Isomorphic Transfer of Syntactic Structures in Cross-Lingual NLP

Edoardo Maria Ponti
LTL, University of Cambridge
ep490@cam.ac.uk

Roi Reichart
Technion, IIT
roiri@ie.technion.ac.il

Anna Korhonen
LTL, University of Cambridge
alk23@cam.ac.uk

Ivan Vulić
LTL, University of Cambridge
iv250@cam.ac.uk

Abstract

The transfer or share of knowledge between languages is a popular solution to resource scarcity in NLP. However, the effectiveness of cross-lingual transfer can be challenged by variation in syntactic structures. Frameworks such as Universal Dependencies (UD) are designed to be cross-lingually consistent, but even in carefully designed resources trees representing equivalent sentences may not always overlap. In this paper, we measure cross-lingual syntactic variation, or *anisomorphism*, in the UD treebank collection, considering both morphological and structural properties. We show that reducing the level of anisomorphism yields consistent gains in cross-lingual transfer tasks. We introduce a source language selection procedure that facilitates effective cross-lingual parser transfer, and propose a typologically driven method for syntactic tree processing which reduces anisomorphism. Our results show the effectiveness of this method for both machine translation and cross-lingual sentence similarity, demonstrating the importance of syntactic structure compatibility for boosting cross-lingual transfer in NLP.

1 Introduction

Linguistic information can be transferred from resource-rich to resource-poor languages using approaches such as annotation projection, model transfer, and/or translation (Agić et al., 2014). Such cross-lingual transfer may rely on syntactic information. Structured and more cross-lingually consistent than linear sequences (Ponti, 2016), syntactic information has proved useful for cross-lingual parsing (Tiedemann, 2015; Rasooli and Collins, 2017), multilingual representation learning (Vulić and Korhonen, 2016; Vulić, 2017), causal relation identification (Ponti and Korhonen, 2017), and neural machine translation (Eriguchi et al., 2016; Aharoni and Goldberg, 2017). It can also guide the generation of synthetic data for multilingual tasks (Wang and Eisner, 2016).

Universal Dependencies (UD) (Nivre et al., 2016) is a collection of treebanks for a variety of languages, annotated with a scheme optimised for knowledge transfer. The tag sets are language-independent and there are direct links between content words. This reduces the variation of dependency trees, because content words are cross-lingually more stable than function words (Croft et al., 2017), and benefits semantically-oriented applications (de Marneffe et al., 2014)[1]. Importantly, although UD is tailored to offer support to cross-lingual transfer, it also supports monolingual applications with a quality comparable to language-specific annotations (Vincze et al., 2017, *inter alia*).

Despite the careful design of this resource, there are still substantial variations in morphological richness and strategies employed to express the same syntactic constructions across languages. These variations posit challenges for syntax-based knowledge transfer. The first challenge is how to match the source and target languages so that differences are minimised. The common criteria are based on the typology of word order (Naseem et al., 2012; Täckström et al., 2013; Zhang and Barzilay, 2015) or part-of-speech n-grams (Rosa and Zabokrtsky, 2015; Agić, 2017). The second one is how to make knowledge transfer effective by harmonising syntactic trees (Smith and Eisner, 2009; Vilares et al., 2016) as to enable a better correspondence between source and target nodes.

[1] It is controversial whether it improves parsing: e.g., Groß and Osborne (2015, *inter alia*) argue against whereas Attardi et al. (2015, *inter alia*) argue in favour.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 1531–1542
Melbourne, Australia, July 15 - 20, 2018.

In this paper we address these two challenges. We propose the concept of *isomorphism* (i.e., identity of shapes: syntactic structures) and its opposite, *anisomorphism*, as a probe to measuring quantitatively the extent to which syntactic tree pairs are cross-lingually compatible. We assess the variation of syntactic constructions by **a)** the average Zhang and Shasha (1989)'s tree edit distance between UD treebanks, and **b)** the variation in morphology by the Jaccard index of morphological feature sets. We show that these metrics are strong indicators for source language selection, and even preferable over widespread metrics such as genealogical language relatedness.

Moreover, the concept of isomorphism facilitates the process of reshaping trees to make them compatible across languages via operations of deletion, addition, and relabeling. To this end, we propose a tree processing method which increases the level of isomorphism between trees of cross-lingually compatible sentences. This method leads to consistent improvements on cross-lingual tasks achieved through transfer.

To verify the relevance of isomorphism for cross-lingual transfer in NLP, we perform experiments on three tasks. Firstly, we use the Jaccard index of morphological feature sets to choose source languages for cross-lingual dependency parsing. Secondly, we use syntactic trees harmonised by our method in syntax-based neural machine translation of two typologically distant language pairs (Arabic to Dutch; Indonesian to Portuguese). Finally, we evaluate cross-lingual sentence similarity in a real-life resource-lean scenario where the target language has no annotated data. In all experiments, we enhance performance compared to baselines where the source shows a lower degree of isomorphism.

In §2, we define the concept of (an)isomorphism, propose novel metrics for measuring it quantitatively, and introduce the tree processing algorithm. We then desribe the data (§3), methods (§4), and experimental results (§5). Related work is summarised in §6 and conclusions are drawn in §7.

