Phylogenetic Multi-Lingual Dependency Parsing
Mathieu Dehouck, Pascal Denis

To cite this version:
Mathieu Dehouck, Pascal Denis. Phylogenetic Multi-Lingual Dependency Parsing. NAACL 2019 - Annual Conference of the North American Chapter of the Association for Computational Linguistics, Jun 2019, Minneapolis, United States. hal-02143747

HAL Id: hal-02143747
https://hal.archives-ouvertes.fr/hal-02143747
Submitted on 29 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

Languages evolve and diverge over time. Their evolutionary history is often depicted in the shape of a phylogenetic tree. Assuming parsing models are representations of their languages grammars, their evolution should follow a structure similar to that of the phylogenetic tree. In this paper, drawing inspiration from multi-task learning, we make use of the phylogenetic tree to guide the learning of multi-lingual dependency parsers leveraging languages structural similarities. Experiments on data from the Universal Dependency project show that phylogenetic training is beneficial to low resourced languages and to well furnished languages families. As a side product of phylogenetic training, our model is able to perform zero-shot parsing of previously unseen languages.

1 Introduction

Languages change and evolve over time. A community that spoke once a single language can be split geographically or politically, and if the separation is long enough their language will diverge in direction different enough so that at some point they might not be intelligible to each other. The most striking differences between related languages are often of lexical and phonological order but grammars also change over time.

Those divergent histories are often depicted in the shape of a tree in which related languages whose common history stopped earlier branch off higher than languages that have shared a longer common trajectory (Jäger, 2015). We hypothesize that building on this shared history is beneficial when learning dependency parsing models. We thus propose to use the phylogenetic structure to guide the training of multi-lingual graph-based neural dependency parsers that will tie parameters between languages according to their common history.

As our phylogenetic learning induces parsing models for every inner node in the phylogenetic tree, it can also perform zero-shot dependency parsing of unseen languages. Indeed, one can use the model of the lowest ancestor (in the tree) of a new language as an approximation of that language grammar.

We assess the potential of phylogenetic training with experiments on data from the Universal Dependencies project version 2.2. Our results show that parsers indeed benefit from this multi-lingual training regime as models trained with the phylogenetic tree outperform independently learned models. The results on zero-shot parsing show that a number of factors such as the genre of the data and the writing system have a significant impact on the quality of the analysis of an unseen language, with morphological analysis being of great help.

The remaining of this paper is organized as follows. Section 2 presents both the neural parsing model as well as the phylogenetic training procedure. Section 3 presents some experiments over data from UD 2.2. Section 4 presents some related works on multi-task learning and multi-lingual parsing. Finally, Section 5 closes the paper and gives some future perspectives.

2 Model

We propose a multi-task learning framework that shares information between tasks using a tree structure. The tree structure allows us to both share model parameters and training samples between related tasks. We instantiate it with a graph-based neural parser and use the language phylogenetic tree to guide the learning process, but it can in principle be used with any tree that encodes tasks relatedness and any learning algorithm that supports fine-tuning.

In this section we first describe the intuition be-
Figure 1: A possible phylogenetic tree for languages in the Slavic family.

2.1 Phylogenetic Hypothesis

Languages evolve from earlier stages and sometimes a language will change differently in different places leading to different languages with a common ancestor. This evolution process is often depicted in the shape of a tree in which leaves are actual languages and inner nodes can be either attested ancestral languages or their idealized reconstruction. Figure 1 gives an example of such a tree for a subset of the Slavic family of Indo-European languages (Simons and Fennig, 2018).

Just as languages evolve and diverge, so do their grammars. Assuming a parsing model is a parameterized representation of a grammar, then we can expect those models to evolve in a similar way. We thus take a multi-task approach to the problem of multi-lingual dependency parsing. What was once a single problem (e.g. parsing sentences in Proto-West-Slavic) becomes a set of distinct but related problems (parsing sentences in Czech, Polish, Slovak and Sorbian) as Proto-West-Slavic was evolving into its modern descendants.

