# Typological Features for Multilingual Delexicalised Dependency Parsing

**Manon Scholivet** **Franck Dary** **Alexis Nasr** **Benoit Favre** **Carlos Ramisch**
Aix-Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
firstname.lastname@lis-lab.fr

## Abstract

The existence of universal models to describe the syntax of languages has been debated for decades. The availability of resources such as the Universal Dependencies treebanks and the World Atlas of Language Structures make it possible to study the plausibility of universal grammar from the perspective of dependency parsing. Our work investigates the use of high-level language descriptions in the form of typological features for multilingual dependency parsing. Our experiments on multilingual parsing for 40 languages show that typological information can indeed guide parsers to share information between similar languages beyond simple language identification.

## 1 Introduction

Human languages may share some syntactic features, but differ on others. For example, some languages tend to place the subject before the verb (e.g., English) whereas others favour the reverse order (e.g., Arabic), and some do not exhibit a clear preference (e.g., Polish). These features can be viewed as the *parameters* of a language's syntax (Greenberg, 1963; Chomsky, 1995).

When training a multilingual parser, it could be interesting to explicitly represent these parameters, and to integrate them into the parsing model. If a successful strategy to do so was found, then, a parser could be trained simultaneously on several languages whose syntactic parameters have been explicitly represented. Such parser could then use a single model to parse texts in any language with known syntactic parameters.

In theory, if we had at our disposal a set of parameters that completely describes the syntax of languages as well as treebanks that explore the whole space of parameters and their values, then such a universal parser could be designed. To make such a program realistic, though, several issues have to be addressed. In this paper, we propose to study the feasibility of learning such multilingual parser by addressing some of these issues.

The first one is the choice of syntactic parameters that will be used (Naseem et al., 2012; Täckström et al., 2013; Zhang and Barzilay, 2015). In our work, we approximate these parameters by extracting syntactic information from the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). [1] A language is represented by a vector containing the values it selects in the WALS. This vector plays the role of the parameters mentioned above.

The second issue is the design of a unified scheme for representing syntax. Our natural choice is the Universal Dependencies (UD) initiative. [2] UD specifically proposes a set of universal dependency relations, part-of-speech tags and morphological features (Nivre et al., 2016). The UD treebanks are available for many languages, annotated according to common guidelines.

The third issue is the lexicon. UD proposes a common language for describing languages' morpho-syntax, but we do not dispose of a "universal lexicon" to which we can map the lexical units of different languages. The solution adopted in this work is to resort to delexicalised parsing (Zeman and Resnik, 2008). This technique consists in ignoring the lexicon when training a parser. Such impoverishment of the data leads to less accurate parsers, but offers a simple solution to the lexicon issue. Using an alternative solution for representing words in different languages, such as multilingual word embeddings, would have introduced in our experimental setting some biases that are difficult to assess and would have prevented

[1] https://wals.info/
[2] http://universaldependencies.org



*Proceedings of NAACL-HLT 2019*, pages 3919–3930
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

to measure the precise influence of the typological features on the behaviour of the parser.

The fourth issue concerns the parser, which must be language independent and produce syntactic trees based on combinations of parameter values and sentential configurations. We use a transition-based parser with a multi-layer perceptron classifier (Chen and Manning, 2014), responsible for proposing how parameter values match observable patterns in the data.

Our research hypotheses are: (a) features derived from the WALS enable cross-lingual sharing in multilingual parsing, and (b) these features do more than acting as mere language identifiers. Our main contributions are to reassess the utility of the WALS as informant of typological features of parsed languages, to evaluate their benefit in a controlled multilingual setting with full supervision, and to perform a set of analyses to better understand how they interact with the parser model. In addition to multilingual parsing, our method is suitable for zero-shot learning for under-resourced languages (Ammar et al., 2016; Guo et al., 2015).

After discussing related work (Sec. 2), we describe UD (Sec. 3), the WALS (Sec. 4) and our parser (Sec. 5). The experimental setup (Sec. 6) precedes our results (Sec. 7), analyses (Sec. 8) and conclusions (Sec. 9).

## 2 Related Work

Our work is at the intersection of three trends in the multilingual dependency parsing literature. The first is *transfer parsing*, when a parser is trained on a language (or a collection of languages) and tested on another one. The second is *delexicalised parsing*, which aims at abstracting away from the lexicon in order to neutralise genre, domain and topic biases which are heavily marked in the treebanks' vocabulary. The third trend is the use of a handcrafted *typological resources*, such as the WALS, in multilingual NLP methods.

Transfer parsing is often a suitable solution when dealing with low-resource languages (McDonald et al., 2011). Projected transfer relies on parallel corpora in which one of the languages does not have labelled training data to learn a parser, but the other does. One commonly employed solution is to use word alignments to project parsed sentences from one side onto the low-resource side of the parallel text, using heuristics (Hwa et al., 2005) or partial annotations (Lacroix et al., 2016). Agić et al. (2016) parse the resource-rich languages in a multi-parallel corpus, proposing a projection method to obtain POS tags and dependency trees for low-resource languages from multiple-language word alignments. The parsing model for the target language can also be obtained in an unsupervised fashion, by optimising a function that combines the likelihood of parallel data and the likelihood of the transferred model on non-annotated data in the low-resource language (Ma and Xia, 2014).

