generate a new output state. In this way, the RNN state is updated incrementally for each input word.

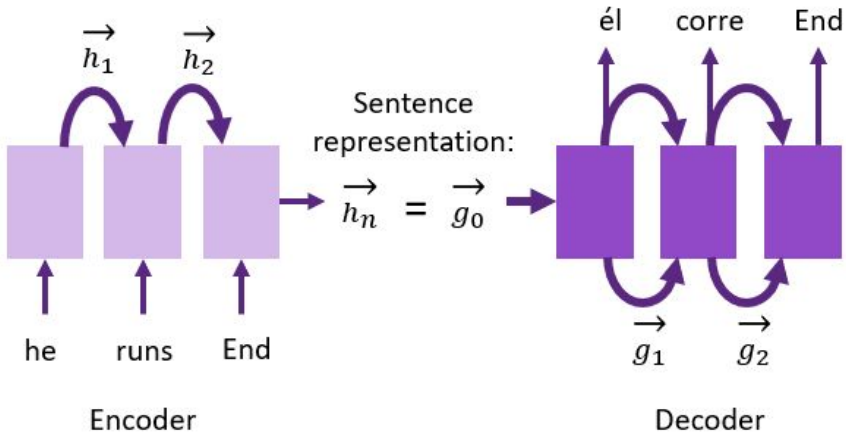


Figure 2. A visualization of a simple RNN encoder-decoder NMT framework, translating the English sentence "he runs" to the Spanish sentence "él corre." Vectors h_i represent encoder states, and vectors g_i represent decoder states.

As shown in Figure 2, the NMT encoder outputs the final RNN state once it reaches the end of the input sentence; this state is considered a vectorized representation of the entire input sentence. The decoder then uses this final encoder state as its initial state. The decoder generates a sentence in the target language by repeatedly producing words based on the previously-generated word and the previous decoder state. The encoder-decoder framework can then be seen as a connectionist model of human language perception and production. The NMT encoder "perceives" language by converting a sentence into a real-valued vector (the sentence representation), and the decoder "produces" a sentence in the target language based on the original sentence's abstract vector representation. Each intermediate encoder and decoder state vector represents a set of activation values over a set of nodes, simulating a connectionist network that updates node activations for each input or output word.

Attention and Working Memory in NMT Models

Building upon the simple RNN encoder-decoder framework, many modern NMT models use an attention mechanism, a computational and connectionist analog of human attention. Specifically, NMT attention frameworks allow the decoder to focus upon specific words in the source sentence when generating each decoder state (Luong, Pham, & Manning, 2015). As shown in Figure 3, this attention mechanism is implemented in the decoder as an added context vector defined as a weighted sum of the states of the encoder at each word; weights are determined by the similarity of the encoder state at the source word to the decoder state at the

target word.[1] Described intuitively, source words contribute more heavily to the next output word if their meaning is similar to the current state of the decoder.

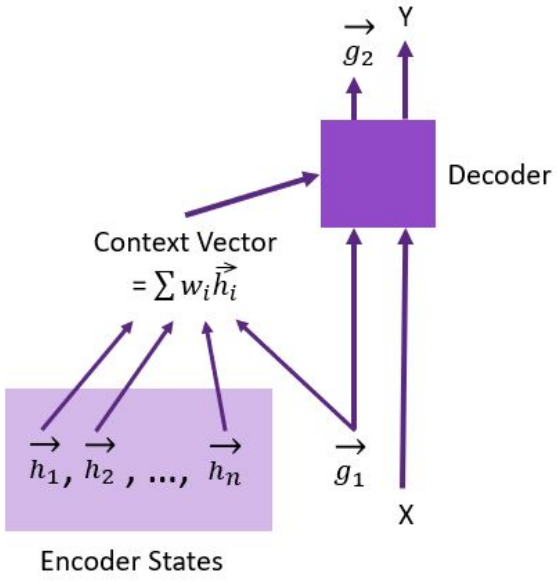


Figure 3. A visualization of the decoder framework with attention, generating one translated word "Y" given the previous translated word "X." This process is repeated for each subsequent target word. Vectors h_i represent encoder states, and vectors g_i represent decoder states.

Word-by-word attention mechanisms in language processing are supported by behavioral studies, although these studies generally do not approach attention from an explicitly connectionist point of view. It has been shown that participants are better able to recall a target word in a sentence when the word is a semantic focus of the sentence, implying that words are

[1] Similarity here is defined as the dot product between the encoder and decoder state vectors. Encoder state weights can be determined by other functions between the encoder and decoder state vectors, but the dot product has been found to perform equally well as or better than more generalized functions (Luong et al., 2015).

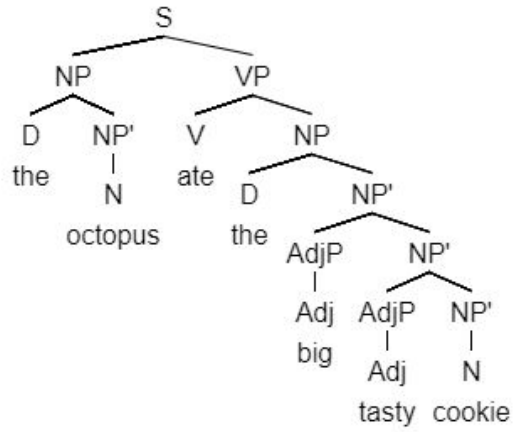
not weighted equally when processing language (Osaka, Nishizaki, Komori, & Osaka, 2002). These results are traditionally explained by theories of working memory, which have found that people can only store about seven items in working memory at any given time (Miller, 1956). It is possible that only the words most relevant to meaning are stored in working memory. Higher working memory capacity has been shown to correlate with performance on tasks such as ambiguous word resolution and with reading comprehension in young second-language-learning students, suggesting that working memory plays an active role in language comprehension (Chang, Wang, Cai, & Wang, 2019; Daneman & Carpenter, 1980).

NMT models with attention can then be viewed as connectionist models of attention and working memory in language tasks. Attention is represented by weighting the encoder states corresponding with each source word; a higher weight for a given word state is analogous to a greater amount of attention given to that word. Each summed context vector is comparable to a mental representation of the source sentence, after factoring in human attentional processes. Similar to the human limitations of working memory, NMT models can use a local attention framework that limits the number of source words that can be considered when generating the weighted context vector (Luong et al., 2015).

Linguistic Syntax and Connectionism

Because this project studies the emergence of linguistic syntax in NMT models, we now turn to a discussion of hierarchical linguistic syntax in connectionist models of language. Because connectionist models emphasize non-hierarchical connections between units, connectionist approaches initially seem at odds with hierarchical linguistic syntax, which models structured relationships between words in sentences. In particular, generative syntax posits that

sentences are generated by grammatical rules, such as a verb phrase (VP) becoming a verb plus a noun phrase (NP) in the predicate "eat the cookie" (Adger, 2015; Carnie, 2013, p. 6).[2] Generative syntax often results in hierarchical tree structures such as in Figure 4. Sentences can be parsed into constituents, strings consisting of all children of some parent node in the tree. Evidence for constituent structure in language can be found in linguistic data; sentence constituents can be nested within one another as relative clauses, as in "The canary (that the cat (that grinned) ate) sang" (Catania, 1972). Verbs agree with the head of the subject noun phrase constituent rather than the nearest noun according to linear distance, as in "The girls from Paris are singing," where "are" agrees in number with "girls" instead of "Paris" (in contrast, the sentence "The girls from Paris is singing" is ungrammatical; Adger, 2015).



[2] Phrases such as "the cookie" are generally represented as determiner phrases (DPs) under the X-bar theory of syntax, which is the focus of Carnie (2013). For the purposes of this paper, simpler grammatical rules will be considered, omitting features such as the voice phrase (little *v*P), the determiner phrase (DP), and many bar levels. The remaining simpler grammatical rules are used in most natural language processing parse tree datasets.

Figure 4. A syntax tree using generative grammar.

Then, generative syntax may seem incompatible with connectionist approaches to cognition. Connectionist approaches emphasize spreading activation with non-hierarchical relationships between units, while generative syntax posits hierarchical structures inherent to human language. Schonbein (2012) formalizes this discrepancy as the thesis of linguistic structuring, which states that any successful form of language processing must recapitulate the hierarchical structures of language. Under the thesis of linguistic structuring, connectionist models of language would need to encode hierarchical structures of language.

Previous studies have shown that connectionist models of language can indeed adopt hierarchical internal structures. Artificial neural networks have been found to separate their internal state space to reflect syntactic categories of the target language when recognizing formal languages (Schonbein, 2012). Furthermore, RNN-based language models of English, which predict the next word in a piece of text, have been shown to store whether the current constituent state is a subordinate or main clause (Futrell, Wilcox, Morita, Ballesteros, & Levy, 2019). In both studies, the neural network models were not explicitly provided with hierarchical structures, suggesting that connectionist models can learn hierarchical syntax based on sequential inputs alone.