We assume that the grammar of the last common ancestor is a good approximation of those languages grammars. Thus it should be easier to learn a language’s grammar starting from its ancestor grammar than from scratch. There are however some issues with this assumption. First, a language grammar can be very different from its ancestor one from two millennia earlier. Consider the difference between modern French and early Classical Latin for example, in two millennia Latin has witnessed the loss of its case system and a complete refoundation of its verbal system. And the "last common" ancestors can have very different age depending on the languages we consider. We expect the common ancestor of Tagalog and Indonesian to be much much older than the common ancestor of Portuguese and Galician. Second, a lot of languages have only started to be recorded very recently thus lacking historical data all together. And when historical records are available, much work still needs to be done to render those data usable by parsers. For example the Universal Dependencies Project (Nivre et al., 2018) only has annotated corpora for Latin, old Greek, old Church Slavonic and Sanskrit. And even for those classical languages, it is not clear to which extent their modern counterparts really descend from them. Thus we need to find another way to access the ancestor language grammar than using historical data.

We propose to use all the data from descendent languages to represent an ancestor language. In principle, one could give more weight to older languages or to languages that are known to be more conservative, but this knowledge is not available for all languages families. Thus we resort to using all the available data from descendent languages without distinction.

Another problem is that the tree view is too simple to represent the complete range of phenomena involved in language evolution, such as language contacts. Furthermore, languages do not evolve completely randomly, but follow some linguistic universals and have to keep a balance between speakability, learnability and understandability. Thus, languages can share grammatical features without necessarily being genetically related, either by contact or by mere chance. However, the tree model is still a good starting point in practice and language families align well with grammatical similarity as recent works on typological analysis of UD treebanks have shown (Chen and Gerdes, 2017; Schluter and Agić, 2017). We thus make the simplifying assumption that a language grammar evolves only from an older stage and can be approximated by that previous stage.

2.2 Neural Model

Our scoring model is an edge factored graph-based neural model in the vein of recent works by Dozat et al. (Dozat et al., 2017). There are two major differences here compared to the parser of Dozat
et al. The first difference is in individual word representation, for which we use only the UPOS\(^1\) tag, morphological information provided by UD tree-banks and a character based word representation, whilst Dozat et al. use also the XPOS\(^2\) tag, holistic word vectors (from Word2Vec (Mikolov et al., 2013) and their own) and they do not use morphological information beside what might already be given by the XPOS. The second difference is the scoring function proper. While they use biaffine scoring functions and decouple edge scoring from label scoring, we use a simple multi-layer perceptron to compute label scores and pick the max over the label as the edge score.

Let \(x = (w_1 w_2 \ldots w_l)\) be a sentence of length \(l\). Each word \(w_i\) is represented as the concatenation of 3 subvectors, one for its part-of-speech, one for its morphological attributes and one for its form:

\[
  w_i = \text{pos}_i \oplus \text{morph}_i \oplus \text{char}_i.
\]

The part-of-speech vector (\(\text{pos}_i\)) is from a lookup table. The morphological vector (\(\text{morph}_i\)) is the sum of the representation \(m_m\) of each morphological attribute \(m\) of the word given by the tree-banks:

\[
  \text{morph}_i = \sum_{m \in \text{morph}_i} m_m.
\]

We add a special dummy attribute representing the absence of morphological attributes.

The form vector (\(\text{char}_i\)) is computed by a character BiLSTM (Hochreiter and Schmidhuber, 1997). Characters are fed one by one to the recurrent neural network in each direction. The actual form vector is then the concatenation of the outputs of the forward character LSTM and of the backward character LSTM as depicted in Figure 2.

---

\(^1\)Universal part-of-speech for a set of 17 tags. Does not encode morphology.

\(^2\)Language specific part-of-speech. Might include morphological information, but is not available for all languages.
For each word, there are two terms. The first term enforces that for all potential governors that are neither the word itself nor its actual governor, their highest score (irrespective of the relation label) should be smaller than the score of the actual governor and actual label by a margin of 1. The second term is similar and enforces that for the actual governor, any label that is not the true label should have a score smaller than the score of the actual label again by a margin of 1.

2.3 Phylogenetic Training

Let $\mathcal{L} = \{l_1, l_2, ..., l_{n_l}\}$ be a set of $n_l$ languages and let $\mathcal{P} = \{p_1, p_2, ..., p_{n_p}\}$ be a set of $n_p$ proto-languages (hypothesized ancestors of languages in $\mathcal{L}$). Let $\mathcal{T}$ be a tree over $\mathcal{L}^* = \mathcal{L} \cup \mathcal{P}$ such that languages of $\mathcal{L}$ are leaves and proto-languages of $\mathcal{P}$ are inner nodes. This means that we assume no two languages in $\mathcal{L}$ share a direct parenthood relation, but they all descend both from a hypothesized parent. We could in principle have data appearing only in inner nodes. Tree $\mathcal{T}$ has a single root, a proto-language from $\mathcal{P}$ that all grammars descend from. This ancestor of all languages shall model linguistic universals\(^\text{3}\) and ensure we deal with a well-formed tree. We use the notation $p > l$ for the fact that language/node $l$ descends from language/node $p$.