2 Anisomorphism

The ideal situation for knowledge transfer from one (syntactic) structure into another is when these structures are equivalent. In graph theory, there is isomorphism between the nodes V_S of graph S and the nodes V_T of graph T if there exists a bijection $f(V_S) \rightarrow V_T$ such that $\forall s_i, s_j \in V_S$, it holds that: $s_i \bullet\!\!-\!\!\bullet s_j \Leftrightarrow f(s_i) \bullet\!\!-\!\!\bullet f(s_j)$, where the symbol $\bullet\!\!-\!\!\bullet$ stands for adjacency between nodes. In simple words, the mapping must preserve adjacencies between corresponding nodes.

Syntactic trees are a special case of such graphs. However, vocabularies (the words in their nodes) are peculiar to each language, making their comparison impractical across languages. In this work, we probe isomorphism on delexicalised trees, where each node is the (cross-lingually consistent) dependency relation of the word in that position. Even so, however, isomorphic bijection is often impossible between trees of equivalent sentences in different languages owing to typological variation (see §2.1).

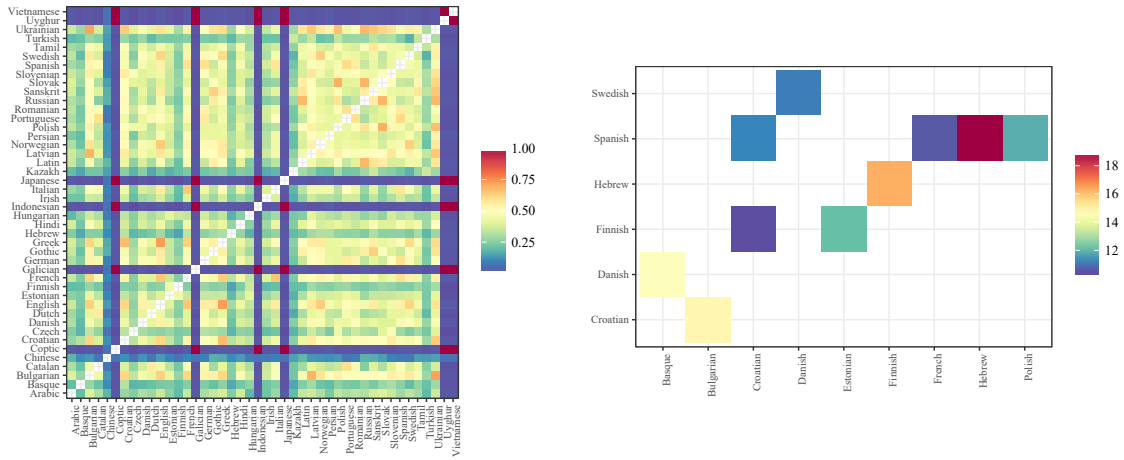
Adopting the term from Ambati (2008), we define this property as *an*isomorphism, which can be quantified as the extent to which two structures differ in their morphological and syntactic properties (§2.2). We present a tree processing method to mitigate anisomorphism in §2.3. Afterwards, in §§3-5 we show how the concepts defined in this section facilitate cross-lingual transfer in three NLP tasks.

2.1 Sources of Anisomorphism

Two main causes underpin anisomorphism. The first cause is the *morphological type* of a language: the same grammatical function may be expressed via morphemes, via separate words (so-called function words), or may not be expressed at all (Bybee, 1985, ch. 2). For instance, consider the following Latin-English example:

(1) *Crimen* *er-it* *super-is* *et*
crime.NOM be-FUT.3SG god-DAT.PL also
me *fec-isse* *nocent-em.*
me make-INF.PST guilty-ACC
'It will be a reproach to the gods, that they have made even me guilty.'

The future tense is expressed by inflecting the verb *erit* in Latin, whereas English has the auxiliary verb *will*. In addition, Latin can express the English preposition *to* with the dative case *-is*. This variation has systematic impact on UD annotation. On one hand, Latin would display the attribute-value pairs TENSE=FUTURE and CASE=DATIVE among the features of *erit* and *superis*. On the other hand, in English the function words (*will* and *to*) add nodes to the dependency structure, modifying the equivalent words (*be* and *gods*). This pattern is not unique to English and Latin: there are similar correspondences between specific function words and morphological features in many other languages.



(a) Jaccard index of the morphological feature sets.

(b) Average tree edit distance.

Figure 1: Heatmaps of anisomorphism metrics for UD language pairs. The colours range from blue (low values) to red (high values).

The other source of anisomorphism are *construction strategies*: the same syntactic construction is expressed through different types of strategies (Croft et al., 2017), which results in different kinds of subtrees in UD. An example construction is *predicative possession*, which conveys the ownership of an item by a possessor through the predicate of a clause (Stassen, 2009). Consider these examples in Dutch and Arabic, respectively:

(2) *Ik* *heb* *een* *filmidee*
I have.1SG a film+idea
'I have an idea for a movie.'

(3) *Laday-himā* *'ašyā-'u* *muštarakat-un*
at-them thing-NOM.PL common-NOM.PL
'They have things in common.'

In Dutch (Example 2), the owner *Ik* is the subject and the item *filmidee* is the object of a transitive verb (*hab*). However, in Arabic (Example 3) the owner is a predicate with a locative prefix (*laday-himā*), the item *'ašyā'u* is the subject, and there is no verb. These are called transitive and locative strategies, respectively. Each strategy results in a different (delexicalised) subtree, as shown in Figure 2b: this simple example with one construction already suggests that the variation in syntactic constructions affects the compatibility of cross-lingual trees pervasively.[2]

[2]Other strategies for predicative possession include *topic*, *conjunctional* and *genitive*. More examples of constructions are available in the supplemental material.