However, there is contrary evidence that these connectionist models are simply using syntax heuristics to produce the expected hierarchical results. For instance, the subsequence heuristic assumes that the truth of any coherent subsequence within a sentence is implied by the truth of the original sentence. The subsequence heuristic is typically accurate for English, but there are many sentences for which the heuristic does not work. For example, the sentence “The

octopus near the dog swam" fails the subsequence heuristic because the subsequence "the dog swam" is not implied by the original sentence. High performance English natural language inference models (identifying whether one sentence implies another) have been found to perform poorly on sentences that fail the subsequence heuristic (McCoy, Pavlick, & Linzen, 2019). Similarly, RNN-based language models of English perform poorly on datasets involving structure-sensitive linguistic phenomena such as long-distance subject-verb agreement and reflexive anaphora (Marvin & Linzen, 2018). These results suggest that neural network models are able to employ syntactic heuristics that lead to high performance on common sentences but poor performance on syntactically complex sentences. Thus it is unclear whether connectionist models of human language can fully reconstruct the hierarchical structures proposed by generative syntax.

Linguistic Syntax in NMT

The studies discussed previously find evidence and counter-evidence for generative syntax in neural network models of language by assessing the models' performance on syntax-related language tasks. Recent studies in NMT have tried to identify encoded syntax by directly assessing representations of source sentences (recall that the encoder updates its internal state vector after each input word). These studies often compare the information contained within encoder state vectors (after reading a given word) with the information contained in a standard word embedding, a commonly-used vector representation of the word in natural language processing tasks. For instance, using an NMT model with attention from English to multiple other languages, Belinkov et al. (2017b) found that the part of speech (POS) tag of a given source word could be predicted more accurately from the encoder state vector after reading

the word than from a standard word embedding. This implies that NMT encoder states contain useful information about the context around a given word, disambiguating word senses such as "play" (verb) versus "play" (noun).

Similar studies have found evidence for more global syntactic properties encoded within NMT models. Shi et al. (2016) found that voice (active or passive), tense, and top level syntactic sequence (overall sentence structure) could be predicted from the final encoder state in NMT models without attention. For example, the sentence "On Tuesdays, the dogs play" has active voice, present tense, and top level syntactic sequence PP-NP-VP (prepositional phrase-noun phrase-verb phrase). Building on these results, Blevins, Levy, and Zettlemoyer (2018) found that a word's parent, grandparent, and great-grandparent constituent label in a syntactic parse tree could be predicted from the state of an NMT encoder after reading the word. These results indicate that NMT models encode hierarchical syntactic information, suggesting that connectionist models of language may implicitly recreate hierarchical linguistic syntax.

The current project seeks to clarify the types of syntax present in NMT encoder states, identifying whether NMT systems encode syntactic information regardless of target language and whether the encoded syntactic information differs across target languages. Similar to Blevins et al. (2018), we tested whether a word's part-of-speech, parent, grandparent, and great-grandparent constituent label could be predicted from NMT encoder states after reading the word. Extending on the previous work, we trained NMT models from English to a variety of target languages. Belinkov et al. (2017b) found small but statistically significant differences in POS tag accuracy depending on an NMT model's target language. Training from an English source to six target languages (including one English-to-English autoencoder model), POS tag

accuracy followed a decreasing trend in the order: Spanish / French, Arabic, Russian, Chinese, then English. Interestingly, the ordering between Spanish and French changed depending on NMT training dataset size. The larger NMT training dataset resulted in smaller overall differences in POS tag accuracy between target language encoder states, suggesting that NMT target language may have little effect on the implicit encoding of syntactic information. Additionally, POS tag accuracy corresponded only loosely with overall translation quality, which followed the decreasing trend: English, Spanish, French, Russian, Arabic, then Chinese (Belinkov et al., 2017b). These results indicate that NMT models do not rely heavily on syntactic information such as POS tags. For instance, an English-to-English model is likely able to translate linearly word-by-word without encoding any syntax.

In the current project (predicting part-of-speech, parent, grandparent, and great-grandparent constituent labels given NMT encoder states), we hypothesized that constituent label prediction accuracy scores would significantly exceed the baseline (most frequent label given the input word) accuracy regardless of the NMT model's target language, replicating results from English-to-German NMT models (Blevins et al., 2018). There is little research on the encoding of hierarchical syntax in NMT models cross-linguistically, so we assumed that encoding POS serves as an important precursor to encoding hierarchical syntax. Therefore, between target languages, we predicted approximately the same decreasing trend in accuracy as was found for POS tag accuracy in Belinkov et al. (2017b): Spanish / French, Arabic, Russian, Chinese, then English. However, based on small effect sizes and inconsistent orderings between target languages depending on NMT dataset size, we hypothesized that there

would only be relatively small differences between target languages in constituent label prediction accuracy scores.

Method

We trained NMT models from English to six target languages: English (an autoencoder), Spanish, French, Russian, Arabic, and Chinese. We then trained simple feedforward neural networks to predict ancestor constituent labels (POS, parent, grandparent, and great-grandparent) of words, given the NMT encoder state after reading the word. We used constituent label prediction accuracy scores to measure the amount of syntactic information encoded by NMT models with different target languages.

Datasets and Computation Resources

NMT models were trained on the United Nations (UN) Parallel Corpus, using the fully aligned subcorpus of approximately 11 million sentences from UN documents translated to all six UN official languages: English, Spanish, French, Russian, Arabic, and Chinese (Ziemski, Junczys-Dowmunt, & Pouliquen, 2016). For the UN dataset, we used the provided separation into training, evaluation, and test datasets.

Syntax evaluation models were trained on tree-parsed sentences from the CoNLL-2012 dataset containing sentences from English news and magazine articles, web data, and transcribed conversational speech (Pradhan, Moschitti, Xue, Uryupina, & Zhang, 2012). As in Blevins et al. (2018), syntax evaluation models were trained on the CoNLL-2012 development dataset and tested on the test dataset. A subset of the CoNLL-2012 training dataset was used as an evaluation dataset; the training, evaluation, and test datasets each contained approximately

160,000 English words.[3] All NMT and syntax evaluation models were trained using Google Colab, which provides one free NVIDIA Tesla P100 graphics processing unit (GPU) to improve computation speed.

NMT Models

NMT models were trained using OpenNMT's PyTorch implementation (Klein et al., 2017). Because not all of the target languages mark separations between tokens (words) using space characters, we used OpenNMT's implementation of byte pair encoding for subword tokenization in all languages.[4] This method identifies tokens by initially separating all sentences in the dataset for the given language into individual characters, then iteratively combining tokens that often occur together (Sennrich, Haddow, & Birch, 2016). The resulting tokens are used as "words" in the given language. Byte pair encoding has been shown to improve translation quality in NMT models (Sennrich et al., 2016).

Each NMT model used a recurrent neural network (RNN) encoder-decoder framework with attention, as described in the introduction. Following the methodology in Belinkov et al. (2017b) and Blevins et al. (2018), each NMT encoder and decoder consisted of a four-layer RNN, where each layer was a 500-dimensional long short-term memory (LSTM) layer. An LSTM layer is a common variant of a simple RNN hidden layer that is designed to recognize long-distance dependencies in sequential data by encoding a cell state (encoding long-term

[3] The evaluation dataset was obtained by taking every eight sentences of the training dataset.

[4] Due to constraints on RAM and virtual memory, byte pair encoding for Chinese was based on a subset (approximately 1.8 million sentences) of the Chinese sentences in the UN training dataset. All other languages used the entire UN training dataset for byte pair encoding.

information) as well as the usual RNN hidden state (Hochreiter & Schmidhuber, 1997; Neubig, 2017). Our NMT models used dot-product global attention as described in the introduction.

Each NMT model was trained for 11 epochs (approximately 2,000,000 steps) using the Adam optimization algorithm (Kingma & Ba, 2015). The first 10 epochs used a learning rate of 0.0002; the learning rate was halved every 30,000 steps during the final epoch.[5] The model with the best performance on the evaluation dataset was used to generate syntax evaluation models. Each NMT model was evaluated on the evaluation dataset every 1000 steps, or every 64,000 sentences.

For reference, we also computed the BLEU score for each NMT model. BLEU scores are commonly used in machine translation to measure how well a predicted translated sentence matches a provided reference translation; BLEU scores have been shown to correlate highly with human evaluations of translation quality (Papineni, Roukos, Ward, & Zhu, 2002). BLEU scores were computed using the UN Parallel Corpus test set, providing a general metric for the translation quality of each NMT model.

