On the Difficulty of Translating Free-Order Case-Marking Languages

Arianna Bisazza **Ahmet Üstün** **Stephan Sportel**
Center for Language and Cognition
University of Groningen, The Netherlands
{a.bisazza, a.ustun}@rug.nl, research@spor.tel

Abstract

Identifying factors that make certain languages harder to model than others is essential to reach language equality in future Natural Language Processing technologies. Free-order case-marking languages, such as Russian, Latin, or Tamil, have proved more challenging than fixed-order languages for the tasks of syntactic parsing and subject-verb agreement prediction. In this work, we investigate whether this class of languages is also more difficult to translate by state-of-the-art Neural Machine Translation (NMT) models. Using a variety of synthetic languages and a newly introduced translation challenge set, we find that word order flexibility in the source language only leads to a very small loss of NMT quality, even though the core verb arguments become impossible to disambiguate in sentences without semantic cues. The latter issue is indeed solved by the addition of case marking. However, in medium- and low-resource settings, the overall NMT quality of *fixed*-order languages remains unmatched.

1 Introduction

Despite the tremendous advances achieved in less than a decade, Natural Language Processing remains a field where language equality is far from being reached (Joshi et al., 2020). In the field of Machine Translation, modern neural models have attained remarkable quality for high-resource language pairs like German-English, Chinese-English, or English-Czech, with a number of studies even claiming human parity (Hassan et al., 2018; Bojar et al., 2018; Barrault et al., 2019; Popel et al., 2020). These results may lead to the unfounded belief that Neural Machine Translation (NMT) methods will perform equally well in any language pair, provided similar amounts of training data. In fact, several studies suggest the opposite (Platanios et al., 2018; Ataman and Federico, 2018; Bugliarello et al., 2020).

Why, then, do some language pairs have lower translation accuracy? And, more specifically: Are certain typological profiles more challenging for current state-of-the-art NMT models? Every language has its own combination of typological properties, including word order, morphosyntactic features, and more (Dryer and Haspelmath, 2013). Identifying language properties (or combinations thereof) that pose major problems to the current modeling paradigms is essential to reach language equality in future MT (and other NLP) technologies (Joshi et al., 2020), in a way that is orthogonal to data collection efforts. Among others, natural languages adopt different mechanisms to disambiguate the role of their constituents: Flexible order typically correlates with the presence of case marking and, vice versa, fixed order is observed in languages with little or no case marking (Comrie, 1981; Sinnemäki, 2008; Futrell et al., 2015b). Morphologically rich languages *in general* are known to be challenging for MT at least since the times of phrase-based statistical MT (Birch et al., 2008) due to their larger and sparser vocabularies, and remain challenging even for modern neural architectures (Ataman and Federico, 2018; Belinkov et al., 2017). By contrast, the relation between word order flexibility and MT quality has not been directly studied to our knowledge.

In this paper, we study this relationship using strictly controlled experimental setups. Specifically, we ask:

- Are current state-of-the-art NMT systems biased towards *fixed-order* languages?
- To what extent does *case marking* compensate for the lack of a fixed order in the source language?

Unfortunately, parallel data are scarce in most of the world languages (Guzmán et al., 2019), and

Transactions of the Association for Computational Linguistics, vol. 9, pp. 1233–1248, 2021. https://doi.org/10.1162/tacl_a_00424
Action Editor: Alexandra Birch. Submission batch: 1/2021; Revision batch: 5/2021; Published 11/2021.

Fixed	vso	follows the little cat the friendly dog
	vos	follows the friendly dog the little cat
Free+Case		follows the little cat#S the friendly dog#O OR follows the friendly dog#O the little cat#S
Translation		de kleine kat volgt de vriendelijke hond

Table 1: Example sentence in different fixed/ flexible-order English-based synthetic languages and their SVO Dutch translation. The subject in each sentence is underlined. Artificial case markers start with #.

corpora in different languages are drawn from different domains. Exceptions exist, like the widely used Europarl (Koehn, 2005), but represent a small fraction of the large variety of typological feature combinations attested in the world. This makes it very difficult to run a large-scale comparative study and isolate the factors of interest from, for example, domain mismatch effects. As a solution, we propose to evaluate NMT on synthetic languages (Gulordava and Merlo, 2016; Wang and Eisner, 2016; Ravfogel et al., 2019) that differ from each other only by specific properties, namely: the order of main constituents, or the presence and nature of case markers (see example in Table 1).

We use this approach to isolate the impact of various source-language typological features on MT quality and to remove the typical confounders of corpus size and domain. Using a variety of synthetic languages and a newly introduced challenge set, we find that state-of-the-art NMT has little to no bias towards fixed-order languages, but only when a sizeable training set is available.

2 Free-order Case-marking Languages

The word order profile of a language is usually represented by the canonical order of its main constituents, (S)ubject, (O)bject, (V)erb. For instance, English and French are SVO languages, while Turkish and Hindi are SOV. Other, less commonly attested, word orders are VSO and VOS, whereas OSV and OVS are extremely rare (Dryer, 2013). Although many other word order features exist (e.g., noun/adjective), they often correlate with the order of main constituents (Greenberg, 1963).

A different, but likewise important dimension is that of word order *freedom* (or *flexibility*). Languages that primarily rely on the position of a word to encode grammatical roles typically display rigid orders (like English or Mandarin Chinese), while languages that rely on case marking can be more flexible allowing word order to express discourse-related factors like topicalization. Examples of highly flexible-order languages include languages as diverse as Russian, Hungarian, Latin, Tamil, and Turkish.[1]

In the field of psycholinguistics, due to the historical influence of English-centered studies, word order has long been considered the primary and most natural device through which children learn to infer syntactic relationships in their language (Slobin, 1966). However, cross-linguistic studies have later revealed that children are equally prepared to acquire both fixed-order and inflectional languages (Slobin and Bever, 1982).