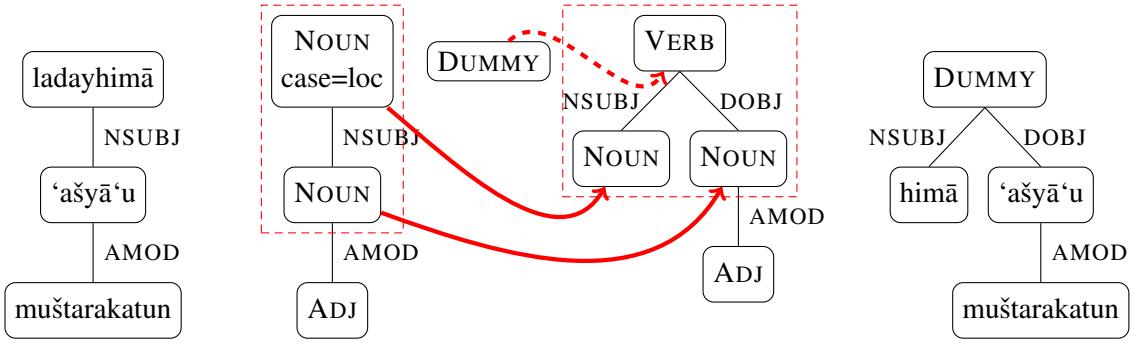
2.2 Measures of Anisomorphism

How can the differences described in §2.1 translate into quantitative metrics of compatibility between sentences in different languages? As the first answer to this question, we propose to measure the affinity in morphological type by considering the sets of morphological features attested within each of the UD treebanks.[3] Particularly, for each pair of a source language set M_S and a target language set M_T, we estimate their Jaccard index, which is defined for two sets as the cardinality of their intersection divided by the cardinality of their union, as shown in Equation (4).

$$J(M_S, M_T) = \frac{||M_S \cap M_T||}{||M_S \cup M_T||} \quad (4)$$

The values of the Jaccard index lie in $[0, 1]$ A heatmap is displayed in Figure 1a: the morphological similarity between language pairs varies considerably, ranging from low (0.07 in Chinese-Uyghur), mild (0.48 in Latvian-Tamil), to high (0.72 in Bulgarian-Ukrainian). Note that the Jaccard index 1 is an artifact for languages with no expression of grammatical function (in Vietnamese, among others) or lacking morphological annotation (in Japanese). This metric exhibits other disadvantages: it does not take into account another source of variation, the construction strategies, and is based on general properties of a grammar rather

[3]The full list of features can be consulted at http://universaldependencies.org/u/feat/



(a) Original (source) tree (b) Match templates (c) Processed (source) tree

Figure 2: Tree processing steps that transform the locative strategy for predicative possession in an Arabic sentence into a transitive strategy. Tree processing is always applied on source language constructions.

than specific individual sentences.

Hence, we propose an approach to measure anisomorphism between individual sentences. We parse the texts of the multi-parallel Bible corpus (Christodouloupoulos and Steedman, 2015) with the SyntaxNet parser (see §4). The language pairs taken into account are limited to those present both in our UD sample and in the Bible corpus, and sentence-aligned by book, chapter and verse indices. For a given language pair, we estimate the tree edit distance between every corresponding pair of sentence trees S and T with the Zhang-Sasha algorithm (Zhang and Shasha, 1989) and then average over the number of trees.[4]

This tree edit distance operates on ordered trees with node (but not edge) labels, hence it is suited for delexicalised dependencies. In particular, it is defined over a map M, which is a list of node pairs where the former belongs to S or ϵ (empty node), and the latter belongs to T or ϵ. If both are non-empty, they trigger an operation of *relabeling*; if the latter is ϵ, it is *deletion*; if the former is ϵ, it is *addition*. The edit distance is the number of operations required for a complete transformation weighted by the factor γ.[5] The following equation summarises the tree edit distance measure:

$$\gamma(M, S, T) = \sum_{i,j \in M} \gamma(S_i \to T_j) + \gamma(S_i \to \epsilon) + \gamma(\epsilon \to T_j)$$

The possible values of this metric are non-negative real numbers. We opted for this metric in particular because it allows the insertion of internal nodes but not transpositions. The former criterion allows to capture complex transformations without rebuilding entire subtrees, the latter is aimed at taking into account also variations in word order. In order to evaluate pure syntactic isomorphism one should allow for transpositions and/or operate on unordered trees.[6]

A heatmap of tree edit distances is shown in Figure 1b. The values reflect the typological affinity of the language pairs: e.g., Spanish is very close to French (both are Romance languages), mildly similar to Polish (Slavic language, but still part of the Indo-European family), but remote from Hebrew (from a different family, Semitic). The values agree in part with the metrics of Figure 1a, where the Jaccard indices of Hebrew (0.26), Polish (0.46), and French (0.59) mirror the same relationships.

In §4, we show how these metrics can benefit the source selection for knowledge transfer, sometimes even outranking established criteria such as genealogical closeness. However, they have also weaknesses: the Jaccard index of feature sets is not reliable for languages with a limited number of morphologically expressed grammatical categories. On the other hand, the tree edit distance measure requires resources (such as treebanks and parallel corpora) that are not available for many languages.

2.3 Reduction of Anisomorphism

The measures of anisomorphism reveal which languages are structurally similar, which is directly useful for source selection. However, the data available for many tasks are often limited to distant languages. Hence, it is necessary to increase their affinity by gearing one towards the other. We propose to process source dependency trees with an algorithm inspired by the same rules of the tree edit

[4] We implement this algorithm with the *zs* Python package, available at https://github.com/timtadh/zhang-shasha.