Instead of assuming the availability of parallel corpora, direct transfer approaches capitalize on language similarities. For instance, Lynn et al. (2014) build parser for Irish by first training a delexicalised parser on another language, and then applying it on Irish. They surprisingly found out that Indonesian was the language providing the best parsing results for Irish, even if they do not belong to the same language family, because long-distance dependencies are better represented in Indonesian than in the other languages tested.

Low-resource languages may have some (insufficient) amount of training material available. One can employ bilingual parsing, concatenating training corpora in two languages, to verify if there is an improvement in the results compared to a monolingual parser (Vilares et al., 2015). Direct transfer and bilingual parsing methods are close to the present article, since we also concatenate training corpora. However, in our case, we combine treebanks from many more sources (around 40 languages) and include typological features.

The combination of corpora in multiple languages for parser training is facilitated by the recent advent of multilingual standards and resources, in particular in Universal Dependencies for dependency syntax (Nivre et al., 2016). This initiative enables the annotation of POS, morphology and syntactic dependencies for all languages with the same guidelines and label sets. The availability of such corpora favours the development of cross-lingual methods (Tiedemann, 2015).

Multilingual parsing research is also encouraged by initiatives such as the CoNLL 2017 and 2018 shared tasks, on highly multilingual dependency parsing from raw text (Zeman et al., 2017, 2018).

Delexicalised parsers ignore the word forms and lemmas when analysing a sentence, usually relying on more abstract features such as word classes

and POS tags. The use of delexicalised parsers is especially relevant when learning multilingual parsers, since languages generally share only a limited amount of lexical units. The approach proposed by Zeman and Resnik (2008) consists in adapting a parser for a new related language using either parallel corpora or delexicalised parsing. This method can be used to quickly construct a parser if the source and target languages are sufficiently related. McDonald et al. (2011) show that delexicalised parsers can be directly transferred between languages, yielding significantly higher accuracy than unsupervised parsers.

Moreover, typological features such as those present in the WALS provide information about the structure of languages (Dryer and Haspelmath, 2013). These could be useful to guide multilingual parsers, informing them about the model parameters that can be shared among languages with similar characteristics. Naseem et al. (2012) and Zhang and Barzilay (2015) use word-order features available for all their languages, while Ponti et al. (2018) used features they judged relevant in many categories (not only word order). The parameters proposed in the WALS are not the only way to represent properties of languages. Methods based on language embeddings (Östling and Tiedemann, 2017; Bjerva et al., 2019) also constitute interesting language representation.

Täckström et al. (2013) use a multilingual delexicalised transfer method, showing how selective parameter sharing, based on typological features and language family membership, can be incorporated in a discriminative graph-based dependency parser. They select the typological features based on those used by Naseem et al. (2012), removing two features not considered useful.

The work closest to ours experimented with concatenating treebanks to train a multilingual parser (Ammar et al., 2016). The authors use an S-LSTM transition-based parser similar to ours (although we do not include recurrent representations) trained on a set of lexicalised features that include multilingual word embeddings, Brown clusters, and fine-grained POS tags, whereas we only use coarse-grained POS and morphological features in a delexicalised setting. They include a one-hot language-ID vector, a set of six word-order features from the WALS (Naseem et al., 2012), or the whole WALS vectors. We use the two former plus a set of 22 selected features from WALS. They perform experiments on seven high-resourced languages while we report results on a larger set of 40 languages. Although Ammar et al. (2016) showed that, in a lexicalised setting, treebank concatenation could perform on par with monolingual parsers, the origins and limits of these improvements are not clear. We explore directions for assessing the benefits of typological features in a delexicalised parser.

## 3 Universal Dependencies

A major issue in multilingual parsing is the consistency of annotation across languages, since most corpora are annotated using different guidelines and tagsets. Universal Dependencies (UD) is an initiative whose goal is to create cross-linguistically consistent treebanks, facilitating cross-lingual analyses for language and parsing studies. Currently at version 2.3, 129 treebanks in 76 languages are available.

We use the UD v2.0 release for training and development, and the CoNLL 2017 shared task test sets for evaluation. For training and development, 64 UD treebanks in 45 languages are available. These treebanks vary in size: some are very small (e.g., 529 words for Kazakh), whereas others can be rather large (e.g., 1,842,867 words for Czech). Test corpora contain at least 10,000 words per language and are available for 49 languages.[3]

We learn delexicalised parsers from the UD treebanks using universal parts of speech (UPOS) and morphological features (FEAT) as input, and predicting labelled dependency trees which include language-specific extensions (e.g., *acl:relcl*). Morphological features are present in almost all treebanks, but exhibit high variability. Therefore, we choose to keep only the 16 most frequent features (e.g., Number, Case, VerbForm), which appear in at least 28 languages. Furthermore, morphology is represented as a list of *key*=*value* pairs, which we split so that each pair is considered separately, yielding a fixed set of 16 morphological features per word.

## 4 World Atlas of Language Structures

The World Atlas of Language Structures (WALS) is a database of structural (phonological, grammatical and lexical) properties of languages gathered by 55 authors from descriptive materials such

[3] 4 languages do not have training sets.

as reference grammars. We have used this resource to associate to every language of UD corpora a set of features describing its properties that are relevant for syntactic parsing.

The WALS describes 2,676 languages with a set of 192 features, organized into 11 feature genus (e.g. Phonology, Word Order). It can therefore be represented as a matrix $W$ of 2,676 rows and 192 columns, in which cell $W(l, f)$ gives the value of feature $f$ for language $l$, and each row $W(l)$ is the feature vector of a language $l$. This matrix has been pruned and completed to match our experimental setup. First, we have kept only the rows corresponding to the 49 languages of our test corpora. Conversely, four UD languages do not appear in the WALS and have been left aside: Old Church Slavonic (cu), Gothic (got), Ancient Greek (grc), and Latin (la). As a result, we obtain a reduced version of $W$ containing 45 rows.

