Abstract

Creating a descriptive grammar of a language is an indispensable step for language documentation and preservation. However, at the same time it is a tedious, time-consuming task. In this paper, we take steps towards automating this process by devising an automated framework for extracting a first-pass grammatical specification from raw text in a concise, human- and machine-readable format. We focus on extracting rules describing agreement, a morphosyntactic phenomenon at the core of the grammars of many of the world’s languages. We apply our framework to all languages included in the Universal Dependencies project, with promising results. Using cross-lingual transfer, even with no expert annotations in the language of interest, our framework extracts a grammatical specification which is nearly equivalent to those created with large amounts of gold-standard annotated data. We confirm this finding with human expert evaluations of the rules that our framework produces, which have an average accuracy of 78%. We release an interface demonstrating the extracted rules at https://neulab.github.io/lase/. The code is publicly available here.^[https://github.com/Aditi138/LASE-Agreement]

† Work done at Carnegie Mellon University.

1 Introduction

While the languages of the world are amazingly diverse, one thing they share in common is their adherence to grammars — sets of morpho-syntactic rules specifying how to create sentences in the language. Hence, an important step in the understanding and documentation of languages is the creation of a grammar sketch, a concise and human-readable description of the unique characteristics of that particular language (e.g. Huddleston (2002) for English, or Brown and Ogilvie (2010) for the world’s languages).

One aspect of morphosyntax that is widely described in such grammatical specifications is agreement, the process wherein a word or morpheme selects morphemes in correspondence with another word or phrase in the sentence (Corbett, 2009). Languages have varying degrees of agreement ranging from none (e.g. Japanese, Malay) to a large amount (e.g. Hindi, Russian, Chichewa). Patterns of agreement also vary across syntactic subcategories. For instance, regular verbs in English agree with their subject in number and person but modal verbs such as “will” show no agreement.

Having a concise description of these rules is of obvious use not only to linguists but also language teachers and learners. Furthermore, having such descriptions in machine-readable format will further enable applications in natural language processing (NLP) such as identifying and mitigating gender stereotypes in morphologically rich languages (Zmigrod et al., 2019).

The notion of describing a language “in its own terms” based solely on raw data has an established tradition in descriptive linguistics (e.g. Harris (1951)). In this work we present a framework (outlined in Figure 1) that automatically creates a first-pass specification of morphological agreement rules for various morphological features (Gender, Number, Person, etc.) from a raw text corpus for the language in question. First, we perform syntactic analysis, predicting part-of-speech (POS) tags, morphological features, and dependency trees. Using this analyzed data, we then learn an agreement prediction model that contains the desired rules. Specifically, we devise a binary classification problem of identifying whether agreement will be observed between a head and its dependent token on a given morphological property. We use decision trees as our classification model because...
they are easy to interpret and we can easily extract the classification rules from the tree leaves to get an initial set of potential agreement rules. Finally, we perform rule labeling of the extracted rules, identifying which tree leaves correspond to probable agreement. This is required because not all agreeing head-dependent token pairs are necessarily due to some underlying rule. For instance, in Figure 1’s example of Greek gender agreement, both the head and its dependent token Ιταλίας/Αλβανίας have feminine gender, but this agreement is purely by-chance, as correctly identified by our framework.

The quality of the learnt rules depends crucially on the quality and quantity of dependency parsed data, which is often not readily available for low-resource languages. Therefore, we experiment with not only gold-standard treebanks, but also trees generated automatically using models trained using cross-lingual transfer learning. This assesses the applicability of the proposed method in a situation where a linguist may want to explore the characteristics of agreement in a language that does not have a large annotated dependency treebank.

We evaluate the correctness of the extracted rules conducting human evaluation with linguists for Greek, Russian, and Catalan. In addition to the manual verification, we also devise a new metric for automatic evaluation of the rules over unseen test data. Our contributions can be summarized to:

1. We propose a framework to automatically extract agreement rules from raw text, and release these rules for 55 languages as part of an interface\(^2\) which visualizes the rules in detail along with examples and counter-examples.

2. We design a human evaluation interface to allow linguists to easily verify the extracted rules. Our framework produces a decent first-pass grammatical specification with the extracted rules having an average accuracy of 78%. We also devise an automated metric to evaluate our framework when human evaluation is infeasible.

3. We evaluate the quality of extracted rules under real zero-shot conditions (on Breton, Buryat, Faroese, Tagalog, and Welsh) as well as low-resource conditions (with simulation experiments on Spanish, Greek, Belarusian and Lithuanian) varying the amount of training data. Using cross-lingual transfer, rules extracted with as few as 50 sentences with gold-standard syntactic analysis are nearly equivalent to the rules extracted when we have hundreds/thousands of gold-standard data available.

2 Problem Formulation

For a head \(h\) and a dependent \(d\) that are in a dependency relation \(r\), we will say that they agree on a morphological property \(f\) if they share the same value for that particular property i.e. \(f_h = f_d\). Some agreements that we observe in parsed data can be attributed to an underlying grammatical rule. For example, in Figure 2 the Spanish A.1 shows an example of where subject (enigmas) and verb (son) need to agree on number. We will refer to such rules as required-agreement. Such a required agreement rule dictates that an example like A.2 is ungrammatical and would not appear in well-formed Spanish sentences, since the subject and

\(^2\)https://neulab.github.io/lase/
which are defined over the dependency relation, we have the following rule for Spanish: “subjects we will observe at most chance-agreement is treated as singular for purposes of agreement. eg. United Nations is, despite United Nations being plural, it is treated as singular for purposes of agreement.

Figure 2: Subject-verb number agreement is required in Spanish, as in example A.1, which renders example A.2 ungrammatical. Object-verb agreement is not required, so both B.1 and B.2 are grammatical. The object and the verb in B.1 only agree by chance.

The verb do not have the same number marking. However, not all word pairs that agree do so because of some underlying rule, and we will refer to such cases as chance-agreement. For example, in Figure 2 the object (perro) and verb (tiene) in B.1 only agree in number by chance, and example B.2 (where the object of a singular verb is plural) is perfectly acceptable.

Our goal is to extract, from textual examples, the set of rules \( \mathcal{R} \) that concisely describe the agreement process for language \( l \). Concretely, this will indicate for which head-dependent pairs the language displays required-agreement and for which we will observe at most chance-agreement. Canonically, agreement rules are defined over syntactic features of a language as seen in Figure 2 where we have the following rule for Spanish: “subjects agree with their verbs on number.”

To formalize this notion, we define a rule to be a set of features which are defined over the dependency relation, head and dependent token types. In this paper, we make the simplifying assumption that head and dependent tokens are represented by only part-of-speech features, as we would like our extracted rules to be concise and easily interpretable downstream, although this assumption could be relaxed in future work.

The rule discovery process consists of two major steps: a rule extraction step followed by a rule labeling and merging step (also see Figure 1).

2.1 Rule Extraction

To create our training data for rule extraction, we first annotate raw text with part-of-speech (POS) tags, morphological analyses, and dependency trees. We then base our training data on these annotations by converting each dependency relation into a triple \( \langle h, d, r \rangle \), indicating the head token, dependent/child token, and dependency relation between \( h \) and \( d \) respectively. From the whole treebank, we now have input features \( X_f = \{ \langle h_1, d_1, r_1 \rangle, \ldots, \langle h_n, d_n, r_n \rangle \} \) and binary output labels \( Y = y_1, \ldots, y_n \), where if the head and the dependent token agree on feature \( f \) (such that \( f_h = f_d \)) we set \( y = 1 \), otherwise \( y = 0 \). We filter out the tuples where either of the linked tokens does not display the morphological feature \( f \).

We train a model for \( p(Y|X) \) using decision trees (Quinlan, 1986) using the CART algorithm (Breiman et al., 1984). A major advantage of decision trees is that they are easy to interpret and we can visualize the exact features used by the decision tree to split nodes. The decision tree induces a distribution of agreement over training samples in each leaf, e.g. 99% agree, 1% not agree in Leaf-3 for gender agreement in Spanish (Figure 3(a)).

2.2 Rule Labeling

Now that we have constructed a decision tree where each tree leaf corresponds to a salient partition of the possible syntactic structures in the language, we then label these tree leaves as required-agreement or chance-agreement. For this we apply a threshold on the ratio of agreeing training samples within a leaf – if the ratio exceeds a certain number the leaf will be judged as required-agreement. We experiment with two types of thresholds:

**Hard Threshold:** We set a hard threshold on the ratio that is identical for all leaves. In all experiments, we set this threshold to 90% based on manually inspecting some resulting trees to find a threshold that limited the number of non-agreeing syntactic structures being labeled as required-agreement.

**Statistical Threshold:** Leaves with very few examples may exceed the hard threshold purely by chance. In order to better determine whether the agreements are indeed due to a true pattern of required agreement, we devise a thresholding strategy based on significance testing. For all agreement-majority leaves, we apply a chi-squared goodness of fit test to compare the observed output distri-
bution with an expected probability distribution specified by a null hypothesis. Our null hypothesis $H_0$ will be that any agreement we observe is due to chance. If we reject the null hypothesis, we will conclude from the alternate hypothesis $H_1$ that there exists a grammatical rule requiring agreement for this leaf’s cases:

$H_0$: The leaf has chance-agreement.

