AUTOLEX: An Automatic Framework for Linguistic Exploration

Anonymous ACL submission

Abstract

Each language has its own complex systems of word, phrase, and sentence construction, the guiding principles of which are often summarized in grammar descriptions for the consumption of linguists or language learners. However, manual creation of such descriptions is a fraught process, as creating descriptions which describe the language in "its own terms" without bias or error requires both a deep understanding of the language at hand and linguistics as a whole. We propose an automatic framework AUTOLEX that aims to ease linguists' discovery and extraction of concise descriptions of linguistic phenomena. Specifically, we apply this framework to extract descriptions for three phenomena: *morphological agreement, case marking,* and *word order*, across several languages. We evaluate the descriptions with the help of language experts and propose a method for automated evaluation when human evaluation is infeasible.[1]

1 Introduction

Languages are amazingly diverse, consisting of different systems for word formation (*morphology*), phrase construction (*syntax*), and meaning (*semantics*). These systems are governed by a set of guiding principles, referred to as *grammar*. Creating a human-readable description that highlights salient points of a language is one of the major endeavors undertaken by linguists. Such descriptions form an indispensable component of language documentation efforts, particularly for endangered or threatened languages (Himmelmann, 1998; Hale et al., 1992; Moseley, 2010). Furthermore, if descriptions can be created in a machine-readable format they can be used for developing language technologies (Pratapa et al., 2021).

Linguists and researchers have undertaken initiatives to collect linguistic properties in a machine-readable format across several languages, WALS (Dryer and Haspelmath, 2013) being a standing example. For instance, WALS can tell us that English objects occur after verbs, or that Turkish pronouns have symmetrical case. However, because WALS presents these properties across many diverse languages, these properties are necessarily defined at a coarse-grained level and cannot capture language-specific nuances. WALS does not inform us of any exceptions to its general rules (e.g. the cases when English objects come before verbs), and there are many aspects that are not even covered (e.g. when a Turkish pronoun takes the accusative marker and when the nominative). There are other challenges to creating detailed descriptions, as for many of the 6,500+ languages, there are few or no formally trained linguists. Even in the ideal case where there is such a linguist, there are a plethora of linguistic phenomena to be covered, and it is hard to enumerate every single one through introspection.

Thanks to the NLP advances, it is now possible to automate some *local* aspects of linguistic analysis such as POS tagging (Toutanvoa and Manning, 2000), dependency parsing (Kiperwasser and Goldberg, 2016) or morphological analysis (Malaviya et al., 2018). Recent advances in transfer learning have shown that this is possible to an extent, even for under-resourced languages (Kondratyuk and Straka, 2019). A small amount of prior work has proposed methods for answering specific questions about language, such as the analysis of word order (Östling, 2015; Wang and Eisner, 2017) and morphological agreement (Chaudhary et al., 2020), or grammar extraction from inter-linear glosses (Bender et al., 2002) (Table 2 in Appendix A compares the questions answered by our and related work).

In this work, we propose AUTOLEX, an automatic framework to aid linguistic exploration and description, with the goal of helping linguists de-

[1] Code and data are released on https://github.com/emnlp-autolex/autolex. Currently, the online web site (https://emnlp-autolex.github.io/autolex) shows all rules for some languages, we are working on adding the rest.

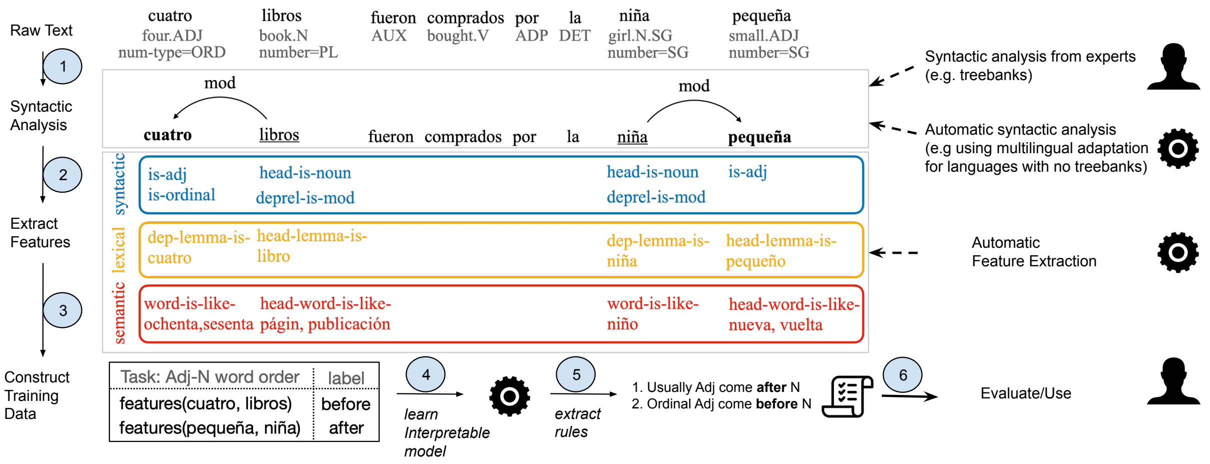


Figure 1: An overview of the AUTOLEX framework (with Adj-N order in Spanish as an example). The example sentence translates to *Four books were bought by the small girl*. First, we formulate a linguistic question (e.g. regarding Adj-N order) as a binary classification task (e.g. "whether the Adj comes before/after the N'). Next, we perform syntactic analysis on the raw text, from which we extract syntactic, lexical, and semantic features to construct the training data. Finally, we learn an interpretable model from which we extract concise rules.

velop fine-grained understanding of different linguistic phenomena. The framework allows the linguist to ask a question such as "what are the rules of object-verb order?", or "when do pronouns take the accusative case in Turkish?", and automatically acquire first-pass answers. AUTOLEX analyses the texts in the corresponding languages and finds answers such as in English "typical declarative constructions show VO, but interrogative sentences can show OV", or in Turkish "object pronouns take the accusative case." Specifically, we follow a multi-step process, as shown in Figure 1. First, we define the linguistic question as a classification task (e.g. "does the adjective come before the noun or not"; § 2). Second, we automatically extract syntactic, semantic, and surface-level features that may be predictive of the answer to this question (§ 3). Next, we construct the training data and train an interpretable classifier such as a decision tree to identify the underlying patterns that answer this question. Finally, we extract and visualize interpretable rules (§ 4). This methodology is inspired by previous work on discovering fine-grained distinctions for individual phenomena (Wang and Eisner, 2017; Chaudhary et al., 2020), but is significantly more general in that we demonstrate its ability to discover interesting features for word order, case marking, and morphological agreement.

We experiment with 61 languages for which we design an automated evaluation protocol which informs us how successful our framework is in discovering valid grammar rules (§ 7.1). We further conduct a user study with linguists to evaluate how correct, readable, and novel the rules are perceived to be (§ 7.2). Finally, we apply this framework to a threatened language variety, Hmong Daw (`mww`), and evaluate how well our framework extracts rules under zero-resource conditions (§ 8).

2 Formalizing Linguistic Questions

The first step in applying AUTOLEX to answer a question is to determine whether we can formulate it as a classification task, with training data $\{\langle \mathbf{x_1}, y_1\rangle, \langle \mathbf{x_2}, y_2\rangle \cdots, \langle \mathbf{x_n}, y_n\rangle\}$, where $\mathbf{x_i} \in X$ are the input features and $y_i \in Y$ are the labels indicating the linguistic phenomenon. Below, we describe how we define Y for each phenomena, and discuss how to construct X in the following section. We use the UD schema (McDonald et al., 2013) for representing the syntax and morphology.

Case Marking (CM) is a system of "marking syntactic dependents for the type of grammatical relation (subject, object, etc.) they bear to their heads" (Blake, 2009). Although there are different theories on how to formalize CM, we commit to the viewpoint that there are two types of cases: *abstract* and *morphological*, where the former is a universal property and the latter is its overt realization (Chomsky, 1993; Halle et al., 1993). Thus, we formulate the explanation of CM as determining *when a word class (e.g. nouns) marks a particular case (e.g. nominative, etc.)*. Formally, for each POS tag t we learn a separate model, where the input examples x_i are the words having POS tag t with the case feature marked (e.g.

Case=Nominative). The model is trained to predict an output label ($y_i \in Y$), where Y is the label set of all observed case values for that language.

Word Order (WO) describes the relative position of the syntactic elements (Dryer., 2007), and is one of the major axes of linguistic description appearing in grammar sketches or databases such as WALS. We consider the following five WALS relations R: subject-verb (82A), object-verb (83A), adjective-noun (87A), adposition-noun (85A) and numeral-noun (89A). In contrast to WALS, which only provides a single canonical order for the entire language, we pose the linguistic question as determining *when does one word in such a relation appear before or after the other*. Formally, the pair of words involved in the syntactic relation $\langle w_i^a, w_i^b \rangle \in r$ form the input example x_i and the output label $y_i \in Y$ where $Y = \{\text{before, after}\}$.

Agreement (AM) is the process where one word or morpheme selects a morphological form that agrees with that of another word/phrase in the sentence (Corbett, 2003). We follow a similar problem formulation as Chaudhary et al. (2020), which asks the question *when is agreement required between a head (w_h) and its dependent (w_d) for a morphological attribute m*. We focus on the morphological attributes $M = \{\text{gender, person, number}\}$, which more often show agreement than other attributes (Corbett, 2009), and train a separate model for each. The pair of head-dependent words which both mark the morphological property m form the input example x_i and the output labels (y_i) are binary denoting if agreement is observed or not between the pair.

3 Feature Extraction

Now that we have provided three examples of converting linguistic questions into classification tasks, we design features to help predict each question's answer. We use linguistic knowledge to design features, but the feature extraction process itself is automatic. For a different question or language, a linguist can begin the process by using these initial features or even design new features as they deem fit. In step-2 of Figure 1, we demonstrate example features extracted from a Spanish sentence for training the adjective-noun WO model. We refer to the words participating in an input x_i as *focus words*. These include the words describing the relation itself (e.g. the adjective *cuatro* and its noun *libros*) and also their respective heads and dependents.

Syntactic Features Prior work (Blake, 2009; Kittilä et al., 2011; Corbett, 2003) has discussed the role of syntax and morphology being important for determining the case and agreement. In Figure 1, we show a subset of features extracted for some of the focus words. For example, for the adjective, we derive features from its POS tag (e.g. "is-adj"), all of its morphological tags (e.g. "is-ordinal") and the dependency relation it is involved in (e.g. "deprel-is-mod"). We extract similar features for the adjective's head, which is *libros* (e.g. "head-is-noun").