We experimented with two language representations obtained from the WALS. The first one, henceforth $W_N$, is based on the work of Naseem et al. (2012). They selected the six Word Order features available for all their 17 target languages, identified by the codes 81A, 85A, 86A, 87A, 88A, 89A[4]. These features cover phenomena such as verb-object and adjective-noun order, and have been widely discussed in the literature (Täckström et al., 2013; Zhang and Barzilay, 2015; Ammar et al., 2016). The resulting matrix has 45 rows (languages) and 6 columns (features). However, the WALS seen as a matrix is sparse, as some features are unspecified for some languages. Therefore, we chose to keep only languages for which at most half of this vector is unspecified, resulting in the removal of 5 more languages: Galician (gl), Upper Sorbian (hsb), Kazakh (kk), Slovak (sk), and Uyghur (ug). All our experiments are carried out on this set of 40 languages.

The second language representation proposed in this work, henceforth $W_{80}$, is a relaxed version of $W_N$. Since the WALS is sparse, we include in $W_{80}$ all features specified for at least 80% of our 40 languages. Furthermore, in addition to features from the Word Order family, we also include features from the Simple Clauses family. This results in a matrix of 40 rows and 22 columns corresponding to 3 features from the Simple Clauses family (101A, 112A, 116A), and 19 from the Word Order family (81A, 82A, 83A, 85A, 86A, 87A, 88A, 89A, 90A, 92A, 94A, 95A, 96A, 97A, 144A, 143A, 143E, 143F, 143G).[5]

[4]The description of the features of the WALS relevant for this paper can be found in appendix A.

| | Romance | Germanic | Slavic | Random |
|---|---|---|---|---|
| $W_N$ | 0.33 | 1.33 | 0.67 | 2.41 |
| $W_{80}$ | 4.13 | 4.47 | 4.19 | 10.15 |

Table 1: MID values of typological language genus compared to Random. Random MID is the average of the MID for 50,000 sets of 6 languages randomly selected.

The final matrices $W_N$ and $W_{80}$ obtained after feature selection are not complete: they contain respectively 4 and 35 unspecified values, which were filled automatically. Each matrix $W$ (shorthand for $W_N$ and $W_{80}$) offers a straightforward way to compare languages $l_1$ and $l_2$ using the Hamming distance[6] between their vectors $W(l_1)$ and $W(l_2)$, noted $d(l_1, l_2)$. To fill in the missing values, we have selected, for every language $l_1$ containing unspecified feature values ('?'), the corresponding value from its closest fully specified language $l_2$, that is, $l_2 = \underset{l_i \mid ? \notin W(l_i)}{\arg\min} d(l_1, l_i)$.

The $W_N$ and $W_{80}$ matrices only provide partial descriptions of languages, heavily biased towards parsing and ignoring other aspects (e.g., phonology). Nevertheless, it is tempting to compare how they relate languages that belong to the same typological genus. In order to do so, we have concentrated on three genus present in our set of 40 languages: Romance (6 languages), Germanic (6 languages) and Slavic (7 languages), and computed how close the vectors of these languages are. We define the *mean internal distance* (MID) of a language set $L = \{l_1, \ldots, l_n\}$, as the average of the distances of every pair of languages in $L$:

$$MID(L) = \frac{1}{n^2 - n} \sum_{\substack{(l_i, l_j) \in L \times L \\ i \neq j}} d(l_i, l_j)$$

We have computed the MID of each language genus, and compared it with the MID of randomly chosen sets of 6 languages (number of languages in the Romance and Germanic genus). The results in Table 1 clearly indicate that WALS vectors capture language genus similarities.

Others methods could have been used to assess whether the language descriptions that we have

[5]We have also considered the Complex Sentences family, but no feature exceeded the 80% threshold.

[6]The number of dimensions for which their values differ.

extracted from the WALS can measure language proximity. It could be interesting, for example, to reproduce the results of (Rabinovich et al., 2017) on reconstructing phylogenetic trees of language from the WALS features.

## 5 Parser

The parser used in our experiments is an arc-eager transition-based parser (Nivre, 2008), trained with a dynamic oracle (Goldberg and Nivre, 2012). The prediction of transitions is performed with a multilayer perceptron (MLP), as in Chen and Manning (2014). The MLP consists of an input layer, one hidden layer, and an output layer. Two sets of delexicalised features have been defined for the prediction: BASIC and EXTENDED. BASIC is a standard set composed of 9 POS features, 7 syntactic features, 32 morphological features, and a distance feature (the distance between the head and the dependent).[7] EXTENDED adds to BASIC new features that correspond to the WALS vectors $W_N$ (6 features) and $W_{80}$ (22 features), and/or the language ID of the sentence's language (1 feature). Each feature, including ID, is associated with a zero-initialized learnable embedding of size 3. The input layer of the MLP corresponds to the concatenation of the embeddings of the different features, with dimensions varying from 396 to 465, depending on the configuration (with or without language vectors $W_N$ and $W_{80}$, or a language identifier ID). The output layer has 263 neurons, corresponding to the number of transitions that the parser can predict. The hidden layer has 1,000 units, the dropout rate used during training is equal to 0.4, the number of epochs is equal to 10, the activation function is a ReLU, the loss function is negative softmax log likelihood, and the learning algorithm is AMSgrad, using default parameters from Dynet (Neubig et al., 2017).[8]