Coming to computational linguistics, data-driven MT and other NLP approaches were also historically developed around languages with remarkably fixed orders and very simple to moderately simple morphological systems, like English or French. Luckily, our community has been giving increasing attention to more and more languages with diverse typologies, especially in the last decade. So far, previous work has found that free-order languages are more challenging for parsing (Gulordava and Merlo, 2015, 2016) and subject-verb agreement prediction (Ravfogel et al., 2019) than their fixed-order counterparts. This raises the question of whether word order flexibility also negatively affects MT quality.

Before the advent of modern NMT, Birch et al. (2008) used the Europarl corpus to study how various language properties affected the quality of phrase-based Statistical MT. Amount of reordering, target morphological complexity, and historical relatedness of source and target languages were identified as strong predictors of MT quality. Recent work by Bugliarello et al. (2020), however, has failed to show a correlation between NMT difficulty (measured by a novel information-theoretic metric) and several linguistic properties of source and target language, including Morphological Counting Complexity (Sagot, 2013) and Average Dependency Length (Futrell et al., 2015a). While that work specifically

[1] See Futrell et al. (2015b) for detailed figures of word order freedom (measured by the entropy of subject and object dependency relation order) in a diverse sample of 34 languages.

aimed at ensuring cross-linguistic comparability, the sample on which the linguistic properties could be computed (Europarl) was rather small and not very typologically diverse, leaving our research questions open to further investigation. In this paper, we therefore opt for a different methodology: namely, synthetic languages.

3 Methodology

Synthetic Languages This paper presents two sets of experiments: In the first (§4), we create parallel corpora using very simple and predictable artificial grammars and small vocabularies (Lupyan and Christiansen, 2002). See an example in Table 1. By varying the position of subject/verb/object and introducing case markers to the source language, we study the biases of two NMT architectures in optimal training data conditions and a fully controlled setup, that is, without any other linguistic cues that may disambiguate constituent roles. In the second set of experiments (§5), we move to a more realistic setup using synthetic versions of the English language that differ from it in only one or few selected typological features (Ravfogel et al., 2019). For instance, the original sentence's order (SVO) is transformed to different orders, like SOV or VSO, based on its syntactic parse tree.

In both cases, typological variations are introduced in the source side of the parallel corpora, while the target language remains fixed. In this way, we avoid the issue of non-comparable BLEU scores across different target languages. Lastly, we make the simplifying assumption that, when verb-argument order varies from the canonical order in a flexible-order language, it does so in a totally arbitrary way. Although this is rarely true in practice, as word order may be predictable given pragmatics or other factors, we focus here on *''the extent to which word order is conditioned on the syntactic and compositional semantic properties of an utterance''* (Futrell et al., 2015b).

Translation Models We consider two widely used NMT architectures that crucially differ in their encoding of positional information: (i) Recurrent sequence-to-sequence BiLSTM with attention (Bahdanau et al., 2015; Luong et al., 2015) processes the input symbols sequentially and has each hidden state directly conditioned on that of the previous (or following, for the backward LSTM) timestep (Elman, 1990; Hochreiter and Schmidhuber, 1997). (ii) The non-recurrent, fully attention-based Transformer (Vaswani et al., 2017) processes all input symbols in parallel relying on dedicated embeddings to encode each input's position.[2] Transformer has nowadays surpassed recurrent encoder-decoder models in terms of generic MT quality. Moreover, Choshen and Abend (2019) have recently shown that Transformer-based NMT models are indifferent to the absolute order of source words, at least when equipped with learned positional embeddings. On the other hand, the lack of recurrence in Transformers has been linked to a limited ability to capture hierarchical structure (Tran et al., 2018; Hahn, 2020). To our knowledge, no previous work has studied the biases of either architectures towards fixed-order languages in a systematic manner.

4 Toy Parallel Grammar

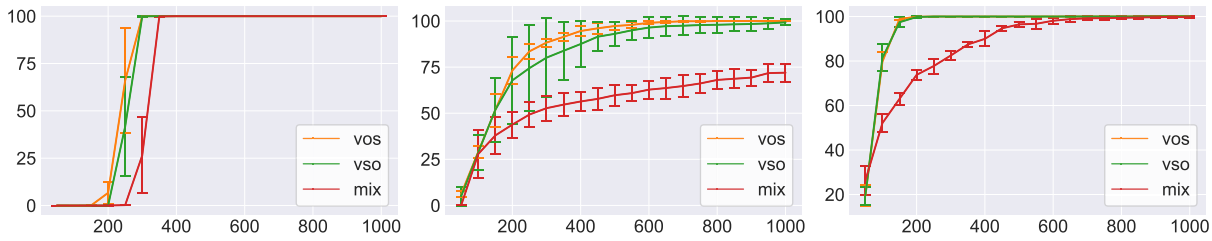
We start by evaluating our models on a pair of toy languages inspired by the English-Dutch pair and created using a Synchronous Context-Free Grammar (Chiang and Knight, 2006). Each sentence consists of a simple clause with a transitive verb, subject, and object. Both arguments are singular and optionally modified by an adjective. The source vocabulary contains 6 nouns, 6 verbs, 6 adjectives, and the complete corpus contains 10k generated sentence pairs. Working with such a small, finite grammar allows us to simulate an otherwise impossible situation where the NMT model can be trained on (almost) the totality of a language's utterances, canceling out data sparsity effects.[3]

Source Language Variants We consider three source language variants, illustrated in Table 1:

- fixed-order VSO;
- fixed-order VOS;
- mixed-order (randomly chosen between VSO or VOS) with nominal case marking.

[2] We use sinusoidal embeddings (Vaswani et al., 2017). All our models are built using OpenNMT: `https://github.com/OpenNMT/OpenNMT-py`.

[3] Data and code to replicate the toy grammar experiments in this section are available at `https://github.com/573phn/cm-vs-wo`.