Lexical Features An influential family of linguistic theories such as lexical functional grammar (Kaplan et al., 1981), head-driven phrase structure grammar (Pollard and Sag, 1994), places most of the explanatory weight for morphosyntax on the lexicon – the properties of the head word (and other words) drive the realization of the rest of the phrase or sentence. Therefore, we add the lemma for the focus words (e.g. "dep-lemma-is-cuatro, head-lemma-is-libro") as features.

Semantic Features There is a strong interaction between semantics and sentence structure. Some well-known examples are of *animacy* or semantic class of a word determining CM (Dahl and Fraurud, 1996) and WO (Thuilier et al., 2021) for some languages. Continuous vectors (Mikolov et al., 2013; Bojanowski et al., 2017a) have been used to capture semantic (and syntactic) similarity across words. However, most vectors are high-dimensional and not easily interpretable, i.e. what semantic/syntactic property each individual vector value represents is not obvious. Since our primary goal is to extract comprehensible descriptions of linguistic phenomena, we first generate sparse non-negative vectors using Subramanian et al. (2018). For each dimension, we extract the top-k words having a high positive value, resulting in features like dim-1={radio,nuclear}, dim-2={hotel,restaurante}. This helps us interpret what property each dimension is capturing, for example, dim-1 refers to words about nuclear technology, while dim-2 refers to accommodations. Now that we can interpret what each feature (dimension) corresponds to, we directly add these vector as features. In Figure 1, a semantic feature (e.g. "dep-word-is-like={ochenta,sesenta}"[2]) extracted for **cuatro** informs us that the adjective denotes a numeric quantity.

[2] This translates to {eight, sixty}

4 Learning and Extracting Rules

Training Data To construct the training data D^p_{train} for each task p, we start with the raw text D of the language in question and perform syntactic analysis, producing POS tags, lemmas, morphological analysis and dependency trees for each sentence. Using this analysis, we then identify the focus word(s) and extract features, forming the input example ($\mathbf{x_i} = \{x_i^0, x_i^1, \cdots, x_i^k\}$).

Model Training Given that the learned model must be interpretable to linguists using the system, we opt to use decision trees (Quinlan, 1986), which split the data into leaves, where each leaf corresponds to a portion of the input examples following common syntactic/semantic/lexical patterns.

Rule Extraction Each leaf in the decision tree is assigned a label based on the distribution of examples within that leaf. For instance, if a leaf of the adjective-noun WO decision tree has 60% of examples with adjectives before their nouns, the leaf is labeled as *before*. However, a majority-based threshold alone is insufficient as it does not account for leaves with very few examples, which may be based on spurious correlations or nonsensical feature divisions. Instead, we use a statistical threshold for leaf labeling, inspired by Chaudhary et al. (2020), performing a chi-squared test to first determine which leaves differ significantly from the base distribution. For this, we first define the null H_0 and test H_1 hypotheses. For instance, for WO we can define that a leaf:

H_0 : takes either *before/after* label

H_1 : takes the label dominant under that leaf

We can design such H_0 as the words participating in the relation can either be *before* or *after* the other. To apply the chi-squared test, we compute the expected probability distribution for H_0 considering a uniform distribution.We then compute the p-value and leaves which are not statistically significant are assigned the label of *cannot decide*, which informs a user that the model was uncertain about the label (details in Appendix B). Leaves that pass this test are then assigned the majority label and correspond to a rule that will be shown to linguists, where the "rule" is described by the syntactic/semantic/lexical features on the branch that lead to that leaf.

Rule Visualization For each rule, we extract illustrative examples from the underlying corpus and visualize them in an interface (Figure 2). We select such examples that are both short and consist of diverse word forms to illustrate the rule usage in different contexts. Along with examples which follow a rule, we also show examples which do not follow the rule, giving a softer, more nuanced view of the data (details in Appendix B). Specifically, to not overwhelm the user, we only present 10 examples for each type.

5 Automated Evaluation Protocol

In the next two sections, we devise protocols for evaluation of the extracted rules using both automatic metrics (for rapid evaluation that can be applied widely across languages), and evaluation by human language experts (as our gold-standard evaluation). We first describe below the process of automatic evaluation per linguistic phenomenon.

Case Marking As noted earlier, we use the UD scheme for deriving the training data. Under this scheme, not every word is labeled with *case*, restricting our training and evaluation to be only on such labeled examples. For such words, we consider *case* to be a universal property i.e. each word marks a particular *case* value and, we evaluate whether our model can correctly predict that value. Thus, we measure the accuracy on a test example $\langle \mathbf{x_i}, y_i \rangle \in D^t_{\text{test}}$, comparing the models prediction $\hat{y}_i$ with the observed case value y_i. We compare our model against a frequency-based baseline which assigns the most frequent case value in the training data to all input examples.

Word Order Similar to CM, we can assume that every input example has a word order value, for example subjects will occur either *before* or *after* the verbs. Therefore, for an input example, we consider the observed order to be the ground truth and compute the accuracy by comparing it with the model's prediction. We compare against a frequency-baseline where the most frequent WO value is assigned to all input examples.

Comparing the model's prediction with the observed order is reasonable for languages which have a dominant WO. There are a considerable set of languages which have a freer WO. WALS labels such relations as "no dominant order" (e.g. subject-verb order for Modern Greek). For such cases, considering accuracy alone might be insufficient as there is no ground truth. Therefore, we also

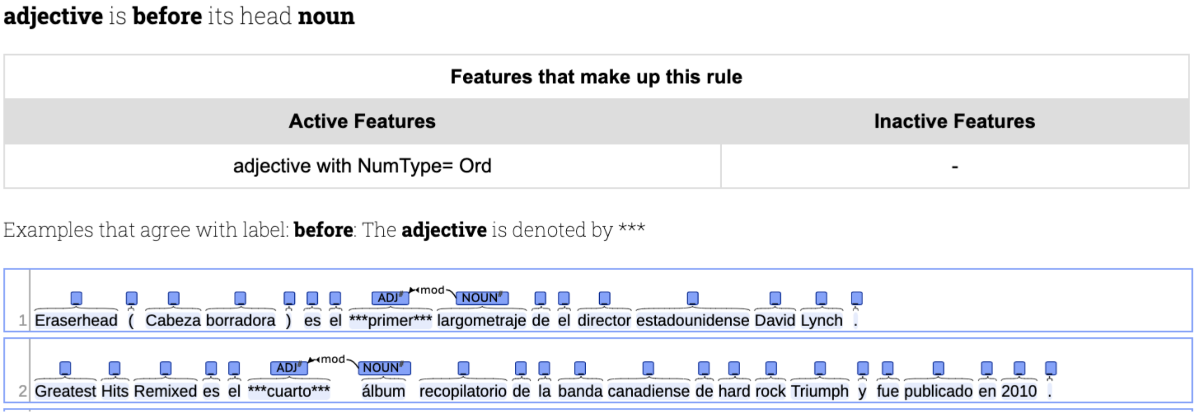


Figure 2: A rule extracted for Spanish adjective-noun word order.

report the entropy over the predicted distribution:

$$H^r_{\text{wo}} = - \sum_{k=\text{before, after}} p_k \log p_k$$

$$p_k = \frac{\sum_{\langle \mathbf{x}^r_i, y_i \rangle \in D^r_{\text{test}}} \mathbb{1} \begin{cases} 1 & \hat{y}_i = k \\ 0 & \text{otherwise} \end{cases}}{|D^r_{\text{test}}|}$$

For languages with no dominant order, the model should be uncertain about the predicted order and we expect the model's entropy to be high. The accuracy computed against the observed order is still useful, as despite there being "no dominant order", speakers tend to prefer one order over the other. A high accuracy would entail that the model was successful in capturing this "preferred order."

Agreement We use the automated rule metric (ARM) proposed by Chaudhary et al. (2020) which computes accuracy by comparing the ground truth label to the predicted label. The ground truth label of an example is decided using a predefined threshold on the leaf to which the example belongs. ARM does not use the observed agreement between the head and its dependent as ground truth because an observed agreement might not necessarily mean *required* agreement. We compare with Chaudhary et al. (2020), which uses simple syntactic features such as POS of the head, the dependent and, the dependency relation between them.

6 Human Expert Evaluation Protocol

Since our primary objective is to extract rules which are human-readable and of assistance to the linguists, we enlist the help of language experts to evaluate the rules on three parameters: *correctness*, *prior knowledge*, *feature correctness*. Before starting with the actual evaluation, we first ask the expert to provide answers regarding the linguistic questions we are evaluating. For example, we ask questions such as "when are subjects after verbs in Greek", and they are required to provide a brief answer (e.g. "for questions or when giving emphasis to a subject"). We then direct them to our interface where we show the extracted features and a few illustrative examples for the rule, then ask questions regarding each of the three parameters (as shown in Figure 6 in the Appendix).

Regarding *correctness*, the expert is asked to annotate whether the illustrative examples, shown for that rule, are governed by some underlying grammar rule. If so, they are then required to judge how precise it is. Consider some rules extracted for Spanish adjective-noun order in Table 1. Looking at the examples and features for the Type-1 rule, it is evident that this rule *precisely defines the linguistic distinction*.[3] Some rules, although valid, may be too general (Type-3) or too specific (Type-4). Finally, a rule *may not correspond to any underlying grammar rule*, like the Type-5 where the model simply discovered a spurious correlation in the data. For *prior knowledge*, if an extracted rule was indeed a valid grammar rule, then we ask the expert whether they were aware of such a rule. This will inform us how useful our framework is in discovering rules which a) align with the expert's prior knowledge and, b) are novel i.e. rules which the expert were not aware of apriori. Finally, for *feature correctness*, we ask whether the features selected by the model accurately describe said rule. For the Type-1 rule, the answer would be *yes*. But for rules like Type-2, the features are not informative even though the corresponding examples do

[3] https://www.thoughtco.com/ordinal-numbers-in-spanish-3079591

Type	Rule Features	Examples	Label
Type-1 (valid)	Adj is a Ordinal	También se utilizaba en las **primeras** grabaciones y arreglos jazzísticos. *It was also used in **early** jazz recordings and arrangements.* Las **primeras** 24 horas son cruciales. *The **first** 24 hours are crucial.*	Before
Type-2 (valid, not informative)	Adj belongs to group: con,como,no,más,lo	Matisyahu piensa editar pronto un **nuevo** disco grabado en estudio. *Matisyahu plans to release a **new** studio-recorded album soon.* Es una experiencia **nueva** estar desempleado. *It's a **new** experience being unemployed*	Before
Type-3 (valid, too general)	Adj is NOT Ordinal	Además de una **gran** variedad de aplicaciones *In addition to a **great** variety of applications.* Una unión **solemnizada** en un país extranjero *An union **solemnized** in a foreign country*	After
Type-4 (valid, too specific)	Adj's lemma is numeroso	En África hay **numerosas** lenguas tonales *In Africa there are **numerous** tonal languages* Ellas poseen **varios** libros *They own **several** books*	Before
Type-5 (invalid)	Adj's head noun is a conjunct	Las consecuencias de cualquier (colapso) de divisa e inflación **masiva** . *The consequences expected from any currency collapse and **massive** inflation.* (Realizan) trabajos de alta calidad , muy **buenos** profesionales *They do high quality work, very **good** professionals*	After

Table 1: Types of rules discovered by the model for Spanish adjective-noun word order. **Adjectives** are highlighted and the nouns they modify are underlined. Illustrative examples under each rule are also shown with their English translation in italics. Label denotes the predicted order.

follow a common pattern.