At every step of the parsing process, the parser predicts an action to perform, which may yield the creation of a new dependency between two words of the sentence. The prediction of the actions is based on the values of the features fed to the MLP. In BASIC mode, these features describe different aspects of the head and the dependent, as well as their neighbourhood. For example, if the head is a verb and the dependent is a noun located before the verb, a subject dependency has high probability in languages that prefer subject-verb ordering (SV). In EXTENDED mode, the information of whether the language is SV is made explicit. The MLP has therefore the possibility to combine a *sentential configuration* (e.g., a noun before a verb) with a *language configuration* (e.g., the language is SV) when predicting an action. All languages that share a common feature in $W$ will therefore be able to perform the same prediction for sentential configurations that are specific to this common feature (e.g., the noun preceding the verb and the language being SV).[9]

## 6 Experimental Settings

**Corpora** Our experiments were performed on the CoNLL 2017 shared task data (Zeman et al., 2017), on gold tokenisation and ignoring contractions (i.e., ranges). We evaluate our models individually on each of the 40 languages for which we have a $W(l)$ vector (section 4), using the original CoNLL 2017 shared task test sets. The test corpora for each language are simply the concatenation of all test treebanks for that language.

Training and development are performed on multilingual corpora (henceforth TRAIN-ML and DEV-ML) derived from the training and development treebanks of 37 UD languages.[10] The UD training and development corpora have different sizes for different languages, ranging from 529 words for Kazakh (kk) to 1,842,867 for Czech (cs). Thus, simply concatenating all corpora to constitute TRAIN-ML and DEV-ML would overrepresent certain languages and possibly bias the parser towards them. This is why we have decided to balance TRAIN-ML and DEV-ML across languages.

First, all available training and development corpora of the 37 languages have been concatenated. From this large corpus, we build two new intermediate corpora, PRE-TRAIN-ML and PRE-DEV-ML, with each sentence having 90% chances to belong to PRE-TRAIN-ML, and 10% chances

[7] Corpora, configuration files and WALS vectors available at: http://pageperso.lis-lab.fr/~carlos.ramisch/?page=downloads/wals-ud-parse

[8] Hyperparameters were tuned in preliminary experiments, in conditions similar to $\Sigma\, W_N$ (see Section 6).

[9] Our parser cannot predict non projective trees, systematically generating a wrong parse at test time. The average non projectivity rate of the test corpora is equal to 1%, with a standard deviation of 1% among the 40 languages. We ran some tests with pseudo projective tree transformation (Nivre and Nilsson, 2005), but it had a negligible impact on the results, so we have decided to keep the original projective algorithm.

[10] Three languages among our 40 target languages have no corresponding training nor development data (bxr, kmr, sme).

to belong to PRE-DEV-ML. Second, we build TRAIN-ML (respectively DEV-ML) by randomly selecting sentences from PRE-TRAIN-ML (resp. PRE-DEV-ML) until the number of tokens exceeds 20,000 (resp. 2,000 for DEV-ML) per language. At the end, we shuffle the selected sentences to obtain the final training and development corpora TRAIN-ML and DEV-ML. Using this procedure, the same sentence can appear several times in a corpus. Nonetheless, this method guarantees a balanced representation of every language in TRAIN-ML and DEV-ML.

**Metrics** The quality of the predicted trees is assessed with a standard measure for dependency parsing: labelled attachment score (LAS).[11, 12] We report LAS per language, as well as MACRO-LAS which is the macro-average of LAS on all languages that have a training set. This measure is therefore independent of the size of the test corpus of each language, and is not biased towards over-represented languages in the test sets.

**Training Configurations** Our experiments on several ⟨training corpus, language vector⟩ pairs are designated by the following codes:

***L*:** Monolingual corpus. The training corpus of a language $l$ consists of the sentences of $l$ in TRAIN-ML. Thirty-seven BASIC delexicalised parsers have been trained, one per language. This configuration corresponds to the standard one in parsing experiments: training and testing on the same language.

**Σ:** Multilingual corpus. A BASIC parser is trained on the whole TRAIN-ML corpus, with no indication of the inputs language. The parsing model is delexicalised, so the corpus contains only POS tags (gold), morphological features (gold) and syntactic relations (to be learned).[13]

**Σ ID:** Multilingual corpus + language ID. An EXTENDED parser is trained on the TRAIN-ML corpus using as extra feature the identifier of the language attached to each word.

[11] For brevity, we omit UAS figures in our experiments, as UAS and LAS are tightly correlated ($r = 0.98$)

[12] Using the CoNLL shared task 2017 evaluation script.

[13] The decision to use gold POS tags and morphological features may seem unrealistic. This article is the first step of a process in which we intend to predict the POS tags and the morphological features in the same fashion.

**Σ $W_N$, Σ $W_{80}$:** Multilingual corpus + WALS. Two EXTENDED parsers are trained on the TRAIN-ML corpus, with $W_N$ (resp. $W_{80}$) vectors derived from the WALS attached to each word.