Syntax Evaluation

The part-of-speech (POS), parent, grandparent, and great-grandparent constituent label for each word in the CoNLL-2012 dataset was predicted based on the state of the deepest LSTM layer in the NMT encoder after reading the given word. Note that regardless of target language, constituent label predictions used sentences in English (the source language). We made constituent label predictions based on the deepest encoder layer because deeper layers have been

[5] A neural network's learning rate governs the amount that weights between nodes are changed for each step when training the network.

shown to perform better on constituent label prediction tasks (Blevins et al., 2018). Furthermore, it is the deepest encoder layer that is sent to the decoder to produce a sentence in the target language. Finally, both the hidden and cell states of the deepest LSTM layer were used to make constituent label predictions, even though only the hidden state of the LSTM is passed to the decoder in NMT. The cell state is still used to generate future hidden states, and the cell state is designed to encode long-distance dependencies in sequences (sentences). It then seems plausible that the cell state would encode syntactic information, so our constituent label prediction models received both hidden and cell states as input, resulting in input vectors of length 1,000 (500 entries from each state).

As in Blevins et al. (2018) and Belinkov et al. (2017b) , we used a simple feedforward neural network to make constituent label predictions. Each feedforward network contained only one hidden layer with 500 nodes. Constituent label prediction models of this type were trained for each type of label prediction (POS, parent, grandparent, and great-grandparent) and for each NMT model, where each NMT model was trained on a different target language. Each constituent label prediction model was trained until convergence, where convergence was defined as 10 consecutive epochs with no improvement on the evaluation dataset. To account for variation between constituent label prediction models based on random initialization of weights and shuffling of the training data, we trained 20 constituent label prediction models for each combination of label type and NMT model. For each constituent label prediction model, we recorded constituent label prediction accuracy on the CoNLL-2012 test dataset.

We computed a baseline score for each type of constituent label prediction by simply predicting the most frequent label given the current input word (e.g. given the current input word

"dog," the most frequent POS tag would be "noun"). This baseline score is the maximum possible accuracy for a deterministic model that only knows the current input word.

Results

Comparison to the Baseline

We first assessed whether constituent label predictions using representations from each NMT model were significantly different from the baseline accuracy score. For each combination of NMT model (trained towards a given target language) and constituent label prediction type (POS, parent, grandparent, or great-grandparent), we conducted a one sample *t*-test comparing the baseline accuracy to the mean accuracy of the 20 constituent label prediction models of the desired type. We adjusted our *p*-value using the Bonferroni adjustment for 24 comparisons, one comparison for each combination of target language and constituent label prediction type. The mean constituent label prediction accuracy was significantly different from the baseline for all combinations of target language and constituent label prediction type (adjusted $p < 0.001$ for all comparisons; see Figure 5 for mean accuracy scores compared to the baselines).

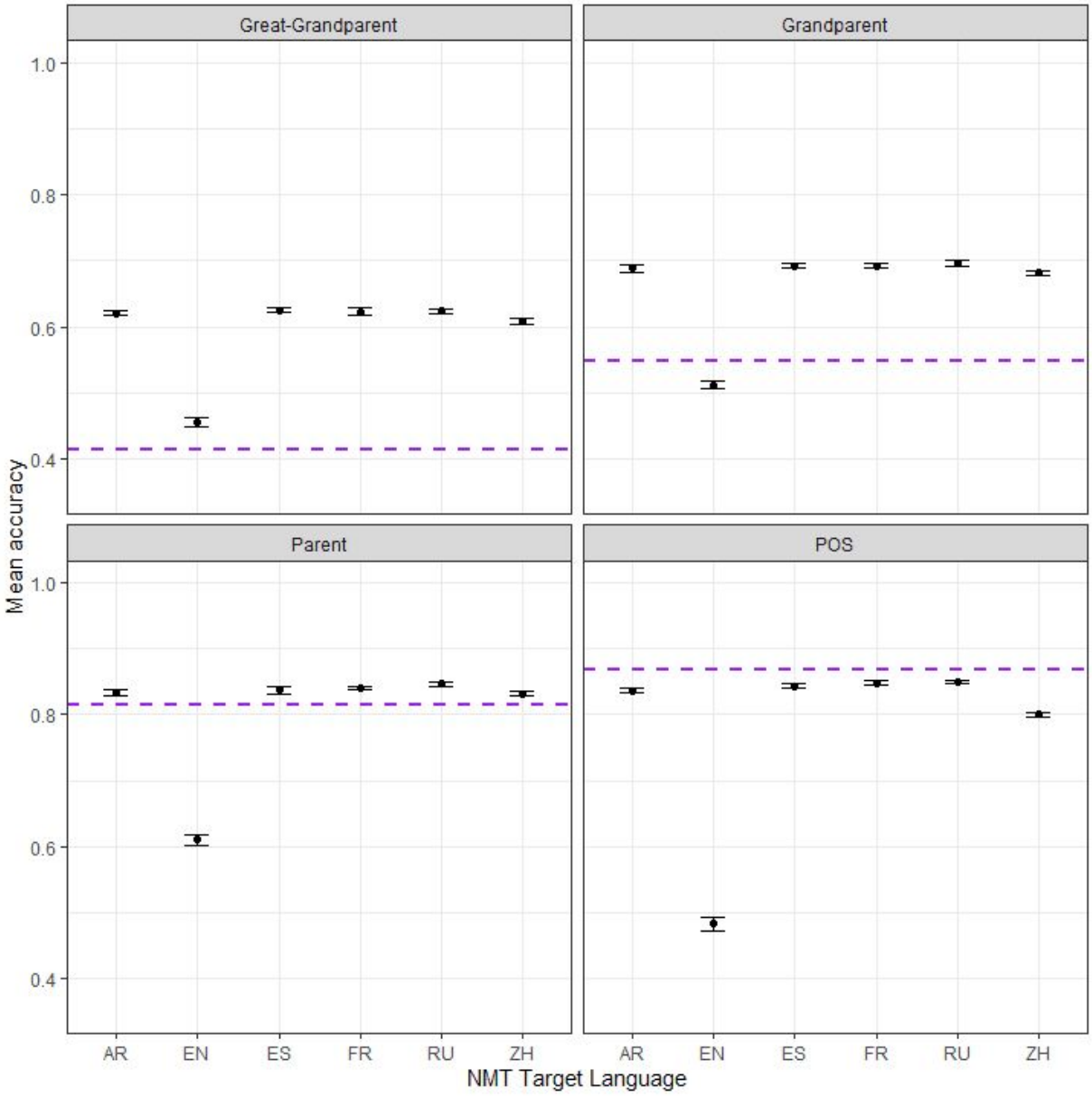


Figure 5. Mean accuracy scores (based on 20 prediction models) with two standard deviations from the mean for the constituent label prediction tasks using representations from different NMT models. Dashed lines represent baseline accuracy scores. NMT models were trained towards six different target languages: Arabic (AR), English (EN; the autoencoder), Spanish (ES), French (FR), Russian (RU), and Chinese (ZH).

As hypothesized, all target languages aside from English resulted in accuracy scores well above the baseline for the grandparent and great-grandparent tasks. All non-English target languages also performed slightly above the baseline for the parent task. In contrast to Blevins et al. (2018) but in line with Belinkov et al. (2017b), all target languages performed slightly below the baseline for the POS task. This result for POS may be because POS encodes less useful information for tasks such as machine translation; for instance, Belinkov et al. (2017b) found that models performed above the baseline if the task was modified to use semantic tags, a variant of POS that separates words into classes based on semantic rather than syntactic function. Furthermore, it should be noted that the baseline score is extremely high (~87%) for the POS prediction task. Because the baseline score uses the majority label given the current word, the baseline model essentially knows which of the ~14,000 possible input words is the current word; each NMT representation is simply a 1000-dimensional vector representing the current encoder state. The below-baseline POS accuracy scores could indicate that precise information about the current word is not encoded in each encoder state. That said, high accuracy scores on the other prediction tasks indicate that the encoder states do encode syntactic information.

Differences Between Target Languages

Next, we assessed differences in constituent label prediction accuracy across NMT models with different target languages. Assessing these differences required examining how variation between target languages compared to variation within different models trained on the same target language. We used a one-way ANOVA for each constituent label prediction type (POS, parent, grandparent, and great-grandparent) to account for variation within models trained on the same target language. While other work such as Belinkov et al. (2017b) used the

approximate randomization test to evaluate differences between models, pilot results suggested that one-way ANOVAs were more appropriate; see Appendix A for details (Padó, 2006). Due to time and computing resource constraints, we could only train one NMT model for each target language, and thus we were unable to account for variance arising from different NMT models trained towards the same target language. In other words, while we could account for variation between constituent label prediction models trained on the same NMT model, small differences between languages might be due to natural variation between NMT models trained on the same dataset.