$H_1$: The leaf has required-agreement.

If there is no rule requiring agreement, we assume that the morphological properties of the head and the dependent token are independent and identically distributed discrete random variables following a categorical distribution. We compute the probability of chance agreement based on the number of values that the specific morphological property $f$ can take. Since morphological feature values are not equally probable, we use a probability proportional to the observed value counts. For a binary number property where 90% of all observed occurrences are singular and 10% are plural, the probability of chance agreement is equal to 0.82 = 0.9 × 0.9 + 0.1 × 0.1, which gives the observed output distribution $p = [0.18, 0.82]$. Using $p$ we compute the expected frequency count $E_i = np_i$ where $n$ is the total number of samples in the given leaf, $i = [0, 1]$ is the output class of the leaf, and $p_i$ is the hypothesized proportion of observations for class $i$. The chi-squared test calculates the test statistic $\chi^2$ as follows:

$$\chi^2 = \sum_{i \in [0, 1]} \frac{(O_i - E_i)^2}{E_i}$$

where $O_i$ is the observed frequency count in the given leaf. The test outputs a $p$-value, which is the probability of observing a sample statistic as extreme as the test statistic. If the $p$-value is smaller than a chosen significance level (we use 0.01) we reject the null hypothesis and label the leaf as required-agreement.

The chi-squared test especially helps in being cautious with leaves with very few examples. However, for leaves with larger number of examples statistical significance alone is insufficient, because there are a large number of cases where there are small but significant differences from the ratio of agreement expected by chance. Therefore, in addition to comparing the $p$-value we also compute the effect size which provides a quantitative measure on the magnitude of an effect (Sullivan and Feinn, 2012). Cramér’s phi $\phi_c$ (Cramér, 1946) is a commonly used method to measure the effect size:

$$\phi_c = \frac{\chi^2}{N(k-1)}$$

where $\chi^2$ is the test statistic computed from the chi-squared test, $N$ is the total number of samples within a leaf, and $k$ is the degree of freedom (which in this case is 2 since we have two output classes). Cohen (1988) provides rules of thumb for interpreting these effect size. For instance, $\phi_c > 0.5$ is considered to be a large effect size and a large effect size suggests that the difference between the two hypotheses is important. Therefore, a leaf is labeled as required-agreement when the $p$-value is less than the significance value and the effect size is greater than 0.5. Now Leaf-1 in Figure 3(b) is correctly identified as chance-agreement.

\footnote{One limitation of this is that rules that show agreement sometimes get incorrectly labeled as either chance-agreement or required-agreement. We consider this in evaluation, but predicting sometimes agreement is relegated to future work.}
**Rule Merging:** Because we are aiming to have a concise, human-readable representation of agreement rules of a language, after labeling the tree leaves we merge sibling leaves with the same label as shown in Figure 3(c). Further, we collapse tree nodes having all leaves with the same label thereby reducing the apparent depth of the tree.

### 3 Experimental Settings and Evaluation

Our experiments aim to answer the following research questions: (1) can our framework extract linguistically plausible agreement rules across diverse languages? and (2) can it do so even if gold-standard syntactic analyses are not available? To answer the first question we evaluate rules extracted from gold-standard syntactic analysis (Sec. §4). For the second question we experiment in low-resource and zero-shot scenarios using cross-lingual transfer to obtain parsers on the languages of interest, and evaluate the effect of noisy parsing results on the quality of rules (Sec. §5).

#### 3.1 Settings

**Data** We use the Surface-Syntactic Universal Dependencies (SUD) treebanks (Gerdes et al., 2018, 2019) as the gold-standard source of complete syntactic analysis. The SUD treebanks are derived from Universal Dependencies (UD) (Nivre et al., 2016, 2018), but unlike the UD treebanks which favor content words as heads, the SUD ones express dependency labels and links using purely syntactic criteria, which is more conducive to our goal of learning syntactic rules. We use the tool of Gerdes et al. (2019) to convert UD v.2.5 (Nivre et al., 2020) into SUD. We only use the training portion of the treebanks for learning our rules.

**Rule Learning** We use sklearn’s (Pedregosa et al., 2011) implementation of decision trees and train a separate model for each morphological feature $f$ for a given language. We experiment with six morphological features (Gender, Person, Number, Mood, Case, Tense) which are most frequently present across several languages. We perform a grid search over the decision tree parameters (detailed in Appendix A.1) and select the model performing best on the validation set. We report results with the Statistical Threshold because on manual inspection we find the trees to be more reliable than the ones learnt from the Hard Threshold (see Appendix A.5 for an example).

#### 3.2 Evaluation

We explore two approaches to evaluate the extracted rules, one based on expert annotations, and an automated proxy evaluation.

**Expert Evaluation** Ideally, we would collect annotations for all head-relation-dependent triples in a treebank, but this would involve annotating hundreds of triples, requiring a large time commitment from linguists in each language we wish to evaluate. Instead, for each language/treebank we extract and evaluate the top 20 most frequent “head POS, dependency relation, dependent POS” triples for the six morphological features amounting to 120 sets of triples to be annotated. We then present these triples with 10 randomly selected illustrative examples and ask a linguist to annotate whether there is a rule in this language governing agreement between the head-dependent pair for this relation. The allowed labels are: Almost always agree if the construction must almost always exhibit agreement on the given feature; Sometimes agree if the linked arguments sometimes must agree, but sometimes do not have to; Need not agree if any agreement on the feature is random. An example of the annotation interface is shown in the Appendix A.2.

For each of the human annotated triples in feature $f$, we extract the label assigned to it by the learnt decision tree $T$. We find the leaf to which the given triple $t$ belongs and assign that leaf’s label to the triple, referred by $l_{\text{tree},f,t}$. The human evaluation score (HS) for each triple marking feature $f$ is given by:

$$HS_{f,t} = 1 \{ \begin{array}{ll} \text{l}_{\text{human},f,t} = l_{\text{tree},f,t} \\ 0 \text{ otherwise} \end{array}$$

where $l_{\text{human},f,t}$ is the label assigned to the triple $t$ by the human annotator. These scores are then averaged across all annotated triples $T_f$ to get the human evaluation metric (HRM) for feature $f$

$$\text{HRM}_f = \frac{\sum_{t \in T_f} HS_{f,t}}{|T_f|}.$$  

**Automated Evaluation** As an alternative to the infeasible manual evaluation of all rules in every language, we propose an automated rule metric (ARM) that evaluates how well the rules extracted from decision tree $T$ fit to unseen gold-annotated test data. For each triple $t$ marking feature $f$, we

---

5The top 20 most frequent triples covered approximately 95% of the triples where this feature was active on average.
first retrieve all examples from the test data corresponding to that triple. Next, we calculate the empirical agreement by counting the fraction of test samples that exhibit agreement, referred to as $q_{f,t}$. For a required-agreement leaf, we expect most test samples satisfying that rule to show agreement. To account for any exceptions to the rule and/or parsing-related errors, we use a threshold that acts as proxy for evaluating whether the given triple denotes required agreement. We use a threshold of 0.95, and if $q_{f,t} > 0.95$ then we assign the test label $l_{test,f,t}$ for that triple as required-agreement, and otherwise choose chance-agreement. Similar to the human evaluation, we compute a score for each triple $t$ marking feature $f$:

$$AS_t = \begin{cases} 1 & l_{test,f,t} = l_{tree,f,t} \\ 0 & \text{otherwise} \end{cases}$$

then average scores across all annotated triples in $T_f$ to get the ARM score for each feature $f$:

$$\text{ARM}_f = \frac{\sum_{t \in T_f} AS_t}{|T_f|}$$

### 4 Experiments with Gold-Standard Data

In this section, we evaluate the quality of the rules induced by our framework, using gold-standard syntactic analyses and learning the decision trees over triples obtained from the training portion of all SUD treebanks. As baseline, we compare with trees predicting all leaves as chance-agreement.

---

6There are exceptions: e.g. when the head of dependent is a multword expression (MWE), in which case dependency parsers might miss or pick only one of its constituents as head/dependent, or if the MWE is syntactically idiosyncratic.

7We keep a 5% margin to account for any exceptions or parsing errors based on the feedback given by the annotators.

---

Figure 4: Difference in the ARM scores of decision trees over gold-standard syntactic analysis with baseline trees where all leaves predict chance-agreement.

The extracted rules have an 0.574 ARM score (averaged across all treebanks and features), outperforming the baseline scores by 0.074 ARM points. Of all the 451 decision trees across all treebanks and features, we find 78% trees outperforming the baseline trees. In Figure 4, we show the improvements over the baseline averaged across language families/genera. In families with extensive agreement systems such as Slavic and Baltic our models clearly outperform the baseline discovering correct rules, as they do for the other Indo-European genera, Indo-Aryan and Germanic. For mood and tense, the chance-agreement baseline performs on par with our method. This is not surprising because there is little agreement observed for these features given that only verbs and auxiliary verbs mark these features. We find that for both tense and mood in the Indo-Aryan family, our model identifies required-agreement primarily for conjoined verbs, which mostly need to agree only if they share the same subject. However, subsequent analysis revealed that in the treebanks nearly 50% of the agreeing verbs do not share the same subject but do agree by chance.