(a) BiLSTM with attention (b) Large Transformer (c) Small Transformer

Figure 1: Toy language NMT sentence-level accuracy on validation set by number of training epochs. Source languages: fixed-order VSO, fixed-order VOS, and mixed-order (VSO/VOS) with case marking. Target language: always fixed SVO. Each experiment is repeated five times, and averaged results are shown.

We choose these word orders so that, in the flexible-order corpus, the only way to disambiguate argument roles is case marking, realized by simple unambiguous suffixes (*#S* and *#O*). The target language is always fixed SVO. The same random split (80/10/10% training/validation/test) is applied to the three corpora.

NMT Setup As recurrent model, we trained a 2-layer BiLSTM with attention (Luong et al., 2015) with 500 hidden layer size. As Transformer models, we trained one using the standard 6-layer configuration (Vaswani et al., 2017) and a smaller one with only 2 layers, given the simplicity of the languages. All models are trained at the word level using the complete vocabulary. More hyperparameters are provided in Appendix A.1. Note that our goal is not to compare LSTM and Transformer accuracy to each other, but rather to observe the different trends across fixed- and flexible-order language variants. Given the small vocabulary, we use sentence-level accuracy instead of BLEU for evaluation.

Results As shown in Figure 1, all models achieve perfect accuracy on all language pairs after 1000 training steps, except for the Large Transformer on the free-order language, likely due to overparametrization (Sankararaman et al., 2020). These results demonstrate that our NMT architectures are equally capable of modeling translation of both types of language, when all other factors of variation are controlled for.

Nonetheless, a pattern emerges when looking at the learning curves within each plot: While the two fixed-order languages have very similar learning curves, the free-order language with case markers always requires slightly more training steps to converge. This is also the case, albeit to a lesser extent, when the mixed-order corpus is pre-processed by splitting all case suffixes from the nouns (extra experiment not shown in the plot). This trend is noteworthy, given the simplicity of our grammars and the transparency of the case system. As our training sets cover a large majority of the languages, this result might suggest that free-order *natural* languages need larger training datasets to reach a similar translation quality than their fixed-order counterparts. In §5 we validate this hypothesis on more naturalistic language data.

5 Synthetic English Variants

Experimenting with toy languages has its shortcomings, like the small vocabulary size and non-realistic distribution of words and structures. In this section, we follow the approach of Ravfogel et al. (2019) to validate our findings in a less controlled but more realistic setup. Specifically, we create several variants of the Europarl English-French parallel corpus where the source sentences are modified by changing word order and adding artificial case markers. We choose French as target language because of its fixed order, SVO, and its relatively simple morphology.[4] As Indo-European languages, English and French are moderately related in terms of syntax and vocabulary while being sufficiently distant to avoid a word-by-word translation strategy in many cases.

Source language variants are obtained by transforming the syntactic tree of the original sentences. While Ravfogel et al. (2019) could rely on the Penn Treebank (Marcus et al., 1993) for their monolingual task of agreement prediction, we

[4]According to the Morphological Counting Complexity (Sagot, 2013) values reported by Cotterell et al. (2018), English scores 6 (least complex), Dutch 26, French 30, Spanish 71, Czech 195, and Finnish 198 (most complex).

Original (no case): The woman says her sisters often invited her for dinner.
SOV (no case): The woman her sisters her often invited for dinner say.
SOV, syncretic case marking (overt): The woman*.arg.sg* her sisters*.arg.pl* she*.arg.sg* often invited*.arg.pl* for dinner say*.arg.sg*.
SOV, unambiguous case marking (overt): The woman*.nsubj.sg* her sisters*.nsubj.pl* she*.dobj.sg* often invited*.dobj.sg.nsubj.pl* for dinner say*.nsubj.sg*.
SOV, unambiguous case (implicit): The woman*kar* her sisters*kon* she*kin* often invited*kinkon* for dinner say*kar*.
SOV, unambiguous case (implicit with declensions): The woman*kar* her sisters*pon* she*kit* often invited*kitpon* for dinner say*kar*.
French translation: La femme dit que ses soeurs l'invitaient souvent à dîrner.

Table 2: Examples of synthetic English variants and their (common) French translation. The full list of suffixes is provided in Appendix A.3.

instead need parallel data. For this reason, we parse the English side of the Europarl v.7 corpus (Koehn, 2005) using the Stanza dependency parser (Qi et al., 2020; Manning et al., 2014). After parsing, we adopt a modified version of the synthetic language generator by Ravfogel et al. (2019) to create the following English variants:[5]

- **Fixed-order**: either SVO, SOV, VSO or VOS;[6]
- **Free-order**: for each sentence in the corpus, one of the six possible orders of (Subject, Object, Verb) is chosen randomly;
- **Shuffled words**: all source words are shuffled regardless of their syntactic role. This is our lower bound, measuring the reordering ability of a model in the total absence of source-side order cues (akin to bag-of-words input).

To allow for a fair comparison with the artificial case-marking languages, we remove number agreement features from verbs in all the above variants (cf. *says* → *say* in Table 2).

To answer our second research question, we experiment with two artificial case systems proposed by Ravfogel et al. (2019) and illustrated in Table 2 (overt suffixes):

- **Unambiguous case system**: suffixes indicating argument role (subject/object/indirect object) and number (singular/plural) are added to the heads of noun and verb phrases;
- **Syncretic case system**: suffixes indicating number but not grammatical function are added to the heads of main arguments, providing only partial disambiguation of argument roles. This system is inspired from subject/object syncretism in Russian.

Syncretic case systems were found to be roughly as common as non-syncretic ones in a large sample of almost 200 world languages (Baerman and Brown, 2013). Case marking is always combined with the fully flexible order of main constituents. As in Ravfogel et al. (2019), English number marking is removed from verbs and their arguments before adding the artificial suffixes.