7 Gold-standard Analysis Experiments

In this section, we present results to demonstrate that our framework can discover the conditions which govern the different linguistic phenomena. Specifically, we experiment with gold-standard syntactic analysis derived from SUD treebanks, and run experiments to answer questions about word order, agreement, and case marking (§ 7.1). Furthermore, we manually verify a subset of these extracted rules (§ 7.2). Experimenting with languages that have been already studied and have annotated treebanks is crucial for verifying the efficacy of our approach before applying it to other true low- or zero-resource languages. Under this setting we not only have clean and expert-annotated data, but we can also quickly compare the effect of data size on the system performance as different languages have treebanks of varying size.

Data and Model We use the Syntactic Universal Dependencies v2.5 (SUD) (Gerdes et al., 2019) treebanks which are based on the Universal Dependencies (UD) (Nivre et al., 2016, 2018) project, the difference being that the former allows function words to be syntactic heads (as opposed to UD's preference for content words), which is more conducive to our goal of learning grammar rules. We experiment with treebanks for 61 languages, which are publicly available with annotations for POS tags, lemmas, dependency parses, and morphological analysis. We use the standard SUD train, validation and test splits. Syntactic and lexical features are directly extracted from these gold syntactic analyses. Semantic features are derived from continuous word vectors: we start with 300-dim pre-trained fasttext word vectors (Bojanowski et al., 2017b) which are transformed into sparse vectors using Subramanian et al. (2018)[4]. Last, we use the `XGBoost` (Chen and Guestrin, 2016) library to learn the decision tree. Further details on the model setup are discussed in Appendix C.

7.1 Automated Evaluation Results

We train models using syntactic features for all languages covered by SUD, wherever the linguistic question is applicable. We find that our models outperform the respective baselines by an (avg.) accuracy of +7.3 for word order, +28.1 for case marking, and +4.0 ARM for agreement.[5] We also report the result breakdown under three resource settings, low, mid, and high, where low-resource refers to the treebanks with <500 sentences, mid-resource has 500–5000 sentences and high-resource has >5000 sentences. Across all three linguistic phenomena, the (avg.) model gains over the baseline are +3.19 for the low-resource, +10.7 for the mid-resource

[4] https://github.com/harsh19/SPINE

[5] We also experimented with Random forests (RF), as suggested by anonymous reviewers, but found the decision trees (DT) to be slightly underperforming ((avg.) -0.12 acc). But given that it is straightforward to extract interpretable rules from DT, which is our primary goal, as compared to RF, we use the former for all experiments, details in Appendix D.

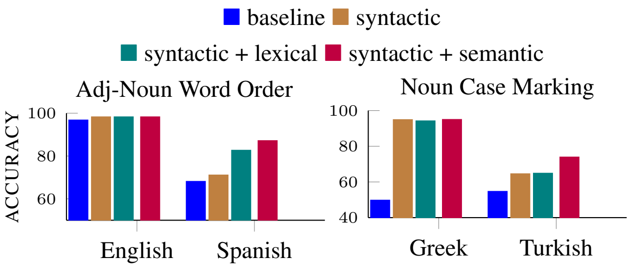


Figure 3: Comparing the effect of different features on the word order and case marking.

and +12.8 for the high-resource. The larger the treebank size, the larger the improvement of our model's performance over the baseline. Even in low-resource settings, a gain over the baseline suggests that our approach is extracting valid rules, which is encouraging for language documentation efforts. We present the result breakdown of individual relations in Appendix (Table 3).

As motivated in § 3, the conditions which govern a linguistic phenomenon vary considerably across languages, which is also reflected in our model's performance. For example, the model trained on syntactic features alone is sufficient to reach a high accuracy (avg94.2%) for predicting the adjective-noun order in Germanic languages. But for Romance languages, using only syntactic features leads to much lower performance (avg74.6%). We experiment with different features and report results for a subset of languages in Figure 3. Observe that for Spanish adjective-noun order adding lexical features improves the performance significantly (+11.57) over syntactic features, and semantic features provide an additional gain of +4.48. Studying the languages marked as having "no dominant order" in WALS, we find our model does show a higher entropy. SUD contains 8 such languages for subject-verb order, and our model produces an (avg.) entropy of 1.09, as opposed to (avg.) 0.75 entropy for all other languages. For noun case marking in Greek, syntactic features already bring the model performance to 94%. For Turkish, the addition of semantic features raises the model performance by +9.38. The model now precisely captures that nouns for locations like ev, oda, kapı, dünya[6] typically take the locative case. This is in-line with Bamyacı and von Heusinger (2016) which outlines the importance of animacy in Turkish differential case marking.

To confirm that these *discovered conditions generalize to the language as a whole and not the specific dataset on which it was trained*, we train a model on one treebank of a language and apply the trained model directly on the test portions of other treebanks of the same language. There are 30 languages in the SUD which fit this requirement. Figure 4 in the Appendix demonstrates one such setting for understanding the word order patterns across different French corpora, where the models have been trained on the largest treebank (`fr-gsd`). For subject-verb order, all treebanks except the `fr-fqb` show similar high test performance (>90% acc.). Interestingly, the model severely underperforms (28% acc.) on `fr-fqb` which is a question-bank corpus comprising of only questions, and questions in French can have varying word order patterns.[7] The model fails to correctly predict the word order because in the training treebank only 1.7% of examples are questions making it challenging for the model to learn word order rules for different question types.

Through this tool, a linguist can potentially inspect and derive insights on how the patterns discovered for a linguistic question vary across different settings, both within a language and across different languages as well.

7.2 Human Evaluation Results

Through the above experiments, we *automatically* evaluated that the extracted rules are predictive (to some extent) and applicable to the language in general. Before applying this framework on an endangered language we first perform a manual evaluation ourselves for English and Greek. We select these languages based on the availability of human annotators, using one expert each for English and Greek. First, we note that the total number of rules for English (29) are much less than that for Greek (161), the latter being more morphologically rich. We find that 80% of the rules (across all phenomena) are valid grammar rules for both languages. A significant portion (40%) of the valid rules are either too specific or too general, which highlights that there is scope of improvement in the feature and/or model design. Interestingly, even for English, there were 7 rules which the expert was not aware of. For example, the following rule for adjective-noun order – "when the nominal is a

[6] house, room, door, world

[7] In questions such as *Que signifie l'* ***acronyme*** *NASA?* ("What does the acronym NASA mean?"), the verb comes before its **subject**, but for questions such as ***Qui*** *produit le logiciel ?* ("Who produces the software?") the **subject** is before the verb.

word like *something,nothing,anything*, the adjective can come after the noun.". For Greek, almost all valid rules were known to the expert, except for one Gender agreement rule[8]. Regarding feature correctness, the Greek expert found 69% of the valid rules to be readable and informative, while the English expert found 58% of such rules. We show the individual results in Appendix (Figure 5).

8 Hmong Daw Study

Finally, to test the applicability of AUTOLEX in a language documentation situation, we experiment with Hmong Daw (mww), a threatened language variety, spoken by roughly 1M people across US, China, Laos, Vietnam and Thailand. It certainly can be categorized as a low-resourced language with respect to computational resources as well as accessible and detailed machine-readable grammatical descriptions. Furthermore, this study presents a realistic setting of language analysis as there is no expert-annotated syntactic analysis available.

We had access to 445k Hmong sentences, which were collected from the `soc.culture.hmong` Usenet group. Since the data was scraped from the web, it was noisy and intermixed with English. Therefore, first we automatically clean the corpus using a character-level language model trained on English. This automatically filtered 61k sentences. Next, we automatically obtain syntactic analyses, for which we train `Udify` (Kondratyuk and Straka, 2019), a multilingual automatic parser that jointly predicts POS tags, lemmas, morphological analysis and dependency parses, on Vietnamese, Chinese and English treebanks and apply it to the Hmong text. We randomly split the parsed data into a train and test set (80:20) and apply our general framework to extract rules (details in Appendix E).

Results Hmong has no inflectional morphology so we only train the model to answer word order questions. We conduct the expert evaluation on four relations where our model outperforms the baseline, albeit slightly (+4.08 for Adj-N, +0.12 for Subj-V, +0.52 for Adp-N, +0.72 for Num-N). For Obj-V relation, our model is on par with the baseline which could indicate that either there were not many examples whose word order deviated from the dominant order or the model needs improvement. First, we ask the expert, a linguist who studies Hmong, to describe the rules (if any) for each relation. Comparing with the expert's provided rules, we find that the model is successful in discovering the dominant pattern for all relations. However, of the 30 rules (across all relations) presented to the expert for annotation, only 5 rules (1 rule for subject-verb, 4 rules for numeral-noun) were found to precisely describe the linguistic distinction. For instance, according to the expert, numerals cannot occur immediately before nouns, rather they occur before classifiers which then occur before nouns ("1 clf-1 noun-1"). Interestingly, one rule captured examples where the numerals were occurring immediately before nouns without the classifiers (e.g. "1 noun-1, 2 noun-2"), which the expert was not aware of. On one hand, this is promising as the model, despite being trained on noisy sentences and syntactic analyses, was able to discover instances of interesting linguistic behavior. However, the expert noted that a large portion of the rules were difficult to evaluate as these referred to examples which were incorrectly parsed, some of which even described the English portion of code-mixed data.

Despite showing the promise of automatically obtaining detailed descriptions on languages with good syntactic analyzers, we can see that it is still challenging to apply methods to such under-resourced languages. This poses a new challenge for zero-shot parsing, even the relatively strong model of Kondratyuk and Straka (2019) resulted in a high enough error rate that it impacted the effectiveness of our method, and methods with higher accuracy may further improve the results of end-to-end generation of grammar descriptions.