For each language $l \in \mathcal{L}$, we assume access to a set of $n$ annotated examples $\mathcal{D}_l$. For each proto-language $p \in \mathcal{P}$, we create an annotated set $\mathcal{D}_p = \bigcup_{p \geq l} \mathcal{D}_l$ as the union of its descendant sets.

For each language $l \in \mathcal{L}^*$, we want to learn a parsing model $\theta_l$.

2.3.1 Model Evolution

The main idea behind phylogenetic training is to initialize a new model with the model of its parent, thus effectively sharing information between languages and letting models diverged and specialize over time. The training procedure is summarized in Algorithm 1.

At the beginning, we initialize a new blank/random model that will be the basic parsing model for all the world languages. Then, we sample sentences (we will discuss sampling issues in next section) randomly from all the available languages, parse them, compute the loss and update the model accordingly. Since the training sentences are sampled from all the available languages, the model will learn to be as good as possible for all the languages at the same time.

When the model $\theta_p$ has reached an optimum (that we defined hereafter), we pass a copy of it to each of its children. Thus, for each child $c$ of $p$, we initialize $\theta_0^p = \theta_p$ to its parent ($p$) final state. Each model $\theta_c$ is then refined on its own data set $\mathcal{D}_c$ which is a subset of $\mathcal{D}_p$, until it reaches its own optimum state and is passed down to its own children. This process is repeated until the model reaches a leaf language, where the model $\theta_c$ is eventually refined over its monolingual data set $\mathcal{D}_c$.

By passing down optimal models from older/larger languages sets to newer/smaller ones, models get the chance to learn relevant information from many different languages while specializing as time goes by.

The question now is to find when to pass down a model to its children. In other words, at which stage has a model learned the most it could from its data and should start to diverge to improve again?

Following the principle of cross-validation, we

---

\(^3\)It does not imply anything about our belief or not in the monoglotto genesis hypothesis.
propose to let held-out data decide when is the right time to pass the model down. Let \(D_p\) be a set of held-out sentences from the same languages as \(D_p\). Then, after every epoch \(i\) of \(k\) training examples, we freeze the model \(\theta_p\), and test it on \(k'\) sentences from \(D_p\). This gives a score \(a_i\) (UAS/LAS) to the current model. If the score is higher than the score of the previous model \(\theta_p^{i-1}\) then training goes on, otherwise we discard it and retrain \(\theta_p^{i-1}\) for another \(k\) sentences. If after having discarded \(r\) epochs in a raw we have not yet found a better one, then we assume we have reached an optimal model \(\theta_p^{i-1}\) and pass it on to its children (unless it is a leaf, in which training is over for that language).

2.3.2 Sentence Sampling
There are a few things we should consider when drawing examples from a proto-language distribution. Beside the question of whether some languages are more conservative than others with respect to their ancestor, which we have decided to simplify saying that all languages are as representative of their ancestors, there is the problem of data unbalance and tree unbalance.

Sampling sentences uniformly across languages is not a viable option for the size of datasets varies a lot across languages and that they do not correlate with how close a language is to its ancestor. For example, there are 260 Belarusian training sentences against 48814 Russian ones. The basic question is thus whether one should draw examples from languages or branches. Basic linguistic intuition tells us that drawing should be performed on branches. Modern languages distribution has no reason to reflect their proximity to their ancestor language. Amongst Indo-European languages, there are one or two Armenian languages as well as one or two Albanian languages (depending on the criteria for being a language), while there are tens of Slavic languages and Romance languages. However, there is no reason to believe that Slavic or Romance languages are better witnesses of proto-Indo-European than Armenian or Albanian.

Drawing examples from languages would bias the intermediate models toward families that have more languages (or more treebanks). It might be a good bias depending on the way one compute the overall accuracy of the system. If one uses the macro-average of the individual language parsers, then biasing models toward families with many members should improve the accuracy overall.