5.1 NMT Setup

Models As recurrent model, we used a 3-layer BiLSTM with hidden size of 512 and MLP attention (Bahdanau et al., 2015). The Transformer model has the standard 6-layer configuration with hidden size of 512, 8 attention heads, and sinusoidal positional encoding (Vaswani et al., 2017). All models use subword representation based on 32k BPE merge operations (Sennrich et al., 2016), except in the low-resource setup where this is reduced to 10k operations. More hyperparameters are provided in Appendix A.1.

Data and Evaluation We train our models on various subsets of the English-French Europarl corpus: 1.9M sentence pairs (high-resource), 100K (medium-resource), and 10K (low-resource). For evaluation, we use 5K sentences randomly held out from the same corpus. Given the importance of word order to assess the correct translation of verb arguments into French, we compute the

[5] Our revised language generator is available at `https://github.com/573phn/rnn_typology`.

[6] To keep the number of experiments manageable, we omit object-initial languages, which are significantly less attested among world languages (Dryer, 2013).

reordering-focused RIBES[7] metric (Isozaki et al., 2010) in addition to the more commonly used BLEU (Papineni et al., 2002). In each experiment, the source side of training and test data is transformed using the same procedure whereas the target side remains unchanged. We repeat each experiment 3 times (or 4 for languages with random order choice) and report the averaged results.

5.2 Challenge Set

Besides syntactic structure, natural language often contains semantic and collocational cues that help disambiguate the role of an argument. Small BLEU/RIBES differences between our language variants may indicate actual robustness of a model to word order flexibility, but may also indicate that a model relies on those cues rather than on syntactic structure (Gulordava et al., 2018). To discern these two hypotheses, we create a challenge set of 7,200 simple affirmative and negative sentences where swapping subject and object leads to another plausible sentence.[8] Each English sentence and its reverse are included in the test set together with the respective translations, as for example:

(1) (a) The president thanks the minister. /
Le président remercie le ministre.

(b) The minister thanks the president. /
Le ministre remercie le président.

The source side is then processed as explained in §5 and translated by the NMT model trained on the corresponding language variant. Thus, translation quality on this set reflects the extent to which NMT models have robustly learned to detect verb arguments and their roles independently from other cues, which we consider an important sign of linguistic generalization ability. For space constraints we only present RIBES scores on the challenge set.[9]

[7]BLEU captures local word-order errors only indirectly (lower precision of higher-order n-grams) and does not capture long-range word-order errors at all. By contrast, RIBES directly measures correlation between the word ranks in the reference and those in the MT output.

[8]More details can be found in Appendix A.2. We release the challenge set at `https://github.com/arianna-bis/freeorder-mt`.

[9]We also computed BLEU scores: They strongly correlate with RIBES but fluctuate more due to the larger effect of lexical choice.

5.3 High-Resource Results

Table 3 reports the high-resource setting results. The first row (original English to French) is given only for reference and shows the overall highest results. The BLEU drop observed when moving to any of the fixed-order variants (including SVO) is likely due to parsing flaws resulting in awkward reorderings. As this issue affects all our synthetic variants, it does not undermine the validity of our findings. For clarity, we center our main discussion on the Transformer results and comment on the BiLSTM results at the end of this section.

Fixed-Order Variants All four tested fixed-order variants obtain very similar BLEU/RIBES scores on the Europarl-test. This is in line with previous work in NMT showing that linguistically motivated pre-ordering leads to small gains (Zhao et al., 2018) or none at all (Du and Way, 2017), and that Transformer-based models are *not* biased towards monotonic translation (Choshen and Abend, 2019). On the challenge set, scores are slightly more variable but a manual inspection reveals that this is due to different lexical choices, while word order is always correct for this group of languages. To sum up, in the high-resource setup, our Transformer models are perfectly able to disambiguate the core argument roles when these are consistently encoded by word order.

Fixed-Order vs Random-Order Somewhat surprisingly, the Transformer results are only marginally affected by the random ordering of verb and core arguments. Recall that in the 'Random' language all six possible permutations of (S,V,O) are equally likely. Thus, Transformer shows an excellent ability to reconstruct the correct constituent order *in the general-purpose test set*. The picture is very different on the challenge set, where RIBES drops severely from 97.6 to 74.1. These low results were to be expected given the challenge set design (it is impossible even for a human to recognize subject from object in the 'Random, no case' challenge set). Nonetheless, they demonstrate that the general-purpose set cannot tell us whether an NMT model has learned to reliably exploit syntactic structure of the source language, because of the abundant non-syntactic cues. In fact, even when *all* source words are shuffled, Transformer still achieves a respectable 25.8/71.2 BLEU/RIBES on the Europarl-test.

English*→French	BI-LSTM			TRANSFORMER		
Large Training (1.9M)	Europarl-Test		Challenge	Europarl-Test		Challenge
	BLEU	RIBES	RIBES	BLEU	RIBES	RIBES
Original English	39.4	85.0	98.0	38.3	84.9	97.7
Fixed Order:						
S-V-O	38.3	84.5	98.1	37.7	84.6	98.0
S-O-V	37.6	84.2	97.7	37.9	84.5	97.2
V-S-O	38.0	84.2	97.8	37.8	84.6	98.0
V-O-S	37.8	84.0	98.0	37.6	84.3	97.2
Average (fixed orders)	37.9±0.4	84.2±0.3	97.9±0.2	37.8±0.1	84.5±0.1	97.6±0.4
Flexible Order:						
Random, no case	37.1	83.7	75.1	37.5	84.2	74.1
Random + syncretic case	36.9	83.6	75.4	37.3	84.2	84.4
Random + unambig. case	37.3	83.9	97.7	37.3	84.4	98.1
Shuffle all words	18.5	65.2	79.4	25.8	71.2	83.2

Table 3: Translation quality from various English-based synthetic languages into standard French, using the largest training data (1.9M sentences). NMT architectures: 3-layer BiLSTM seq-to-seq with attention; 6-layer Transformer. Europarl-Test: 5K held-out Europarl sentences; Challenge set: see §5.2. All scores are averaged over three training runs.