## 7 Results

The detail of the LAS obtained for every language, as well as the macro-averaged LAS (MACRO) are displayed in Table 2. We comment below the results for *L*, and compare the results of meaningful pairs of experiments, summarised in Table 3.

| *L* | Σ | Σ ID | Σ $W_N$ | Σ $W_{80}$ | Lang. | |
|---|---|---|---|---|---|---|
| 65.89 | 60.59 | 62.97 | 63.15 | 64.38 | ar | |
| 78.59 | 74.32 | 76.43 | 76.26 | 77.47 | bg | S |
| 77.18 | 72.76 | 74.11 | 73.03 | 76.27 | ca | R |
| 68.92 | 68.01 | 68.91 | 68.72 | 69.61 | cs | S |
| 73.62 | 67.38 | 70.25 | 70.19 | 70.25 | da | G |
| 71.07 | 63.76 | 66.47 | 69.18 | 69.22 | de | G |
| 77.11 | 71.26 | 72.72 | 73.29 | 75.84 | el | |
| 70.05 | 66.02 | 69.3 | 69.91 | 70.19 | en | G |
| 71.47 | 71.98 | 72.38 | 72.29 | 73.22 | es | R |
| 66.98 | 63.76 | 66.89 | 65.75 | 67.79 | et | |
| 63.26 | 55.76 | 59.54 | 60.22 | 59.39 | eu | |
| 72.85 | 66.02 | 67.23 | 69.63 | 70 | fa | |
| 60.97 | 56.29 | 58.86 | 57.37 | 59.28 | fi | |
| 75.74 | 74.25 | 75.44 | 74.79 | 75.82 | fr | R |
| 66.55 | 60.41 | 63.12 | 64.68 | 65.96 | ga | |
| 70.21 | 63.03 | 65.69 | 66.09 | 67.45 | he | |
| 78.91 | 73.86 | 75.81 | 75.77 | 74.45 | hi | |
| 71.03 | 67.49 | 68.39 | 70 | 70.4 | hr | S |
| 67.08 | 62.55 | 66.89 | 67.19 | 67.51 | hu | |
| 68.64 | 58.38 | 63.99 | 62.61 | 64.57 | id | |
| 81.44 | 76.45 | 78.03 | 76.97 | 79.83 | it | R |
| 78.26 | 68.22 | 76.21 | 74.85 | 75.56 | ja | |
| 47.68 | 37.17 | 40.03 | 38.07 | 39.66 | ko | |
| 59.89 | 54.11 | 58.45 | 58.23 | 60.17 | lv | |
| 62.56 | 57.21 | 57.15 | 58.13 | 58.59 | nl | G |
| 74.59 | 73.19 | 75.13 | 73.51 | 75.93 | no | G |
| 81.24 | 74.24 | 77.29 | 74.78 | 79.02 | pl | S |
| 72 | 65.74 | 68.1 | 68.74 | 69.86 | pt | R |
| 70.99 | 67.72 | 69.37 | 70.38 | 70.61 | ro | R |
| 74.06 | 61.35 | 74.12 | 68.65 | 74.45 | ru | S |
| 67.1 | 64.12 | 65.41 | 64.75 | 66.36 | sl | S |
| 72.05 | 69.97 | 71.6 | 70.71 | 71.88 | sv | G |
| 46.78 | 41.01 | 45.41 | 43.26 | 41.62 | tr | |
| 71.6 | 69.4 | 71.96 | 69.77 | 72.81 | uk | S |
| 74.35 | 69.15 | 72.07 | 70.71 | 70.76 | ur | |
| 54.4 | 42.42 | 51.28 | 51.94 | 51.72 | vi | |
| 59.83 | 45.46 | 58.94 | 53.42 | 54.87 | zh | |
| 69.32 | 63.64 | 66.92 | 66.41 | 67.64 | MACRO | |
| - | 33.32 | 31.99 | 30.37 | 28.49 | bxr | |
| - | 40.34 | 38.14 | 41.41 | 44.04 | kmr | |
| - | 47.34 | 46.5 | 47.63 | 42.38 | sme | |

Table 2: LAS for each language and MACRO LAS, for the five configurations. Languages followed by a S belong to the Slavic genus, G belong to the German genus and R belong to the Romance genus.

***L*:** The results obtained in the *L* experiment show an important variation of performances for different languages. LAS ranges from 46.78 for

| $X$ | $Y$ | $\overline{X-Y}$ | $\sigma$ | min | max |
|---|---|---|---|---|---|
| $L$ | $\Sigma$ | 5.68 | 3.32 | -0.51 es | 14.37 zh |
| $\Sigma\, W_{80}$ | $\Sigma$ | 4.00 | 2.58 | 0.59 hi | 13.10 ru |
| $\Sigma\, W_{80}$ | $\Sigma\, W_N$ | 1.24 | 1.45 | -1.64 tr | 5.80 ru |
| $\Sigma$ ID | $\Sigma$ | 3.27 | 3.00 | -0.06 nl | 13.48 zh |
| $L$ | $\Sigma$ ID | 2.41 | 1.99 | -0.91 es | 7.65 ko |
| $\Sigma\, W_{80}$ | $\Sigma$ ID | 0.73 | 1.54 | -4.07 zh | 3.12 el |

Table 3: Differences between $X$ and $Y$ configurations: average ($\overline{X-Y}$), standard deviation ($\sigma$), minimum and maximum with corresponding languages.