Agreement for Indo-European languages like Hindi and Russian is well documented (Comrie, 1984; Crockett, 1976) and is reflected in our large improvements over the baseline (Figure 5). Similarly, Arabic exhibits extensive agreement on noun phrases including determiners and adjectives (Aoun et al., 1994). We find that for Arabic gender the lower ARM scores of our method are an artifact of the small test data.

North Sami is an interesting test bed: as a Uralic language, case agreement would be somewhat unexpected and indeed our model’s predictions are not better than the baseline. Nevertheless, with our interface we find patterns of rare positive paratactic constructions with required agree-
Where demonstrative pronouns overwhelmingly agree with their heads. The case decision tree also uncovers interesting patterns of 100% agreement on Tamil constructions with nominalized verbs (Gerunds) where the markings propagate to the whole phrase.

Conciseness of Extracted Rules We further analyze the decision trees learnt by our framework for conciseness and find that the trees grow more complex with increasing morphological complexity of languages as seen in Figure 6. To compute the morphological complexity of a language, we use the word entropy measure proposed by Bentz et al. (2016) which measures the average information content of words and is computed as follows:

$$H(D) = -\sum_{i \in V} p(w_i) \log p(w_i)$$

where $V$ is the vocabulary, $D$ is the monolingual text extracted from the training portion of the respective treebank, $p(w_i)$ is the word type frequency normalized by the total tokens. Since this entropy doesn’t account for unseen word types, Bentz et al. (2016) use the James-Stein shrinkage estimator (Hauser and Strimmer, 2009) to calculate $p(w_i)$:

$$p(w_i) = \lambda p^{\text{target}}(w_i) + (1 - \lambda)p^{\text{ML}}(w_i)$$

where $\lambda \in [0, 1]$, $p^{\text{target}}$ denotes the maximum entropy case given by the uniform distribution $\frac{1}{v}$ and $p^{\text{ML}}$ is the maximum likelihood estimator which is given by the normalized word type frequency. Languages with a larger word entropy are considered to be morphologically rich as they pack more information into the words. In Figure 6 we plot the morphological richness with the average number of leaves across all features and find these to be highly correlated.

Manual Evaluation Results We conduct an expert evaluation for Greek (el), Russian (ru) and Catalan (ca) as described in Section §3.2. For a strict setting, we consider both Sometimes agree and Need not agree as chance-agreement and report the human evaluation metric (HRM) in Figure 7. Overall, our method extracts first-pass grammar rules achieving 89% accuracy for Greek, 78% for Russian and 66% for Catalan.

In most error cases, like person in Russian, our model produces required-agreement labels, which we can attribute to skewed data statistics in the treebanks. In Russian and Greek, for instance, conjoined verbs only need to agree in person and number if they share the same subject (in which case they implicitly agree because they both must agree with the same subject phrase). In the treebanks, though, only 15% of the agreeing verbs do indeed share the same subject – the rest agree by chance. In a reverse example from Catalan, the overwhelming majority (92%) of 8650 tokens are in the third-person, causing our model to label all leaves as chance agreement despite the fact that person/number agreement is required in such cases. Similarly for tense in Catalan, our framework predicts chance-agreement for auxiliary verbs with verbs as their dependent because of overwhelming majority of disagreeing examples. We believe this is because of both annotation artifact and the way past tense is realized.

To demonstrate how well the automated evaluation correlates with the human evaluation protocol, we compute the Pearson’s correlation ($r$) between the ARM and HRM for each language under four model settings: simulate-50, simulate-100, baseline and gold. simulate-x is a simulated low-resource setting where the model is trained using $x$ gold-standard syntactically analysed data. The baseline setting is the one where all leaves predict chance-agreement and under the gold setting we train using the entire gold-standard data. We compute the ARM and HRM scores for the rules learnt under each of the four settings and report the Pearson’s correlation, averaged across all features. Overall, we observe a moderate correlation for all three languages, with $r=0.59$ for Greek, $r=0.41$ for Russian and $r=0.38$ for Catalan. The correla-
tions are very strong for some features such as Gender \((r_d=0.97, r_u=0.82, r_c=0.98)\) and Number \((r_n=0.97, r_u=0.69, r_c=0.96)\) where we expect to see extensive agreement.

5 Low-Resource Experiments

5.1 Simulated Zero-/Few-Shot Experiments

It is not always possible to have access to gold-standard syntactic analyses. Therefore, in order to investigate how the quality of rules are affected by the quality of syntactic analysis, we conduct simulation experiments by varying the amount of gold-standard syntactically analysed training data. For each language, we sample \(x\) fully parsed sentences from its treebank out of \(L\) training sentences available. For the remaining \(L-x\) sentences, we use silver syntactic analysis i.e., we train a syntactic analysis model on \(x\) sentences and use the model predictions for the \(L-x\) sentences.

**Data and Setup:** We experiment with Spanish, Greek, Belarusian and Lithuanian. For transfer learning, we use Portuguese, Ancient Greek, Ukrainian and Latvian treebanks respectively. The data statistics and details are in Appendix A.2.

We train **Udify** (Kondratyuk and Straka, 2019), a parser that jointly predict POS tags, morphological features, and dependency trees, using the \(x\) gold-standard sentences as our training data. We generate model predictions on the remaining \(L-x\) sentences. Finally, we concatenate the \(x\) gold data with the \(L-x\) automatically parsed data from which we extract the training data for learning the decision tree. We experiment with \(x=[50, 100, 500]\) gold-standard sentences. To account of sampling randomness, we repeat the process 5 times and report averages across runs.

To further improve the quality of the automatically obtained syntactic analysis, we use cross-lingual transfer learning where we train the Udify model by concatenating \(x\) sentences of the target language with the entire treebank of the related language. We also conduct zero-shot experiments under this setting where we directly use the Udify model trained only on the related language and get the model predictions on \(L\) sentences. As before, we train five decision trees for each \(x\) setting and report the average ARM over the test data.

**Results** We report the results for Number agreement in Figure 8. Similar plots for other languages and features can be found in the Appendix A.5. We observe that using cross-lingual transfer learning (CLTL) already leads to high scores across all languages even in zero-shot settings where we do not use any data from the gold-standard treebank. Taking Spanish gender as an example, 93% of the rule-triples extracted from the gold-standard tree (which are overwhelmingly correct) are also extracted by the zero-shot tree. The zero-shot tree only makes a few mistakes (shown in Table 1 and reflected in its overall ARM score) on certain proper noun and auxiliary verb constructions. Interestingly, using CLTL, training with just 50 gold-standard target language sentences is almost equivalent to

![Graph](image-url)  
**Figure 7:** Annotation accuracy for Greek, Russian and Catalan per each morphological feature.

![Graph](image-url)  
**Figure 8:** Comparing the (avg.) ARM score for Number agreement with and without cross-lingual transfer learning (transfer language in parenthesis). x-axis in log space. The higher the ARM the better.

| Relation, Head, Dependent | correct label | gold zero-shot |
|--------------------------|---------------|----------------|
| det, NOUN, DET,          | almost always | required        |
| mod, NOUN, ADJ           | almost always | required        |
| flat, PROPN, PROPN       | almost always | required        |
| mod, PROPN, PROPN        | almost always | required        |
| appos, PROPN, PROPN      | sometimes    | required        |
| comp:aux@pass, AUX, VERB | need not      | required        |
| conj, PROPN, PROPN       | required      | chance          |
| ARM score over the test set: | 0.644 | 0.632 |

Table 1: The Spanish gender rules extracted in a zero-shot setting are generally similar to the ones extracted from the gold data (93%). We highlight the few mistakes that the zero-shot tree makes.
training with 100 or 500 gold-standard sentences. This opens new avenues for language documentation: with as few as 50 expertly-annotated syntactic analysis of a new language and CLTL our framework can produce decent first-pass agreement rules. Needless to say, in most cases the extracted rules improve as we increase the number of gold-standard sentences and CLTL further helps bridge the data availability gap for low-resource settings.

5.2 Real Zero-Shot Experiments

Some languages like Breton, Buryat, Faroese, Tagalog and Welsh have test data only; there is no gold-standard training data available, which presents a true zero-shot setting. In such cases, we can still extract grammar rules with our framework using zero-shot dependency parsing.

Data and Setup: We collect raw text for the above languages from the Leipzig corpora (Goldhahn et al., 2012). Data statistics are listed in Appendix A.2. We parse these sentences using the “universal” Udify model that has been pre-trained on all of the UD treebanks, as released by (Kondratyuk and Straka, 2019). As before, we use these automatically parsed syntactic analyses to extract the rules which we evaluate with ARM over the gold standard test data of the corresponding SUD treebanks.