The one-way ANOVAs found significant differences between target languages for all four constituent label prediction tasks ($p < 0.001$ for all four label prediction types). We used Tukey's HSD post-hoc test to identify language pairs that differed significantly. Only seven out of the 60 pairwise comparisons did not differ significantly (see Appendix B for mean accuracy scores and all reported significance levels between languages). As predicted, English performed poorly on all four tasks (15–30% lower accuracy scores than all other target languages), supporting the hypothesis that English-to-English autoencoders do not rely heavily on syntactic information. Consistent with Belinkov et al. (2017b), Chinese consistently performed worse than the other non-English target languages on all four tasks. Loosely, non-English target languages tended to perform better on constituent label prediction tasks when the target language was more similar to English; for instance, Spanish and French performed better than Chinese and Arabic on all tasks.

However, while there were significant differences between target languages on all four constituent label prediction tasks, the non-English target languages exhibited surprisingly similar

accuracy scores, particularly considering the wide range of BLEU scores (see Table 1 for BLEU scores for each target language). The non-English target languages varied by less than 2% within each of the parent, grandparent, and great-grandparent prediction tasks. Non-significant pairwise differences between target languages were unpredictable across tasks; for instance, in the great-grandparent task, French did not perform significantly differently from Russian or Arabic, and Russian did not perform significantly differently from Spanish. These results indicate that the NMT models might encode similar syntactic information regardless of target language. Even though significant differences between target languages often exist (which is unsurprising given that the languages are indeed different), effect sizes based on target language are small.

Table 1

BLEU Scores

	AR	EN	ES	FR	RU	ZH
BLEU	37.3	99.9	56.3	44.8	37.8	24.9
Detokenized BLEU	38.0	100.0	56.3	44.5	37.4	

Note. BLEU scores were computed both before and after detokenizing the predicted translations.[6] Before detokenization, each token is treated as a separate word. BLEU scores are on a 0–100 scale. Differences from the raw BLEU scores reported by Belinkov et al. (2017b) are likely due

[6] The detokenized BLEU score was not computed for Chinese because words were generally not separated by spaces in the Chinese dataset. After detokenization, Chinese word boundaries could not be identified.

to the byte pair encoding methodology used for subword tokenization in this study (see Method section).

Similarities Between Target Languages

To further test the hypothesis of similar syntactic information encoded across target languages, we considered performance when predicting individual constituent labels (e.g. noun phrase) for each NMT model, considering the constituent label predictions as the result of binary classification tasks. For instance, when considering the noun POS tag, all POS tags would be separated into two categories: noun and not noun. This operation (converting into a binary task) was performed on the existing constituent label predictions; new constituent label prediction models were not trained specifically for each binary classification task. We computed F1 scores on each individual constituent label for each NMT model. F1 scores are the harmonic mean of precision (fewer false positives) and recall (more true positives correctly identified). If NMT models encoded similar syntactic information regardless of target language, then we would expect similar F1 scores for individual constituent labels in addition to the similar overall accuracy scores observed already.

Indeed, individual constituent label F1 scores correlated extremely highly between non-English target languages (all pairwise Pearson correlations $r[47] > 0.93$, $p < 0.001$ for the POS task; $r[27] > 0.98$, $p < 0.001$ for the parent task; $r[25]$, $r[22] > 0.99$, $p < 0.001$ for the grandparent and great-grandparent tasks). In other words, the models performed well or poorly on the same individual labels regardless of target language. Figure 6 shows mean F1 scores for the five most common labels in each constituent label prediction task and for each NMT model. Significant differences in F1 scores were found between non-English target languages for nearly

every constituent label displayed in the figure (using one-way ANOVAs and the Bonferroni adjustment for 20 comparisons; see Figure 6 for significance levels for individual constituent labels). However, similar to the overall accuracy scores, effect sizes were small between non-English target languages, and post-hoc Tukey's HSD tests indicated no clear trends for which pairwise comparisons were significant.

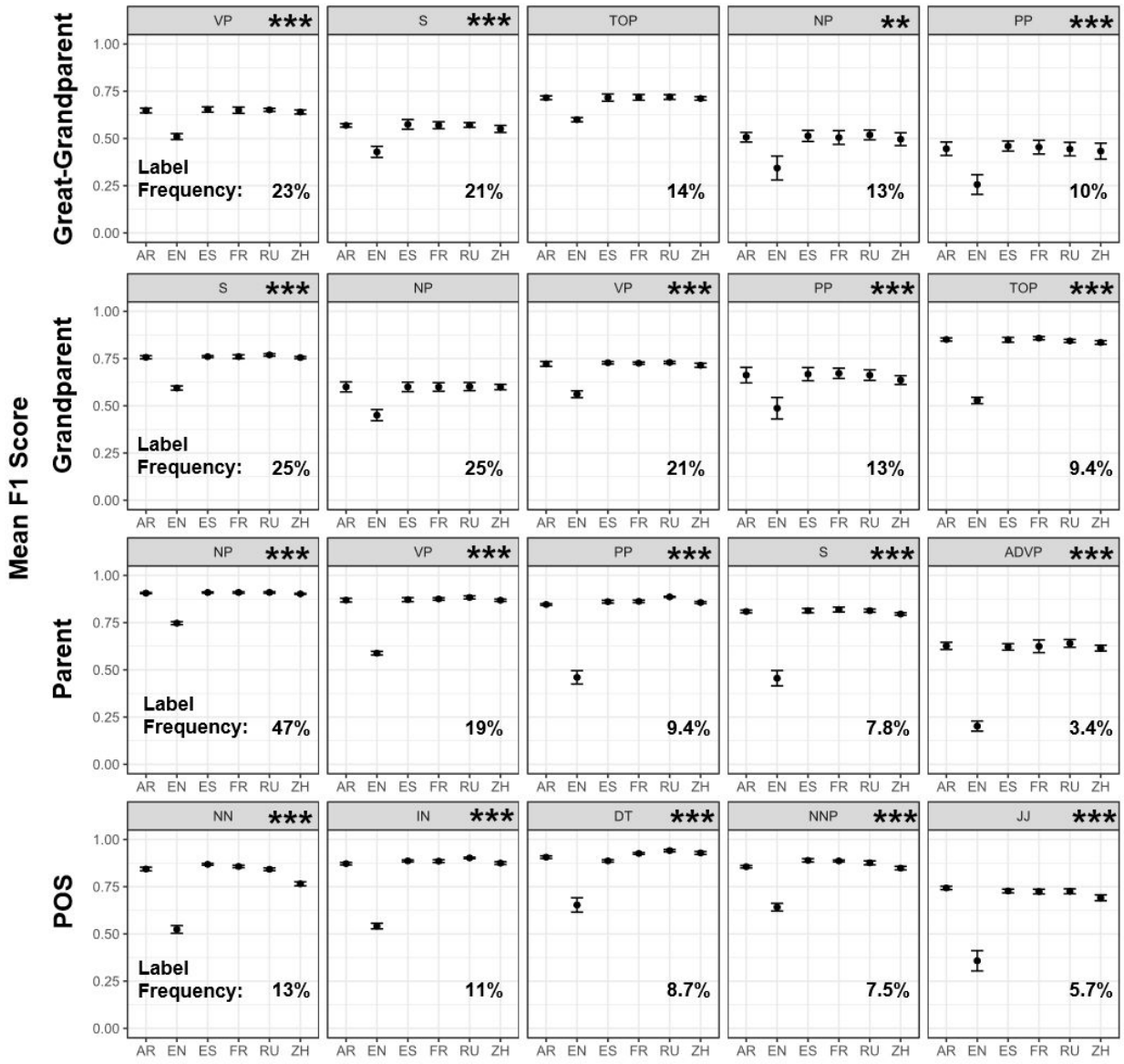


Figure 6. Mean F1 scores (based on 20 models) for individual constituent label predictions, treating the constituent label prediction as a binary classification task. Bars indicate two standard deviations from the mean. Rows indicate different constituent label prediction tasks (POS, parent, grandparent, and great-grandparent). Columns indicate different individual constituent labels, sorted by decreasing frequency in the CoNLL-2012 test set; the five most frequent labels are displayed in the figure. Each label's frequency in the test set is displayed on its corresponding plot. Asterisks indicate significant differences between non-English target languages (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$).

In particular, the similar F1 scores did not appear to simply be the result of different label frequencies; there was only a loose relationship between F1 scores and label frequencies. For instance, all non-English target languages performed similarly well when identifying noun phrase grandparent labels (25% of grandparent labels, F1 scores 0.59–0.60) and question-sentence grandparent labels (0.6% of grandparent labels, F1 scores 0.55–0.61), despite over a 20% difference in corresponding label frequencies. The models performed well on several rare constituent labels, such as WH-prepositional phrase grandparent labels (0.04% of grandparent labels, F1 scores 0.80–0.94). Mean F1 scores (averaged across non-English target languages) for all constituent labels and corresponding label frequencies are plotted in Figure 7. Because the correlation between F1 scores and label frequencies was only loose, effects of label frequency were not sufficient to explain the similarity of F1 scores across non-English target languages. These results support the hypothesis that while NMT models trained towards different non-English target languages might encode statistically-significantly different syntactic information, these NMT models still generally encode very similar syntactic information.