[5] For simplicity, we set $\gamma = 1$.

[6] For a survey on tree edit distances, see Bille (2005).

distance described in §2.2.

We leverage the readily available documentation in typological databases (e.g., World Atlas of Language Structures: WALS) (Dryer and Haspelmath, 2013).[7] Given a source and a target language, the documentation informs about their respective strategies. For each strategy, we manually define a 'template', i.e. the subtree it corresponds to, in terms of morpho-syntactic features. For instance, see the dashed circles in Figure 2b: note that templates are limited to a head and its immediate dependents.

Then we explore source trees in a top-down breadth-first fashion, and if a template for a source strategy is identified, it is mapped to the corresponding target template. In order to preserve semantic information, contrary to Zhang and Shasha (1989), the mapping operates on lexicalised edge-labeled trees. Hence, ADD and CHANGE affect both words (nodes) and edges (dependency relations). The whole process is summarised in Algorithm 1.

Algorithm 1 Tree processing with rules

```
strategies_s ← WALS_s                                   ▷ Define templates
strategies_t ← WALS_t
function CHANGE(s, t^(l))                               ▷ Define operations
    s ← t^(l)
function DELETE(s)
    s ← ϵ
function ADD(t^(l))
    ϵ ← t^(l)
function MAPPING(r_s, strategies_t)                     ▷ Define mapping
    assert(r_s ∈ strategies_s)
      return {CHANGE, DELETE, ADD}*
for subtree in tree_s do                                ▷ Explore tree
    if subtree ∈ strategies_s then
        list ← MAPPING(subtree)
        for n_s, n_t in list do                         ▷ Perform operations
            if n_s ≠ ϵ ∧ n_t ≠ ϵ then
                CHANGE(n_s, n_t)
            else if n_t = ϵ then
                DELETE(n_s)
            else if n_s = ϵ then
                ADD(n_t)
```

For instance, consider the transformation from the locative strategy for predicative possession in Arabic from Example 3 into a transitive strategy. By exploring its dependency graph (Figure 2a), the Algorithm identifies a subtree corresponding to one of the source strategies (left side of Figure 2b). This subtree is mapped to the target template (right side of Figure 2b) with the following operations: it CHANGEs the root noun *ladayhimā* (the possessor) with a dummy node (the verb). The same noun is re-ADDed as a dependent with a new label *nsubj*. Finally, the dependency relation of the other noun *'ašyā-'u* is CHANGEd from *nsubj* to *dobj*. The resulting tree uses the source language vocabulary, but target language construction strategies, as shown in Figure 2c.

3 Data

In order to validate the usefulness of anisomorphism reduction through guided source selection and tree processing, we experiment with three different cross-lingual tasks: *cross-lingual dependency parsing*, *neural machine translation (NMT)*, and *cross-lingual sentence similarity (STS)*. In this section, we present the data used in these tasks.

The data for dependency parsing are sourced from Universal Dependencies v1.4.[8] We sample a group of 21 treebanks ensuring their representativeness by balancing them by family. We filter out all languages but two belonging to same branches of the Indo-European family, and keep those of all the other families.[9] We take into account only the language-independent components of the annotation: coarse POS tags, morphological features, and dependency relations.

Regarding NMT data, English is ubiquitous in the current datasets, overshadowing the wide spectrum of existing morphological types and syntactic strategies. To address this limitation, we create a *new NMT dataset* that matches typologically distant languages directly without the need of a bridge/pivot language. We extract aligned sentences from the Open Subtitles 2016 tokenised corpus (Tiedemann, 2009)[10] for Arabic-Dutch and Indonesian-Portuguese. This choice was made based on their volume of parallel data in order to produce evaluation data similar in size to those of NMT datasets in shared tasks such as WMT16 (Bojar et al., 2016). Training and test sets consist of 3M and 5K sentences, respectively. These sentences come automatically annotated by SyntaxNet.

The data for cross-lingual STS are chosen to resemble a real-world scenario with a resource-poor target language. The training data (9,709 sentence

[7] In particular, we take into account the following relevant WALS features: 116 (polar questions), 122-123 (relativisation on subjects and obliques), 117 (predicative possession), 113-115 (negation), 107 (passive), 37-38 (articles), and 85 (prepositions).

[8] http://universaldependencies.org/

[9] Language names are substituted in this work by their corresponding ISO 639-1 codes. A table of names and codes is provided in the supplemental material.

[10] http://opus.nlpl.eu/OpenSubtitles.php

pairs) are in English, taken from the STS benchmark, the ensemble of all the datasets from SemEval 2012-2017 STS tasks. The test data (250 sentence pairs) come from Task 1 of SemEval 2017 (Cer et al., 2017); target language is Arabic.[11] All the sentence pairs are associated with a label ranging from 0 (dissimilarity) to 5 (equivalence).

4 Methodology

Cross-lingual Dependency Parsing. To assess if the anisomorphism metrics devised in §2.2 are reliable in finding compatible languages for knowledge transfer, we use the Jaccard index of the morphological feature sets as a criterion to choose source languages for cross-lingual parser transfer. We adopt the variant of delexicalised model transfer (Zeman and Resnik, 2008) for this task. This technique ignores lexicalised features and leverages only language-independent features instead.