Results: We report the ARM scores in Figure 9. Averaged over all rules, our approach obtains a ARM of 0.566, while the naive all-chance baseline only achieves 0.506. The difference appears to be small, but we still consider it significant, because these languages do not actually require agreement for many grammatical features. Tagalog and Buryat are the most distant languages that we test on (no Philippine and Mongolic language is present in our training data) and yet we observe our method being at par with the baseline and even outperforming in case of Tagalog. Breton and Welsh, on the other hand, are an interesting test bed: Celtic languages are to some degree outliers among Indo-European languages (Borsley and Roberts, 2005), and we suspect that as a result the parser performs generally worse. Despite that, our approach has an ARM of 0.730 for Welsh gender agreement, as opposed to the mere 0.615 that the baseline achieves.

6 Related Work

Bender et al. (2014) use interlinear glossed text (IGT) to extract lexical entities and morphological rules for an endangered language. They experiment with different systems which individually extract lemmas, lexical rules, word order and the case system, some of which use hand-specified rules. Howell et al. (2017) extend this to work to predict case system on additional languages. Zamaraeva (2016) also infer morphotactics from IGT using $k$-means clustering. To the best of our knowledge, our work is the first to propose a framework to extract first-pass grammatical agreement rules directly from raw text in a statistically-informed objective way. A parallel line of work (Hellan, 2010) extracts a construction profile of a language by having templates that define how sentences are constructed.

7 Future Work

While we have demonstrated that our approach is effective in extracting a first-pass set of agreement rules directly from raw text, it focuses only on agreement between a pair of words and hence might fail to capture more complex phenomena that require broader context or operate at the phrase level. Consider this simple English example: “John and Mary love their dog”. Under both UD and SUD formalisms, the coordinating conjunction “and” is a dependent, hence the verb will not agree with either of the (singular) nouns (“John” or “Mary”). Also, deciding agreement based on only POS tags is insufficient to capture all phenomena that may influence agreement for e.g. mass nouns such as ‘rice’ do not follow the standard number agreement rules in English. We leave a more expressive model and evaluation on more languages as future work. We also plan to expand our methodology for extracting grammar rules from raw text to other aspects of morphosyntax, such as argument structure and word order phenomena.
Acknowledgments

The authors are grateful to the anonymous reviewers who took the time to provide many interesting comments that made the paper significantly better, and to Josep Quer, Ekaterina Vylomova and Maria Ryskina, for participating in the human annotation experiments. This work is sponsored by the DARPA grant FA8750-18-2-0018 and by the National Science Foundation under grant 1761548.

References

Joseph Aoun, Elabas Benmamoun, and Dominique Sportiche. 1994. Agreement, word order, and conjunction in some varieties of arabic. Linguistic inquiry, pages 195–220.

Emily M. Bender, Joshua Crowgey, Michael Wayne Goodman, and Fei Xia. 2014. Learning grammar specifications from IGT: A case study of chintang. In Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 43–53, Baltimore, Maryland, USA. Association for Computational Linguistics.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CLALC), pages 142–153, Osaka, Japan. The COLING 2016 Organizing Committee.

Robert D Borsley and Ian Roberts. 2005. The syntax of the celtic languages: a comparative perspective. Cambridge University Press.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. Classification and regression trees. CRC press.

Keith Brown and Sarah Ogilvie. 2010. Concise encyclopedia of languages of the world. Elsevier.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pages 108–122.

Robin Cohen. 1988. Book reviews: Reasoning and discourse processes. Computational Linguistics, 14(4).

Bernard Comrie. 1984. Reflections on verb agreement in hindi and related languages.

Greville G Corbett. 2009. Agreement. In Die slavischen Sprachen/The Slavic Languages.

Harald Cramér. 1946. Mathematical methods of statistics. Princeton U. Press, Princeton, page 500.

Dina B Crockett. 1976. Agreement in contemporary standard Russian. Slavica Publishers Inc.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. In Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019. Improving surface-syntactic universal dependencies (SUD): MWEs and deep syntactic features. In Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019), pages 126–132, Paris, France. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In LREC, volume 29, pages 31–43.

Zellig S Harris. 1951. Methods in structural linguistics.

Jean Haussler and Korbinian Strimmer. 2009. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. Journal of Machine Learning Research, 10(7).

Lars Hellan. 2010. From descriptive annotation to grammar specification. In Proceedings of the Fourth Linguistic Annotation Workshop, pages 172–176, Uppsala, Sweden. Association for Computational Linguistics.

Kristen Howell, Emily M Bender, Michel Lockwood, Fei Xia, and Olga Zamarayaeva. 2017. Inferring case systems from igt: Enriching the enrichment. In Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 67–75.

Rodney D Huddleston. 2002. The Cambridge grammar of the English language. Cambridge, UK; New York: Cambridge University Press.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Joakim Nivre, Rogier Blokland, Niko Partanen, Michael Rießler, and Jack Rueter. 2018. Universal Dependencies 2.3.
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 the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 1659–1666.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 4034–4043, Marseille, France. European Language Resources Association.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

J. Ross Quinlan. 1986. Induction of decision trees. Machine learning, 1(1):81–106.

Gail M Sullivan and Richard Feinn. 2012. Using effect size—or why the p value is not enough. Journal of graduate medical education, 4(3):279–282.

Olga Zamaraeva. 2016. Inferring morphotactics from interlinear glossed text: Combining clustering and precision grammars. In Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 141–150, Berlin, Germany. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.
A Appendix

A.1 Decision Tree Hyperparameters

We perform a grid search over the following hyperparameters of the decision tree:

- \( \text{criterion} = [\text{gini, entropy}] \)
- \( \text{max depth} = [6, 15] \)
- \( \text{min impurity decrease} = 1e^{-3} \)

The best parameters are selected based on the validation set performance. For some treebanks which have no validation set we use the default cross-validation provided by sklearn (Buitinck et al., 2013). Average model runtime for a treebank is 5-10 mins depending on the size of the treebank.

A.2 Dataset Statistics

For the true low-resource experiments, the dataset details are in Table 2.

| Language      | Train / Test |
|---------------|--------------|
| Breton-KEB    | 30000 / 888  |
| Buryat-BXR    | 10000 / 908  |
| Faroese-OFT   | 50000 / 1208 |
| Tagalog-TRG   | 30000 / 1208 |
| Welsh-CCG     | 30000 / 956  |

Table 2: Dataset statistics. Training data is obtained by parsing the Liepzig corpora (Goldhahn et al., 2012) and test data is obtained from the respective treebank. Each cell denotes the number of sentences in train/test.

A.3 Evaluation

A.4 Annotation Interface for Expert Evaluation

In Figure 10, we show the annotation interface used for verifying Gender agreement rules in Catalan. For each triple, we display 10 randomly selected examples from the training portion of the treebank.

A.5 Low-resource Experiment Results

For the simulation experiments, the dataset details are in Table 3.

A.5.1 Udify (Kondratyuk and Straka, 2019) Model Details

We used the Udify model for automatically annotating the raw text with part-of-speech (POS), dependency links and morphological features. For each of the simulation experiment we report the udify parsing performance on the test data in Table 4. We used the same hyperparameters for training with a related languages as specified by the authors.\(^{11}\) In the configuration file, we only change the parameters \texttt{warmup steps= 100} and \texttt{start-step= 100}, as recommended by the authors for low-resource languages.

A.5.2 Results and Discussion

For each language and feature, we plot the ARM score with and without transfer learning in Figure 12-14. Similar to our findings for Gender in Figure 5, we find that cross-lingual transfer leads to a better score across all languages in the zero-shot setting. As we increase the number of gold-standard sentences, the quality of extracted rules improve. Although, for Belarusian we observe the opposite trend for Person agreement. On closer inspection we find that it is because person applies only to non-past finite verb forms (VERB and AUX) as an inflectional feature and to pronouns (PRON) as a lexical feature which means that in many cases person is not explicitly marked, even though it implicitly exists\(^{12}\).

A.6 Experiments with Gold-Standard Data

We present the ARM scores for all treebanks and features in Tables 5-11. We also report the validation results in the same tables for our best setting which uses the Statistical Threshold. In Section 2.2, we proposed using two types of thresholds for retaining the high probability agreement rules. In order to compare which threshold is the best for all treebanks, we manually inspect some of the learnt decision trees. We find that for the trees learnt from the hard threshold often over-fit on the training data causing to produce leaves with very few examples. In Figure 15 we compare the trees constructed for number agreement with the two thresholds for Marathi. One reason why Statistical-Threshold performs better for low-resource languages is because there are more leaves with fewer samples overall causing the Hard Threshold to have more false positives. Whereas the Statistical Threshold uses effect size with the significance test which takes into account the sample size within a leaf leading to better leaves. Therefore, we choose to use Statistical-Threshold for all our simulation experiments.

In Figure 11, we report that (avg.) number of leaves in the decision trees grouped by language

\(^{11}\)https://github.com/Hyperparticle/udify

\(^{12}\)https://universaldependencies.org/be/
family. Overall, Gender and Case tend to have more complex trees. For Case, it is probably because languages have more number of cases making it harder for the decision tree to model them.