9 Next Steps

While we have demonstrated that our automatic framework can answer linguistic questions across different languages, the rules we discover are limited by the SUD annotation decisions. For example, several nouns in German are not annotated for the default case, which means these nouns get ignored by our model in the current setting. Possibly, using language-specific annotations or heuristics could help alleviate this problem. As noted in the Hmong study, the quality of rules depends on the quality of the underlying parses. We plan to devise an iterative process where a linguist, assisted by an automatic parser, can improve syntactic parsing. The model extracts rules using improved analyses, which the linguist can inspect and provide more inputs to further improve.

[8]The rule was, "proper-nouns modifiers do not need to necessarily agree with their head nouns".

Statement of Ethics

We acknowledge that there are several ethical concerns while working with endangered or threatened languages, in particular, that we include and take guidance from community members when designing any technology using their data. Secondly, that any data we collect is appropriately used, without causing any detrimental effect or bias on the community. In adherence to that, this work is done in collaboration with a Hmong linguist who is in close collaboration and consultation with the community. We will release any tools that we build for Hmong in consultation with them and the community.

References

Elif Bamyacı and Klaus von Heusinger. 2016. Animacy effects on differential object marking in turkish. *Poster presentation at linguistic evidence*.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *COLING-02: Grammar Engineering and Evaluation*.

Emily M. Bender, Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 74–83, Sofia, Bulgaria. Association for Computational Linguistics.

Barry J Blake. 2009. History of the research on case. In *The Oxford handbook of case*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017a. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017b. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. In *COLING-02: Grammar Engineering and Evaluation*.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5212–5236, Online. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Noam Chomsky. 1993. *Lectures on government and binding: The Pisa lectures*. Walter de Gruyter.

Greville G Corbett. 2003. Agreement: Terms and Boundaries. In *Texas Linguistic Society Conference*.

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

Osten Dahl and Kari Fraurud. 1996. Animacy in grammar and discourse. *Pragmatics and Beyond New Series*.

Matthew S. Dryer. 2007. Word order. In *Language Typology and Syntactic Description*.

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

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.

Ken Hale, Michael Krauss, Lucille J Watahomigie, Akira Y Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C England. 1992. Endangered Languages. *Language*.

Morris Halle, Alec Marantz, Kenneth Hale, and Samuel Jay Keyser. 1993. Distributed morphology and the pieces of inflection. *The view from Building 20*.

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.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*.

Kristen Howell, Emily M. Bender, Michel Lockwood, Fei Xia, and Olga Zamaraeva. 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, Honolulu. Association for Computational Linguistics.

Ronald M Kaplan, Joan Bresnan, et al. 1981. *Lexical-functional grammar: A formal system for grammatical representation*. Citeseer.

Tracy Holloway King, Martin Forst, Jonas Kuhn, and Miriam Butt. 2005. The feature space in parallel grammar writing. *Research on Language and Computation*, 3(2-3):139–163.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Seppo Kittilä, Katja Västi, and Jussi Ylikoski. 2011. Introduction to case, animacy and semantic roles. *John Benjamins Publishing*.

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.

Haley Lepp, Olga Zamaraeva, and Emily M. Bender. 2019. Visualizing inferred morphotactic systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 127–131, Minneapolis, Minnesota. Association for Computational Linguistics.

William D. Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Chaitanya Malaviya, Matthew R. Gormley, and Graham Neubig. 2018. Neural factor graph models for cross-lingual morphological tagging. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2663, Melbourne, Australia. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. UNESCO.

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.

Robert Östling. 2015. Word Order Typology through Multilingual Word Alignment. In *ACL*.

Carl Pollard and Ivan A Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

Adithya Pratapa, Antonios Anastasopoulos, Shruti Rijhwani, Aditi Chaudhary, David R. Mortensen, Graham Neubig, and Yulia Tsvetkov. 2021. Evaluating the morphosyntactic well-formedness of generated texts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7131–7150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2018. Spine: Sparse interpretable neural embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Juliette Thuilier, Margaret Grant, Benoît Crabbé, and Anne Abeillé. 2021. Word order in french: the role of animacy. *Glossa: a journal of general linguistics*, 6(1).

Kristina Toutanvoa and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 63–70, Hong Kong, China. Association for Computational Linguistics.

Dingquan Wang and Jason Eisner. 2017. Fine-grained prediction of syntactic typology: Discovering latent structure with supervised learning. *Transactions of the Association for Computational Linguistics*, 5:147–161.

A Related Work

Prior work (Lewis and Xia, 2008; Hellan, 2010; Bender et al., 2013; Howell et al., 2017) have proposed methods to map descriptive grammars, present in the form of inter-linear glossed text (IGT), to existing head-phrase structure grammar (HPSG) based grammar system which is machine-readable. Lewis and Xia (2008) enrich IGT data with syntactic structures to determine canonical word order and case marking observed in the language. They do note that, while a linguist carefully chooses the examples to create the IGT corpus such that they are representative of the linguistic phenomena of interest, insights derived from IGT may suffer from this bias as the data doesn't encompass many of the naturally-occurring examples. Hellan (2010) present a sentence-level annotation code which maps the properties of the sentence to discrete labels. These discrete labels form a template which are then mapped to in a mixed to HPSG or LFG format (Pollard and Sag, 1994; Kaplan et al., 1981). Bender et al. (2013) extract major-constituent word order and case marking properties from the IGT for a diverse set of languages. Potentially, grammar rules can also be derived from existing projects such as the LinGO Grammar Matrix (Bender et al., 2002), ParGram (Butt et al., 2002; King et al., 2005). These are grammar development tools designed to write and create grammar specifications that support a wide range of languages, in a unified format. They focus on mapping simple description of languages, obtained from existing IGT-annotated data or input from a linguist, to precision grammar fragments, grounded in a grammar formalism such as HPSG, LFG. Our work differs in that, 1) we attempt to discover and explain the local linguistic behaviors for the language in general, 2) we do not extract rules for an individual sentence in isolation, as some of the HPSG/LFG-based approaches do, 3) we discover these behaviors from naturally occurring sentences. We do note that the rules we present in this work are based on the SUD annotation scheme, but the current framework can be easily extended to any other such scheme. In Table 2, we outline the different linguistic questions answered by our work and the related work.

There has also been work on developing toolkits to visualize some aspects of language structure – Lepp et al. (2019) present a web-based system to explore different morphological analyses. They also allow a user to improve the analyses thereby also improving the grammar specification which relies on those analyses.

B Learning and Extracting Rules

Statistical Threshold for Rule Extraction Similar to Chaudhary et al. (2020), we apply statistical testing to label leaves. For morphological agreement, we use the same hypothesis definition where the null hypothesis H_0 states that *each leaf denotes chance-agreement*. This means that there is no required agreement between a head and its dependent on the morphological attribute m. The hypothesis to be tested for is H_1 which states that *the leaf denotes required-agreement*. For case marking, we follow a similar approach as explained for word order. We can design H_0 as word order, because under the abstract case viewpoint (§ 3), case is a universal property for each word. We use a $\mathrm{p-value} = 0.01$ based on the recommendation of Chaudhary et al. (2020).

Rule Visualization Under each rule, we present a subset of examples from the training portion of the treebank to illustrate the rule. Positive examples refer to the examples which have features (from that rule) and follow the label as predicted by that leaf. However, there could be examples in the training data which have the same features as defined under that rule, but these example do not follow the predicted label. We refer to these examples as negative examples.

Since we only show a small set of examples, we select these examples to be concise and representative. We first group the examples under the rule with the lemmatized forms of the focus words. For example, under the Type-1 rule (Table 1) extracted for Spanish adjective-noun word order, the focus words are the **adjective** (w_a) and the <u>noun</u> (w_b). We group these examples by the lemmatized forms of the adjective and noun $\langle l_a, l_b \rangle$. The examples grouped under a lemmatized pair $\langle l_a, l_b \rangle$ are then sorted by their lengths. For each lemmatized pair $\langle l_a, l_b \rangle$, we select the top shortest examples. Finally, all selected examples are shuffled and we randomly select 10 examples.

C Experimental Setup

Data Below we describe the license details of the datasets we used:

- `SUD treebanks`: No specific license is specified, but the data is released as part of research

Linguistic Phenomena	Work	Rule-Type	Corpus Type
WordOrder	Ours	C+FG	Raw text
	Grammar Matrix (Bender et al., 2002)	C+FG	IGT text*
	Lewis and Xia (2008)	C	IGT text
	Bender et al. (2013)	C	IGT text
	Östling (2015)	C	Raw text
	Wang and Eisner (2017)	C	Raw text
	WALS Dryer and Haspelmath (2013)	C	Reference grammar*
Case Marking	Ours	C+FG	Raw text
	Grammar Matrix (Bender et al., 2002)	C+FG	IGT text*
	WALS Dryer and Haspelmath (2013)	C	Reference grammar*
	Howell et al. (2017)	C	IGT text
Agreement	Ours	C+FG	Raw text
	Grammar Matrix (Bender et al., 2002)	C+FG	IGT text*
	Chaudhary et al. (2020)	C+FG	Raw text
Sentence construction	Hellan (2010)	FG	IGT text*

Table 2: An overview of linguistic questions *automatically* answered by our current work and existing related work. Some of them combine semi-automatic approaches with manually annotated resources, there are marked with *. Rule-Type denotes the type of rule extracted for a language, C refers to coarse-grained such as rules for canonical word order, FG refers to fine-grained i.e. rules extracted at a local level.

work (Gerdes et al., 2019). We have used this data as intended which is for academic research purposes.

- `Fasttext embeddings`: Released[9] under the Creative Commons Attribution-Share-Alike License 3.0. We have used this data as intended, which is for academic research purposes.

- `Hmong Daw`: This dataset was collected by one of the co-authors from the Usenet group `soc.culture.hmong` and is currently in submission to LREC. The data used will be released as part of the Creative Commons Zero v1.0 Universal license. Accordingly, we will also release the train/test split for better reproducibility.

  The obvious identifying information has been removed from the data, although it would be possible to recover that information by going back to the original Usenet posts.

Model As described in the main text, we use the `XGBOOST` to learn a decision tree. For each language, the running time of the model is approximately 2-5 mins. We perform a grid search over a set of hyperparameters and select the best performing model based on the validation set performance. Here the hyperparameters we use:

- `criterion`: {gini, entropy}
- `max-depth`: {3, 4, 5, 6, 7, 8, 9, 10, 15, 20}
- `n-estimators`: 1
- `learning-rate`: 0.1
- `objective`: multi:softprob

D Gold-standard Experiments

D.1 Automated Evaluation Results

In the main text, we reported the average improvement for the word order, agreement and case marking models. In Table 3 we present the breakdown per each question. The word order results are reported over 56 languages, agreement over 38 and case marking over 35 languages.[10]

We also report results under three resource settings as shown in Table 5. We also show individual results per each language for word order (Table 7, Table 8), agreement (Table 9), case marking (Table 10, Table 11). We note that we report these results on a single run of the experiment.