For each language from a sample of 7 (typologically diverse) targets, we report LAS scores using three different source languages: (1) the highest-ranked source according to the Jaccard index; (2) a source sampled from the middle of the list ranked by the Jaccard indices; (3) a very dissimilar language sampled from the bottom of the ranked list. The total number of sentences used for training corresponds to the smallest of the three source language treebanks in order to isolate the effect of treebank size on the final transfer results.

We conduct experiments with two well-known transition-based parsers (Nivre, 2006): (1) *DeSR* (Attardi et al., 2007) and (2) *SyntaxNet* (Andor et al., 2016; Alberti et al., 2017). The two were selected as they represent two different architectures: the former is an SVM-based model with a polynomial kernel, whereas the latter is a feed-forward neural network with beam search based on conditional random fields. The results are evaluated in terms of LAS and UAS scores.

Neural Machine Translation. For NMT, we examine whether the tree processing procedure from §2.3 can reduce anisomorphism between source and target language syntactic structures. We thus run NMT models in two settings: with and without the anisomorphism reduction procedure.

For this experiment we rely on a state-of-the-art syntax-aware NMT architecture. We report its performance by BLEU scores (Papineni et al., 2002). In particular, we use an attentional encoder-decoder network that jointly learns to translate and align words (Bahdanau et al., 2015) implemented in the Nematus suite[12] (Sennrich et al., 2017). The encoder is a bidirectional gated recurrent network. For each step *i*, the decoder predicts the next word in output by taking as input the current hidden state h_i, the previous word w_{i-1} and a context vector, i.e., a weighted sum of all the hidden states $\sum_{j=1}^{n} w_j \cdot h_1$. The weights are learned by a multi-layer perceptron that estimates the likelihood of the alignment between the predicted word and each of the input words: $w_{i,j} = P(a|y_i, x_j)$.

This model is enriched with additional linguistic features on input, as proposed by Sennrich and Haddow (2016). In particular, we select the following which are proven as useful in prior work, and also relevant to our experiment: word form, POS tag, and dependency relations. These features are concatenated and fed to the encoder. Tree processing from §2.3 affects these features (and consequently the sentence representation) by changing the initial tree structure. For instance, the original tree in Figure 2a and the processed one in Figure 2c would correspond to these feature sets:

Original	**Preprocessed**
ladayhimā ⊕ N ⊕ ROOT	*himā* ⊕ N ⊕ NSUBJ
	DUMMY ⊕ V ⊕ ROOT
ʻašyāʻu ⊕ N ⊕ NSUBJ	*ʻašyāʻu* ⊕ N ⊕ DOBJ
muštarakatun ⊕ A ⊕ Amod	*muštarakatun* ⊕ A ⊕ Amod

Cross-lingual STS. We use cross-lingual STS as another evaluation task to validate if the anisomorphism reduction algorithm from §2.3 generalises beyond the initial application in NMT. The state-of-art approach to this task in the monolingual setting encodes trees of sentence pairs with a TreeLSTM architecture (Tai et al., 2015). The hidden representations of the tree roots of both sentences in each pair are then concatenated and fed to a multi-layer perceptron classifier, which yields a probability distribution over the six classes (from 0=dissimilarity to 5=equivalence).

The following TreeLSTM has been implemented in PyTorch. The parameters of an LSTM model are the matrix weights W_q for inputs and U_q for hidden representations, and a bias b_q. q corresponds to an input gate i_t, a forget gate f_t, an output gate o_t, or a memory cell c_t at time step t. The hidden state h_t

[11] http://alt.qcri.org/semeval2017/task1/

[12] https://github.com/EdinburghNLP/nematus

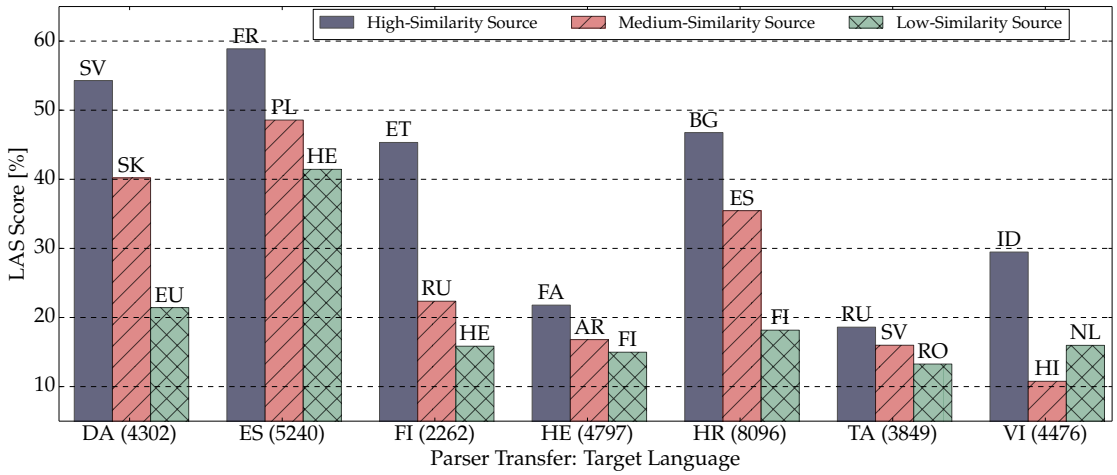


Figure 3: Results of delexicalised cross-lingual transfer using DeSR. Results with SyntaxNet are omitted as they show very similar patterns. The numbers in parentheses denote the amount of training sentences.