### A.7 SUD treebanks

Figure 16 presents a comparison of UD and SUD-style trees for the German sentence, "Ich werde lange Bücher lesen.". The SUD tree has the function word ‘werde’ as the syntactic head to the content word ‘lesen’.
Table 4: `udify` model performance on the test data for each low-resource setting. The scores are averaged across five runs of each setting.

| LANGUAGE | #TRAINING | W/O TRANSFER | SETTING  |
|----------|-----------|--------------|----------|
| Greek    | 0         | -            | upos:0.661, ufeats:0.392, uas:0.632, las:0.465 |
|          | 50        | upos:0.507, ufeats:0.330, uas:0.309, las:0.203 |
|          | 100       | upos:0.915, ufeats:0.664, uas:0.755, las:0.691 |
|          | 500       | upos:0.970, ufeats:0.891, uas:0.891, las:0.866 |
| Spanish  | 0         | -            | upos:0.922, ufeats:0.764, uas:0.855, las:0.776 |
|          | 50        | upos:0.529, ufeats:0.463, uas:0.289, las:0.152 |
|          | 100       | upos:0.920, ufeats:0.832, uas:0.755, las:0.690 |
|          | 500       | upos:0.952, ufeats:0.919, uas:0.860, las:0.820 |
| Belarusian| 0         | -            | upos:0.941, ufeats:0.520, uas:0.863, las:0.797 |
|          | 50        | upos:0.570, ufeats:0.323, uas:0.217, las:0.141 |
|          | 100       | upos:0.919, ufeats:0.446, uas:0.521, las:0.482 |
| Lithuanian| 0         | -            | upos:0.869, ufeats:0.528, uas:0.752, las:0.610 |
|          | 50        | upos:0.566, ufeats:0.371, uas:0.346, las:0.211 |
|          | 100       | upos:0.813, ufeats:0.453, uas:0.551, las:0.421 |
|          | 500       | upos:0.925, ufeats:0.744, uas:0.757, las:0.697 |

Figure 11: (Avg.) number of leaves for each feature grouped by language family.
Figure 12: Comparing the (avg.) ARM score for Gender agreement with and without cross-lingual transfer learning (transfer language in parenthesis). Note: the higher the ARM the better.

Figure 13: Comparing the (avg.) ARM score for Person agreement with and without cross-lingual transfer learning (transfer language in parenthesis). Note: the higher the ARM the better.
Figure 14: Comparing the (avg.) ARM score for Case agreement with and without cross-lingual transfer learning (transfer language in parenthesis). Note: the higher the ARM the better. For Spanish, there was < 10 data points with Case annotated hence we do not report results for it.
Figure 15: Comparing the learnt trees for Number agreement extracted using (a) Hard Threshold and (b) Statistical Threshold. Hard Threshold overfits on the training data resulting in leaves with very few samples.
Figure 16: Comparing the UD (a) tree with the SUD (b) tree for the German sentence “Ich werde lange Bücher lesen.”.
| TREEBANK        | FEATURE | STATISTICAL | HARD | BASELINE | DEV   |
|----------------|---------|-------------|------|----------|-------|
| ru-gsd         | Gender  | 0.678       | 0.51 | 0.623    |       |
| ru-gsd         | Person  | 0.125       | 0.125| 0.286    |       |
| ru-gsd         | Number  | 0.628       | 0.578| 0.384    | 0.62  |
| ru-gsd         | Tense   | 0.667       | 0.667| 0.667    | 0.571 |
| ru-gsd         | Mood    | 0.0         | 1.0  | 0.0      | 0.1   |
| ru-gsd         | Case    | 0.649       | 0.614| 0.395    | 0.537 |
| id-gsd         | Number  | 0.047       | 0.961| 0.047    | 0.045 |
| it-isdt        | Gender  | 0.816       | 0.816| 0.289    | 0.738 |
| it-isdt        | Person  | 0.304       | 0.87 | 0.304    | 0.619 |
| it-isdt        | Number  | 0.615       | 0.603| 0.41     | 0.588 |
| it-isdt        | Tense   | 0.765       | 0.765| 0.647    | 0.611 |
| it-isdt        | Mood    | 0.25        | 0.75 | 0.25     | 0.273 |
| la-proiel      | Gender  | 0.538       | 0.568| 0.636    | 0.496 |
| la-proiel      | Person  | 0.56        | 0.56 | 0.54     | 0.653 |
| la-proiel      | Number  | 0.648       | 0.574| 0.452    | 0.553 |
| la-proiel      | Tense   | 0.818       | 0.879| 0.879    | 0.824 |
| la-proiel      | Mood    | 0.6         | 0.52 | 0.44     | 0.667 |
| la-proiel      | Case    | 0.759       | 0.782| 0.466    | 0.691 |
| no-nynorsk     | Gender  | 0.644       | 0.567| 0.407    | 0.75  |
| no-nynorsk     | Person  | 0.636       | 0.606| 0.606    | 0.683 |
| no-nynorsk     | Number  | 0.626       | 0.626| 0.586    | 0.693 |
| no-nynorsk     | Tense   | 0.452       | 0.839| 0.645    | 0.467 |
| no-nynorsk     | Mood    | 0.676       | 0.765| 0.676    | 0.4   |
| no-nynorsk     | Case    | 0.694       | 0.702| 0.636    | 0.704 |
| he-htb         | Gender  | 0.747       | 0.747| 0.663    | 0.629 |
| he-htb         | Person  | 0.737       | 0.789| 0.789    | 0.769 |
| he-htb         | Number  | 0.585       | 0.585| 0.415    | 0.505 |
| he-htb         | Tense   | 0.3         | 0.3  | 0.1      | 0.545 |
| no-bokmaal     | Gender  | 0.477       | 0.568| 0.545    | 0.675 |
| no-bokmaal     | Person  | 1.0         | 1.0  | 0.5      | 1.0   |
| no-bokmaal     | Number  | 0.655       | 0.673| 0.364    | 0.733 |
| no-bokmaal     | Tense   | 0.55        | 0.55 | 0.55     | 0.55  |
| no-bokmaal     | Mood    | 0.0         | 1.0  | 0.0      | 0.1   |
| no-bokmaal     | Case    | 0.0         | 0.333| 0.0      | 0.0   |
| no-nynorsk     | Gender  | 0.464       | 0.536| 0.536    | 0.514 |
| no-nynorsk     | Person  | 0.0         | 0.0  | 0.0      | 0.667 |
| no-nynorsk     | Number  | 0.702       | 0.702| 0.511    | 0.596 |
| no-nynorsk     | Tense   | 0.368       | 0.368| 0.684    | 0.429 |
| no-nynorsk     | Mood    | 0.0         | 1.0  | 0.0      | 0.048 |
| pl-tdt         | Person  | 0.387       | 0.677| 0.677    | 0.607 |
| pl-tdt         | Number  | 0.502       | 0.493| 0.511    | 0.559 |
| pl-tdt         | Tense   | 0.474       | 0.368| 0.474    | 0.5   |
| pl-tdt         | Mood    | 0.75        | 0.75 | 0.75     | 0.471 |
| pl-tdt         | Case    | 0.786       | 0.828| 0.821    | 0.781 |
| pl-ifg         | Gender  | 0.646       | 0.646| 0.463    | 0.641 |
| pl-ifg         | Person  | 0.688       | 0.688| 0.562    | 0.714 |
| pl-ifg         | Number  | 0.691       | 0.68 | 0.412    | 0.624 |
| pl-ifg         | Tense   | 0.556       | 0.667| 0.667    | 0.6   |
| pl-ifg         | Mood    | 0.333       | 0.667| 0.333    | 0.4   |
| pl-ifg         | Case    | 0.744       | 0.667| 0.41     | 0.617 |
| grc-perseus    | Gender  | 0.62        | 0.718| 0.563    | 0.699 |
| grc-perseus    | Person  | 0.8         | 0.8  | 0.7      | 0.636 |
| grc-perseus    | Number  | 0.531       | 0.63 | 0.605    | 0.537 |
| grc-perseus    | Tense   | 0.889       | 1.0  | 1.0      | 0.778 |
| grc-perseus    | Mood    | 0.833       | 0.833| 0.667    | 0.429 |
| grc-perseus    | Case    | 0.708       | 0.792| 0.556    | 0.712 |
| fi-ftb         | Person  | 0.56        | 0.76 | 0.6      | 0.63  |
| fi-ftb         | Number  | 0.524       | 0.441| 0.524    | 0.54  |
| fi-ftb         | Tense   | 0.846       | 0.769| 0.308    | 0.538 |
| fi-ftb         | Mood    | 0.429       | 0.5  | 0.429    | 0.529 |
| fi-ftb         | Case    | 0.724       | 0.848| 0.781    | 0.748 |
| wo-wtb         | Gender  | 0.5         | 0.5  | 0.5      | 0.0   |
| wo-wtb         | Person  | 0.55        | 0.45 | 0.4      | 0.609 |
| wo-wtb         | Number  | 0.486       | 0.6  | 0.6      | 0.632 |
| wo-wtb         | Tense   | 0.5         | 0.625| 0.375    | 0.625 |
| wo-wtb         | Mood    | 0.143       | 0.143| 0.143    | 0.364 |
| en-partut      | Person  | 0.5         | 0.5  | 0.417    | 0.857 |
| en-partut      | Number  | 0.559       | 0.559| 0.441    | 0.676 |
| en-partut      | Tense   | 0.667       | 0.733| 0.667    | 0.583 |
| en-partut      | Mood    | 0.091       | 0.818| 0.091    | 0.1   |