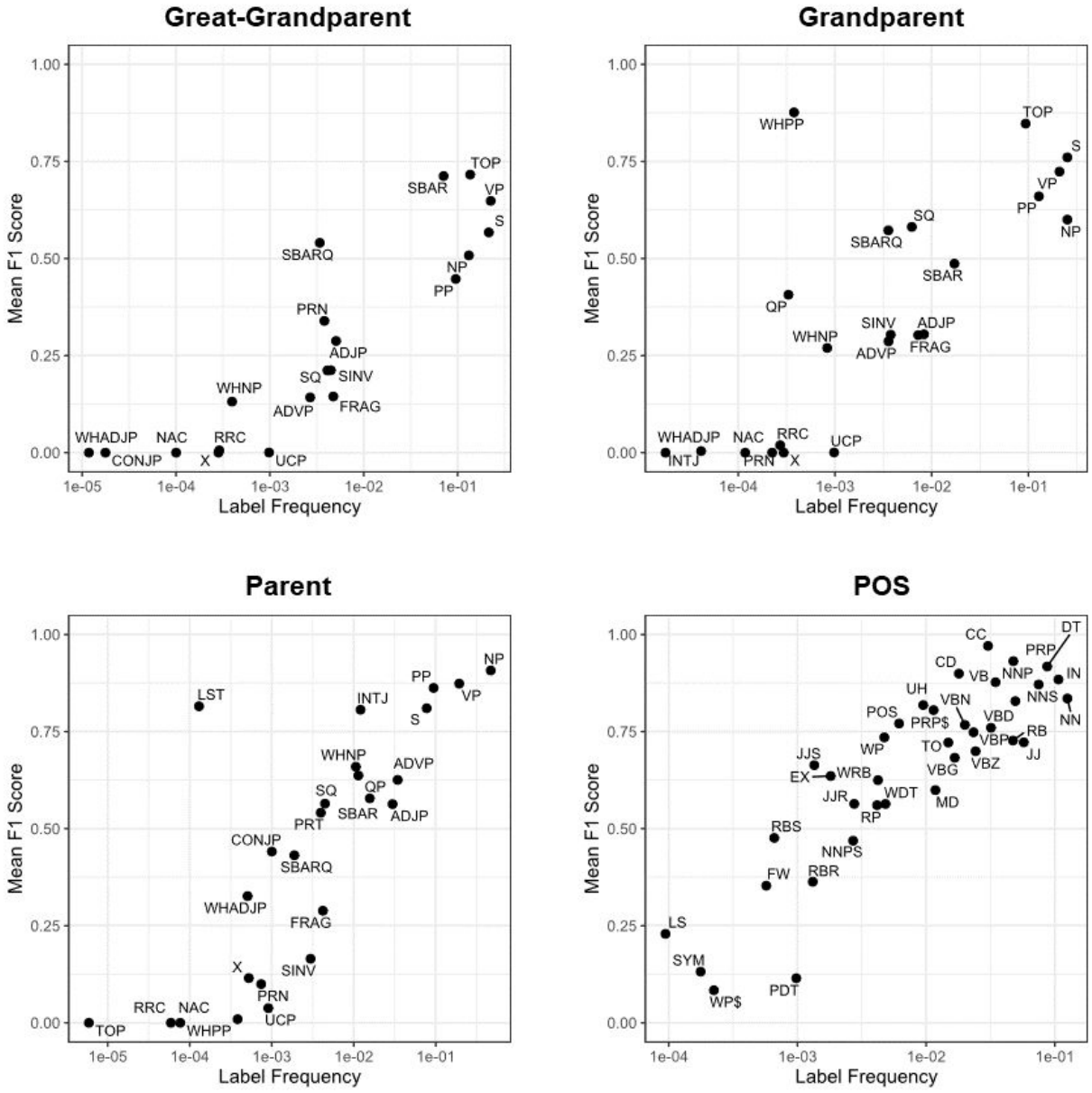


Figure 7. Mean F1 scores for individual constituent labels (averaged across non-English target languages) compared to label frequencies. Each point represents an individual constituent label.

Qualitative Analysis of Syntax Errors

Because it seems that NMT models encode similar syntactic information regardless of target language, it would be helpful to gain a better understanding of the types of sentences for

which the NMT models encoded a low amount of syntax. To do this, we turn to a qualitative analysis of sentences for which the great-grandparent constituent label prediction models exhibited high error rates. We selected the great-grandparent constituent label prediction task because prediction models had the highest accuracy scores above the baseline for this task, indicating a large amount of learned syntax. When computing average accuracy for each sentence (average of the great-grandparent constituent label accuracy scores for each word in the sentence), we took the average across the five non-English target languages; we used this average because there were high pairwise correlation scores for per-sentence accuracy between all non-English target languages (all correlations $r[9472] > 0.88$, $p < 0.001$; see Appendix C for the computed Pearson correlations in per-sentence accuracy scores for each pair of non-English target languages).

Then, we considered the 50 sentences with the highest average great-grandparent accuracy scores and the 50 sentences with the lowest average great-grandparent accuracy scores. When identifying these sentences, we counted only sentences that we considered "complete" sentences, and we considered only sentences from the written news and magazine sources in the CoNLL-2012 dataset (e.g. we excluded transcribed conversational speech, to maximize similarity with the United Nations NMT training dataset).[7] The top 50 sentences all had average great-grandparent constituent label accuracy scores above 90%, and the bottom 50 sentences all had scores below 35%. Sample sentences from the top and bottom 50 sentences are shown in Table 2. Linguistic patterns found in the top and bottom 50 sentences are compiled in Table 3.

[7] All complete sentence judgments were made by native speakers of English. Excluded "sentences" included page numbers, document headers, and some titles of news articles.

Table 2

Sample Sentences from the Top and Bottom 50 Sentences

Sample Top 50 Sentences	**Mean Great-Grandparent Accuracy**
The Justice Department announced that the FBI has been given the authority to seize U.S. fugitives overseas without the permission of foreign governments.	0.933
If these riches didn't exist, people would not be able to describe the course of history.	0.944
Southmark said [that] it plans to amend its 10K to provide financial results as soon as its audit is completed.	0.947
If you want to open a plant on the mainland, you're going to have to sacrifice family intimacy.	0.954
Of course, stories don't have to be old to be interesting.	0.964

Sample Bottom 50 Sentences	**Mean Great-Grandparent Accuracy**
What is so remarkable about the enterovirus detection chip when [it is] compared with traditional test methods?	0.279
"Applications and commercialization are our goal," says Johnsee Lee.	0.281
Canadian crude production averaged about 1.69 million barrels a day during 1989's first half, [which is] about 1% below the 1988 level.	0.281
How has Dongguan, [which is] this rising economic star, done it?	0.313
NASA pronounced [that] the space shuttle Atlantis [is] ready for launch tomorrow following a five-day postponement of the flight because of a faulty engine computer.	0.347

Note. Mean great-grandparent constituent label accuracies are averaged across non-English target languages. Null complementizers, null copulas, and null subjects are inserted and indicated by brackets.[8]

Table 3

[8] Null subjects implied by infinitives and relative clauses are not included in Table 2.

Linguistic Features in the Top and Bottom 50 Sentences

	Top 50 Sentences	Bottom 50 sentences
Average length	9.3 words	21.7 words
Average great-grandparent constituent label accuracy	0.949	0.310
Question sentences	2	10
Infinitive phrases	26	5
Sentences with negation	13	4
Sentences containing a null copula or appositive	0	16
Embedded sentences (excluding infinitives) • Head before • Head after	 9 0	 5 10

Note. The top and bottom 50 sentences were selected based on great-grandparent constituent label prediction accuracy scores (averaged across non-English target languages).

In general, the constituent label prediction models performed better on shorter sentences, which is unsurprising given that longer sentences are more likely to contain complex syntactic structures (e.g. compound and embedded sentences). Additionally, the constituent label prediction models performed poorly on questions, which may be due to the relatively low number of questions in the United Nations NMT training dataset.

More notably, the bottom 50 sentences contained a disproportionate number of null features. These features omit words or morphemes that would indicate syntactic structure in a sentence. For instance, null copulas omit forms of the verb "to be," as in the sentence "He pronounced the homework [was] finished." Appositives, where two noun phrases are placed one after another to describe the same entity (e.g. "Grant, the star baker"), serve as relative clauses

with the usual explicit syntactic cues omitted (e.g. “Grant, [who is] the star baker”). Of the bottom 50 sentences, 16 contained at least one null copula or appositive; the top 50 sentences contained none of either feature. This suggests that when generating encoder states, NMT models typically do not identify syntactic structures based on non-explicit cues.