We experiment with Random Forest, which is a better classifier, in comparison to decision trees,

[9] https://fasttext.cc/docs/en/pretrained-vectors.html

[10] Some languages have very little training data on which we couldn't fit a model while for some languages the linguistic questions was not applicable.

Linguistic Phenomena	Model	Gain
Word Order	adjective-noun	2.61
	subject-verb	6.95
	object-verb	10.78
	numeral-noun	9.88
	noun-adposition	2.31
Agreement	Gender	4.02
	Person	1.08
	Number	4.95
Case Marking	NOUN	30.03
	PRON	32.66
	DET	47.33
	PROPN	29.77
	ADJ	35.59
	VERB	18.76
	ADP	15.4
	NUM	25.81

Table 3: Breakdown of the performance gain (over the baseline) for each linguistic question. The performance of the agreement models is compared with the models trained over simple syntactic features in Chaudhary et al. (2020).

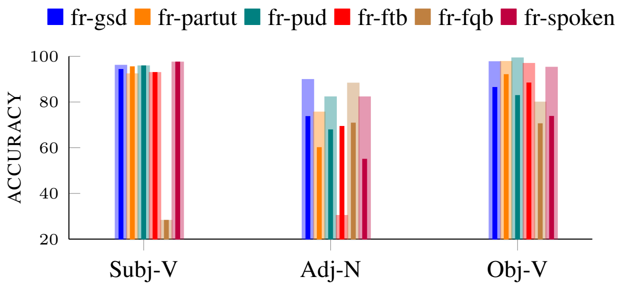


Figure 4: Comparing the accuracy of the model across different treebanks. Each model is trained on the fr-gsd treebank and directly applied on the other treebanks. Shaded bars denote the best model performance trained using all features while solid bars denote the most-frequent baseline for that treebank.

but it is not as interpretable as the latter. Nevertheless, as requested by the anonymous reviewers, we compare how decision trees fare against Random forest in Table 4. We train models for answering word order questions, across 15 languages from the SUD treebanks. Overall, we observe that decision trees slightly underperform the Random forest, but by only (avg.) -0.12 acc. points, where the range of accuracy is 0-100. Given our primary goal is to extract comprehensible descriptions, we opt to use decision trees.

D.2 Human Evaluation Results

We conduct expert evaluation for English and Greek. Both the English and Greek language expert are co-authors of the paper. For English, a total of 15 rules were evaluated for agreement, 11 for word order and 3 for case marking. For Greek, a total of 35 rules were evaluated for agreement, 11 for word order and 115 for case marking. We discussed the results in the main text, here we present the figures for English and Greek (Figure 5). For English, there were some rules which the expert was not aware of. We discussed one example for word order in the main text, we show an example for agreement and case marking in Table 6.

Model	Language	Random forest (acc.)	Decision tree (acc.)	Baseline
adjective-noun	el	99.29	99.29	99.29
	es	**73.32**	71.46	68.1
	ur	99.04	99.04	99.04
	fi	98.37	**99.09**	98.37
	lv	98.84	98.84	98.84
	it	**70.4**	69.26	66.02
	no	97.92	97.92	97.76
	fr	**74.01**	73.6	73.6
	ro	**95.83**	95.19	92.95
	bg	97.98	**98.49**	97.23
	gl	79.2	79.2	79.2
subject-verb	en	98.81	98.81	94.15
	el	**85.52**	83.45	73.56
	es	**83.5**	82.52	71.52
	tr	92.96	92.96	92.96
	hi	99.56	99.56	99.56
	fi	87.14	**90.36**	79.16
	lv	**79.79**	77.73	73.99
	it	**82.37**	81.44	71.76
	no	**86.28**	85.33	70.34
	fr	94.21	94.21	94.21
	ug	95.13	95.13	95.13
	ro	**75.62**	73.49	54.36
	bg	**81.67**	79.22	72.73
	gl	**86.26**	85.5	82.14
object-verb	en	**98.8**	98.66	97.26
	el	96.2	96.2	86.0
	es	95.99	95.99	90.4
	tr	96.64	96.64	96.64
	hi	**99.78**	99.61	74.71
	ur	99.45	**99.59**	79.5
	fi	85.21	**86.36**	74.83
	lv	**83.31**	82.95	75.24
	it	**94.97**	94.79	84.92
	no	98.68	98.68	95.86
	fr	**96.96**	96.53	86.33
	ro	86.99	**87.79**	65.06
	bg	**92.53**	92.22	80.66
	gl	**94.48**	94.17	82.2
noun-adposition	en	99.42	99.42	99.42
	es	100.0	100.0	98.83
	ur	98.91	98.91	98.91
	fi	89.35	**98.12**	89.35
	lv	97.78	97.78	97.78
	no	**99.3**	99.26	99.14
	gl	99.32	99.32	99.18
numeral-noun	en	88.06	88.06	82.09
	el	80.6	80.6	80.6
	es	88.62	88.62	75.61
	ur	95.63	95.63	95.63
	fi	**92.14**	87.25	90.71
	it	**82.33**	79.32	79.32
	no	85.78	**88.44**	88.44
	fr	81.16	**81.88**	60.87
	ro	84.14	**84.83**	62.07
	bg	88.24	88.24	88.24

Table 4: Comparing the accuracy of Random forest classifier with the Decision Tree for different word order relations.

E Hmong Daw Study

Data We experimented with the Hmong Daw variety in this setting. One of co-authors of the paper is a Hmong linguist who is in close collaboration and consultation with the community, and is the expert who provided us with the Hmong data and

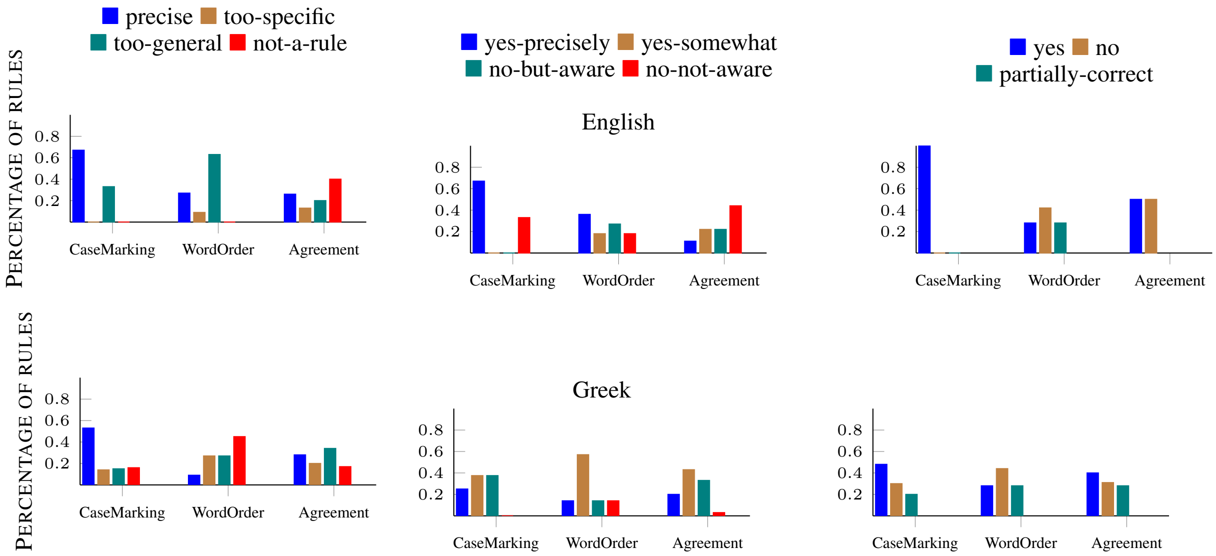
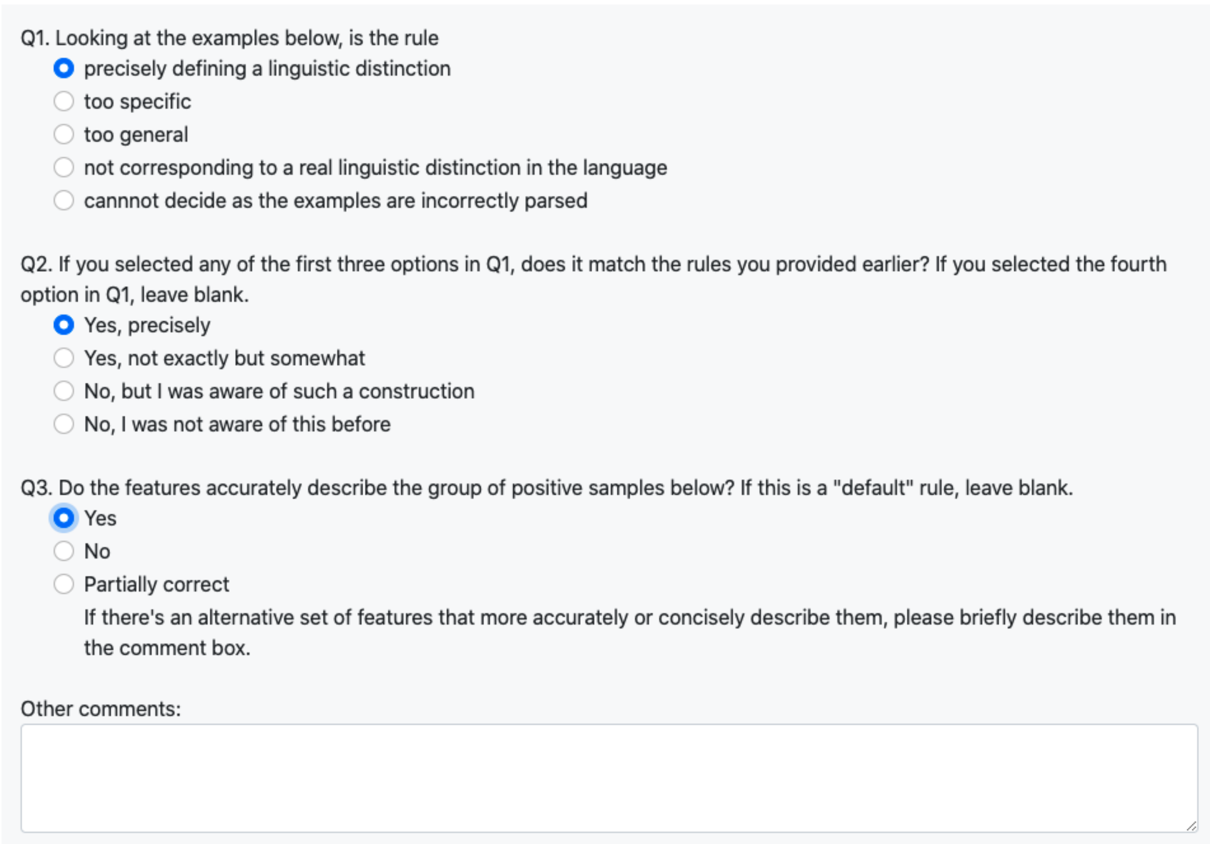


Figure 5: Evaluating rule correctness (left), prior knowledge (middle) and feature correctness (right). Top plot shows the results for English while the bottom plot shows for Greek.