Table 5: Comparing the ARM scores for SUD treebanks across both Statistical and Hard thresholding.
| TREEBANK       | FEATURE | STATISTICAL | HARD  | BASELINE | DEV  |
|---------------|---------|-------------|-------|----------|------|
| fr-ftb        | Gender  | 0.631       | 0.631 | 0.477    | 0.621|
| fr-ftb        | Person  | 0.14        | 0.14  | 0.14     | 0.171|
| fr-ftb        | Number  | 0.635       | 0.635 | 0.502    | 0.634|
| fr-ftb        | Tense   | 0.714       | 0.857 | 0.857    | 0.833|
| lv-lvtb       | Mood    | 0.409       | 0.591 | 0.409    | 0.6  |
| lv-lvtb       | Person  | 0.5         | 0.632 | 0.579    | 0.583|
| lv-lvtb       | Number  | 0.688       | 0.688 | 0.429    | 0.706|
| lv-lvtb       | Tense   | 0.667       | 0.815 | 0.889    | 0.741|
| lv-lvtb       | Mood    | 0.476       | 0.619 | 0.476    | 0.333|
| lv-lvtb       | Case    | 0.719       | 0.734 | 0.489    | 0.772|
| ro-rtt        | Gender  | 0.583       | 0.583 | 0.51     | 0.591|
| ro-rtt        | Person  | 0.327       | 0.755 | 0.347    | 0.304|
| ro-rtt        | Number  | 0.535       | 0.585 | 0.528    | 0.56 |
| ro-rtt        | Tense   | 0.421       | 0.684 | 0.789    | 0.526|
| ro-rtt        | Mood    | 0.931       | 1.0   | 0.448    | 0.867|
| ro-rtt        | Case    | 0.862       | 0.788 | 0.588    | 0.854|
| it-vit        | Gender  | 0.672       | 0.672 | 0.375    | 0.678|
| it-vit        | Person  | 0.625       | 0.625 | 0.792    | 0.667|
| it-vit        | Number  | 0.712       | 0.728 | 0.528    | 0.61 |
| it-vit        | Tense   | 0.773       | 0.955 | 0.955    | 0.75 |
| it-vit        | Mood    | 0.4         | 0.6   | 0.4      | 0.231|
| fr-partut     | Gender  | 0.579       | 0.632 | 0.421    | 0.615|
| fr-partut     | Person  | 0.818       | 0.727 | 0.273    | 0.75 |
| fr-partut     | Number  | 0.771       | 0.542 | 0.292    | 0.542|
| fr-partut     | Tense   | 0.857       | 0.857 | 0.714    | 0.6 |
| fr-partut     | Mood    | 0.333       | 0.667 | 0.333    | 0.167|
| en-ewt        | Person  | 0.812       | 0.812 | 0.25     | 0.85 |
| en-ewt        | Number  | 0.357       | 0.643 | 0.357    | 0.304|
| en-ewt        | Tense   | 0.591       | 0.773 | 0.773    | 0.593|
| en-ewt        | Mood    | 0.4         | 0.733 | 0.4      | 0.333|
| ru-syntagrus  | Gender  | 0.697       | 0.747 | 0.624    | 0.673|
| ru-syntagrus  | Person  | 0.625       | 0.667 | 0.667    | 0.72 |
| ru-syntagrus  | Number  | 0.591       | 0.661 | 0.562    | 0.576|
| ru-syntagrus  | Tense   | 0.727       | 0.818 | 0.818    | 0.667|
| ru-syntagrus  | Mood    | 0.4         | 0.8   | 0.44     | 0.407|
| ru-syntagrus  | Case    | 0.649       | 0.707 | 0.575    | 0.681|
| sv-talbanken  | Gender  | 0.719       | 0.719 | 0.438    | 0.643|
| sv-talbanken  | Person  | 0.659       | 0.634 | 0.463    | 0.571|
| sv-talbanken  | Number  | 0.559       | 0.588 | 0.5      | 0.607|
| sv-talbanken  | Tense   | 0.623       | 0.952 | 0.048    | 0.056|
| sv-talbanken  | Case    | 0.189       | 0.189 | 0.286    | 0.143|
| olo-kkpp      | Person  | 0.286       | 0.571 | 0.286    | -    |
| olo-kkpp      | Number  | 0.667       | 0.692 | 0.667    | -    |
| olo-kkpp      | Tense   | 0.75        | 0.75  | 0.75     | -    |
| olo-kkpp      | Mood    | 0.7         | 0.75  | 0.7      | -    |
| olo-kkpp      | Case    | 0.7         | 0.7   | 0.7      | -    |
| cs-cac        | Gender  | 0.663       | 0.673 | 0.602    | 0.678|
| cs-cac        | Person  | 0.562       | 0.562 | 0.5      | 0.583|
| cs-cac        | Number  | 0.636       | 0.531 | 0.469    | 0.575|
| cs-cac        | Tense   | 0.467       | 0.667 | 0.6      | 0.333|
| cs-cac        | Mood    | 0.2         | 0.4   | 0.2      | 0.111|
| cs-cac        | Case    | 0.81        | 0.84  | 0.46     | 0.833|
| ur-udtdb      | Gender  | 0.567       | 0.567 | 0.536    | 0.576|
| ur-udtdb      | Person  | 0.152       | 0.946 | 0.065    | 0.195|
| ur-udtdb      | Number  | 0.485       | 0.583 | 0.485    | 0.496|
| ur-udtdb      | Tense   | 0.333       | 0.5   | 0.5      | 0.667|
| ur-udtdb      | Mood    | 0.714       | 0.714 | 0.143    | 0.714|
| ur-udtdb      | Case    | 0.685       | 0.696 | 0.696    | 0.7 |
| et-ewt        | Person  | 0.609       | 0.696 | 0.609    | -    |
| et-ewt        | Number  | 0.551       | 0.551 | 0.48     | -    |
| et-ewt        | Tense   | 0.409       | 0.682 | 0.636    | -    |
| et-ewt        | Mood    | 0.533       | 0.4   | 0.533    | -    |
| et-ewt        | Case    | 0.7         | 0.754 | 0.657    | -    |
| fro-srnl0mf   | Tense   | 0.5         | 1.0   | 0.5      | 1.0  |
| es-gsd        | Gender  | 0.718       | 0.718 | 0.366    | 0.736|
| es-gsd        | Person  | 0.591       | 0.545 | 0.591    | 0.355|
| es-gsd        | Number  | 0.644       | 0.644 | 0.424    | 0.567|
| es-gsd        | Tense   | 0.529       | 0.824 | 0.824    | 0.409|
| es-gsd        | Mood    | 0.533       | 0.467 | 0.533    | 0.474|
| es-gsd        | Case    | 0.0         | 1.0   | 0.0      | 0.0  |

