On Difficulties of Cross-Lingual Transfer with Order Differences: A Case Study on Dependency Parsing

Zhao Li

2019.4.3

- Investigate cross-lingual transfer and posit that an order-agnostic model will perform better when transferring to distant foreign languages.

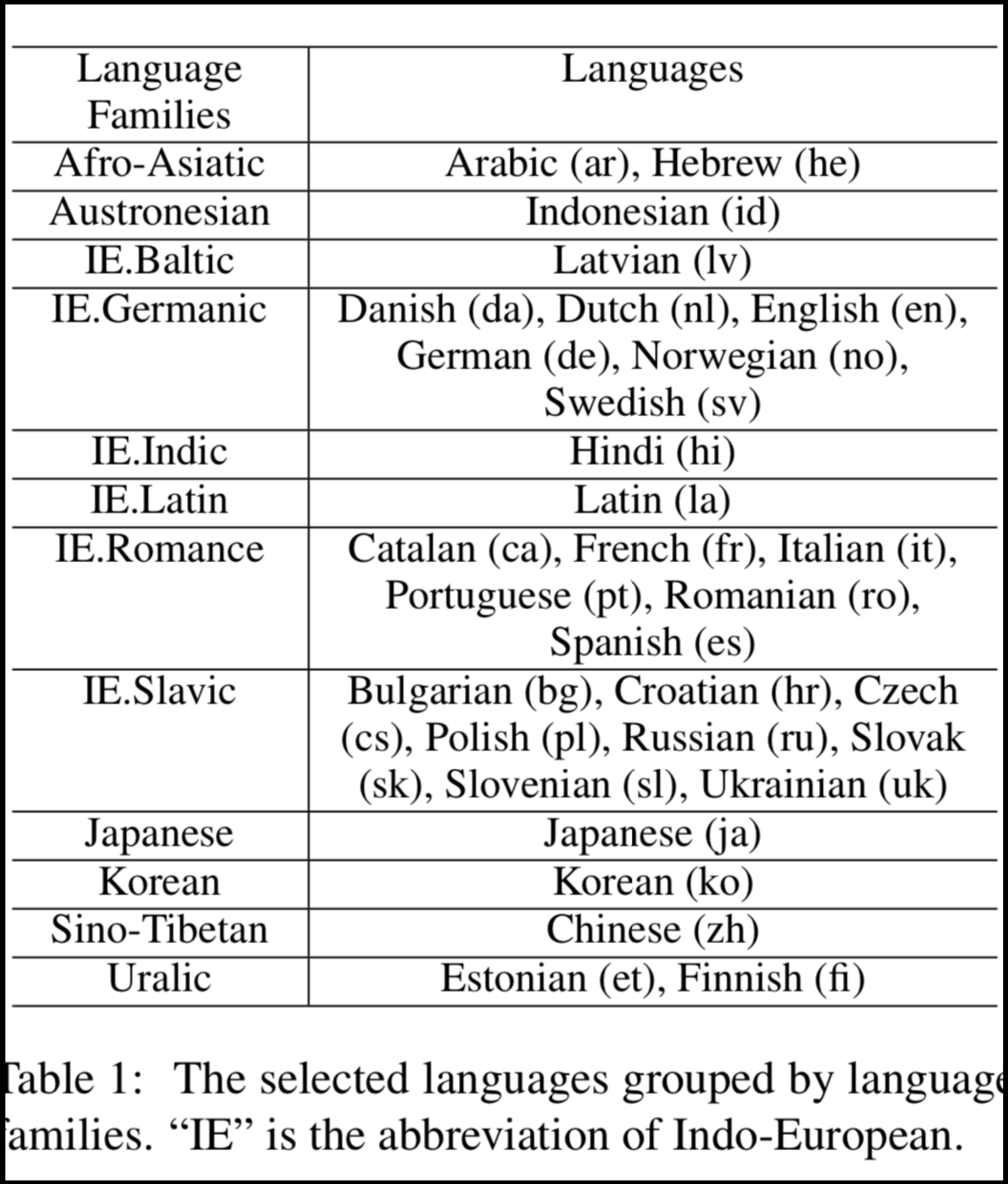
- Train dependency parsers on an English corpus and evaluate their transfer performance on 30 other languages.

- Compare encoders and decoders based on Recurrent Neural Networks (RNNs) and modified self-attentive architectures. The former rely on sequential information while the latter are more flexible at modeling token order.

- RNN-based architectures transfer well to languages that are close to English, while self-attentive models have better overall cross-lingual transferability and perform especially well on distant languages.

Quantifying Language Distance

- Measure “language distance” based on word order.
- 31 languages are selected, the total token number of a language is over 100K.
- Augment dependency labels: “(ModifierPOS, HeadPOS, DependencyLabel)”
- Use the relative frequency of the left-direction (modifier before head) as the directional feature.