In this work, at a given inner node, we decided to sample uniformly at random over branches spanning from this node, then uniformly at random over languages and then uniformly at random over sentences. It boils down to flattening the subtree below an inner node to have a maximum depth of 2. For example in Figure 1, at the root (Proto-Slavic) we pick a branch at random (e.g. Proto-South-Slavic), then a language at random (e.g. Croatian) then a sentence at random. Given that we have picked the Proto-South-Slavic branch, all South-Slavic languages are then as likely to be chosen. This biases slightly the model toward bigger sub-families. In our example, Croatian and Serbian have the same chances to be sampled than Slovenian, therefore their family, Proto-Serbocroatian is twice as likely to be chosen as Slovenian is, while being at the same depth in the tree.

We could otherwise sample over branches, then over sub-branches again and again until we reach a leaf and only then pick a sentence. In this case, Proto-Serbocroatian and Slovenian would have the same probability to be chosen. This would give much more weight to languages high in the tree than languages low in the tree. While this would give more balance to the actual model, it could be detrimental to the averaged results since the data distribution is itself unbalanced. It would of course be possible to try any variation between those two, picking sub-branches according to a probability that would depend on the number of languages in that family for example, therefore mitigating the unbalance problem.

2.4 Zero-Shot Parsing
An interesting property of the phylogenetic training procedure is that it provides a model for each inner node of the tree and thus each intermediary grammar. If one were to bring a new language with its position in the tree, then we can use the pre-trained model of its direct ancestor as an initialization instead of learning a new model from scratch. Similarly, one can use this ancestor model directly to parse the new language, effectively performing zero-shot dependency parsing. We investigate this possibility in the experiment section.

3 Experiments
To assess the potential of phylogenetic training both in terms of multi-task learning and zero-shot parsing capabilities, we experimented with data from the Universal Dependencies project version
When several corpora are available for a language, we chose one to keep a good balance between morphological annotation and number of sentences. For example, the Portuguese GSD treebank has slightly more sentences than the Bosque treebank but it is not well morphologically annotated. The zero-shot parsing models have been directly tested on languages that lack of training set. The treebanks names are given in the tree 4 and the result table 1.

3.1 Setting

As some languages have no training data and unique writing systems making the character model inefficient for them, we resorted to use gold parts-of-speech and morphological attributes rather than predicted ones. For example, Thai has no training data, no language relative and a unique script, which altogether make it really hard to parse (from a phylogenetic perspective).

The phylogenetic tree used for the experiment is adapted from the Ethnologue (2018). For space reasons, it is reported in the appendix in Figures 4 and 5. We tried to have a tree as consensual as possible, but there are still a few disputable choices, mostly about granularity and consistency. Sanskrit could have its own branch in the Indic family just as Latin in the Romance family, but because Sanskrit has no training data, that would not actually change the results. Likewise, as Czechoslovak and Dutch-Afrikaans have their own branches, Scandinavian languages could also distributed between east and west Scandinavian. As an English based Creole, Naija could as well be put in the Germanic family, but we kept it as a separate (Creole) family.

Regarding model training proper, we used $k = 500$ training sentences per iteration, $k^{'prime} = 500$ held-out sentences from the development set to compute running LAS and a maximum number of reboot $r = 5$. Following Dozat et al (2017), we use Eisner algorithm (Eisner, 1996) at test time to ensure outputs are well formed trees. The neural model is implemented in Dynet (Neubig et al., 2017) and we use Adadelta with default parameters as our trainer. We averaged the results over 5 random initializations. Independent models are trained in the same manner but with mono-lingual data only. We report both labeled and unlabeled edge prediction accuracy (UAS/LAS). In the appendix we also report results averaged per family.

3.2 Multi-Task Learning

Table 1 reports parsing results for languages that have a training set. Note that a few languages do not have a separate development set, then we used the training set for both training and validation. The training set size of those languages is reported in square brackets. This has low to no impact on other languages results but it can be problematic for the language itself as it can over-fit its training data especially when they are very few as is the case of Buryat for example. To be fair, we report two different averages. Avg is the average over languages that have a separate development set, and Avg No Dev is the average over languages that do not have a separate development set. For each language, the best UAS/LAS are reported in bold.