is derived from the equations below. To extend this model to dependency trees, we consider h_{t-1} to equal the sum of the hidden states of the children of a node $\sum_{k \in C(x_t)} h_k$, and provide a different forget gate f_{tk} for each child.

$$q_t = \sigma\left(W_q x_t + U_q h_{t-1} + b_q\right) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \quad (6)$$

$$h_t = o_t \odot \tanh(c_t) \quad (7)$$

In our resource-lean cross-lingual scenario the language of the training data (English) differs from that of the target (Arabic). Since TreeLSTM is a lexicalised model, we employ multilingual word embeddings, such that the words of both languages lie in the shared cross-lingual semantic space. In particular, we map English into Arabic through the iterative Procustes method devised by Artetxe et al. (2017). The results are evaluated through the Pearson correlation and the Mean Squared Error (MSE) between predicted and golden labels.

Hyperparameters. DeSR has degree 2, γ 0.18, C 0.4, $coef_0$ 0.4, and ϵ 1.0. The hyper-parameters for the deep models are shown in Table 1: we have followed the training setup suggestions from prior work for all the models used in our experiments.

	SyntaxNet (Parsing)	Nematus (NMT)	TreeLSTM (STS)
Hidden layers	2	2	1
Hidden size	512	1000	300
Input size	160	280	512
Batch size	256	80	25
Epochs	12 (greed); 10 (beam)	Early stopping	5
Learning rate	0.8	1^{-4}	1^{-2}
Optimiser	Adam	AdaDelta	SGD
Dropout	0.2 / 0.3	0.1 / 0.2	0

Table 1: Hyper-parameters of the models.

5 Results and Discussion

Source Selection. The results for cross-lingual parser transfer with the DeSR parser are provided in Figure 3, while the results with SyntaxNet are provided as supplemental material as they follow the same trends. The selection of the source for delexicalised cross-lingual parsing based on the proposed Jaccard index measure shows than selecting a source language with a lower degree of anisomorphism is crucial for knowledge transfer. The values for the selected languages are listed in Table 2.

Target	High	Mid	Low
Danish	0.49	0.39	0.19
Spanish	0.59	0.46	0.26
Finnish	0.44	0.23	0.15
Hebrew	0.31	0.24	0.15
Croatian	0.62	0.46	0.25
Tamil	0.48	0.43	0.38
Vietnamese	1.00	0.02	0.01

Table 2: Jaccard indices of source-target pairs.

The high-similarity source always outperforms the alternatives with both DeSR and SyntaxNet, and with respect to both LAS and UAS scores. For instance, Swedish is the best source for Danish, Estonian for Finnish, and Bulgarian for Croatian. Similarly, the preference for medium- over low-

	AR-NL	ID-PT
Baseline	7.01	14.79
+Syntax	14.40	23.70
++Preprocessing	**15.40**	**24.12**

Table 3: NMT results: BLEU scores of a joint translator and aligner *(Baseline)*, fed with linguistic features *(+Syntax)*, and with processed trees to reduce anisomorphism *(++Preprocessing)*.

	Pearson	MSE
Mono-lingual	77.9	0.94
Cross-lingual	44.7	1.82
+Preprocessing	**48.0**	**1.64**

Table 4: Cross-lingual STS results: Pearson and MSE scores of the TreeLSTM architecture with original and processed trees.

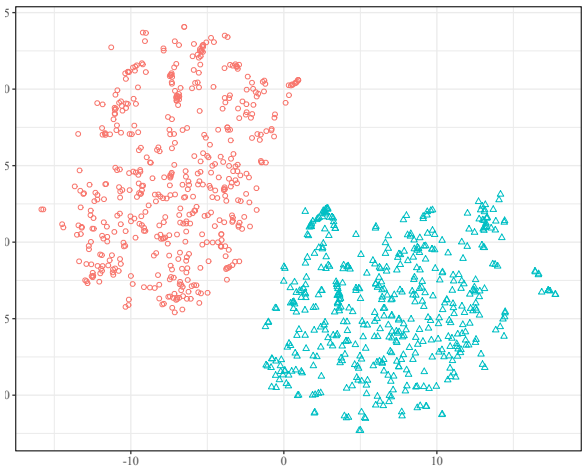


Figure 4: Hidden representations of original (red circles) and processed (blue triangles) sentences.

ranking languages is pronounced, too, as it holds for 6 groups out of 7. For instance, Slovak is a better source choice for Danish than Basque, Polish is a better source choice for Spanish than Hebrew.

Most notably, our findings generalise even to cases when the top-ranking language (e.g. Farsi) does not belong to the language family of the target (e.g. Hebrew) whereas the language with a medium overlap does (e.g. Arabic).

Tree Processing. The results of the experiments also corroborate the idea that tree harmonisation informed by linguistic typology, and implemented through our anisomorphism reduction procedure can assist model transfer in cross-lingual tasks. The BLEU scores for Neural Machine Translation, shown in Table 3, reveal consistent improvements. The model enriched with syntactic features outperforms the baseline with joint translation and alignment without syntactic features by 7.39 BLEU points in Arabic-Dutch and 8.91 BLEU points in Indonesian-Portuguese. Importantly, our extension which reduces anisomorphism by processing syntactic trees in the source language leads to further improvements for both language pairs: it surpasses the model with syntactic features by 1.0 BLEU points in Arabic-Dutch, and 0.42 BLEU points in Indonesian-Portuguese.