Q1. Looking at the examples below, is the rule

- precisely defining a linguistic distinction
- too specific
- too general
- not corresponding to a real linguistic distinction in the language
- cannnot decide as the examples are incorrectly parsed

Q2. If you selected any of the first three options in Q1, does it match the rules you provided earlier? If you selected the fourth option in Q1, leave blank.

- Yes, precisely
- Yes, not exactly but somewhat
- No, but I was aware of such a construction
- No, I was not aware of this before

Q3. Do the features accurately describe the group of positive samples below? If this is a "default" rule, leave blank.

- Yes
- No
- Partially correct

If there's an alternative set of features that more accurately or concisely describe them, please briefly describe them in the comment box.

Other comments:

Figure 6: Rule evaluation form presented to the language expert.

Linguistic Phenomena	Resource-Setting	Gain (number of models)
Agreement	low	-3.39 (10)
	mid	1.07 (25)
	high	5.89 (55)
Case Marking	low	12.14 (11)
	mid	28.17 (56)
	high	37.17 (56)

Table 5: Breakdown of the performance gain (over the baseline) for each linguistic question by resource setting.

also helped evaluate the extracted grammar rules. We chose Vietnamese, Chinese and English to train `udify` model as they share syntactic and lexical similarity with Hmong. We use the same hyperparameter setting as specified in the code[11].

[11] https://github.com/Hyperparticle/udify

Linguistic Phenomena	Rule	Examples	Label
Number Agreement	dependent's head is a NOUN	**Kids** fun games are added to the building. **Nationalist** groups are coming to the conference.	Not-required-agreement
Object Case Marking	Pronoun is a oblique	Because Large Fries give **you** FOUR PIECES ! Give **him** a call tommorow	Accusative

Table 6: Some example of rules for agreement and case marking, which the expert annotator was not aware of. The **focus word** is highlighted, for agreement we also underline the head with which the dependent's agreement is checked. The examples under number agreement demonstrate that when dependent's head is a noun the **dependent** need not agree with its head. We show one example where the first example shows the dependent matches the number of the head, and the second example shows that it didn't not match.

Type	Lang	Train - Test - Baseline	Type	Lang	Train - Test - Baseline
adjective-noun	it-vit	70.71 - **69.51** - 66.02	object-verb	cu-proiel	80.37 - **82.72** - 76.03
adjective-noun	no-nynorsk	97.68 - **97.92** - 97.76	object-verb	be-hse	87.79 - 95.38 - **95.38**
adjective-noun	ro-nonstandard	87.46 - **95.19** - 92.95	object-verb	sv-lines	96.75 - **96.79** - 95.31
adjective-noun	bg-btb	97.27 - **98.49** - 97.23	object-verb	uk-iu	82.77 - **87.16** - 83.16
adjective-noun	gl-ctg	79.02 - 79.2 - **79.2**	object-verb	ga-idt	94.89 - **91.55** - 82.8
adjective-noun	cs-pdt	94.69 - **94.36** - 93.69	object-verb	sk-snk	81.84 - **86.17** - 80.91
adjective-noun	fi-tdt	98.56 - 99.09 - **99.09**	object-verb	hu-szeged	73.23 - **68.26** - 53.73
adjective-noun	pl-pdb	65.61 - **68.0** - 61.84	object-verb	got-proiel	74.58 - **80.15** - 72.44
adjective-noun	la-ittb	63.64 - **59.65** - 40.2	object-verb	hr-set	89.27 - **92.2** - 83.32
adjective-noun	nl-alpino	98.38 - 98.65 - **98.65**	object-verb	lzh-kyoto	97.86 - **98.01** - 95.7
adjective-noun	mt-mudt	78.91 - 82.84 - **82.84**	object-verb	lv-lvtb	85.03 - **82.95** - 75.24
adjective-noun	ja-bccwj	99.4 - 98.69 - **98.69**	object-verb	et-edt	76.03 - **79.51** - 69.67
adjective-noun	orv-torot	71.39 - **65.76** - 53.48	object-verb	fro-srcmf	79.62 - **81.82** - 48.25
adjective-noun	pt-gsd	70.31 - **74.54** - 71.63	object-verb	af-afribooms	82.72 - **96.19** - 86.03
adjective-noun	cu-proiel	84.96 - 84.98 - **84.98**	object-verb	hy-armtdp	71.47 - **74.58** - 44.92
adjective-noun	sv-lines	98.3 - **98.29** - 95.67	object-verb	en-ewt	98.33 - **98.94** - 97.26
adjective-noun	uk-iu	94.68 - 95.19 - **95.19**	object-verb	fr-gsd	98.89 - **97.18** - 86.33
adjective-noun	sk-snk	96.11 - 95.17 - **95.17**	object-verb	el-gdt	97.18 - **96.2** - 86.0
adjective-noun	got-proiel	79.51 - **79.51** - 72.48	object-verb	es-gsd	97.47 - **95.99** - 90.4
adjective-noun	hr-set	96.24 - **96.78** - 96.36	object-verb	tr-imst	95.38 - 96.64 - **96.64**
adjective-noun	lv-lvtb	98.93 - 98.84 - **98.84**	object-verb	ru-syntagrus	87.47 - **88.33** - 85.63
adjective-noun	et-edt	99.57 - **99.36** - 99.01	object-verb	sl-ssj	84.16 - **88.24** - 72.92
adjective-noun	fro-srcmf	73.84 - **74.42** - 73.26	object-verb	id-gsd	99.33 - **98.99** - 95.97
adjective-noun	en-ewt	97.84 - **98.25** - 96.77	object-verb	lt-alksnis	80.76 - **79.02** - 69.73
adjective-noun	fr-gsd	71.04 - **73.8** - 73.6	object-verb	ar-nyuad	96.27 - **95.91** - 95.63
adjective-noun	el-gdt	97.34 - 99.29 - **99.29**	object-verb	grc-proiel	72.98 - **75.87** - 67.05
adjective-noun	es-gsd	76.27 - **71.46** - 68.1	subject-verb	it-vit	82.95 - **82.53** - 71.76
adjective-noun	ru-syntagrus	97.84 - **98.0** - 96.54	subject-verb	no-nynorsk	83.42 - **85.33** - 70.34
adjective-noun	sl-ssj	98.22 - **98.27** - 97.78	subject-verb	ug-udt	95.32 - 95.13 - **95.13**
adjective-noun	id-gsd	93.41 - 92.79 - **92.79**	subject-verb	ro-nonstandard	69.06 - **74.27** - 54.36
adjective-noun	lt-alksnis	98.61 - 98.3 - **98.3**	subject-verb	bg-btb	78.86 - **79.65** - 72.73
adjective-noun	ar-nyuad	99.65 - 99.64 - **99.64**	subject-verb	gl-ctg	84.54 - **85.5** - 82.14
adjective-noun	grc-proiel	65.23 - **72.33** - 64.82	subject-verb	cs-pdt	67.13 - **73.18** - 63.33
adjective-noun	de-hdt	99.47 - **99.66** - 99.26	subject-verb	fi-tdt	88.11 - **90.57** - 88.19
object-verb	it-vit	96.28 - **94.88** - 84.92	subject-verb	pl-pdb	78.19 - **80.6** - 72.1
object-verb	no-nynorsk	97.73 - **98.68** - 95.86	subject-verb	la-ittb	80.29 - **82.69** - 72.54
object-verb	ro-nonstandard	86.05 - **87.79** - 65.06	subject-verb	zh-gsd	99.78 - **99.44** - 97.39
object-verb	bg-btb	92.18 - **92.43** - 80.66	subject-verb	nl-alpino	70.62 - **72.11** - 67.12
object-verb	gl-ctg	92.71 - **94.17** - 82.2	subject-verb	mt-mudt	83.91 - **84.96** - 72.03
object-verb	cs-pdt	82.35 - **83.91** - 73.97	subject-verb	orv-torot	72.38 - **66.07** - 60.46
object-verb	fi-tdt	84.21 - **86.62** - 77.98	subject-verb	he-htb	73.43 - **70.7** - 63.44
object-verb	pl-pdb	88.89 - **90.28** - 81.07	subject-verb	pt-gsd	89.4 - **93.15** - 87.47
object-verb	la-ittb	65.96 - **65.36** - 52.63	subject-verb	cu-proiel	73.88 - **76.31** - 62.48
object-verb	zh-gsd	93.4 - **94.12** - 87.75	subject-verb	be-hse	82.86 - **83.33** - 81.11
object-verb	nl-alpino	90.32 - **94.69** - 47.48	subject-verb	sv-lines	80.17 - **80.72** - 73.06
object-verb	mt-mudt	95.66 - 94.96 - **94.96**	subject-verb	uk-iu	76.89 - **77.14** - 74.56
object-verb	wo-wtb	91.6 - **91.81** - 75.11	subject-verb	ga-idt	99.33 - **99.28** - 85.25
object-verb	orv-torot	76.71 - **72.56** - 65.51	subject-verb	sk-snk	63.43 - 73.69 - **73.69**
object-verb	he-htb	97.87 - 98.03 - **98.03**	subject-verb	hu-szeged	75.91 - **74.59** - 72.43
object-verb	pt-gsd	95.17 - **95.02** - 88.45	subject-verb	got-proiel	67.56 - **73.2** - 66.17

Table 7: Accuracy results for all relations across different languages. Baseline is the most frequent order in the training data.