On average, phylogenetic training improves parsing accuracy, both labeled and unlabeled. This is especially true for languages that have very small training sets (50 sentences or less) and lack of development set. Those languages show an averaged 7 points improvement and up to 15 points (hsb, kmr). Since independent mono-lingual models follow the exact same training procedure but without phylogenetic initialization and that every sentence will be seen several times both at training and validation, the sampling method cannot explain such a difference. This shows that the ancestor’s model is a good initialization and acts as a form of regularization, slowing down over-fitting.

Phylogenetic training is also beneficial as one gains information from related languages. Indo-European languages gain from sharing information. This is especially true for Balto-Slavic (sk +5.82, lt +5.07 UAS) and Indo-Iranian languages (mr +2.05 UAS). It is less consistent for Romance and Germanic languages. This might be due to the tree not representing well typology for those families. Typically, English tends to group syntactically with Scandinavian languages more than with West-Germanic. Turkic and Uralic languages show the same benefits overall (ug +2.67, fi +3.39 UAS).

Dravidian and Afro-Asiatic languages are not as consistent. While Telugu seems to gain from Tamil data, the reverse is not true. Result variation for Arabic, Hebrew and Coptic are marginal. This is likely due to the fact that we only have three quite different languages from that family and that they all have their own script.

Similarly, phylogenetic training is not consistently useful for languages that do not have rela-
<table>
<thead>
<tr>
<th>Lang</th>
<th>Model</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>am</td>
<td>Semitic</td>
<td>57.27</td>
<td>26.25</td>
</tr>
<tr>
<td>br</td>
<td>Celtic</td>
<td>61.36</td>
<td>43.89</td>
</tr>
<tr>
<td>fo</td>
<td>North-Germanic</td>
<td>52.40</td>
<td>46.52</td>
</tr>
<tr>
<td>sa</td>
<td>Indic</td>
<td>56.18</td>
<td>40.46</td>
</tr>
<tr>
<td>pcm</td>
<td>World</td>
<td>60.43</td>
<td>43.80</td>
</tr>
<tr>
<td>th</td>
<td>World</td>
<td>29.14</td>
<td>17.61</td>
</tr>
<tr>
<td>tl</td>
<td>Austronesian*</td>
<td>70.89</td>
<td>50.38</td>
</tr>
<tr>
<td>wbp</td>
<td>World</td>
<td>87.67</td>
<td>65.66</td>
</tr>
<tr>
<td>yo</td>
<td>World</td>
<td>56.16</td>
<td>37.51</td>
</tr>
<tr>
<td>yue</td>
<td>Sino-Tibetan*</td>
<td>41.68</td>
<td>25.02</td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td>58.04</td>
<td>40.83</td>
</tr>
</tbody>
</table>

Overall results are a bit below the state of the art, but the model is very simple and relies on gold morphology, so it is not really comparable.

### 3.3 Zero-Shot Parsing

Table 2 reports parsing results for languages that do not have a training set. Because of phylogenetic training and the tree structure that guides it, it can happen that a language ancestor’s model is in fact trained on data only accounting for a narrow range of later stages. For example, while Faroese uses the North-Germanic model refined on both Norwegians, Swedish and Danish data, Tagalog uses the Austronesian model only refined with Indonesian data thus making it more an Indonesian model than an actual Austronesian model. Those cases are marked by an asterisk in the table. Komi (kpv) model is refined on Finno-Samic data, Breton (br) model on Irish data, Cantonese (yue) model on Mandarin data.

Looking at Table 2, we make the following observations. As expected scores are on average lower than for languages with training data, however the UAS/LAS gap is substantially bigger from 6.781 to 17.08 points. It is hard to compare to other works on zero-shot parsing since they use different data and scores span a big range, but our results are comparable to those of Aufrant et al. (2016) and Naseem et al. (2012), while our zero-shot models are given for free by the phylogenetic training method.
On a language per language basis, we see that there are a few important factors, the most striking being genre. Tagalog (tl) and more surprisingly Warlpiri (wbp) have relatively high parsing accuracy despite being either completely isolated or having only one relative (Indonesian). This is likely because their data are well annotated stereotypical sentences extracted from grammars, thus making them easy to parse.