Case Marking The key comparison in our study lies between fixed-order and free-order case-marking languages. Here, we find that case marking can indeed restore near-perfect accuracy on the challenge set (98.1 RIBES). However, this only happens when the marking system is completely unambiguous, which, as already mentioned, is true for only about a half of the real case-marking languages (Baerman and Brown, 2013). Indeed, the syncretic system visibly improves quality on the challenge set (74.1 to 84.4 RIBES) but remains far behind the fixed-order score (97.6). In terms of overall NMT quality (Europarl-test), fixed-order languages score only marginally higher than the free-order case-marking ones, regardless of the unambiguous/syncretic distinction. Thus our finding that Transformer NMT systems are equally capable of modeling the two types of languages (§4) is also confirmed with more naturalistic language data. That said, we will show in Section 5.4 that this positive finding is conditional on the availability of large amounts of training samples.

BiLSTM vs Transformer The LSTM-based results generally correlate with the Transformer results discussed above, however our recurrent models appear to be slightly more sensitive to changes in the source-side order, in line with previous findings (Choshen and Abend, 2019). Specifically, translation quality on Europarl-test fluctuates slightly more than Transformer among different fixed orders, with the most monotonic order (SVO) leading to the best results. When *all* words are randomly shuffled, BiLSTM scores drop much more than Transformer. However, when comparing the fixed-order variants to the ones with free order of main constituents, BiLSTM shows only a slightly stronger preference for fixed-order, compared to Transformer. This suggests that, by experimenting with arbitrary permutations, Choshen and Abend (2019) might have overestimated the bias of recurrent NMT towards more monotonic translation, whereas the more realistic combination of constituent-level reordering with case marking used in our study is not so problematic for this type of model.

Interestingly, on the challenge set, BiLSTM and Transformer perform on par, with the notable exception that syncretic case is much more difficult for the BiLSTM model. Our results agree with the large drop of subject-verb agreement prediction accuracy observed by Ravfogel et al. (2019) when experimenting with the random order of main

constituents. However, their scores were also low for SOV and VOS, which is not the case in our NMT experiments. Besides the fact that our challenge set only contains short sentences (hence no long dependencies and few agreement attractors), our task is considerably different in that agreement only needs to be predicted in the target language, which is fixed-order SVO.

Summary Our results so far suggest that state-of-the-art NMT models, especially if Transformer-based, have little or no bias towards fixed-order languages. In what follows, we study whether this finding is robust to differences in data size, type of morphology, and target language.

5.4 Effect of Data Size and Morphological Features

Data Size The results shown in Table 3 represent a high-resource setting (almost 2M training sentences). While recent successes in cross-lingual transfer learning alleviate the need for labeled data (Liu et al., 2020), their success still depends on the availability of large unlabeled data as well as other, yet to be explained, language properties (Joshi et al., 2020). We then ask: Do free-order case-marking languages need more data than fixed-order non-case-marking ones to reach similar NMT quality? We simulate a medium- and low-resource scenario by sampling 100K and 10K training sentences, respectively, from the full Europarl data. To reduce the number of experiments, we only consider Transformer with one fixed-order language variant (SOV)[10] and exclude syncretic case marking. To disentagle the effect of word order from that of case marking on low-resource translation quality, we also experiment with a language variant combining fixed-order (SOV) and case marking. Results are shown in Figure 2 and discussed below.

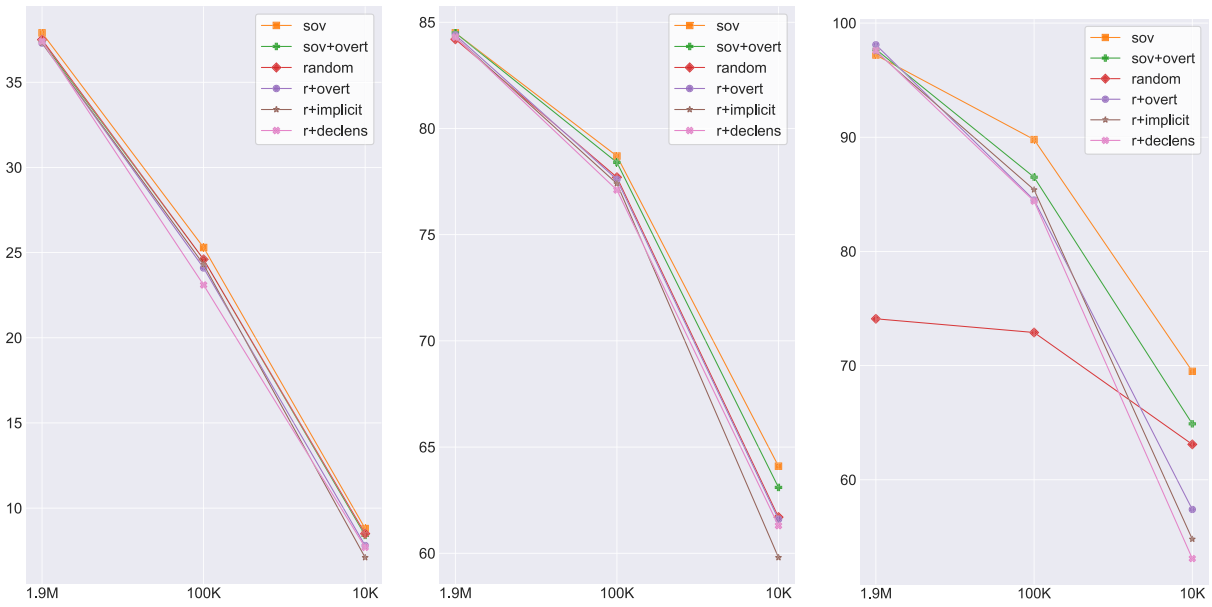
Morphological Features The artificial case systems used so far included easily separable suffixes with a 1:1 mapping between grammatical categories and morphemes (e.g., *.nsubj.sg, .dobj.pl*) reminiscent of agglutinative morphologies. Many world languages, however, do not comply to this 1:1 mapping principle but display exponence (multiple categories conveyed by one morpheme) and/or flexivity (the same category expressed by various, lexically determined, morphemes). Well-studied examples of languages with case+number exponence include Russian and Finnish, while flexive languages include, again, Russian and Latin. Motivated by previous findings on the impact of fine-grained morphological features on language modeling difficulty (Gerz et al., 2018), we experiment with three types of suffixes (see examples in Table 2):