Language Families	Languages
Afro-Asiatic	Arabic (ar), Hebrew (he)
Austronesian	Indonesian (id)
IE.Baltic	Latvian (lv)
IE.Germanic	Danish (da), Dutch (nl), English (en), German (de), Norwegian (no), Swedish (sv)
IE.Indic	Hindi (hi)
IE.Latin	Latin (la)
IE.Romance	Catalan (ca), French (fr), Italian (it), Portuguese (pt), Romanian (ro), Spanish (es)
IE.Slavic	Bulgarian (bg), Croatian (hr), Czech (cs), Polish (pl), Russian (ru), Slovak (sk), Slovenian (sl), Ukrainian (uk)
Japanese	Japanese (ja)
Korean	Korean (ko)
Sino-Tibetan	Chinese (zh)
Uralic	Estonian (et), Finnish (fi)

able 1: The selected languages grouped by languag amilies. “IE” is the abbreviation of Indo-European.

Quantifying Language Distance

- Concatenate the directional features of all triples, obtain a word-ordering feature vector for each language.

- Calculate the word-ordering distance using these vectors(Manhattan distance), then perform hierarchical clustering based on the word-ordering vectors.

- Outliers: de and nl adopt a larger portion of Object-Verb order in embedded clauses.

- Result: Word ordering is a major feature to characterize distance between languages

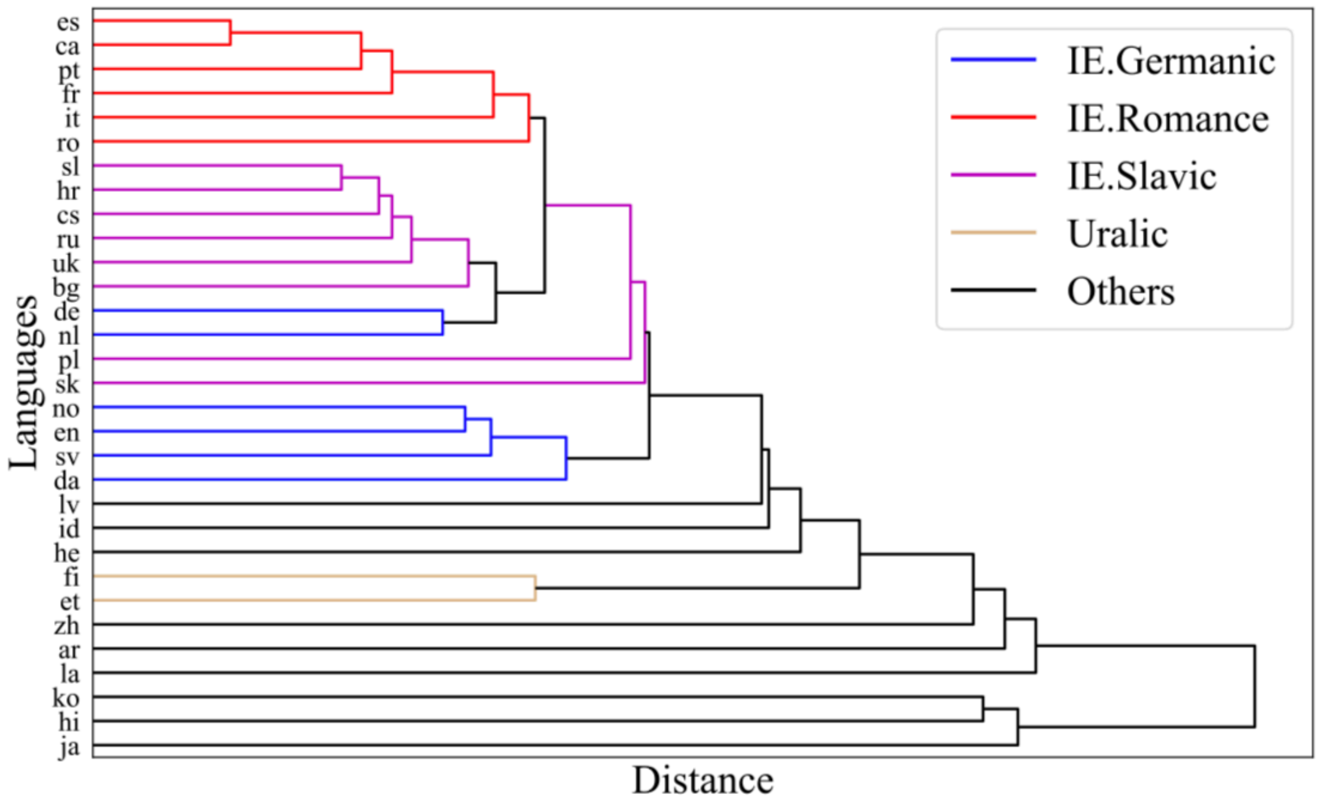


Figure 1: Hierarchical clustering (with the Nearest Point Algorithm) dendrogram of the languages by their word-ordering vectors.