Table 6: Comparing the ARM scores for SUD treebanks across both Statistical and Hard thresholding.
| TREEBANK   | FEATURE  | STATISTICAL | HARD  | BASELINE | DEV  |
|-----------|----------|-------------|-------|----------|------|
| sl-ssj    | Gender   | 0.818       | 0.8   | 0.527    | 0.772|
| sl-ssj    | Person   | 0.667       | 0.722 | 0.722    | 0.706|
| sl-ssj    | Number   | 0.683       | 0.683 | 0.564    | 0.712|
| sl-ssj    | Tense    | 0.333       | 0.583 | 0.333    | 0.364|
| sl-ssj    | Mood     | 0.5         | 0.75  | 0.5      | 0.667|
| sl-ssj    | Case     | 0.607       | 0.721 | 0.557    | 0.61 |
| cs-pdt    | Gender   | 0.564       | 0.788 | 0.75     | 0.58 |
| cs-pdt    | Person   | 0.591       | 0.705 | 0.614    | 0.541|
| cs-pdt    | Number   | 0.477       | 0.64  | 0.629    | 0.481|
| cs-pdt    | Tense    | 0.667       | 0.786 | 0.786    | 0.658|
| cs-pdt    | Mood     | 0.538       | 0.538 | 0.538    | 0.48 |
| cs-pdt    | Case     | 0.646       | 0.675 | 0.545    | 0.633|
| hsb-ufal  | Gender   | 0.857       | 0.714 | 0.786    | -    |
| hsb-ufal  | Number   | 0.692       | 0.538 | 0.692    | -    |
| hsb-ufal  | Tense    | 0.667       | 0.667 | 0.667    | -    |
| ga-idt    | Gender   | 0.546       | 0.846 | 0.462    | -    |
| ga-idt    | Person   | 0.625       | 0.875 | 0.5      | 1.0  |
| ga-idt    | Number   | 0.468       | 0.571 | 0.468    | 0.446|
| ga-idt    | Tense    | 0.714       | 0.571 | 0.429    | 0.5  |
| ga-idt    | Mood     | 0.833       | 0.833 | 0.667    | 0.714|
| ga-idt    | Case     | 0.69       | 0.724 | 0.724    | 0.667|
| gl-treegal| Gender   | 0.722       | 0.685 | 0.333    | -    |
| gl-treegal| Person   | 0.522       | 0.565 | 0.522    | -    |
| gl-treegal| Number   | 0.546       | 0.546 | 0.546    | -    |
| gl-treegal| Tense    | 0.462       | 0.538 | 0.538    | 0.692|
| gl-treegal| Mood     | 0.462       | 0.692 | 0.462    | -    |
| fa-seraji | Person   | 0.667       | 0.667 | 0.381    | 0.842|
| fa-seraji | Number   | 0.514       | 0.514 | 0.514    | 0.556|
| fa-seraji | Tense    | 0.455       | 0.545 | 0.636    | 0.545|
| fa-seraji | Mood     | 0.333       | 0.667 | 0.333    | 0.0  |
| et-edt    | Person   | 0.613       | 0.613 | 0.71     | 0.714|
| et-edt    | Number   | 0.648       | 0.644 | 0.539    | 0.676|
| et-edt    | Tense    | 0.579       | 0.632 | 0.763    | 0.537|
| et-edt    | Mood     | 0.524       | 0.571 | 0.571    | 0.667|
| et-edt    | Case     | 0.565       | 0.756 | 0.786    | 0.614|
| la-perseus| Gender   | 0.692       | 0.585 | 0.538    | -    |
| la-perseus| Person   | 0.667       | 0.667 | 0.833    | -    |
| la-perseus| Number   | 0.544       | 0.662 | 0.603    | -    |
| la-perseus| Tense    | 0.75       | 1.0    | 1.0      | -    |
| la-perseus| Mood     | 0.667       | 0.667 | 0.833    | -    |
| la-perseus| Case     | 0.717       | 0.66  | 0.528    | -    |
| ug-udt    | Person   | 0.526       | 0.526 | 0.579    | 0.611|
| ug-udt    | Number   | 0.767       | 0.6   | 0.533    | 0.697|
| ug-udt    | Tense    | 0.625       | 0.75  | 0.5      | 0.778|
| ug-udt    | Mood     | 0.692       | 0.923 | 0.769    | 0.833|
| ug-udt    | Case     | 0.683       | 0.683 | 0.683    | 0.671|
| es-ancora | Gender   | 0.754       | 0.754 | 0.431    | 0.759|
| es-ancora | Person   | 0.429       | 0.429 | 0.429    | 0.526|
| es-ancora | Number   | 0.664       | 0.664 | 0.539    | 0.651|
| es-ancora | Mood     | 0.625       | 0.833 | 0.833    | 0.63 |
| es-ancora | Case     | 0.652       | 0.348 | 0.652    | 0.5  |
| de-hdt    | Gender   | 0.541       | 0.607 | 0.607    | 0.603|
| de-hdt    | Person   | 0.071       | 0.929 | 0.071    | 0.085|
| de-hdt    | Number   | 0.561       | 0.595 | 0.59     | 0.533|
| de-hdt    | Tense    | 0.8         | 0.88  | 0.88     | 0.692|
| de-hdt    | Mood     | 0.0         | 1.0   | 0.0      | 0.077|
| de-hdt    | Case     | 0.738       | 0.836 | 0.574    | 0.7  |
| kk-ktb    | Person   | 0.636       | 0.545 | 0.636    | -    |
| kk-ktb    | Number   | 0.538       | 0.615 | 0.538    | -    |
| kk-ktb    | Mood     | 1.0         | 1.0   | 0.6      | -    |
| de-gsd    | Gender   | 0.699       | 0.781 | 0.397    | 0.641|
| de-gsd    | Person   | 0.567       | 0.433 | 0.567    | 0.667|
| de-gsd    | Number   | 0.638       | 0.638 | 0.35     | 0.619|
| de-gsd    | Tense    | 0.455       | 0.636 | 0.591    | 0.526|
| de-gsd    | Mood     | 0.5         | 0.455 | 0.455    | 0.421|
| de-gsd    | Case     | 0.55        | 0.588 | 0.362    | 0.603|
| nl-alpino | Gender   | 0.667       | 0.8   | 0.8      | 0.562|
| nl-alpino | Number   | 0.548       | 0.548 | 0.565    | 0.625|
| nl-alpino | Tense    | 0.562       | 0.5   | 0.375    | 0.529|
| af-afrribooms | Number | 0.6       | 0.667 | 0.533    | 0.667|
| af-afrribooms | Tense | 0.842     | 0.842 | 0.842    | 0.588|
| af-afrribooms | Case | 0.0       | 1.0   | 0.0      | 0.0  |

Table 7: Comparing the ARM scores for SUD treebanks across both Statistical and Hard thresholding.
| TREEBANK  | FEATURE | STATISTICAL | HARD  | BASELINE | DEV  |
|----------|---------|-------------|-------|----------|------|
| uk-iu    | Gender  | 0.701       | 0.693 | 0.559    | 0.771|
| uk-iu    | Person  | 0.7         | 0.5   | 0.35     | 0.9  |
| uk-iu    | Number  | 0.647       | 0.659 | 0.479    | 0.656|
| uk-iu    | Tense   | 0.476       | 0.476 | 0.571    | 0.615|
| uk-iu    | Mood    | 0.318       | 0.409 | 0.318    | 0.357|
| uk-iu    | Case    | 0.741       | 0.741 | 0.504    | 0.732|
| cs-cltt  | Gender  | 0.857       | 0.929 | 0.75     | 0.806|
| cs-cltt  | Number  | 0.646       | 0.688 | 0.479    | 0.576|
| cs-cltt  | Tense   | 0.167       | 0.5   | 0.167    | 0.143|
| cs-cltt  | Mood    | 0.0         | 1.0   | 0.0      | 0.0  |
| cop-scriptorium | Gender   | 0.714     | 0.857 | 0.143   | 0.8  |
| cop-scriptorium | Number | 0.4   | 0.6   | 0.2     | 0.714|
| ru-taiga | Gender  | 0.648       | 0.724 | 0.638    | 0.667|
| ru-taiga | Person  | 0.667       | 0.75  | 0.583    | 0.786|
| ru-taiga | Number  | 0.662       | 0.601 | 0.459    | 0.646|
| ru-taiga | Tense   | 0.538       | 0.615 | 0.615    | 0.583|
| ru-taiga | Mood    | 0.611       | 0.667 | 0.611    | 0.5  |
| ru-taiga | Case    | 0.557       | 0.696 | 0.633    | 0.593|
| hu-szeged| Person  | 0.444       | 0.556 | 0.444    | 0.138|
| hu-szeged| Number  | 0.396       | 0.64  | 0.396    | 0.434|
| hu-szeged| Tense   | 0.6         | 0.8   | 0.8      | 0.769|
| hu-szeged| Mood    | 0.714       | 0.714 | 0.714    | 0.5  |
| sr-set   | Gender  | 0.803       | 0.817 | 0.479    | 0.622|
| sr-set   | Person  | 0.35        | 0.75  | 0.35     | 0.4  |
| sr-set   | Number  | 0.64        | 0.64  | 0.509    | 0.615|
| sr-set   | Tense   | 0.474       | 0.684 | 0.684    | 0.444|
| sr-set   | Mood    | 0.286       | 0.714 | 0.286    | 0.2  |
| sr-set   | Case    | 0.704       | 0.765 | 0.531    | 0.651|
| en-lines | Person  | 0.625       | 0.688 | 0.562    | 0.789|
| en-lines | Number  | 0.319       | 0.783 | 0.319    | 0.325|
| en-lines | Tense   | 0.704       | 0.778 | 0.704    | 0.636|
| en-lines | Mood    | 0.211       | 0.789 | 0.211    | 0.207|
| en-lines | Case    | 0.778       | 0.778 | 0.444    | 0.833|
| sk-snk   | Gender  | 0.692       | 0.776 | 0.533    | 0.638|
| sk-snk   | Person  | 0.778       | 0.333 | 0.222    | 0.625|
| sk-snk   | Number  | 0.558       | 0.558 | 0.5      | 0.571|
| sk-snk   | Tense   | 0.667       | 0.556 | 0.444    | 0.8  |
| sk-snk   | Mood    | 1.0         | 1.0   | 0.25     | 0.857|
| sk-snk   | Case    | 0.731       | 0.756 | 0.526    | 0.833|
| pl-pdb   | Gender  | 0.645       | 0.779 | 0.529    | 0.661|
| pl-pdb   | Person  | 0.556       | 0.778 | 0.704    | 0.72 |
| pl-pdb   | Number  | 0.637       | 0.613 | 0.481    | 0.644|
| pl-pdb   | Tense   | 0.5         | 0.6   | 0.7      | 0.6  |
| pl-pdb   | Mood    | 0.25        | 0.75  | 0.25     | 0.05 |
| pl-pdb   | Case    | 0.72        | 0.748 | 0.514    | 0.679|
| la-ittb  | Gender  | 0.735       | 0.725 | 0.48     | 0.805|
| la-ittb  | Person  | 0.19        | 0.81  | 0.19     | 0.273|
| la-ittb  | Number  | 0.579       | 0.579 | 0.386    | 0.562|
| la-ittb  | Tense   | 0.5         | 0.6   | 0.6      | 0.414|
| la-ittb  | Mood    | 0.476       | 0.476 | 0.571    | 0.591|
| la-ittb  | Case    | 0.757       | 0.796 | 0.495    | 0.792|