These results support our hypotheses: **a)** syntax is pivotal in NMT, confirming findings from prior work (Sennrich et al., 2017); **b)** the tree processing algorithm from §2.3 facilitates the alignment between source and target words, and also grants the encoder-decoder architecture a better leverage of dependency features. This lends support to our argument that anisomorphism limits the ability of models to generalise beyond single languages, and reducing it can help cross-lingual syntax-aware NLP tasks.

A similar conclusion can be reached by comparing the performance of TreeLSTM-based models on the cross-lingual STS task, reported in Table 4. In particular, the Pearson correlation score increases by 3.3 points and MSE decreases by 0.18 points when our tree processing algorithm is applied. We inspect the hidden representations of both original and processed sentences with t-SNE dimensionality reduction in Figure 4. The impact of the algorithm becomes evident as their clusters are completely separate. However, the comparison against the monolingual STS score obtained on the English test set shows that there is still a wide gap to be bridged by cross-lingual knowledge transfer.

Note that our tree processing algorithm is guided by typological knowledge in WALS. The results of the NMT and cross-lingual STS tasks suggest that existing knowledge in such large typological databases (O'Horan et al., 2016; Bender, 2016) can be readily used to support cross-lingual transfer tasks in NLP, as well as the interpretation of polyglot neural models (Ponti et al., 2017). We hope that our work will spark further research on the use of typology in cross-lingual NLP applications.

6 Related Work

The need to account for discrepancies in tree structures emerged early in the domain of Information Theory: in particular, the tree edit distance turned out to be useful for correcting programming scripts (Tai, 1979), evolution studies, and most notably accounting for transformations in constituency trees (Selkow, 1977). Although previous works were aware of the problem of anisomorphism in the context of syntax-based NLP applications (Ambati, 2008), to our knowledge we are the first to quantify it formally and to leverage it in cross-lingual NLP.

For source selection, similarity metrics from prior work mostly relied on information stored in typological databases (Naseem et al., 2012; Täckström et al., 2013; Zhang and Barzilay, 2015; Deri and Knight, 2016). Otherwise, the metrics were derived empirically: they mostly concerned linear-order properties such as part-of-speech n-grams (Rosa and Zabokrtsky, 2015; Agić, 2017). In domain adaptation, the selection also hinges upon topic models (Plank and Van Noord, 2011) or Bayesian Optimisation (Ruder and Plank, 2017). The metrics we defined in §2.2 are instead based on configurational properties of languages, and add another piece to the puzzle of source selection.

The idea of tree processing dates back to the attempts to steer source towards target syntactic structures in statistical MT, although they were mostly limited to simple reordering steps.

Gildea (2003) proposed cloning operations to relocate subtrees. Other works learned rewrite patterns in an automatic fashion to minimize differences in the order of chunks (Zhang et al., 2007) or labeled dependencies (Habash, 2007). Instead, Smith and Eisner (2009) proposed to learn jointly a translation and a loose alignment of nodes, in order to avoid enforcing the bias of the source structure. Reviving these approaches within the framework of deep learning seems crucial as far as state-of-art models depend on syntactic information (Eriguchi et al., 2016; Dyer et al., 2016).

In general, our approach aims at developing and evaluating models focused on specific constructions rather than languages as a whole (Rimell et al., 2009; Bender, 2011; Rimell et al., 2016). The gist is that current models have reached a plateau in performance because they excel with frequent and simple phenomena, but they still lag behind with respect to rarer or more complex constructions.

7 Conclusions and Future Work

We have demonstrated that syntactic structures differ across languages even in well-developed annotation schemes such as Universal Dependencies. This variation stems from morphological and syntactic differences across languages. This phenomenon, which we have labeled as anismorphism, can challenge the transfer of knowledge from one language to another. We have proposed novel methodology which reduces the degree of anisomorphism cross-lingually 1) by selecting the most compatible languages for transfer, and 2) by editing the syntactic structures (i.e., trees) themselves.

First, we have provided two measures of anisomorphism based on Jaccard distance of morphological feature sets, as well as average tree edit distance of parallel sentences. These can provide reliable indicators for language compatibility for source selection in cross-lingual parsing.

Second, we have proposed a new method for fine-tuning source dependency trees to resemble target language trees in order to reduce anisomorphism. The method does not depend on parallel data, and it leverages readily available information in typological databases. It boosts the performance of standard frameworks in two downstream applications, obtaining competitive or state-of-art results for 1) NMT on a new dataset of Arabic-Dutch and Indonesian-Portuguese and 2) cross-lingual sentence similarity.

Future work will look into automating the tree processing procedure. A parametrised model could be trained to imitate the operations performed by Zhang and Shasha (1989)'s algorithm on multi-parallel texts, conditioned on the tree features and previous operations. Another possible research direction is learning the mapping between structures from parallel texts jointly with a main task, in the spirit of quasi-synchronous grammars (Smith and Eisner, 2009). Finally, a wider range of syntactic constructions could be covered by inferring typological strategies from texts (Östling, 2015; Coke et al., 2016).

The data for NMT, and the code for our cross-lingual STS are available at the following link: `github.com/ducdauge/isotransf`.

Acknowledgements

This work is supported by the ERC Consolidator Grant LEXICAL (no 648909). The authors would like to thank the anonymous reviewers.

References

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10.

Željko Agić, Jörg Tiedemann, Kaja Dobrovoljc, Simon Krek, Danijela Merkler, and Sara Može. 2014. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *Proceedings of the EMNLP 2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 13–24.