Models

- Goal: Conduct cross-lingual transfer of syntactic dependencies without any annotation in the target languages.
- Model structure: embedding layer(word+POS) -> encoder -> decoder.
- Hypothesis: The models capturing less language-specific information of the source language will have better transfer ability.
- Focus on the word order information, and explore different encoders and decoders that are considered as *order-sensitive* and *order-free*, respectively.

Contextual Encoders

- Modeling words one by one with RNN in the sequence inevitably encodes word order information, which may be specific to the source language.
- Self-attention based encoder is RNN Encoder: k-layer bi-LSTMs (*order-sensitive*).
- Self-Attention Encoder: utilize relative position representations without directional information instead of absolute positional embedding (*order-free*). Less sensitive to word order.

Structured Decoders

- Stack-Pointer Decoder: transition-based, RNN is utilized to record the decoding trajectory (*order-sensitive*).

- Biaffine Decoder: graph-based, self-attentive output layer (*order-free*).

Experiments

- Take English as the source language and 30 other languages as target languages.

- “SelfAtt-Graph (OF-OF)”: oder-free encoder and order-free decoder.

- Baseline: shift-reduce transition-based parser, which gave previous SOTA results for single-source zero-resource cross-lingual parsing (Guo et al., 2015).

- Supervised learning results using the “RNN-Graph” model on each language as a reference of the upper-line for cross-lingual parsing.

Results (test sets)

- Lexicalized models of zh and ja performed poorly because their embeddings were not well aligned to English, so delexicalized models are used (POS tags).
- “SelfAtt-Graph” model performs the best.
- Compared with the baseline, the superior results show the importance of the contextual encoder.
- Compared with the supervised models, the cross-lingual results are still lower by a large gap, indicating space for improvements.

Lang	Dist. to English	SelfAtt-Graph (OF-OF)	RNN-Graph (OS-OF)	SelfAtt-Stack (OF-OS)	RNN-Stack (OS-OS)	Baseline (Guo et al., 2015)	Supervised (RNN-Graph)
en	0.00	90.35/88.40	90.44/88.31	90.18/88.06	**91.82**†/**89.89**†	87.25/85.04	90.44/88.31
no	0.06	80.80/72.81	80.67/72.83	80.25/72.07	**81.75**†/**73.30**†	74.76/65.16	94.52/92.88
sv	0.07	80.98/73.17	81.23/73.49	80.56/72.77	**82.57**†/**74.25**†	71.84/63.52	89.79/86.60
fr	0.09	77.87/72.78	**78.35**†/**73.46**†	76.79/71.77	75.46/70.49	73.02/64.67	91.90/89.14
pt	0.09	**76.61**†/67.75	76.46/**67.98**	75.39/66.67	74.64/66.11	70.36/60.11	93.14/90.82
da	0.10	76.64/67.87	77.36/68.81	76.39/67.48	**78.22**†/**68.83**	71.34/61.45	87.16/84.23
es	0.12	74.49/66.44	**74.92**†/**66.91**†	73.15/65.14	73.11/64.81	68.75/59.59	93.17/90.80
it	0.12	80.80/75.82	**81.10**/**76.23**†	79.13/74.16	80.35/75.32	75.06/67.37	94.21/92.38
hr	0.13	**61.91**†/**52.86**†	60.09/50.67	60.58/51.07	60.80/51.12	52.92/42.19	89.66/83.81
ca	0.13	73.83/65.13	**74.24**†/**65.57**†	72.39/63.72	72.03/63.02	68.23/58.15	93.98/91.64
pl	0.13	**74.56**†/**62.23**†	71.89/58.59	73.46/60.49	72.09/59.75	66.74/53.40	94.96/90.68
uk	0.13	**60.05**/**52.28**†	58.49/51.14	57.43/49.66	59.67/51.85	54.10/45.26	85.98/82.21
sl	0.13	**68.21**†/**56.54**†	66.27/54.57	66.55/54.58	67.76/55.68	60.86/48.06	86.79/82.76
nl	0.14	68.55/60.26	67.88/60.11	67.88/59.46	**69.55**†/**61.55**†	63.31/53.79	90.59/87.52
bg	0.14	**79.40**†/**68.21**†	78.05/66.68	78.16/66.95	78.83/67.57	73.08/61.23	93.74/89.61
ru	0.14	60.63/51.63	59.99/50.81	59.36/50.25	**60.87**/**51.96**	55.03/45.09	94.11/92.56
de	0.14	**71.34**†/**61.62**†	69.49/59.31	69.94/60.09	69.58/59.64	65.14/54.13	88.58/83.68
he	0.14	**55.29**/**48.00**†	54.55/46.93	53.23/45.69	54.89/40.95	46.03/26.57	89.34/84.49
cs	0.14	**63.10**†/**53.80**†	61.88/52.80	61.26/51.86	62.26/52.32	56.15/44.77	94.03/91.87
ro	0.15	**65.05**†/**54.10**†	63.23/52.11	62.54/51.46	60.98/49.79	56.01/44.04	90.07/84.50
sk	0.17	**66.65**/**58.15**†	65.41/56.98	65.34/56.68	66.56/57.48	57.75/47.73	90.19/86.38
id	0.17	**49.20**†/**43.52**†	47.05/42.09	47.32/41.70	46.77/41.28	40.84/33.67	87.19/82.60
lv	0.18	70.78/49.30	**71.43**†/**49.59**	69.04/47.80	70.56/48.53	62.33/41.42	83.67/78.13
fi	0.20	66.27/48.69	**66.36**/**48.74**	64.82/47.50	66.25/48.28	58.51/38.65	88.04/85.04
et	0.20	**65.72**†/**44.87**†	65.25/44.40	64.12/43.26	64.30/43.50	56.13/34.86	86.76/83.28
zh*	0.23	**42.48**†/**25.10**†	41.53/24.32	40.56/23.32	40.92/23.45	40.03/20.97	73.62/67.67
ar	0.26	**38.12**†/**28.04**†	32.97/25.48	32.56/23.70	32.85/24.99	32.69/22.68	86.17/81.83
la	0.28	**47.96**†/**35.21**†	45.96/33.91	45.49/33.19	43.85/31.25	39.08/26.17	81.05/76.33
ko	0.33	**34.48**†/**16.40**†	33.66/15.40	32.75/15.04	33.11/14.25	31.39/12.70	85.05/80.76
hi	0.40	**35.50**†/**26.52**†	29.32/21.41	31.38/23.09	25.91/18.07	25.74/16.77	95.63/92.93
ja*	0.49	**28.18**†/**20.91**†	18.41/11.99	20.72/13.19	15.16/9.32	15.39/08.41	89.06/78.74
Average	0.17	**64.06**†/**53.82**†	62.71/52.63	62.22/52.00	62.37/51.89	57.09/45.41	89.44/85.62