Type	Lang	Train - Test - Baseline	Type	Lang	Train - Test - Baseline
subject-verb	hr-set	81.87 - **86.62** - 77.44	noun-adposition	fi-tdt	97.88 - **98.12** - 89.47
subject-verb	cop-scriptorium	85.92 - **83.84** - 76.71	noun-adposition	pl-pdb	99.97 - **99.97** - 99.83
subject-verb	lv-lvtb	76.96 - **77.98** - 73.99	noun-adposition	nl-alpino	99.28 - **99.57** - 99.23
subject-verb	et-edt	68.13 - **71.93** - 61.02	noun-adposition	orv-torot	97.92 - **97.54** - 96.83
subject-verb	fro-srcmf	79.21 - **80.69** - 78.1	noun-adposition	he-htb	99.71 - **99.77** - 99.55
subject-verb	hy-armtdp	81.25 - 80.25 - **80.25**	noun-adposition	cu-proiel	98.06 - 98.4 - **98.4**
subject-verb	en-ewt	98.92 - **98.81** - 94.15	noun-adposition	sv-lines	98.6 - 98.11 - **98.11**
subject-verb	fr-gsd	96.7 - 94.21 - **94.21**	noun-adposition	uk-iu	99.74 - **99.8** - 99.54
subject-verb	el-gdt	77.04 - **77.93** - 73.56	noun-adposition	lzh-kyoto	95.58 - 96.61 - **96.61**
subject-verb	es-gsd	79.15 - **84.14** - 71.52	noun-adposition	cop-scriptorium	99.92 - **99.78** - 99.18
subject-verb	tr-imst	91.12 - 92.96 - **92.96**	noun-adposition	lv-lvtb	98.56 - 97.78 - **97.78**
subject-verb	ru-syntagrus	72.33 - **80.49** - 72.94	noun-adposition	et-edt	98.92 - **98.77** - 81.84
subject-verb	sl-ssj	70.95 - **74.66** - 63.01	noun-adposition	fro-srcmf	99.75 - 99.42 - **99.42**
subject-verb	id-gsd	99.09 - 99.34 - **99.34**	noun-adposition	hy-armtdp	97.22 - **96.83** - 85.71
subject-verb	lt-alksnis	74.44 - **78.39** - 75.33	noun-adposition	en-ewt	99.67 - 99.42 - **99.42**
subject-verb	ar-nyuad	91.01 - **91.32** - 87.82	noun-adposition	es-gsd	99.81 - **100.0** - 98.83
subject-verb	grc-proiel	69.46 - **72.23** - 65.71	noun-adposition	ru-syntagrus	99.24 - **99.41** - 99.13
subject-verb	de-hdt	68.1 - **76.23** - 61.84	noun-adposition	id-gsd	97.67 - **97.81** - 96.81
numeral-noun	it-vit	73.17 - 79.32 - **79.32**	noun-adposition	ar-nyuad	99.84 - **99.87** - 99.48
numeral-noun	no-nynorsk	88.49 - 88.44 - **88.44**	noun-adposition	grc-proiel	99.03 - 98.92 - **98.92**
numeral-noun	ro-nonstandard	87.27 - **84.83** - 62.07	noun-adposition	de-hdt	99.98 - **99.98** - 99.37
numeral-noun	bg-btb	92.22 - 88.24 - **88.24**			
numeral-noun	cs-pdt	84.4 - **88.65** - 69.59			
numeral-noun	fi-tdt	82.35 - **87.25** - 68.3			
numeral-noun	pl-pdb	97.27 - 97.27 - **97.27**			
numeral-noun	la-ittb	88.0 - **87.16** - 53.21			
numeral-noun	nl-alpino	95.03 - **98.7** - 89.61			
numeral-noun	mt-mudt	69.77 - 70.77 - **70.77**			
numeral-noun	wo-wtb	74.63 - **82.5** - 73.75			
numeral-noun	ja-bccwj	99.05 - 98.71 - **98.71**			
numeral-noun	orv-torot	86.64 - **79.8** - 72.73			
numeral-noun	he-htb	85.21 - **80.0** - 64.0			
numeral-noun	pt-gsd	92.18 - **89.42** - 73.56			
numeral-noun	sv-lines	81.3 - 85.48 - **85.48**			
numeral-noun	ga-idt	73.2 - **62.86** - 57.14			
numeral-noun	sk-snk	88.01 - **75.36** - 43.12			
numeral-noun	hr-set	95.39 - **97.28** - 96.94			
numeral-noun	et-edt	91.63 - **91.54** - 83.65			
numeral-noun	en-ewt	85.33 - **89.05** - 82.09			
numeral-noun	fr-gsd	79.7 - **81.88** - 60.87			
numeral-noun	el-gdt	88.2 - 80.6 - **80.6**			
numeral-noun	es-gsd	87.17 - **89.43** - 75.61			
numeral-noun	ru-syntagrus	93.48 - **95.01** - 85.15			
numeral-noun	sl-ssj	84.08 - 78.45 - **78.45**			
numeral-noun	id-gsd	61.04 - **68.12** - 53.44			
numeral-noun	ar-nyuad	88.96 - **91.79** - 47.9			
numeral-noun	grc-proiel	68.76 - 62.9 - **62.9**			
noun-adposition	no-nynorsk	99.31 - **99.26** - 99.14			
noun-adposition	gl-ctg	99.33 - **99.32** - 99.18			
noun-adposition	cs-pdt	99.98 - **99.98** - 99.94			

Table 8: Accuracy results for all relations across different languages. Baseline is the most frequent order in the training data.

Type	Lang	Test - Baseline	Type	Lang	Test - Baseline
Gender	it-vit	**71.01** - 67.19	Gender	hr-set	71.32 - **72.5**
Person	it-vit	**70.83** - 62.5	Person	hr-set	63.33 - **76.92**
Number	it-vit	59.56 - **71.2**	Number	hr-set	64.25 - **67.53**
Gender	no-nynorsk	**70.0** - 46.43	Gender	lv-lvtb	**74.48** - 72.66
Number	no-nynorsk	70.0 - **70.21**	Person	lv-lvtb	**60.98** - 50.0
Gender	ro-nonstandard	61.05 - **63.95**	Number	lv-lvtb	**72.78** - 68.83
Person	ro-nonstandard	55.22 - **63.64**	Gender	hsb-ufal	60.87 - **85.71**
Number	ro-nonstandard	**62.86** - 62.63	Number	hsb-ufal	46.72 - **69.23**
Gender	bg-btb	**66.0** - 63.83	Gender	ru-syntagrus	64.96 - **69.68**
Person	bg-btb	**64.0** - 62.5	Person	ru-syntagrus	**64.0** - 62.5
Number	bg-btb	**73.17** - 63.93	Number	ru-syntagrus	**60.24** - 59.09
Gender	cs-pdt	**75.09** - 56.44	Gender	el-gdt	**73.58** - 63.83
Person	cs-pdt	57.78 - **59.09**	Person	el-gdt	65.0 - **66.67**
Number	cs-pdt	**63.35** - 47.66	Number	el-gdt	**76.54** - 62.73
Gender	pl-pdb	**71.11** - 64.53	Gender	hi-hdtb	**69.11** - 58.59
Person	pl-pdb	**60.71** - 55.56	Number	hi-hdtb	**71.77** - 41.61
Number	pl-pdb	**66.06** - 63.68	Gender	es-gsd	**84.31** - 71.83
Gender	la-ittb	**77.78** - 73.53	Person	es-gsd	**91.67** - 59.09
Person	la-ittb	19.05 - **19.05**	Number	es-gsd	**88.89** - 64.39
Number	la-ittb	**65.14** - 57.89	Gender	ta-ttb	**100.0** - 68.18
Gender	nl-alpino	56.25 - **66.67**	Number	ta-ttb	**77.78** - 52.27
Number	nl-alpino	**60.94** - 54.84	Person	ug-udt	37.93 - **52.63**
Gender	orv-torot	64.52 - **65.54**	Number	ug-udt	47.73 - **76.67**
Person	orv-torot	**66.67** - 60.0	Person	fi-tdt	**58.06** - 38.71
Number	orv-torot	**64.04** - 62.12	Number	fi-tdt	**60.0** - 50.23
Gender	he-htb	**78.16** - 74.7	Person	wo-wtb	52.17 - **55.0**
Person	he-htb	**78.95** - 73.68	Number	wo-wtb	**57.14** - 48.57
Number	he-htb	58.14 - **58.54**	Person	hu-szeged	39.39 - **44.44**
Gender	cu-proiel	58.26 - **61.0**	Number	hu-szeged	38.34 - **39.63**
Person	cu-proiel	61.54 - **66.67**	Person	et-edt	**68.75** - 61.29
Number	cu-proiel	60.4 - **67.21**	Number	et-edt	61.21 - **64.84**
Gender	mr-ufal	53.57 - **60.87**	Person	hy-armtdp	**57.14** - 44.44
Person	mr-ufal	28.57 - **72.73**	Number	hy-armtdp	58.49 - **59.18**
Number	mr-ufal	**66.67** - 39.39	Person	en-ewt	**100.0** - 81.25
Gender	be-hse	**61.29** - 59.57	Number	en-ewt	**69.0** - 35.71
Number	be-hse	**65.82** - 64.62	Person	tr-imst	32.69 - **35.91**
Gender	sv-lines	**65.52** - 53.85	Number	tr-imst	**84.62** - 46.96
Number	sv-lines	60.0 - **64.29**	Number	kmr-mg	55.56 - **78.26**
Gender	uk-iu	68.92 - **70.08**	Number	af-afribooms	**68.75** - 60.0
Person	uk-iu	**72.73** - 70.0	Number	fr-gsd	**75.0** - 62.37
Number	uk-iu	**65.78** - 64.67			
Gender	ga-idt	**73.77** - 64.0			
Person	ga-idt	42.86 - **62.5**			
Number	ga-idt	43.16 - **46.75**			
Gender	sk-snk	**71.9** - 69.16			
Person	sk-snk	**88.89** - 77.78			
Number	sk-snk	**63.36** - 55.83			
Gender	got-proiel	**62.02** - 55.86			
Person	got-proiel	**62.16** - 57.14			
Number	got-proiel	**67.51** - 64.0			

Table 9: Accuracy results for all relations across different languages. Baseline is Chaudhary et al. (2020)