Then we see that Naija (pcm) and Yoruba (yo) are about 25 points higher than Thai (th) despite them three having low morphology (in the treebanks). As they have different genres (spoken, bible, news and wiki), without a deeper look at the trees themselves, our best guess is that this is due to Thai having a different script. Naija and Yoruba both use the Latin alphabet, and as such they can rely to some extent on the character model to share information with other languages, to at least organise the character space. This analysis would also carry for Cantonese (yue). It is a morphologically simple language, and despite having a relative (Mandarin), its score is rather low. The genre alone (spoken) would not explain everything as Naija has also a spoken treebank and a higher score. The writing system might be at blame once again. Indeed, Chinese characters are very different from alphabetic characters and are much harder to use in character models because of sparsity. Comparing Mandarin and Cantonese test sets with Mandarin train set, the amount of out-of-vocabulary words is 32.47% of types (11.90% of tokens) for Mandarin and 54.88% of types (56.50% of tokens) for Cantonese. The results for out-of-vocabulary characters are even more striking with 3.73% of types (0.49% of tokens) for Mandarin and 12.97% of types (34.29% of tokens) for Cantonese. This shows that not only there are a lot of OOV in Cantonese test set, but that those words/characters are common ones as 12.97% of character types missing make up for more than a third of all character tokens missing, where on the contrary Mandarin OOV are seldom and account for less tokens percentage than types. This is one more argument supporting the importance of the character vector.

Other important factors are typology and morphology. Amharic (am) despite its unique script has a higher score than Cantonese that actually shares its scripts (to some extent as we have seen) with Mandarin. The key point for Amharic score, is that all its relatives (Hebrew, Arabic and Coptic) have their own scripts and are morphologically rich, thus the model learns to use morphological information. The analysis is similar for Komi which on top of sharing morphology with its relatives also share the writing system which provides it an extra gain. However, this might work in the opposite direction as well, as we can see with Faroese, Breton and Sanskrit. Faroese (fo) is morphologically rich and that should help, however its North-Germanic relatives are morphologically much simpler. Thus the model does not learn to rely on morphological attributes nor on word endings for the character model as much. The same is true for Sanskrit (sa), which is morphologically richer than its modern Indic relatives, with an extra layer of specific writing systems. Eventually, Breton model (br) is refined over Irish data only and while Irish is a typological outlier amongst Indo-European languages because of its Verb-Subject-Object word order, Breton has the standard Subject-Verb-Object, thus using Irish data might actually be detrimental.

These arguments show the respective importance of the writing system, the genre of the data, the morphological analysis and the typology in phylogenetic zero-shot dependency parsing. Those factors can either work together positively (Komi) or negatively (Cantonese) or cancel each other out (Amharic, Faroese).

4 Related Work

The goal of multi-task learning is to learn related tasks (either sharing their input and/or output space of participating of the same pipeline) jointly in order to improve their models over independently learned one (Caruana, 1997). In Søgaard et al. (2016), task hierarchy is directly encoded in the neural model allowing tasks with different output space to share parts of their parameters (POS tagging comes at a lower level than CCG parsing and only back-propagates to lower layers). Likewise, in Johnson et al. (2017), the encoder/decoder architecture allows to learn encoders that target several output languages and decoders than handle data from various input languages. However, in multi-task learning literature, task relationships are often fixed. In Cavallanti et al. (2010) tasks with the same output spaces share parameter updates through a fixed similarity graph. In this work, changing level in the tree can be seen as splitting the similarity graph into disjoint sub graphs. It is
a way to have tasks relationships evolving during
training and to encode information about task evo-
lution that lacks in other multi-task methods.

In multi-lingual parsing, Ammar et al. (2016) propose to train a single model to parse many lan-
guages using both typological information, cross-
lingual word representations and language specific
information. While their model gives good results,
they only apply it to 7 Germanic and Romance lan-
guages. It would be worth doing the experiment
with 50+ languages and see how the results would
change. However, because of language specific
information their model would probably become
very big. In this work, language specific informa-
tion is not added on the top of the model, but is
just language generic information that refines over
time.

Che et al. (2017; 2018) and Stymne et al. (2018)
propose to train parsers on several concatenated
treebanks either from the same language or from
related languages and to fine-tune the parsers on
individual treebanks afterward to fit specific lan-
guages/domains. The main difference with our
method, is that instead of one step of fine-tuning,
we perform as many fine-tuning as there are ances-
tors in the tree, each time targeting more and more
specific data. This in turn requires that we han-
dle data imbalance therefore using sampling rather
than plain concatenation.

Aufrant et al. (2016) propose to tackle zero-
shot parsing by rewriting source treebanks to bet-
ter fit target language typology. Assuming that ty-
ology is homogeneous in a language family, the
phylogeny should drive models to be typologically
aware. However, as we have seen for Breton and
Irish, that assumption might not always hold.