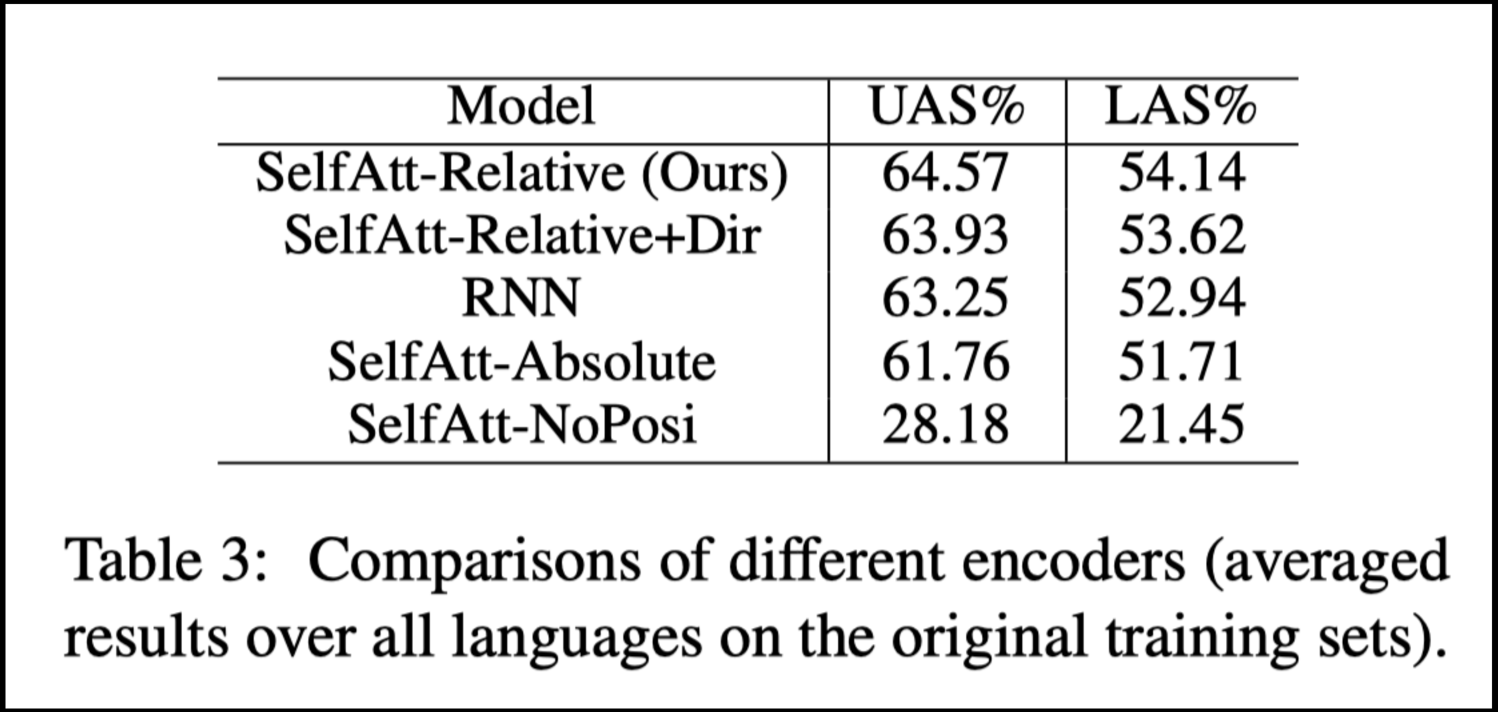
Take a closer look

- RNN-based models perform better at languages that are near English (upper rows in the table), while for languages that are “distant” from English, the “SelfAtt-Graph” performs much better.
- The patterns correspond well with the motivation: the design of models considering word order information is crucial in cross-lingual transfer.

Analysis

- We hypothesize that models that are less sensitive to word order can be better at cross-lingual transfer.

- We conduct controlled comparisons on various encoders with the same graph-based decoder.



Model	UAS%	LAS%
SelfAtt-Relative (Ours)	64.57	54.14
SelfAtt-Relative+Dir	63.93	53.62
RNN	63.25	52.94