Turkish to 81.44 for Italian. A detailed investigation for the reasons of such a variability is beyond the scope of this paper. Let us just mention a few hypotheses. Some are language specific, such as the balance between morphological and syntactic marking of linguistic constructions (i.e., morphologically rich languages are probably favoured in our setting, since the morphological analysis is given as input to the parser). Others are genre specific: the corpora for different languages pertain to different genres. Although delexicalisation neutralises some genre biases (some genres can feature a moderate lexical variability which can ease parsing) genres can also influence syntax, through sentence length (longer sentences are generally harder to parse), or the ratio of error-prone constructions, such as ambiguous prepositional phrases and coordination. Finally, annotation quality is heterogeneous across languages, potentially explaining the variability in LAS.

***L* vs $\Sigma$:** An expected drop in performances is observed when switching from *L* to $\Sigma$. The MACRO LAS loses 5.68 points. The main hypothesis to explain such a drop is the noise introduced when mixing different languages. This noise takes the form of contradictory information seen by the parser during training. For example, the sentential configuration associated to a subject dependency in SV and VS languages are very different, yet the parser is unaware of this distinction and will see contradictory examples. The variation of the LAS drop is different across languages. In the case of Spanish, switching from *L* to $\Sigma$ even increases LAS (+0.51 points). We do not have a conclusive explanation for this result. The intuitive explanation is that $\Sigma$ is a (noisy) language which on average is closer to Spanish than it is to Chinese (which performance drops by 14.37 points). This fact itself is the consequence that, on average, languages that compose $\Sigma$ are closer to Spanish than they are to Chinese.

**$\Sigma$ vs $\Sigma\, W_{80}$:** This is our first major result: when adding $W_{80}$ to the parser, the MACRO LAS increases by 4 points when compared to $\Sigma$. LAS increases for all languages. There are two interpretations of this result. The optimistic one is that $W_{80}$ helps decreasing the noise introduced by mixing languages in $\Sigma$ by "explaining" some apparently contradictory information in the data through the use of linguistic features encoded in the WALS. The pessimistic interpretation is that the WALS vectors are merely an arbitrary encoding of the languages. In this case, the parser's MLP would be associating sentential configurations to specific languages, thus learning different models for different languages. Figuring out what the model is actually learning is not an easy task. We propose in section 8 some clues to answer this question. Moreover, there is not a clear tendency to increase or decrease when using the WALS vector in the case of the 3 languages without training data. More experiments are required to study the performances when the language is not in the training corpus.

**$\Sigma\, W_N$ vs $\Sigma\, W_{80}$:** When added to $\Sigma$, vectors $W_N$ and $W_{80}$ do not have the same impact on the performances. Adding $W_{80}$ to $\Sigma$ yields an increase of 4 points while adding $W_N$ increases the performance by 2.77 points only. The parser is therefore able to take advantage from a richer description of languages when learning the model. This result indicates that the disappointing parsing results reported by Ammar et al. (2016), who adopted the $W_N$ vector, are probably due to the fact that the features extracted from the WALS were not rich enough to explain differences between languages that are important for a parser.

**$\Sigma$ vs $\Sigma$ ID:** Adding the ID vector to $\Sigma$ yields an improvement of 3.28 MACRO points. This increase was expected since, in this setting, sentential configurations are associated to a language ID, which helps decreasing the noise in the data.

***L* vs $\Sigma$ ID:** One could expect that $\Sigma$ ID would reach the result obtained by *L* since in both configurations the same amount of data is available and languages are unambiguously identified. This is not the case: the performance of $\Sigma$ ID is 2.41 points behind *L*. The difference in performances is due to the MLP architecture (in particular the size of the hidden layer), which is the same for $\Sigma$ ID and for each of the *L* models. Each language is de-

scribed with more parameters in an $L$ model than it is in the $\Sigma$ ID model.

**$\Sigma$ ID vs $\Sigma$ $W_{80}$:** This is our second major result: adding $W_{80}$ to $\Sigma$ yields better results than adding ID to $\Sigma$. This result indicates that it is more interesting, in our setting, to describe a language as a vector of typological features, allowing to identify features that are common to several languages, than describing a language by an arbitrary code. As mentioned above, such a conclusion is valid for models of a fixed size only, which is the case here. It could be the case that, when increasing the number of parameters of the models, $\Sigma$ ID gets better results than $\Sigma$ $W_{80}$.

We do not report here a series of experiences combining ID and $W_{80}$. We observed a slight improvement (MACRO=67.86) when adding ID in the input of the parser. This effect indicates that the information contained in ID and $W_{80}$ vectors are complementary and the parser has the opportunity to rely on both of them. Figuring out exactly how the parser uses this information is a complex issue that we address in the following section.

## 8 How does the parser use $W$?

As already conjectured, one hypothesis for explaining the behaviour of the parser in the presence of $W$ is that it uses the additional features to identify a language, not to better generalise on the syntactic phenomena that the features address. Table 4 shows the accuracy of a logistic regression classifier trained to predict the language ID based on either the input features of the parser's MLP, or on the activations after the hidden layer, with $\Sigma$, $\Sigma$ $W_N$ and $\Sigma$ $W_{80}$. The table shows that indeed, WALS features, especially $W_{80}$, greatly improve the capability of the language classifier, suggesting that the parser can use language identity in its predictions. The fact that this information is still available just before the decision layer means that it can be used for predicting parsing actions.