- **Overt**: number and case are denoted by easily separable suffixes (e.g., *.nsubj.sg, .dobj.pl*) similar to agglutinative languages (1:1);
- **Implicit**: the combination of number and case is expressed by unique suffixes without internal structure (e.g., *kar* for *.nsubj.sg*, *ker* for *.dobj.pl*) similar to fusional languages. This system displays exponence (many:1);
- **Implicit with declensions**: like the previous, but with three different paradigms each arbitrarily assigned to a different subset of the lexicon. This system displays exponence *and* flexivity (many:many).

A complete overview of our morphological paradigms is provided in Appendix A.3 All our languages have moderate inflectional synthesis and, in terms of fusion, are exclusively concatenative. Despite this, the effect on vocabulary size is substantial: 180% increase by overt and implicit case marking, 250% by implicit marking with declensions (in the full data setting).

Results Results are shown in the plots of Figure 2 (detailed numerical scores are given in Appendix A.4). We find that reducing training size has, not surprisingly, a major effect on translation quality. Among source language variants, fixed-order obtains the highest quality across all setups. In terms of BLEU (2(a)), the spread among variants increases somewhat with less data however differences are small. A clearer picture emerges from RIBES (2(b)), whereby less data clearly leads to more disparity. This is already visible in the 100k setup, with the fixed SOV language dominating the others. Case marking, despite being necessary to disambiguate argument roles in the absence of semantic cues, does not improve translation quality and

[10] We choose SOV because it is a commonly attested word order and is different from that of the target language, thereby requiring some non-trivial reorderings during translation.

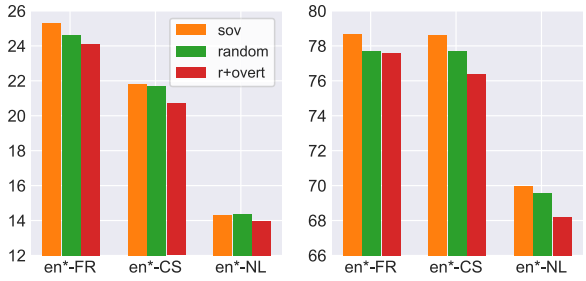


(a) Europarl-Test BLEU (b) Europarl-Test RIBES (c) Challenge RIBES

Figure 2: EN*-FR Transformer NMT quality versus training data size (x-axis). Source language variants: Fixed-order (SOV) and free-order (random) with different case systems (r+overt/implicit/declens). Scores averaged over three training runs. Detailed numerical results are provided in Appendix A.4.

even degrades it in the low-resource setup. Looking at the challenge set results (2(c)) we see that the free-order case-marking languages are clearly disadvantaged: In the mid-resource setup, case marking improves substantially over the underspecified *random,no-case* language but remains far behind fixed-order. In low-resource, case marking notably hurts quality even in comparison with the underspecified language. These results thus demonstrate that free-order case-marking languages require more data than their fixed-order counterparts to be accurately translated by state-of-the-art NMT.[11] Our experiments also show that this greater learning difficulty is not only due to case marking (and subsequent data sparsity), but also to word order flexibility (compare *sov+overt* to *r+overt* in Figure 2).

Regarding different morphology types, we do not observe a consistent trend in terms of overall translation quality (Europarl-test): in some cases, the richest morphology (with declensions) slightly outperforms the one without declensions—a result that would deserve further exploration. On the other hand, results on the challenge set, where



(a) Europarl BLEU (b) Europarl RIBES

Figure 3: Transformer results for more target languages (100k training size). Scores averaged over 2 runs.

most words are case-marked, show that morphological richness inversely correlates with translation quality when data is scarce. We postulate that our artificial morphologies may be too limited in scope (only 3-way case and number marking) to impact overall translation quality and leave the investigation of richer inflectional synthesis to future work.

5.5 Effect of Target Language

All results so far involved translation *into* a fixed-order (SVO) language without case marking. To verify the generality of our findings, we repeat a subset of experiments with the same synthetic English variants, but using Czech or Dutch as target

[11] In the light of this finding, it would be interesting to revisit the evaluation of Bugliarello et al. (2020) in relation to varying data sizes.

languages. Czech has rich fusional morphology including case marking, and very flexible order. Dutch has simple morphology (no case marking) and moderately flexible, syntactically determined order.[12]

Figure 3 shows the results with 100k training sentences. In terms of BLEU, differences are even smaller than in English-French. In terms of RIBES, trends are similar across target languages, with the fixed SOV source language obtaining best results and the case-marked source language obtaining worst results. This suggests that the major findings of our study are not due to the specific choice of French as the target language.

6 Related Work

The effect of word order flexibility on NLP model performance has been mostly studied in the field of **syntactic parsing**, for instance, using Average Dependency Length (Gildea and Temperley, 2010; Futrell et al., 2015a) or head-dependent order entropy (Futrell et al., 2015b; Gulordava and Merlo, 2016) as syntactic correlates of word order freedom. Related work in **language modeling** has shown that certain languages are intrinsically more difficult to model than others (Cotterell et al., 2018; Mielke et al., 2019) and has furthermore studied the impact of fine-grained morphology features (Gerz et al., 2018) on LM perplexity.