SelfAtt-Absolute	61.76	51.71
SelfAtt-NoPosi	28.18	21.45

Table 3: Comparisons of different encoders (averaged results over all languages on the original training sets).

- “SelfAttNoPosi” is the self-attention model without any positional information. Although it is most insensitive to word order, it performs poorly possibly because of the lack of access to the locality of contexts.

- The self-attention model with absolute positional embeddings (“SelfAtt-Absolute”) also does not perform well.

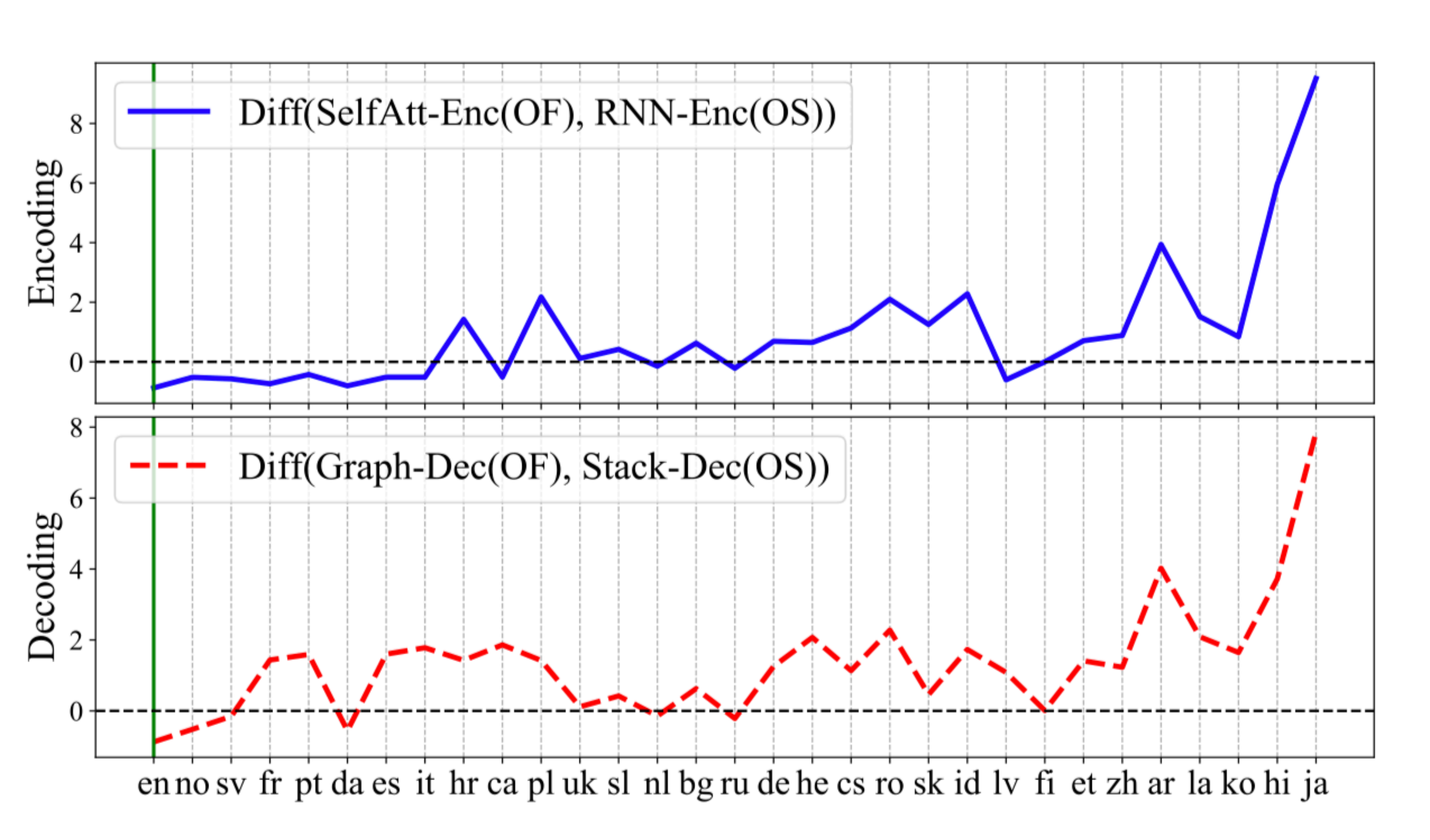
- In the case of parsing, relative positional representations may be more useful as indicated by the improvements bring by the directional relative position representations (“SelfAttRelative+Dir”)