Roee Aharoni and Yoav Goldberg. 2017. Towards string-to-tree neural machine translation. In *Proceedings of ACL*, pages 132–140.

Chris Alberti, Daniel Andor, Ivan Bogatyy, Michael Collins, Dan Gillick, Lingpeng Kong, Terry Koo, Ji Ma, Mark Omernick, Slav Petrov, et al. 2017. SyntaxNet models for the CoNLL 2017 shared task. *arXiv preprint arXiv:1703.04929*.

Vamshi Ambati. 2008. Dependency structure trees in syntax based machine translation. In *Adv. MT Seminar Course Report*, volume 137.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of ACL*, pages 2442—2452.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*, pages 451–462.

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Atanas Chanev, and Massimiliano Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using DeSR. In *Proceedings of EMNLP-CoNLL*, pages 1112–1118.

Giuseppe Attardi, Simone Saletti, and Maria Simi. 2015. Evolution of Italian treebank and dependency parsing towards Universal Dependencies. In *Proceedings of the Second Italian Conference in Computational Linguistics (CLiC-it)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR (Conference Papers)*.

Emily M Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Emily M. Bender. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3).

Philip Bille. 2005. A survey on tree edit distance and related problems. *Theoretical Computer Science*, 337(1-3):217–239.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of WMT*, volume 2, pages 131–198.

Joan L Bybee. 1985. *Morphology: A study of the relation between meaning and form*, volume 9. John Benjamins Publishing.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SEMEVAL*, pages 1–14.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Reed Coke, Ben King, and Dragomir R. Radev. 2016. Classifying syntactic regularities for hundreds of languages. *CoRR*, abs/1603.08016.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets Universal Dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of ACL*, pages 399–408.

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

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of NAACL-HLT*, pages 199–209.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-sequence attentional neural machine translation. In *Proceedings of ACL*, pages 823—-833.

Daniel Gildea. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of ACL*, pages 80–87.

Thomas Groß and Timothy Osborne. 2015. The dependency status of function words: Auxiliaries. In *Proceedings of the International Conference on Dependency Linguistics (DepLing)*, pages 111–120.

Nizar Habash. 2007. Syntactic preprocessing for statistical machine translation. *Proceedings of MT SUMMIT*.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of LREC*, pages 4585–4592.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of ACL*, pages 629–637.

Joakim Nivre. 2006. *Inductive dependency parsing*. Springer.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, pages 1659–1666.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING*, pages 1297–1308.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of ACL*, pages 205–211.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Barbara Plank and Gertjan Van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of ACL*, pages 1566–1576.

Edoardo Maria Ponti. 2016. Divergence from syntax to linear order in Ancient Greek lexical networks. In *Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 541–547.

Edoardo Maria Ponti and Anna Korhonen. 2017. Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. In *Proceedings of the EACL 2017 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 25–30.

Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2017. Decoding sentiment from distributed representations of sentences. In *Proceedings of *SEM*, pages 22–32.

Mohammad Sadegh Rasooli and Michael Collins. 2017. Cross-lingual syntactic transfer with limited resources. *Transactions of the ACL*, 5:279–293.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of EMNLP*, pages 813–821.

Laura Rimell, Jean Maillard, Tamara Polajnar, and Stephen Clark. 2016. RELPRON: A relative clause evaluation data set for compositional distributional semantics. *Computational Linguistics*, 42(4):661–701.

Rudolf Rosa and Zdenek Zabokrtsky. 2015. KLcpos3 - a language similarity measure for delexicalized parser transfer. In *Proceedings of ACL*, pages 243–249.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with bayesian optimization. In *Proceedings of EMNLP*, pages 372–382.

Stanley M. Selkow. 1977. The tree-to-tree editing problem. *Information Processing Letters*, 6(6):184–186.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: A toolkit for neural machine translation. In *Proceedings of EACL*, pages 65–68.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of WMT*, pages 83–91.

David A. Smith and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of EMNLP*, pages 822–831.

Leon Stassen. 2009. *Predicative possession*. Oxford University Press.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of NAACL-HLT*, pages 1061–1071.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of ACL*, pages 1556–1566.

Kuo-Chung Tai. 1979. The tree-to-tree correction problem. *Journal of the ACM*, 26(3):422–433.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of RANLP*, pages 237–248.

Jörg Tiedemann. 2015. Cross-lingual dependency parsing with universal dependencies and predicted POS labels. In *Proceedings of the International Conference on Dependency Linguistics (DepLing)*, pages 340–349.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2016. One model, two languages: Training bilingual parsers with harmonized treebanks. In *Proceedings of ACL*, pages 425–431.

Veronika Vincze, Katalin Ilona Simkó, Zsolt Szántó, and Richárd Farkas. 2017. Universal Dependencies and morphology for Hungarian–and on the price of universality. In *Proceedings of EACL*, pages 356–365.

Ivan Vulić. 2017. Cross-lingual syntactically informed distributed word representations. In *Proceedings of EACL*, pages 408–414.

Ivan Vulić and Anna Korhonen. 2016. Is "universal syntax" universally useful for learning distributed word representations? In *Proceedings of ACL*, pages 518–524.

Dingquan Wang and Jason Eisner. 2016. The Galactic Dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the ACL*, 4:491–505.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of IJCNLP*, pages 35–42.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18(6):1245–1262.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proceedings of EMNLP*, pages 1857–1867.

Yuqi Zhang, Richard Zens, and Hermann Ney. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 1–8.