Another interesting analysis consists in comparing the distribution of activations for two languages. In the following, the activations are measured at the hidden layer before the ReLU nonlinearity, and are assumed to follow normal distributions at the neuron level. We compute the Jensen-Shannon Divergence (JSD) between the activations of a given neuron for a pair of languages. Table 5 shows the mean, maximum and minimum neuron-level JSD between cherry-picked language pairs. We selected three language pairs with increasing distance. Dutch and German (nl-de) belong to the same typological genus and have identical $W_{80}$ vectors. Portuguese and French (pt-fr) also belong to the same genus but their vectors differ in six features (e.g. 101A pronominal subject, 143E postverbal negation). On the other extreme, Russian Buriat and Irish (brx-ga) have very different $W_{80}$ vectors, with only two shared values out of 22.

For nl-de, the average difference between the activation distributions in $\Sigma$ $W_{80}$ (0.854) is lower than in $\Sigma$ (0.86), suggesting that $W_{80}$ helps leveraging the similarity between those languages, which is also confirmed by an increase in LAS (Table 2). For pt-fr, however, the addition of $W_{80}$ results in an increase in the average distance between the activation distributions (0.912) when compared to $\Sigma$ (0.878). Analogously, this difference also increases by a larger margin (from 0.89 to 1.16) for the most distant pair bxr-ga. Overall, these observations indicate that $W_{80}$ reinforces parameter sharing between similar languages and increases contrast between dissimilar ones. As an example, Figure 1 shows that the distributions for the neuron with highest JSD are very similar for nl-de while they are different for bxr-ga.

| Configuration | Features | Accuracy |
|---|---|---|
| $\Sigma$ | input | 0.432 |
| $\Sigma$ $W_N$ | input | 0.678 |
| $\Sigma$ $W_{80}$ | input | 0.954 |
| $\Sigma$ | hidden | 0.436 |
| $\Sigma$ $W_N$ | hidden | 0.682 |
| $\Sigma$ $W_{80}$ | hidden | 0.956 |

Table 4: Language identification accuracy for a logistic regression classifier trained on the activations after the hidden layer, or at the input. The classifier is trained on the development set, results are reported on the test set.

| L1 | L2 | Model | Mean | Max | Min |
|---|---|---|---|---|---|
| nl | de | $\Sigma$ | 0.860 | 1.027 | 0.798 |
| | | $\Sigma$ $W_{80}$ | 0.854 | 0.940 | 0.793 |
| pt | fr | $\Sigma$ | 0.878 | 1.335 | 0.794 |
| | | $\Sigma$ $W_{80}$ | 0.912 | 3.550 | 0.700 |
| bxr | ga | $\Sigma$ | 0.890 | 1.600 | 0.782 |
| | | $\Sigma$ $W_{80}$ | 1.160 | 4.888 | 0.757 |

Table 5: Neuron-level JSD statistics between activations at the hidden layer of the parser models for selected pairs of languages.

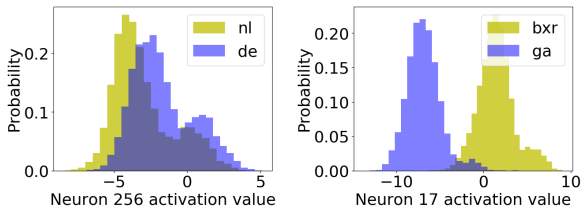


Figure 1: Activation distributions for the neuron with highest JSD on $\Sigma$ for (nl, de) and (bxr, ga) pairs.

## 9 Conclusions and Future Work

This paper has studied how high-level typological language descriptions coming from the WALS can guide a multilingual parser to learn cross-language generalisations. Two interpretations of what the parser is doing in the light of such information have been opposed. In the first (optimistic) one, the parser uses the high-level descriptions to cluster coherent observable patterns across languages. In the second (pessimistic) one, the parser uses the high-level descriptions given as input to figure out the identity of the language and uses this ID to trigger parts of the model that are language specific. Our results and parsing model analyses hint that, although it is difficult to draw definitive conclusions, the model indeed uses information in the WALS vectors as language identifiers, but some extra gain is observed, favouring the cross-lingual sharing hypothesis.

As future work, we plan to study the influence of typological features on each dependency type. Whereas a delexicalised parser offers a simple experimental setup, it impacts parsing performance. Thus, we would like to use multilingual word embeddings to make lexical information accessible to the parser, making it more realistic. The results in section 8 suggest that the parser struggles between two behaviours. One way to intervene would be to penalise the parser when it correctly identifies the language, using adversarial learning (Ganin et al., 2016). Our experiments on the three languages with no training corpus are not conclusive on the usefulness of the WALS vector in zero-shot setting, and we plan to make more tests in this setting.

## References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Sgaard. 2016. Multilingual Projection for Parsing Truly Low-Resource Languages. *Transactions of the Association for Computational Linguistics*, 4(0):301–312.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many Languages, One Parser. *arXiv:1602.01595 [cs]*. ArXiv: 1602.01595.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *arXiv preprint arXiv:1901.02646*.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750. Association for Computational Linguistics.

Noam Chomsky. 1995. *The Minimalist Program*. Current studies in linguistics series. MIT Press, Cambridge, MA, USA.

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

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Yoav Goldberg and Joakim Nivre. 2012. A dynamic oracle for arc-eager dependency parsing. *Proceedings of COLING 2012*, pages 959–976.

Joseph H. Greenberg. 1963. *Universals of Human Language*. MIT Press, Cambridge, MA, USA.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual Dependency Parsing Based on Distributed Representations. In *ACL (1)*, pages 1234–1244.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Ophélie Lacroix, Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Frustratingly easy cross-lingual transfer for transition-based dependency parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1058–1063, San Diego, California. Association for Computational Linguistics.

Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. 2014. Cross-lingual Transfer Parsing for Low-Resourced Languages: An Irish Case Study.

In *Proceedings of the First Celtic Language Technology Workshop*, pages 41–49, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Xuezhe Ma and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1337–1348, Baltimore, Maryland. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea. Association for Computational Linguistics.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 99–106. Association for Computational Linguistics.

Robert Östling and Jrg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic Transfer of Syntactic Structures in Cross-Lingual NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1531–1542, Melbourne, Australia. Association for Computational Linguistics.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 530–540.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071. Association for Computational Linguistics.

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

David Vilares, Carlos Gmez-Rodrguez, and Miguel A. Alonso. 2015. One model, two languages: training bilingual parsers with harmonized treebanks. *arXiv:1507.08449 [cs]*. ArXiv: 1507.08449.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, and Martin Potthast. 2017. CoNLL 2017 shared task: multilingual parsing from raw text to universal dependencies. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Vclava Kettnerov, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missil, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Hctor Martnez Alonso, ar ltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadov, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2018. CoNLL 2018 Shared Task: Multilingual

Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*. Association for Computational Linguistics.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1857–1867, Lisbon, Portugal. Association for Computational Linguistics.

## A Appendix: description of the WALS features used in our work

1. **81A**: Order of Subject, Object and Verb (*SOV; SVO; VSO; VOS; OVS; OSV; No dominant order*)

2. **82A**: Order of Subject and Verb (*SV; VS; No dominant order*)

3. **83A**: Order of Object and Verb (*OV; VO; No dominant order*)

4. **85A**: Order of Adposition and Noun Phrase (*Postpositions; Prepositions; Inpositions; No dominant order; No adpositions*)

5. **86A**: Order of Genitive and Noun (*Genitive-Noun; Noun-Genitive; No dominant order*)

6. **87A**: Order of Adjective and Noun (*Adjective-Noun; Noun-Adjective; No dominant order; Only internally-headed relative clauses*)

7. **88A**: Order of Demonstrative and Noun (*Demonstrative-Noun; Noun-Demonstrative; Demonstrative prefix; Demonstrative suffix; Demonstrative before and after Noun; Mixed*)

8. **89A**: Order of Numeral and Noun (*NumN; NNum; Both orders of numeral and noun with neither order dominant; Numeral only modifies verb*)

9. **90A**: Order of Relative Clause and Noun (*NRel; RelN; Internally-headed relative clause; Correlative relative clause; Adjoined relative clause; Double-headed relative clause; Mixed types of relative clause with none dominant*)

10. **92A**: Position of Polar Question Particles (*Initial; Final; Second position; Other position; In either of two positions; No Question particle*)

11. **95A**: Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase (*OV & Postpositions; OV & Prepositions; VO & Postpositions; VO & Prepositions; Other*)

12. **96A**: Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun (*OV & RelN; OV & NRel; VO & RelN; VO & NRel; Other*)

13. **97A**: Relationship between the Order of Object and Verb and the Order of Adjective and Noun (*OV & AdjN; OV & NAdj; VO & AdjN; VO & NAdj; Other*)

14. **101A**: Expression of Pronominal Subjects (*Pronominal subjects are expressed by pronouns in subject position that are normally if not obligatorily present; Pronominal subjects are expressed by affixes on verbs; Pronominal subjects are expressed by clitics with variable host; Pronominal subjects are expressed by subject pronouns that occur in a different syntactic position from full noun phrase subjects; Pronominal subjects are expressed only by pronouns in subject position, but these pronouns are often left out; More than one of the above types with none dominant*)

15. **112A**: Negative Morphemes (*Negative affix; Negative particle; Negative auxiliary verb; Negative word, unclear if verb or particle; Variation between negative word and affix; Double negation*)

16. **116A**: Polar Question (*Question particle; Interrogative verb morphology; Question particle and interrogative verb morphology; Interrogative word order; Absence of declarative morphemes; Interrogative intonation only; No interrogative-declarative distinction*)

17. **143A**: Order of Negative Morpheme and Verb (*NegV; VNeg; [Neg-V]; [V-Neg]; Negative Tone; Type 1 / Type 2; Type 1 / Type 3; Type 1 / Type 4; Type 2 / Type 3; Type 2 / Type 4; Type 3 / Type 4; Type 3 / Negative Infix; Optional Single Negation; Obligatory Double Negation; Optional Double Negation; Optional Triple Negation with Obligatory Double Negation; Optional Triple Negation with Optional Double Negation*)

18. **143E**: Preverbal Negative Morphemes (*Preverbal negative word; Negative prefix; Both preverbal negative word and negative prefix; No preverbal negative morpheme*)

19. **143F**: Postverbal Negative Morphemes (*Postverbal negative word; Negative suffix; Both postverbal negative word and negative suffix; No postverbal negative morpheme*)

20. **143G**: Minor morphological means of signaling negation (*Negative tone; Negative infix; Negative stem change; No negative tone, infix or stem change*)

21. **144A**: Position of Negative Word With Respect to Subject, Object, and Verb (*NegSVO; SNegVO; SVNegO; SVONeg; NegSOV; SNegOV; SONegV; SOVNeg; NegVSO; VSNegO; VSONeg; NegVOS; ONegVS; ONegVS; OSVNeg; More than one position for negative morpheme, with none dominant; Optional single negation; Obligatory double negation; Optional double negation; Morphological negation only (but not double negation); Other language*)