On The Overall Pattern

- For each target language, we collect two types of distances when comparing it to English: one is the **word-ordering distance** as described before, the other is **performance distance**, which is the gap of evaluation scores between the target language and English.

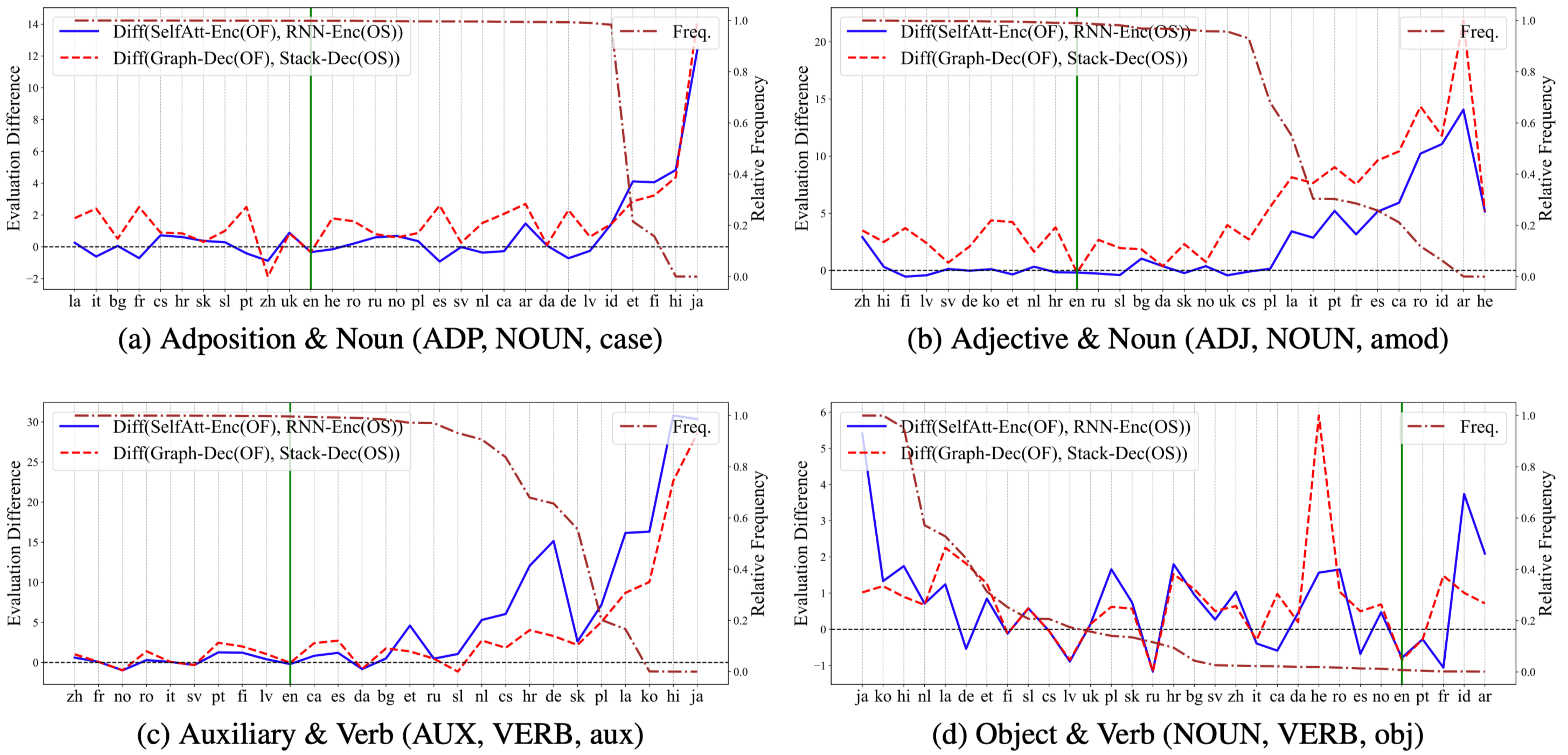
- We calculate the correlation of these two distances on all the concerned languages, and the results turn to be quite high: the Pearson and Spearman correlations are around 0.90 and 0.87 respectively, using the evaluations of any of our four cross-lingual transfer models.