Table 8: Comparing the ARM scores for SUD treebanks across both Statistical and Hard thresholding.
| TREEBANK | FEATURE | STATISTICAL | HARD | BASELINE | DEV |
|----------|---------|-------------|------|----------|-----|
| da-ddt   | Gender  | 0.818       | 0.818| 0.364    | 0.889|
| da-ddt   | Number  | 0.667       | 0.286| 0.672    | -    |
| da-ddt   | Tense   | 0.737       | 0.842| 0.562    | -    |
| da-ddt   | Mood    | 0.2         | 0.8  | 0.2      | 0.077|
| it-postwita | Gender | 0.702       | 0.702| 0.364    | 0.674|
| it-postwita | Number | 0.744       | 0.558| 0.642    | -    |
| it-postwita | Tense  | 0.481       | 0.704| 0.607    | -    |
| eu-bdt   | Number  | 0.792       | 0.792| 0.556    | -    |
| eu-bdt   | Mood    | 0.508       | 0.415| 0.473    | -    |
| eu-bdt   | Case    | 0.421       | 0.421| 0.529    | -    |
| sl-sst   | Gender  | 0.724       | 0.618| 0.513    | -    |
| sl-sst   | Person  | 0.688       | 0.719| -        | -    |
| sl-sst   | Number  | 0.678       | 0.483| -        | -    |
| sl-sst   | Tense   | 0.48        | 0.48 | -        | -    |
| sl-sst   | Mood    | 0.56        | 0.56 | -        | -    |
| sl-sst   | Case    | 0.615       | 0.549| -        | -    |
| be-hse   | Gender  | 0.596       | 0.553| 0.404    | 0.692|
| be-hse   | Person  | 0.5         | 0.5  | 0.0      | 0.75 |
| be-hse   | Number  | 0.646       | 0.431| 0.596    | -    |
| be-hse   | Tense   | 0.429       | 0.571| 0.333    | -    |
| be-hse   | Mood    | 0.286       | 0.286| 0.2      | -    |
| be-hse   | Case    | 0.725       | 0.45 | 0.733    | -    |
| fr-sequoia | Gender | 0.8        | 0.371| 0.647    | -    |
| fr-sequoia | Person | 0.667       | 0.4  | 0.857    | -    |
| fr-sequoia | Number | 0.56        | 0.45 | 0.68     | -    |
| fr-sequoia | Tense  | 0.529       | 0.765| 0.684    | -    |
| sv-lines | Gender  | 0.538       | 0.308| 0.64     | -    |
| sv-lines | Number  | 0.643       | 0.452| 0.529    | -    |
| sv-lines | Tense   | 0.429       | 0.429| 0.655    | -    |
| sv-lines | Mood    | 0.231       | 0.231| 0.161    | -    |
| sv-lines | Case    | 0.583       | 0.25 | 0.51     | -    |
| ta-ttb   | Gender  | 0.682       | 0.659| 0.5      | -    |
| ta-ttb   | Person  | 0.091       | 0.091| 0.167    | -    |
| ta-ttb   | Number  | 0.523       | 0.545| 0.533    | -    |
| ta-ttb   | Tense   | 0.625       | 0.625| 0.667    | -    |
| ta-ttb   | Mood    | 0.5         | 1.0  | 0.5      | -    |
| it-partut | Gender | 0.846       | 0.692| 1.0      | -    |
| it-partut | Person | 0.786       | 0.25 | 0.846    | -    |
| it-partut | Number | 0.833       | 0.25 | 0.615    | -    |
| it-partut | Tense  | 0.714       | 0.286| 0.576    | -    |
| it-partut | Mood   | 0.9         | 0.6  | 0.583    | -    |
| ar-padt  | Gender  | 0.592       | 0.549| 0.712    | -    |
| ar-padt  | Person  | 0.0         | 0.0  | 0.263    | -    |
| ar-padt  | Number  | 0.512       | 0.512| 0.593    | -    |
| ar-padt  | Mood    | 0.571       | 0.571| 0.6      | -    |
| ar-padt  | Case    | 0.871       | 0.753| 0.824    | -    |
| bg-hbt   | Gender  | 0.638       | 0.404| 0.585    | -    |
| bg-hbt   | Person  | 0.625       | 0.625| 0.625    | -    |
| bg-hbt   | Number  | 0.639       | 0.533| 0.679    | -    |
| bg-hbt   | Tense   | 0.6         | 0.6  | 0.579    | -    |
| bg-hbt   | Mood    | 0.056       | 0.056| 0.176    | -    |

Table 9: Comparing the ARM scores for SUD treebanks across both Statistical and Hard thresholding.
| TREEBANK   | FEATURE  | STATISTICAL | HARD | BASELINE | DEV |
|-----------|----------|-------------|------|----------|-----|
| pt-bosque | Gender   | 0.656       | 0.721| 0.41     | 0.792|
| pt-bosque | Person   | 0.25        | 0.75 | 0.25     | 0.455|
| pt-bosque | Number   | 0.669       | 0.669| 0.378    | 0.698|
| pt-bosque | Tense    | 0.5         | 0.438| 0.438    | 0.692|
| lt-alksnis| Gender   | 0.711       | 0.711| 0.41     | 0.671|
| lt-alksnis| Person   | 0.667       | 0.8   | 0.667    | 0.571|
| lt-alksnis| Number   | 0.625       | 0.625| 0.531    | 0.595|
| lt-alksnis| Tense    | 0.667       | 0.667| 0.667    | 0.6 |
| lt-alksnis| Mood     | 0.667       | 0.333| 0.667    | 0.375|
| ar-nyuad  | Gender   | 0.606       | 0.718| 0.718    | 0.536|
| ar-nyuad  | Person   | 0.469       | 0.562| 0.469    | 0.343|
| ar-nyuad  | Number   | 0.502       | 0.554| 0.502    | 0.468|
| ar-nyuad  | Mood     | 0.438       | 0.562| 0.438    | 0.5 |
| ca-ancora | Gender   | 0.804       | 0.786| 0.464    | 0.77 |
| ca-ancora | Person   | 0.389       | 0.611| 0.389    | 0.219|
| ca-ancora | Number   | 0.652       | 0.652| 0.511    | 0.616|
| ca-ancora | Tense    | 0.5         | 0.731| 0.692    | 0.56 |
| grc-proiel| Gender   | 0.605       | 0.516| 0.535    | 0.588|
| grc-proiel| Person   | 0.543       | 0.543| 0.6      | 0.737|
| grc-proiel| Number   | 0.533       | 0.61  | 0.538    | 0.585|
| grc-proiel| Tense    | 0.643       | 0.786| 0.786    | 0.774|
| it-twittiro| Gender    | 0.808      | 0.808| 0.385    | 0.65 |
| it-twittiro| Person   | 0.591       | 0.318| 0.682    | 0.579|
| it-twittiro| Number   | 0.568       | 0.568| 0.419    | 0.634|
| it-twittiro| Tense    | 0.25        | 0.75 | 0.75     | 0.462|
| it-twittiro| Mood     | 0.627       | 0.747| 0.747    | 0.551|
| mr-ufal   | Gender   | 0.609       | 0.652| 0.565    | 0.52 |
| mr-ufal   | Person   | 0.727       | 0.727| 0.364    | 0.889|
| mr-ufal   | Number   | 0.394       | 0.794| 0.242    | 0.514|
| mr-ufal   | Tense    | 0.5         | 0.731| 0.692    | 0.56 |
| hr-set    | Gender   | 0.725       | 0.717| 0.525    | 0.643|
| hr-set    | Person   | 0.769       | 0.769| 0.577    | 0.692|
| hr-set    | Number   | 0.675       | 0.675| 0.51     | 0.658|
| hr-set    | Tense    | 0.429       | 0.714| 0.714    | 0.542|
| hr-set    | Mood     | 0.412       | 0.588| 0.412    | 0.158|
| hr-set    | Case     | 0.669       | 0.725| 0.577    | 0.659|
| kmr-mg    | Gender   | 1.0         | 0.818| 1.0      | -   |
| kmr-mg    | Number   | 0.783       | 0.739| 0.783    | -   |
| kmr-mg    | Case     | 0.909       | 0.727| 0.909    | -   |
| nl-lassysmall | Gender | 0.85     | 0.85 | 0.9     | 0.81 |
| nl-lassysmall | Number | 0.646   | 0.646| 0.523    | 0.646|
| nl-lassysmall | Tense   | 0.6       | 0.6   | 0.4     | 0.364|
| fr-gsd    | Gender   | 0.727       | 0.727| 0.485    | 0.807|
| fr-gsd    | Person   | 0.375       | 0.719| 0.375    | 0.312|
| fr-gsd    | Number   | 0.624       | 0.624| 0.441    | 0.593|
| fr-gsd    | Tense    | 0.706       | 0.706| 0.765    | 0.81 |
| fr-gsd    | Mood     | 0.273       | 0.727| 0.273    | 0.25 |

Table 10: Comparing the ARM scores for SUD treebanks across both Statistical and Hard thresholding.
Table 11: Comparing the ARM scores for SUD treebanks across both Statistical and Hard thresholding.