Type	Lang	Train - Test - Baseline	Type	Lang	Train - Test - Baseline
PRON	no-nynorsk	98.55 - **99.55** - 78.28	VERB	ug-udt	76.0 - **75.64** - 71.37
PRON	ug-udt	92.22 - **94.87** - 73.68	VERB	got-proiel	85.51 - **86.15** - 81.15
PRON	ro-nonstandard	89.77 - **91.2** - 38.33	VERB	lv-lvtb	96.43 - **95.61** - 75.58
PRON	sk-snk	83.19 - **83.9** - 34.75	VERB	tr-imst	67.53 - **66.58** - 46.13
PRON	hu-szeged	73.94 - **79.15** - 59.46	VERB	et-edt	86.95 - **86.08** - 82.91
PRON	got-proiel	87.97 - **91.05** - 36.79	VERB	hy-armtdp	86.63 - **94.34** - 39.62
PRON	hr-set	88.6 - **89.54** - 68.79	VERB	ur-udtb	96.01 - 98.95 - **98.95**
PRON	lv-lvtb	90.64 - **90.85** - 54.03	VERB	lt-alksnis	94.86 - **95.0** - 52.5
PRON	en-ewt	97.74 - **96.76** - 81.48	ADP	ro-nonstandard	98.5 - 98.85 - **98.85**
PRON	el-gdt	93.5 - **93.35** - 36.8	ADP	sk-snk	41.74 - **44.46** - 40.74
PRON	tr-imst	71.0 - **73.33** - 42.5	ADP	hr-set	45.85 - **48.42** - 37.96
PRON	sme-giella	85.31 - **76.82** - 47.05	ADP	hi-hdtb	85.57 - **86.99** - 52.34
PRON	es-gsd	95.89 - **96.14** - 53.71	ADP	ur-udtb	82.06 - **96.59** - 63.54
PRON	da-ddt	84.38 - **82.53** - 54.7	ADP	uk-iu	45.85 - **43.39** - 32.85
PRON	et-edt	79.75 - **81.58** - 45.26	ADJ	ro-nonstandard	98.14 - **96.9** - 96.42
PRON	af-afribooms	58.86 - **53.07** - 31.2	ADJ	ga-idt	95.47 - **93.25** - 90.18
PRON	hy-armtdp	78.1 - **79.05** - 63.81	ADJ	sk-snk	99.03 - **98.71** - 35.01
PRON	mr-ufal	71.58 - 78.95 - **78.95**	ADJ	hu-szeged	98.73 - **98.25** - 92.58
PRON	be-hse	81.3 - **76.12** - 65.67	ADJ	got-proiel	88.48 - **92.33** - 38.36
PRON	ur-udtb	87.78 - **90.53** - 54.73	ADJ	hr-set	97.75 - **98.3** - 37.5
PRON	lt-alksnis	82.8 - **80.28** - 30.28	ADJ	lv-lvtb	93.85 - **94.37** - 39.59
PRON	bg-btb	95.73 - **95.78** - 46.78	ADJ	et-edt	94.13 - **95.16** - 41.07
PRON	sv-lines	99.32 - **99.41** - 58.02	ADJ	el-gdt	85.61 - **89.49** - 48.6
PRON	uk-iu	88.52 - **90.98** - 48.82	ADJ	hi-hdtb	84.36 - **84.34** - 70.48
NOUN	ug-udt	78.31 - **77.13** - 63.46	ADJ	tr-imst	56.45 - **60.22** - 51.88
NOUN	ro-nonstandard	96.68 - **97.53** - 87.8	ADJ	sme-giella	86.09 - 90.55 - **90.55**
NOUN	kmr-mg	53.09 - 47.21 - **47.21**	ADJ	ar-nyuad	94.15 - **96.94** - 62.54
NOUN	ga-idt	93.26 - **95.62** - 80.28	ADJ	be-hse	89.59 - **95.06** - 43.83
NOUN	sk-snk	91.27 - **92.27** - 20.61	ADJ	ur-udtb	99.04 - **98.81** - 62.02
NOUN	hu-szeged	72.25 - **72.1** - 47.6	ADJ	lt-alksnis	96.89 - **96.72** - 25.5
NOUN	got-proiel	84.89 - **87.12** - 27.72	ADJ	uk-iu	97.38 - **98.15** - 46.39
NOUN	hr-set	88.35 - **92.21** - 34.41	DET	ro-nonstandard	97.01 - **95.87** - 75.58
NOUN	lzh-kyoto	89.86 - **93.72** - 76.61	DET	sk-snk	95.7 - **93.24** - 43.74
NOUN	lv-lvtb	81.94 - **83.87** - 31.62	DET	got-proiel	94.78 - **96.25** - 32.29
NOUN	kk-ktb	49.24 - 53.32 - **53.32**	DET	hr-set	94.87 - **95.64** - 42.81
NOUN	et-edt	62.6 - **66.51** - 27.65	DET	lv-lvtb	96.59 - **97.12** - 30.15
NOUN	el-gdt	91.02 - **94.69** - 49.72	DET	et-edt	96.73 - **96.19** - 34.25
NOUN	hi-hdtb	96.1 - **97.35** - 54.72	DET	el-gdt	91.42 - **93.64** - 47.48
NOUN	tr-imst	59.03 - **64.52** - 54.65	DET	hi-hdtb	88.89 - **92.87** - 76.1
NOUN	ta-ttb	77.24 - **76.49** - 68.02	DET	ur-udtb	95.79 - **95.91** - 64.33
NOUN	sme-giella	76.51 - **78.78** - 30.32	DET	lt-alksnis	79.46 - **83.65** - 39.92
NOUN	ar-nyuad	87.66 - **94.66** - 67.49	DET	uk-iu	94.31 - **94.86** - 27.93
NOUN	hsb-ufal	24.07 - 19.53 - **19.53**	PROPN	ro-nonstandard	97.35 - **96.77** - 92.98
NOUN	hy-armtdp	78.17 - **80.2** - 46.08	PROPN	ga-idt	79.87 - **85.78** - 73.28
NOUN	mr-ufal	81.38 - **75.0** - 42.65	PROPN	sk-snk	90.24 - **88.9** - 46.39
NOUN	be-hse	69.27 - **75.95** - 46.1	PROPN	hu-szeged	91.47 - 89.36 - **89.36**
NOUN	ur-udtb	92.1 - **96.81** - 51.25	PROPN	got-proiel	85.91 - **86.89** - 50.91
NOUN	lt-alksnis	85.08 - **82.93** - 39.07	PROPN	hr-set	92.42 - **94.67** - 48.27
NOUN	sv-lines	99.6 - **99.86** - 97.47	PROPN	lv-lvtb	88.64 - **90.13** - 39.91
NOUN	uk-iu	94.1 - **94.73** - 43.79	PROPN	el-gdt	91.44 - **90.32** - 32.58

Table 10: Accuracy results for all relations across different languages. Baseline is the most frequent case value in the training data.

Type	Lang	Train - Test - Baseline	Type	Lang	Train - Test - Baseline
VERB	ug-udt	76.0 - **75.64** - 71.37	PROPN	hi-hdtb	94.91 - **96.49** - 48.51
VERB	got-proiel	85.51 - **86.15** - 81.15	PROPN	tr-imst	73.55 - **71.73** - 68.0
VERB	lv-lvtb	96.43 - **95.61** - 75.58	PROPN	ta-ttb	97.99 - **94.84** - 93.55
VERB	tr-imst	67.53 - **66.58** - 46.13	PROPN	sme-giella	84.23 - **82.9** - 35.81
VERB	et-edt	86.95 - **86.08** - 82.91	PROPN	ar-nyuad	78.68 - **84.27** - 59.85
VERB	hy-armtdp	86.63 - **94.34** - 39.62	PROPN	et-edt	75.05 - **83.18** - 51.24
VERB	ur-udtb	96.01 - 98.95 - **98.95**	PROPN	hy-armtdp	82.28 - **89.13** - 54.89
VERB	lt-alksnis	94.86 - **95.0** - 52.5	PROPN	be-hse	86.43 - 72.68 - **72.68**
ADP	ro-nonstandard	98.5 - 98.85 - **98.85**	PROPN	ur-udtb	92.7 - **97.65** - 59.77
ADP	sk-snk	41.74 - **44.46** - 40.74	PROPN	sv-lines	97.21 - **96.6** - 91.23
ADP	hr-set	45.85 - **48.42** - 37.96	PROPN	uk-iu	93.76 - **95.14** - 36.14
ADP	hi-hdtb	85.57 - **86.99** - 52.34	NUM	sk-snk	81.47 - **77.38** - 39.29
ADP	ur-udtb	82.06 - **96.59** - 63.54	NUM	got-proiel	44.0 - **45.83** - 33.33
ADP	uk-iu	45.85 - **43.39** - 32.85	NUM	hr-set	90.27 - **94.26** - 41.8
ADJ	ro-nonstandard	98.14 - **96.9** - 96.42	NUM	lv-lvtb	88.07 - **85.44** - 38.61
ADJ	ga-idt	95.47 - **93.25** - 90.18	NUM	el-gdt	75.75 - **73.17** - 58.54
ADJ	sk-snk	99.03 - **98.71** - 35.01	NUM	tr-imst	76.55 - **82.22** - 77.78
ADJ	hu-szeged	98.73 - **98.25** - 92.58	NUM	sme-giella	47.8 - 41.84 - **41.84**
ADJ	got-proiel	88.48 - **92.33** - 38.36	NUM	et-edt	88.9 - **93.51** - 70.3
ADJ	hr-set	97.75 - **98.3** - 37.5	NUM	uk-iu	90.46 - **92.48** - 52.29
ADJ	lv-lvtb	93.85 - **94.37** - 39.59	ADV	fa-seraji	85.35 - 81.36 - **81.36**
ADJ	et-edt	94.13 - **95.16** - 41.07			
ADJ	el-gdt	85.61 - **89.49** - 48.6			
ADJ	hi-hdtb	84.36 - **84.34** - 70.48			
ADJ	tr-imst	56.45 - **60.22** - 51.88			
ADJ	sme-giella	86.09 - 90.55 - **90.55**			
ADJ	ar-nyuad	94.15 - **96.94** - 62.54			
ADJ	be-hse	89.59 - **95.06** - 43.83			
ADJ	ur-udtb	99.04 - **98.81** - 62.02			
ADJ	lt-alksnis	96.89 - **96.72** - 25.5			
ADJ	uk-iu	97.38 - **98.15** - 46.39			
DET	ro-nonstandard	97.01 - **95.87** - 75.58			
DET	sk-snk	95.7 - **93.24** - 43.74			
DET	got-proiel	94.78 - **96.25** - 32.29			
DET	hr-set	94.87 - **95.64** - 42.81			
DET	lv-lvtb	96.59 - **97.12** - 30.15			
DET	et-edt	96.73 - **96.19** - 34.25			
DET	el-gdt	91.42 - **93.64** - 47.48			
DET	hi-hdtb	88.89 - **92.87** - 76.1			
DET	ur-udtb	95.79 - **95.91** - 64.33			
DET	lt-alksnis	79.46 - **83.65** - 39.92			
DET	uk-iu	94.31 - **94.86** - 27.93			
PROPN	ro-nonstandard	97.35 - **96.77** - 92.98			
PROPN	ga-idt	79.87 - **85.78** - 73.28			
PROPN	sk-snk	90.24 - **88.9** - 46.39			
PROPN	hu-szeged	91.47 - 89.36 - **89.36**			
PROPN	got-proiel	85.91 - **86.89** - 50.91			
PROPN	hr-set	92.42 - **94.67** - 48.27			
PROPN	lv-lvtb	88.64 - **90.13** - 39.91			
PROPN	el-gdt	91.44 - **90.32** - 32.58			

Table 11: Accuracy results for all relations across different languages. Baseline is the most frequent value training data.