- The word order is indeed an essential factor of cross-lingual transferability.



- We individually analyze the encoders and decoders of the dependency parsers. Since we have two architectures for each of the modules, when examining one, we take the highest scores obtained by any of the other module.

- The order-free models in general perform better than order-sensitive ones in the languages that are distant from the source language English.



(a) Adposition & Noun (ADP, NOUN, case)

(b) Adjective & Noun (ADJ, NOUN, amod)

(c) Auxiliary & Verb (AUX, VERB, aux)

(d) Object & Verb (NOUN, VERB, obj)

Dependency Types

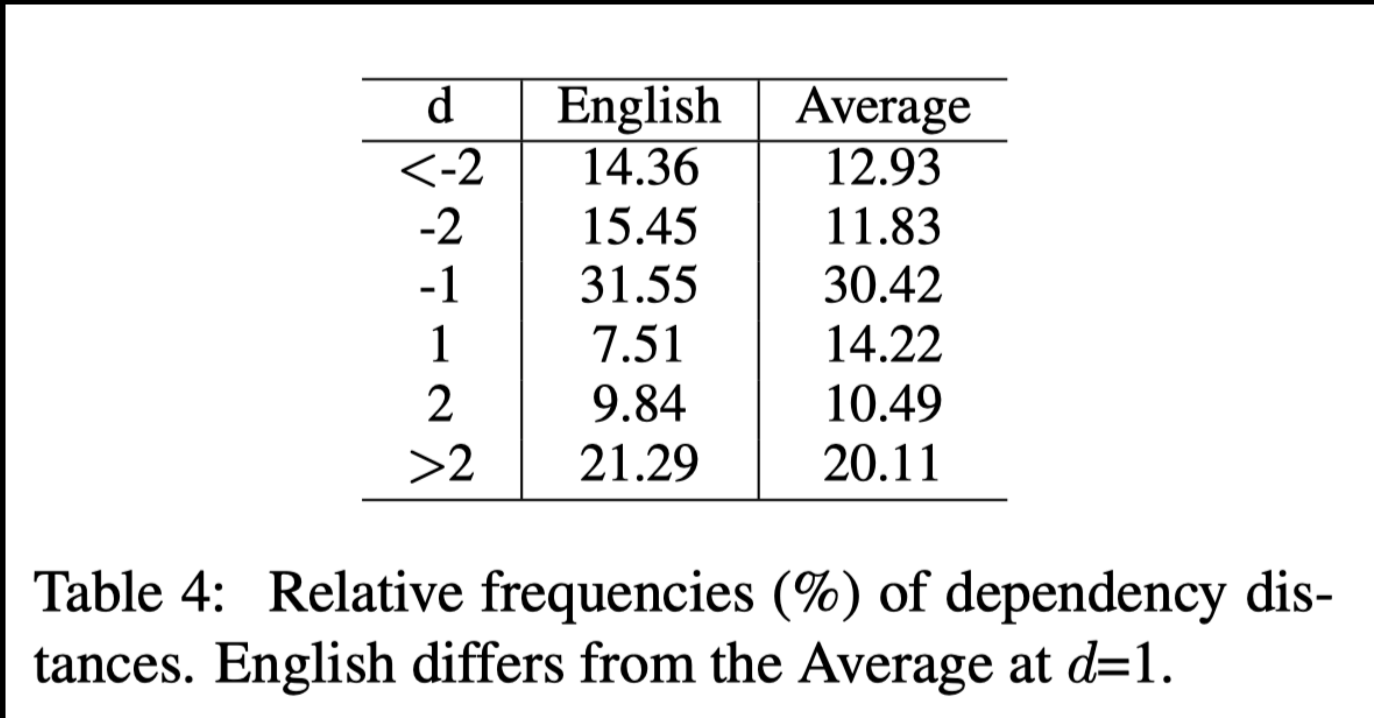
The brown curve and right y-axis represents the relative frequency of left-direction (modifier before head) on this type. The languages (x-axis) are sorted by this relative frequency from high to low.