Regarding the **word order biases of seq-to-seq models**, Chaabouni et al. (2019) use miniature languages similar to those of Section 4 to study the evolution of LSTM-based agents in a simulated iterated learning setup. Their results in a standard ''individual learning'' setup show, like ours, that a free-order case-marking toy language can be learned just as well as a fixed-order one, confirming earlier results obtained by simple Elman networks trained for grammatical role classification (Lupyan and Christiansen, 2002). Transformer was not included in these studies. Choshen and Abend (2019) measure the ability of LSTM- and Transformer-based NMT to model a language pair where the same arbitrary (non-syntactically motivated) permutation is applied to all source sentences. They find that Transformer is largely indifferent to the order of source words (provided this is fixed and consistent across training and test set) but nonetheless struggles to translate long dependencies actually occurring in natural data. They do not directly study the effect of order flexibility.

The idea of permuting dependency trees to generate synthetic languages was introduced independently by Gulordava and Merlo (2016) (discussed above) and by Wang and Eisner (2016), the latter with the aim of diversifying the set of treebanks currently available for language adaptation.

7 Conclusions

We have presented an in-depth analysis of how Neural Machine Translation difficulty is affected by word order flexibility and case marking in the source language. Although these common language properties were previously shown to negatively affect parsing and agreement prediction accuracy, our main results show that state-of-the-art NMT models, especially Transformer-based ones, have little or no bias towards fixed-order languages. Our simulated low-resource experiments, however, reveal a different picture, that is: Free-order case-marking languages require more data to be translated as accurately as their fixed-order counterparts. Because parallel data (like labeled data in general) are scarce for most of the world languages (Guzmán et al., 2019; Joshi et al., 2020), we believe this should be considered as a further obstacle to language equality in future NLP technologies.

In future work, our analysis should be extended to target language variants using principled alternatives to BLEU (Bugliarello et al., 2020), and to other typological features that are likely to affect MT performance, such as inflectional synthesis and degree of fusion (Gerz et al., 2018). Finally, the synthetic languages and challenge set proposed in this paper could be used to evaluate syntax-aware NMT models (Eriguchi et al., 2016; Bisk and Tran, 2018; Currey and Heafield, 2019), which promise to better capture linguistic structure, especially in low-resource scenarios.

Acknowledgments

Arianna Bisazza was partly funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.021.646. We would like to thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine HPC cluster,

[12]Dutch word order is very similar to German, with the position of S, V, and O depending on the type of clause.

and the anonymous reviewers for their helpful comments.

References

Duygu Ataman and Marcello Federico. 2018. An evaluation of two vocabulary reduction methods for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 97–110.

Matthew Baerman and Dunstan Brown. 2013. Case syncretism. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5301

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. https://doi.org/10.18653/v1/P17-1080

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613809

Yonatan Bisk and Ke Tran. 2018. Inducing grammars with and for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 25–35, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2704

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6401

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.149

Rahma Chaabouni, Eugene Kharitonov, Alessandro Lazaric, Emmanuel Dupoux, and Marco Baroni. 2019. Word-order biases in deep-agent emergent communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5166–5175, Florence, Italy. Association for Computational Linguistics.

David Chiang and Kevin Knight. 2006. An introduction to synchronous grammars. *Tutorial available at* http://www.isi.edu/~chiang/papers/synchtut.pdf.

Leshem Choshen and Omri Abend. 2019. Automatically extracting challenge sets for non-local phenomena in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*,

pages 291–303, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1509

Benrard Comrie. 1981. *Language Universals and Linguistic Typology*. Blackwell. Book.

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

Anna Currey and Kenneth Heafield. 2019. Incorporating source syntax into transformer-based neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 24–33, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5203

Matthew S. Dryer. 2013. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://wals.info/

Jinhua Du and Andy Way. 2017. Pre-reordering for neural machine translation: Helpful or harmful? *The Prague Bulletin of Mathematical Linguistics*, 108(1):171–182. https://doi.org/10.1515/pralin-2017-0018

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211. https://doi.org/10.1207/s15516709cog1402_1

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-sequence attentional neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 823–833, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1078

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015a. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341. https://doi.org/10.1073/pnas.1502134112, PubMed: 26240370

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015b. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1029

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310. https://doi.org/10.1111/j.1551-6709.2009.01073.x, PubMed: 21564213

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, MA.

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.

Kristina Gulordava and Paola Merlo. 2015. Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 121–130, Uppsala, Sweden. Uppsala University, Uppsala, Sweden. https://doi.org/10.18653/v1/N18-1108

Kristina Gulordava and Paola Merlo. 2016. Multi-lingual dependency parsing evaluation: A large-scale analysis of word order properties using artificial data. *Transactions of the Association for Computational Linguistics*, 4:343–356. https://doi.org/10.1162/tacl_a_00103

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English. 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 6100–6113. https://doi.org/10.18653/v1/D19-1632

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171. https://doi.org/10.1162/tacl_a_00306

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit Proceedings of Conference*, pages 79–86. International Association for Machine Translation.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. https://doi.org/10.1162/tacl_a_00343

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1166

Gary Lupyan and Morten H. Christiansen. 2002. Case, word order, and language learnability: Insights from connectionist modeling. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP

natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60. https://doi.org/10.3115/v1/P14-5010

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. https://doi.org/10.21236/ADA273556

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1491

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435. https://doi.org/10.18653/v1/D18-1039

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(1):4381. https://doi.org/10.1038/s41467-020-18073-9, PubMed: 32873773

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of RNNs with synthetic variations of natural languages. 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 3532–3542, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1356

Benoît Sagot. 2013. Comparing complexity measures. In *Computational Approaches to Morphological Complexity*. Paris, France. Surrey Morphology Group.