However, the models performed well on complex syntactic structures containing explicit morphosyntactic cues. They performed well on sentences containing infinitives (e.g. “to eat” or “to pillage”) and negation (e.g. “I did not eat”), exhibiting far more of these features in the top 50 sentences than in the bottom 50 sentences (see Table 3). Both infinitives and negation have clear morphosyntactic cues indicating sentence structure. The “to” in each infinitive clearly introduces the infinitized verb, and the word “not” before a verb clearly indicates a negated clause. These results suggest that NMT encoders rely on explicit morphosyntactic cues to extract syntactic structure from sentences.

In fact, the NMT encoders were able to use morphosyntactic cues to identify embedded sentences. An embedded sentence appears within another phrase (e.g. within the verb phrase “said that [sentence]”). The phrase head which introduces an embedded sentence can appear before or after the embedded sentence (e.g. “Alex said [sentence]” versus “[sentence], said Alex”). Because the NMT encoders were provided only with sentences stopping at a given word, they could not be expected to recognize embedded sentences where the corresponding phrase head appeared after the embedded sentence. However, the models performed well on many sentences where the phrase head appeared before the embedded sentence, exhibiting nine such structures in the top 50 sentences (see Table 3). In many of these sentences, the head and

complementizer (e.g. “said that” or “dogs that”) clearly indicate the beginning of an embedded sentence.

Interestingly, the NMT encoders were often able to recognize embedded sentences even when there was a null complementizer introducing the embedded sentence, such as “that” omitted in “The dog wished [that] he was taller.” Of the nine embedded sentences in the top 50 sentences, six had a null complementizer. This result may partially be explained by verb bias, the tendency for certain verbs to be followed by particular types of phrases (Garnsey, Pearlmutter, Myers, & Lotocky, 1997). For instance, the verb “prove” is more often followed by a sentence complement (e.g. “proved [that] the criminal was lying”) than a direct object (e.g. “proved the theorem”). People are more likely to omit complementizers when the head verb biases heavily towards a sentence complement (Ferreira & Schotter, 2013); in these cases, the verb itself serves as a syntactic cue for the upcoming embedded sentence. Of the six null complementizers in the top 50 sentences, five followed a sentence-complement-biased verb. Thus, it appears that NMT encoders are able to recognize embedded sentences using a combination of verb bias and explicit complementizers.

Discussion

In this study, we found that NMT models implicitly encode hierarchical syntax regardless of target language, as long as the target language differs from the source language. Given the NMT encoder state after reading a given word, feedforward neural network models were able to predict the parent, grandparent, and great-grandparent constituent label of the word, with accuracy scores well above our provided baseline. This extends the results of Blevins et al. (2018) for German NMT models to Arabic, Spanish, French, Russian, and Chinese NMT models

(all with source language English). By considering NMT encoders as connectionist models of language perception and comprehension, high constituent label prediction scores indicate that connectionist models of language can implicitly encode similar information to hierarchical syntactic models of language.

Importantly, the NMT models received no explicitly syntactic information in the training data; the models were trained only on raw sentences translated to both the source and target language. As outlined in the introduction, this supports the thesis of linguistic structuring, which states that any successful model of language must recapitulate the hierarchical structures of language (Schonbein, 2012). We found that connectionist models (NMT models) naturally encoded hierarchical structures despite their seemingly non-hierarchical network architectures. This suggests that generative syntax in the source language contains useful information for translation tasks; this syntactic information is recognized and encoded by the NMT model. We found that NMT models rely on explicit morphosyntactic cues (e.g. infinitives) when encoding syntactic information.

Effects of Target Language

All five non-English target languages encoded a similar amount of syntax, even performing similarly to one another for predictions of individual constituent labels (e.g. noun, verb, and adjective phrases). It is important to note that this does not necessarily indicate that the different target languages share an underlying syntactic structure; the decoder can use the encoded syntactic information in a variety of different ways, resulting in different syntactic structures for different target languages.

That said, although the encoded syntactic information might be used in different ways for different target languages, similar English syntactic information was encoded across target languages. For instance, the English-to-Chinese model did not identify English verb phrases any better or worse than the NMT models for the other target languages. This similarity in encoded syntactic information across target languages could suggest that there is some shared upper bound on the amount of syntactic information that is useful in translation tasks; however, this would not align with results finding that explicitly added syntactic information provides improvements to NMT systems (Chen, Huang, Chiang, & Chen, 2017; Wu, Zhou, & Zhang, 2017). Alternatively, the similar syntactic information in the NMT encoders could be a limitation of the encoders' RNN (recurrent neural network) architectures. Across target languages, the NMT encoders could be hitting an upper bound on the amount of syntactic information that they could extract from raw sentences. Of course, this upper bound would depend on the encoder architecture used; optimal translation models would likely extract more syntactic information.

Regardless of the cause of similar syntactic information in our NMT encoders, these similarities are particularly interesting considering the variance in actual translation performance between target languages. Our NMT models ranged from exhibiting "significant grammatical errors" in Chinese (tokenized BLEU: 24.9) to "very high quality, adequate, and fluent translations" in Spanish (tokenized BLEU: 56.3; "Evaluating models," 2020). The wide range in translation quality paired with the similar constituent label prediction scores indicates that morphological and non-syntactic features have large impacts on translation performance.

For instance, a more direct mapping between language vocabularies (e.g. between English and Spanish compared to English and Chinese) likely increases translation quality between two languages. Inflectional morphology (e.g. verb conjugation or noun pluralization) has been found to account for differences in performance between languages in language modeling tasks (predicting the next word in a sentence), but these results vary depending on the metric used for morphological complexity (Cotterell, Mielke, Eisner, & Roark, 2018; Mielke, Cotterell, Gorman, Roark, & Eisner, 2019). It is then possible that morphological features play a significant role in translation tasks; NMT encoder states may then contain significant amounts of morphological information.

Alternatively, semantic content may play the most prominent role in translation tasks; this would seem plausible given that the goal of most translation tasks is to convert semantic or pragmatic information from one language to another. Indeed, Schwenk and Douze (2017) found that multilingual NMT encoder states clustered more based on semantic than syntactic similarity, indicating that semantic information may be more important than syntax in machine translation.[9] Then, NMT encoder states for different target languages may encode different semantic information, leading to differences in translation quality. However, across target languages, Poliak et al. (2018) found inconsistencies for which target language's encoder states resulted in the best performance on various semantic understanding tasks. These results suggest that, like syntactic information, semantic information in NMT encoder representations may be similar across target languages. Future research could investigate specific types of morphological and

[9] Multilingual NMT models train either from multiple source languages or to multiple target languages.

semantic information learned by NMT encoders, providing a better understanding of information contained in NMT encoder states cross-linguistically. Future research could also investigate the impact of the NMT decoder in mediating translation quality across target languages; it is possible that the decoder accounts for more of the divergence in translation quality than the encoder.

Limitations

While this study found similar encoded syntactic information across NMT models trained towards different target languages, there are several limitations to our conclusions. While our five non-English target languages (Arabic, Chinese, French, Russian, and Spanish) came from a relatively wide variety of language families, they had notable syntactic similarities. Five of the six languages had default SVO (subject-verb-object) word order. The only exception, Arabic (VSO default word order), still uses SVO in many sentences; SVO sentences were found to account for 48% of sentences in Arabic political speeches and 30% of sentences in Arabic magazine articles (Parkinson, 1981). Similar word orders among target languages could lead to an overestimate of the similarity of syntactic information in NMT encoder states. A consistently non-SVO language such as Japanese or Korean (both SOV) could lead the NMT models to encode different syntactic information.

Additionally, our translation dataset (the UN Parallel Corpus) contained dramatically different types of sentences from our syntax evaluation dataset (the CoNLL-2012 dataset). The UN Parallel Corpus consists of official records and documents from the United Nations, while the CoNLL-2012 dataset has sources ranging from the New Testament and printed news to web blogs and transcribed telephone conversations (Ziemski et al., 2016; Pradhan et al., 2012). More similar datasets would likely increase overall constituent label prediction accuracy scores,

leading to increased amounts of observed syntax in NMT encoder states. While it is possible that more similar datasets could affect similarities across target languages, there is little evidence that these similarities would entirely disappear. Furthermore, our study demonstrates that syntax learned implicitly by NMT models trained in one domain (official government documents) can at least partially be transferred when the NMT model is applied to a different domain (e.g. conversational speech).

Finally, our NMT encoders specifically used RNN architectures, which may have limited the amount of syntactic information that the NMT encoders could extract from raw sentences. For instance, we found that the NMT encoders relied on explicit morphosyntactic cues to identify features such as negation, embedded sentences, and infinitive phrases. In some ways, this reliance on explicit syntactic cues is similar to sentence processing in people. Many sentences are syntactically ambiguous before they are completed (notably garden-path sentences such as "The horse raced past the barn fell"), and people generally re-evaluate upon reading the disambiguating feature (Frazier & Rayner, 1982; Qian, Garnsey, & Christianson, 2018). Thus, it may be implausible for NMT systems to identify non-explicit syntactic features given only partial sentences. Compounding this problem, RNNs are unable to re-evaluate past inputs and hidden states upon reading disambiguating words. In contrast, more recent NMT architectures called Transformers repeatedly process all words in the source sentence, allowing complex interactions between distant words (Vaswani et al., 2017). The recent successes of Transformer models may be due partially to their ability to combine later information with representations of earlier words. Then, Transformers would likely encode significantly more syntactic information than RNN-based NMT architectures. That said, the Transformers' repeated processing of all

words in every sentence has relatively few analogies to human cognitive processing; storing entire sentences word-by-word in working memory is unfeasible according to current theories of working memory.