Dependency Distances

- Dependency distances |d|=1: for all transfer models, evaluation scores on d=-1 can reach about 80, but on d=1, the scores are only around 40.

- About 80% of the dependencies with |d|=1 in English is the left direction (modifier before head), while overall other languages have more right directions at |d|=1.

- This suggests an interesting future direction of training on more source languages with different dependency distance distributions.



d	English	Average
<-2	14.36	12.93
-2	15.45	11.83
-1	31.55	30.42
1	7.51	14.22
2	9.84	10.49
>2	21.29	20.11

Table 4: Relative frequencies (%) of dependency distances. English differs from the Average at *d*=1.

Dependency Distances

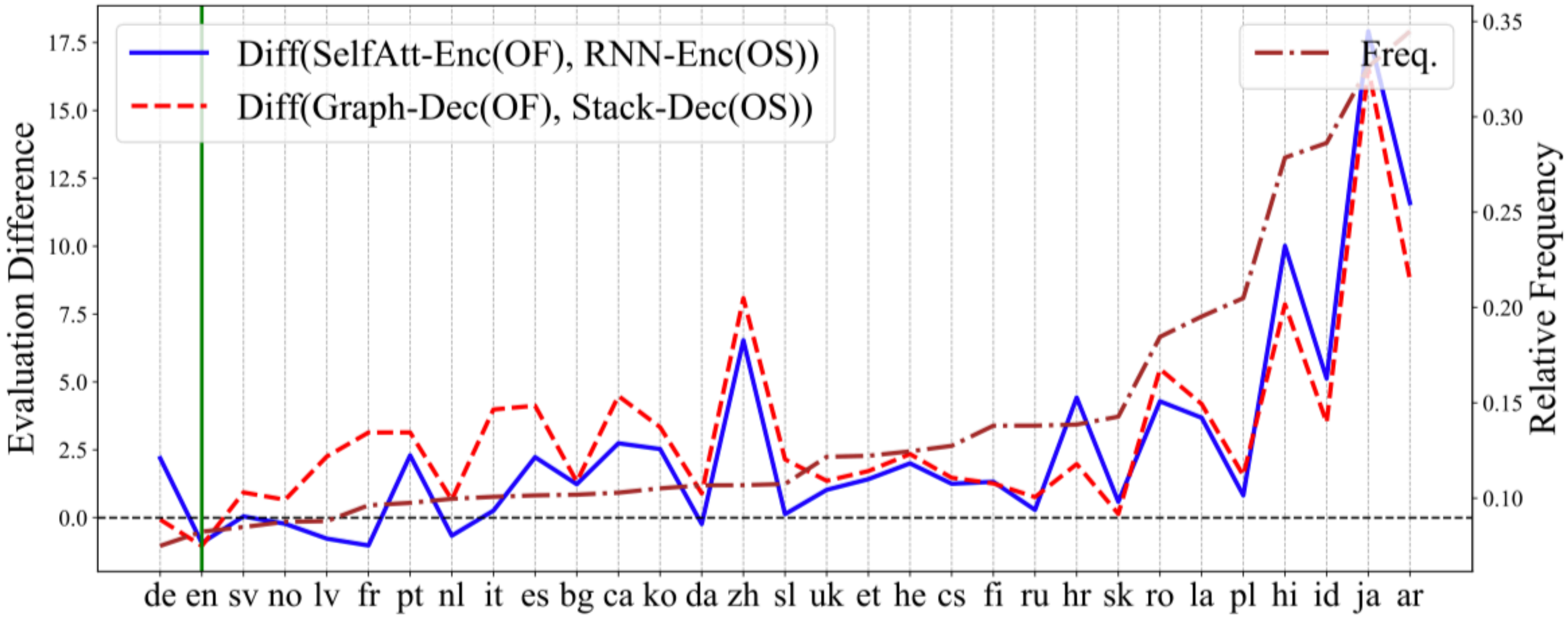


Figure 4: Evaluation differences of models on *d*=1 dependencies. Annotations are the same as in Figure 3, languages are sorted by percentages (represented by the brown curve and right *y*-axis) of *d*=1 dependencies.