Karthik Abinav Sankararaman, Soham De, Zheng Xu, W. Ronny Huang, and Tom Goldstein. 2020. Analyzing the effect of neural network architecture on training performance. In *Proceedings of Machine Learning and Systems 2020*, pages 9834–9845.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Kaius Sinnemäki. 2008. Complexity trade-offs in core argument marking. *Language Complexity*, pages 67–88. John Benjamins. https://doi.org/10.1075/slcs.94.06sin

Dan I. Slobin. 1966. The acquisition of Russian as a native language. *The Genesis of Language: A Psycholinguistic Approach*, pages 129–148.

Dan I. Slobin and Thomas G. Bever. 1982. Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition*, 12(3):229–265. https://doi.org/10.1016/0010-0277(82)90033-6

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for

modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1503

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505. https://doi.org/10.1162/tacl_a_00113

Adina Williams, Tiago Pimentel, Hagen Blix, Arya D. McCarthy, Eleanor Chodroff, and Ryan Cotterell. 2020. Predicting declension class from form and meaning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6682–6695, Online. Association for Computational Linguistics.

Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2018. Exploiting pre-ordering for neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. https://doi.org/10.18653/v1/2020.acl-main.597

A Appendices

A.1 NMT Hyperparameters

In the toy parallel grammar experiments (§4), batch size of 64 (sentences) and 1K max update steps are used for all models. We train BiLSTM with learning rate 1, and Transformer with learning rate of 2 together with 40 warm-up steps by using noam learning rate decay. Dropout ratio of 0.3 and 0.1 are used in BiLSTM and Transformer models respectively. In the synthetic English variants experiments (§5), we set a constant learning rate of 0.001 for BiLSTM. We also increased batch size to 128, number of warm-up steps to 80K and update steps to 2M for all models. Finally, for 100k and 10k datasize experiments, we decreased the warm-up steps to 4K. During evaluation we chose the best performing model on validation set.

Nouns	**Verbs**
president / président	thank / remercier
man / homme	support / soutenir
woman / femme	represent / représenter
minister / ministre	defend / défendre
candidate / candidat	welcome / saluer
secretary / secrétaire	invite / inviter
commissioner / commissaire	attack / attaquer
child / enfant	respect / respecter
teacher / enseignant	replace / remplacer
student / étudiant	exploit / exploiter

Table 4: The English/French vocabulary used to generate the challenge set. Both singular and plural forms are used for each noun.

A.2 Challenge Set

The English-French challenge set used in this paper, and available at https://github.com/arianna-bis/freeorder-mt, is generated by a small synchronous context-free grammar and contains 7,200 simple sentences consisting of a subject, a transitive verb, and an object (see Table 4). All sentences are in the present tense; half are affirmative, and half negative. All nouns in the grammar can plausibly act as both subject and object of the verbs, so that an MT system must rely on sentence structure to get perfect translation accuracy. The sentences are from a general domain, but we specifically choose nouns and verbs with little translation ambiguity that are well represented in the Europarl corpus: Most have thousands of occurrences, while the rarest word has about 80. Sentence example (English side): *'The teacher does not respect the student.'* and its reverse: *'The student does not respect the teacher.'*

A.3 Morphological Paradigms

The complete list of morphological paradigms used in this work is shown in Table 5. The implicit language with exponence (many:1) uses only the suffixes of the 1^{st} (default) declension. The implicit language with exponence and flexivity (many:many) uses three declensions, assigned as

	Overt	Implicit		
		1st(default)	**2**nd	**3**rd
Unambiguous	.nsubj.sg	kar	par	pa
	.nsubj.pl	kon	pon	po
	.dobj.sg	kin	it	kit
	.dobj.pl	ker	et	ket
	.iobj.sg	ken	kez	ke
	.iobj.pl	kre	kr	re
Syncretic	.arg.sg	–	–	–
	.arg.pl	–	–	–

Table 5: The artificial morphological paradigms used in this work, extended from Ravfogel et al. (2019). 1st, 2nd and 3rd are the declensions in the flexive language.

follows: First, the list of lemmas extracted from the training set is randomly split into three classes,[13] with distribution 1st:60%, 2nd:30%, 3rd:10%. Then, each core verb argument occurring in the corpus is marked with the suffix corresponding to its lemma's declension.

A.4 Effect of Data Size and Morphological Features: Detailed Results

Table 6 shows the detailed numerical results corresponding to the plots of Figure 2 in the main text.

Eparl-BLEU	1.9M	100k	10k
original	38.3	26.9	11.0
SOV	37.9	25.3	8.8
SOV+overt	37.4	24.6	8.4
random	37.5	24.6	8.5
random+overt	37.3	24.1	7.8
random+implicit	37.3	24.3	7.1
random+declens	37.4	23.1	7.7
Eparl-RIBES	1.9M	100k	10k
original	84.9	80.1	67.5
SOV	84.5	78.7	64.1
SOV+overt	84.5	78.4	63.1
random	84.2	77.7	61.7
random+overt	84.4	77.6	61.6
random+implicit	84.3	77.4	59.8
random+declens	84.3	77.1	61.3
Challenge-RIBES	1.9M	100k	10k
original	97.7	92.2	74.2
SOV	97.2	89.8	69.5
SOV+overt	97.7	86.5	64.9
random	74.1	72.9	63.1
random+overt	98.1	84.5	57.4
random+implicit	97.5	85.4	54.8
random+declens	97.6	84.4	53.1

Table 6: Detailed results corresponding to the plots of Figure 2: EN*-FR Transformer NMT quality versus training data size (1.9M, 100K, or 10K sentence pairs). Source language variants: Fixed-order (SOV) and free-order (random) with different case systems (+overt/implicit/declens). Scores averaged over three training runs.

[13] See Williams et al. (2020) for an interesting account of how declension classes are actually partly predictable from form and meaning.