Implications and Future Work

This study provides evidence that connectionist models of language implicitly encode syntactic information; it also provides preliminary evidence that similar syntactic information is encoded by NMT models regardless of target language. Further research is required to assess whether the similarity of syntactic information in NMT models has deeper implications for cross-linguistic language cognition in general. For instance, it could be found that similar syntactic structures are used in language comprehension and production across languages; some linguists propose a Universal Grammar that limits the possible structures in human language (e.g. Carnie, 2013, p. 23–27). Such universal structures could include differentiation between word classes (e.g. nouns versus verbs) or hierarchically nested embedded sentences. If all languages shared a substantial underlying syntactic framework, then we would expect NMT encoders to encode similar syntactic information regardless of target language. However, our study does not necessarily provide strong evidence for Universal Grammar. Our study demonstrates that similar English syntactic information is useful when representing sentences to be translated to a variety of target languages; this does not reflect the syntactic structure of the target languages themselves. The NMT decoder simply uses the English syntactic information to make better translations in the target language.

While our study provides limited insight into Universal Grammar structures, the fact that connectionist models can extract syntactic information from raw sentences alone has important

implications in a variety of disciplines. Our results suggest that during language acquisition, people might be able to learn syntactic rules without any genetically-encoded linguistic capabilities. Future research could evaluate whether the syntactic errors made by children acquiring multiple languages are similar to the syntactic errors made by NMT systems trained between those languages. This would extend previous results finding that artificial neural network models learning verbs' past tense forms follow similar patterns of development to children learning past tense forms (Rumelhart & McClelland, 1985). Furthermore, similar syntactic performance across different target languages in NMT suggests that children might exhibit similar syntactic errors in one language regardless of the other language being acquired. This hypothesis is supported by findings that errors caused by first language interference account for less than 5% of syntactic errors in children learning a second language (although notably, these children were not learning the second language simultaneously with the first language; Dulay & Burt, 1974).

Implicit syntax-learning in NMT also suggests that computers may be able to identify syntactic rules given only raw sentences. Unsupervised syntactic parsing (extracting syntactic features from raw sentences alone) is an active area of research in natural language processing (He, Neubig, & Berg-Kirkpatrick, 2018; Reichart & Rappoport, 2009). Researchers have successfully used unsupervised syntactic parsing to identify words' POS tags in text and to identify constituent phrases in sentences (Hänig, Bordag, & Quasthoff, 2008).

Finally, future research could investigate new ways of evaluating general information encoded within sentence representations cross-linguistically. Our study specifically assessed syntactic information in NMT encoder states, but there has also been significant research into

cross-lingual sentence representations encoding all relevant information in a given sentence regardless of language (Chidambaram et al., 2019). Oftentimes, these cross-lingual representations are generated using NMT representations (Schwenk & Douze, 2017; Eriguchi, Johnson, Firat, Kazawa, & Macherey, 2018). Interestingly, NMT representations were found to cluster based on target language family when sentence representations were aligned in a shared space (using correlation analysis; Kudugunta et al., 2019). These results indicate that while NMT encoder states may contain similar syntactic information across target languages, general information content may differ. These comparisons of sentence representations across languages and tasks have implications for transfer learning research, which uses models trained on one language or task to improve another model's performance on a different language or task (Ruder, Peters, Swayamdipta, & Wolf, 2019). For instance, using representations from a pre-trained English model for some task could decrease the training time or training data required to train a model for a different language (oftentimes a low-resource language).

Outside of computer science, similar or different overall mutual information between sentence representations across languages has implications for the hypothesis of linguistic relativity, which claims that people's cognition (e.g. physical conceptions of time) is influenced by the language they speak (Boroditsky, Fuhrman, & Mccormick, 2011). Differences between sentence representations across languages would suggest that different information tends to be considered "relevant" when a sentence is translated to different languages. Of course, these results would be difficult to disentangle from inherent difficulties in preserving meaning when translating sentences between languages.

Conclusions

This study has provided an initial exploration into implicit syntax encodings in NMT models cross-linguistically, providing evidence for the emergence of hierarchical structures in connectionist models of language. Evidence for implicit syntax-learning can inform current research in natural language processing (particularly syntactic parsing) and language acquisition. Furthermore, similar syntactic information encoded by NMT models across target languages opens up new directions of research in cross-lingual natural language processing and the existence of universal grammar structures in language.

References

Abbott, J., Austerweil, J., & Griffiths, T. (2012). Human memory search as a random walk in a semantic network. *Advances in Neural Information Processing Systems 25 (NIPS),* 3041–3049.

Adger, D. (2015). Syntax. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 131–147.

Balduzzi, D., Vanchinathan, H., & Buhmann, J. (2015). Kickback cuts backprop's red-tape: Biologically plausible credit assignment in neural networks. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 485–491.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2017a). What do neural machine translation models learn about morphology? *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 861–872.

Belinkov, Y., Màrquez, L., Durrani, N., Dalvi, F., & Glass, J. (2017b). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP),* 1–10.

Bengio, Y., Lee, D., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards biologically plausible deep learning. *ArXiv.org*.

Blevins, T., Levy, O., & Zettlemoyer, L. (2018). Deep RNNs encode soft hierarchical syntax. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 14–19.

Boroditsky, L., Fuhrman, O., & Mccormick, K. (2011). Do English and Mandarin speakers think about time differently? *Cognition*, *118*(1), 126–132.

Carnie, A. (2013). *Syntax: a generative introduction* (3rd ed.). Malden, Mass: Blackwell Publishing.

Catania, A. (1972). Chomsky's formal analysis of natural languages: A behavioral translation. *Behaviorism*, *1*(1), 1–15.

Chang, X., Wang, P., Cai, M., & Wang, M. (2019). The predictive power of working memory on Chinese middle school students' English reading comprehension. *Reading & Writing Quarterly*, *35*(5), 458–472.

Chen, H., Huang, S., Chiang, D., & Chen, J. (2017). Improved neural machine translation with a syntax-aware encoder and decoder. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1936–1945.

Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Strope, B., & Kurzweil, R. (2019). Learning cross-lingual sentence representations via a multi-task dual-encoder model. *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP, ACL),* 250–259.

Clark, A. (2015). Radical predictive processing. *Southern Journal of Philosophy, 53*(1), 3–27.

Cotterell, R., Mielke, S., Eisner, J., & Roark, B. (2018). Are all languages equally hard to language-model? *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL),* 536–541.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450–466.

Daumé III, H. (2012). Neural networks. In *A course in machine learning* (pp. 129–140).

Dulay, H., & Burt, M. (1974). Errors and strategies in child second language acquisition. *TESOL Quarterly, 8*(2), 129–136.

Eriguchi, A., Johnson, M., Firat, O., Kazawa, H., & Macherey, W. (2018). Zero-shot cross-lingual classification using multilingual neural machine translation. *ArXiv.org*.

Evaluating models (2020). *Google AutoML Translation Documentation*. Retrieved March 15, 2020, from https://cloud.google.com/translate/automl/docs/evaluate.

Ferreira, V., & Schotter, E. (2013). Do verb bias effects on sentence production reflect sensitivity to comprehension or production factors? *The Quarterly Journal of Experimental Psychology, 66*(8), 1548–1571.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14*(2), 178–210.

Futrell, R., Wilcox, E., Morita, T., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL),* 32–42.

Garnsey, S., Pearlmutter, N., Myers, E., & Lotocky, M. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*(1), 58–93.

Hänig, C., Bordag, S., & Quasthoff, U. (2008). UnsuParse: Unsupervised parsing with unsupervised part of speech tagging. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, 1109–1114.

He, J., Neubig, G., & Berg-Kirkpatrick, T. (2018). Unsupervised learning of syntactic structure with invertible neural projections. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1292–1302.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735–1780.

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Klein, G., Yoon, K., Deng, Y., Crego, J., Senellart, J., & Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. *Proceedings of ACL 2017, System Demonstrations,* 67–72.

Kudugunta, S., Bapna, A., Caswell, I., & Firat, O. (2019). Investigating multilingual NMT representations at scale. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1565–1575.

Luong, T., Pham, H., & Manning, C. (2015). Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1412–1421.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1192–1202.

McClelland, J. L. (2000). Connectionist models of memory. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 583–596). New York: Oxford University Press.

McCoy, R., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL),* 3428–3448.