Eventually, the closest work from our in spirit is
the one of Berg-Kirkpatrick et al. (2010). They
use a phylogenetic tree to guide the training of un-
supervised dependency parsing models of several
languages, using ancestor models to tie descendent
ones. The main difference here beside supervision,
is that we do not use ancestor models as biases but
rather as initialization of descendent models.

5 Conclusion

We have presented a multi-task learning frame-
work that allows one to train models for several
tasks that have diverged over time. Leveraging
their common evolutionary history through a phy-
logenetic tree, models share parameters and train-

ing samples until they need to diverge. As a by
product of this phylogenetic training, we are pro-
vided with intermediary models that can be used
to zero-shot a new related task, given its position
in the evolutionary history.

We have applied this framework to dependency
parsing using a graph-based neural parser and the
phylogenetic tree of the languages from UD 2.2 to
guide the training process. Our results show that
phylogenetic training is beneficial for well popu-
lated families such as Indo-European and Uralic. It
also helps generalization and prevents over-fitting
when very few data are available. For zero-shot
parsing, genre, writing system and morphology are
crucial factors for the quality of parse predictions.

Some works have been done on automatically
learning task relationship in multi-task setting. It
would be interesting to see how the algorithm
could figure out when and how to cluster languages
automatically as phylogenetic trees do not directly
depict grammar evolution.

Our model does not know that Latin came before
Old French and before modern French, or that de-
spite being Germanic, English underwent a heavy
Romance influence. It would be worth investigat-
ing softening the tree constraints and instigating
more evolutionary information in the structure.

Another important point is that we use gold part-
of-speech and morphological information which is
unlikely to be available in real scenarios. However,
our new training procedure can be applied to any
task, so a future work would be to use it to perform
phylogenetic POS tagging.

Other directions for the future are designing bet-
ter sampling methods as well as better ways to mea-
sure training convergence at each level.

Acknowledgement

This work was supported by ANR Grant GRASP
No. ANR-16-CE33-0011-01 and Grant from
CPER Nord-Pas de Calais/FEDER DATA Ad-
We also thank the reviewers for their valuable feed-
back.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros,
Chris Dyer, and Noah Smith. 2016. Many languages,
one parser. Transactions of the Association for Com-


<table>
<thead>
<tr>
<th>Language family</th>
<th>Phylogenetic</th>
<th>Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UAS</td>
<td>LAS</td>
</tr>
<tr>
<td>Afro-Asiatic</td>
<td>80.73</td>
<td>74.99</td>
</tr>
<tr>
<td>Indo-European</td>
<td>80.41</td>
<td>73.73</td>
</tr>
<tr>
<td>Germanic</td>
<td>80.64</td>
<td>74.58</td>
</tr>
<tr>
<td>Slavic</td>
<td>80.83</td>
<td>74.37</td>
</tr>
<tr>
<td>Romance</td>
<td>82.03</td>
<td>75.40</td>
</tr>
<tr>
<td>Indo-Iranian</td>
<td>80.03</td>
<td>72.29</td>
</tr>
<tr>
<td>Greek</td>
<td>80.13</td>
<td>75.59</td>
</tr>
<tr>
<td>Turkic</td>
<td>65.60</td>
<td>51.43</td>
</tr>
<tr>
<td>Uralic</td>
<td>78.25</td>
<td>72.40</td>
</tr>
<tr>
<td>Dravidian</td>
<td>81.97</td>
<td>70.59</td>
</tr>
<tr>
<td>Avg</td>
<td>80.05</td>
<td>73.35</td>
</tr>
<tr>
<td>Avg No Dev</td>
<td>70.97</td>
<td>60.69</td>
</tr>
</tbody>
</table>

Table 3: Parsing results for phylogenetic and independent neural models averaged by language family. Families are sorted in the same order as they appear in Table 1. Indo-European averages include Armenian (hy) and Irish (ga). Global averages are repeated for completeness. Best results are reported in bold.
Figures 4 and 5 represent the phylogenetic tree used for guiding the training process. As we only use data from the UD project 2.2, we collapse unique child so that Vietnamese is not an Austro-Asiatic language, it is just Vietnamese. We also only use well attested families, thus Buryat, a Mongolic language, is alone and not linked to Turkic languages. Maybe, the most disputable choice is to put Naija in its own Creole family instead of the Germanic family.