Mielke, S., Cotterell, R., Gorman, K., Roark, B., & Eisner, J. (2019). What kind of language is hard to language-model? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4975–4989.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition.* (pp. 264–336). Hillsdale, NJ: Lawrence Erlbaum Associates.

Neubig, G. (2017). Neural machine translation and sequence-to-sequence models: A tutorial. *ArXiv.org*.

Osaka, M., Nishizaki, Y., Komori, M., & Osaka, N. (2002). Effect of focus on verbal working memory: Critical role of the focus word in reading. *Memory & Cognition*, *30*(4), 562–571.

Padó, S. (2006). User's guide to sigf: Significance testing by approximate randomisation. Retrieved from https://nlpado.de/~sebastian/software/sigf.shtml.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318.

Parkinson, D. (1981). VSO to SVO in Modern Standard Arabic: A study in diglossia syntax. *Al-'Arabiyya, 14*(1), 24–37.

Poliak, A., Belinkov, Y., Glass, J., & Van Durme, B. (2018). On the evaluation of semantic phenomena in neural machine translation using natural language inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 513–523.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. *Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task*, 1–40.

Qian, Z., Garnsey, S., & Christianson, K. (2018). A comparison of online and offline measures of good-enough processing in garden-path sentences. *Language, Cognition and Neuroscience, 33*(2), 227–254.

Reichart, R., & Rappoport, A. (2009). Automatic selection of high quality parses created by a fully unsupervised parser. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, 156–164.

Ruder, S., Peters, M., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Tutorials*, 15–18.

Rumelhart, D., & McClelland, J. (1985). On learning the past tenses of English verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, Mass: MIT Press.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Schonbein, W. (2012). The linguistic subversion of mental representation. *Minds and Machines*, *22*(3), 235–262.

Schwenk, H., & Douze, M. (2017). Learning joint multilingual sentence representations with neural machine translation. *Proceedings of the 2nd Workshop on Representation Learning for NLP (RepL4NLP, ACL),* 157–167.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL),* 1715–1725.

Shi, X., Padhi, I., & Knight, K. (2016). Does string-based neural MT learn source syntax. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1526–1534.

Sutskever, I., Vinyals, O., & Le, Q. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems 27 (NIPS)*, 3104–3112.

Turovsky, B. (2016). Found in translation: More accurate, fluent sentences in Google Translate. *Google Translate*. Retrieved from https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS)*, 5998–6008.

Wu, S., Zhou, M., & Zhang, D. (2017). Improved neural machine translation with source syntax. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, 4179–4185.

Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M., Macherey, W., … Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv.org*.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations Parallel Corpus v1.0. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 3530–3534.

Appendix A

Approximate Randomization Test for Significance Testing Between Languages

Belinkov et al. (2017b) used the approximate randomization test, a type of permutation test, to assess the statistical significance of differences in constituent label prediction accuracy across target languages (Padó, 2006). Given a pair of models, the approximate randomization test repeatedly and randomly shuffles the models' per-sentence accuracy scores, determining statistical significance based on the number of trials in which the random shuffle results in a larger performance difference than the original empirical difference between the two models. Approximate randomization does not require independence assumptions between the two compared models or between individual (word-level) constituent label predictions. The approximate randomization test only assumes approximate independence between sentence-level prediction accuracy scores. Note that when using the approximate randomization test, accuracy scores are computed for each sentence in the test set, and these per-sentence accuracy scores are averaged to obtain an overall accuracy score. In general, sentence-averaged accuracy scores computed in this study tended to be higher than the raw accuracy scores reported; this indicates that models exhibited higher accuracy scores on shorter sentences.

However, the approximate randomization test can only make pairwise comparisons between models, so it cannot account for variance between constituent label prediction models trained on the same NMT representations (i.e. variance based on random initialization of weights for the feedforward neural networks and random shuffles of the training data). The approximate randomization test often reported significant differences between constituent label prediction models trained on the same NMT representations; analogous to significant differences between

individual participants in behavioral studies, these differences do not reflect differences between the participant groups (in this case, different NMT representations).

Appendix B

Constituent Label Accuracy Scores and Differences Between Target Languages

Mean constituent label prediction accuracy scores are shown in Table B1. All constituent label prediction models performed significantly differently from their corresponding baseline score (adjusted $p < 0.001$ for all comparisons). One-way ANOVAs found significant differences between target languages for all four constituent label prediction tasks ($F(5, 114) > 14{,}000$, $p < 0.001$ for all four label prediction types). Using Tukey's HSD (honestly significant difference) post-hoc test to identify language pairs that differed significantly for each task, 52 of the 60 comparisons were significant with $p < 0.001$. The eight remaining language pairs (with corresponding significance levels) are shown in Table B2.

Table B1

Mean Constituent Label Prediction Scores

	Part-of-speech	Parent	Grandparent	Great-grandparent
Spanish	0.843	0.837	0.692	0.625
French	0.848	0.839	0.692	0.622
Russian	0.849	0.846	0.696	0.623
Arabic	0.836	0.833	0.688	0.620
Chinese	0.800	0.832	0.681	0.609
English	0.484	0.610	0.512	0.456
Baseline	0.868	0.814	0.547	0.413

Table B2

Pairwise Significance Levels Between Target Languages

Task	Language Pair	Significance (adjusted)
POS	FR-RU	$p = 0.734$
Parent	AR-ZH	$p = 0.558$
Parent	ES-FR	$p = 0.117$
Grandparent	ES-FR	$p = 0.997$
Great-Grandparent	AR-FR	$p = 0.121$
Great-Grandparent	ES-RU	$p = 0.407$
Great-Grandparent	FR-RU	$p = 0.606$
Great-Grandparent	ES-FR	$p = 0.009$ **

Note. Asterisks indicate significance. All language pairs not listed differed significantly with adjusted $p < 0.001$.

Appendix C

Correlations Between Target Languages in Great-Grandparent Accuracy Scores per Sentence

Pearson correlations between non-English target languages in great-grandparent constituent label prediction accuracy scores for each sentence are shown in Table C1. For reference, the plot for the least correlated language pair is shown in Figure C1.

Table C1

Pearson Correlations for Great-Grandparent Accuracy per Sentence

	ES	FR	RU	ZH
AR	0.890	0.892	0.887	0.885
ES		0.902	0.872	0.878
FR			0.884	0.883
RU				0.878

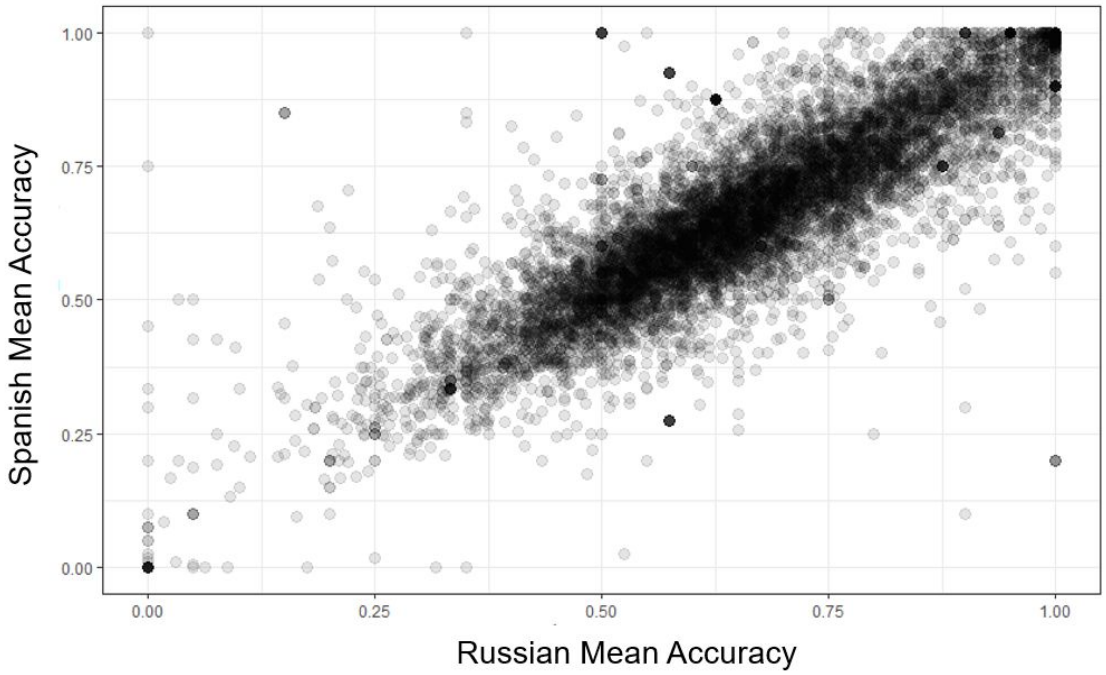


Figure C1. Mean Russian and Spanish great-grandparent accuracy scores for each sentence, where each dot represents a sentence.