Extraction of Contrastive Rules from Syntactic Treebanks: A Case Study in Romance Languages

Santiago Herrera[1], Ioana-Madalina Silai[1], Bruno Guillaume[2], Sylvain Kahane[1,3]
[1]Modyco, Université Paris Nanterre, CNRS
[2]Université de Lorraine, CNRS, Inria, LORIA
[3]IUF - Institut Universitaire de France
{s.herrera, 43016143, skahane}@parisnanterre.fr, bruno.guillaume@loria.fr

Abstract

In this paper, we develop a data-driven contrastive framework to extract common and distinctive linguistic descriptions from syntactic treebanks. The extracted contrastive rules are defined by a statistically significant difference in frequency and precision, and classified as common and distinctive rules across the set of treebanks. We illustrate our method by working on object word order using Universal Dependencies (UD) treebanks in 6 Romance languages: Brazilian Portuguese, Catalan, French, Italian, Romanian and Spanish. We discuss the limitations faced due to inconsistent annotation and the feasibility of conducting contrastive studies using the UD collection.

1 Introduction

Cross-lingual corpus-based studies normally focus on finding common and distinctive structural features or tendencies among languages, language families, or typological balanced samples. Word order tendencies and their correlation with other language formal properties are an example of typological high-level descriptions. However, one might be also interested in comparing fine-grained patterns that explain the variation or the similarity between the compared corpora. This is a common goal in translation, second language teaching, textual genre research, and more generally, in corpus-based contrastive linguistics.

Comparing languages becomes more challenging the closer the languages are to each other. For example, syntactic objects vary considerably among Romance languages, even though they also exhibit some shared properties. Nominal objects often follow their verb, and personal pronominal objects tend to be in preverbal position. Both word order rules are common and dominant (in terms of frequency) in all Romance languages.

However, personal pronouns are enclitics of infinitives (or gerunds where they exist) in languages such as Spanish, Catalan or Italian, among others, while this is not possible in French. And even if the exact syntactic configuration exists in all languages, the relative frequency may vary. In addition, fine-grained differences within a language family are not always shared by the same group of languages. As mentioned, Spanish and French do not have the same word order between infinitive verbs and personal pronominal objects, but they do when the verb is an imperative (see Figure 1).

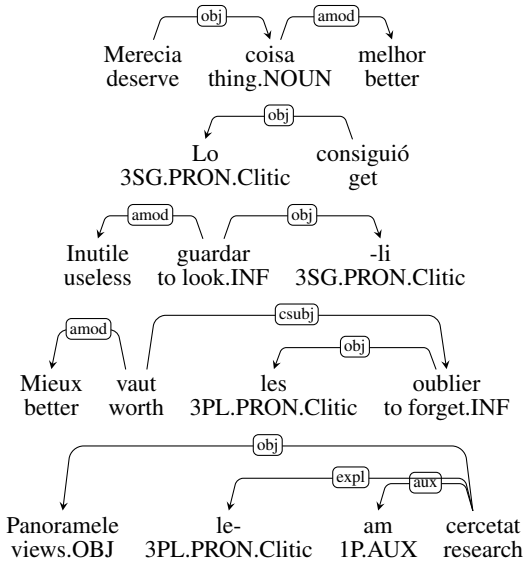

Figure 1: Different nominal and pronominal object word orders in Portuguese, Spanish, Italian, French and Romanian.

From this perspective, corpus-based contrastive analysis should be able to identify absolute differences across a set of languages, such as unique linguistic features, and be flexible enough to capture significant differences in frequency for fine-grained phenomena. Moreover, a contrastive approach should focus not only on distinctive patterns, but also on common ones. This approach allows a more detailed analysis of languages and corpora, focusing on variation or similarity in a specific lin-

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 26–38

guistic feature rather than on language profiles.

To address this issue, we develop a data-driven framework to extract fine-grained common and distinctive linguistic descriptions from syntactic treebanks. The extracted contrastive rules are defined by a statistically significant difference in precision and classified as common and distinctive rules across the set of treebanks. For each distinctive rule, we regroup languages having the same behavior. To test our method, we analyse the word order of objects in six Romance languages using Universal Dependencies (UD) treebanks. Object word order, especially object clitic order, varies considerably across Romance languages (Roberts, 2016). The present study is limited to this specific phenomenon. Further studies are left for future exploration.

Our approach follows and contributes to the core principles of most recent descriptive grammar extraction works. We seek descriptive but systematic descriptions of linguistic phenomena (Chaudhary et al., 2020, 2022), where extracted rules are overlapping rules, more or less fine-grained, and associated to quantitative data (Herrera et al., 2024a,b). The rules should be few and easy to interpret.

The contributions of this study can be summarised as follows:

- We adapt and extend current grammar extraction methods to extract concise and systematic contrastive descriptions for a set of corpora.
- We take a linguistically and statistically motivated approach to understanding common and distinctive patterns.
- We discuss the limitations of the use of universality based datasets (like UD) for automatic contrastive grammar description.
- We provide the UD community (de Marneffe et al., 2021) with a different perspective on annotation at the language family level.

2 Related Works

Corpus-based approaches to cross-linguistic analysis are widespread in both typological and contrastive linguistics, although in recent years they have become much more prominent in the latter. In either case, multilingual corpora have offered the possibility to capture and compare quantitative and gradient properties on a larger scale (Levshina, 2022).

Corpus-based Typological Studies

Typological studies based on parallel and comparable treebanks have mainly focused on high-level structural properties, typically word order or linguistic complexity. For example, Choi et al. (2021) explore the main word order categories in UD treebanks using a rewriting graph tool, and Gerdes et al. (2021) study quantitative word order and implicational universals on the same corpora, generalising to a larger variety of word order patterns. Similar studies have also been conducted using massive parallel corpora in more languages (Östling, 2015).

A common way to compare languages is to cluster them using different syntactic representations to see if phylogenetic groups are reconstructed according to typological databases (Alves et al., 2023). To the best of our knowledge, there are no studies using UD treebanks that cluster languages according to fine-grained syntactic patterns as we do in this paper.

Corpus-based Contrastive Studies

Contrastive linguistics lies somewhere between single-corpus studies and multi-corpus comparative studies, where the interest is in capturing fine-grained similarities and differences of internal linguistic properties in a small set of languages (Hasselgård, 2020). Contrastive approaches include frequency and statistical modeling (Gries et al., 2020) and studies based on information theory (see for example Alves (2025) for comparison between two Portuguese varieties). Normally, contrastive studies favour the use of parallel corpora (Nikolaev et al., 2020; Alves et al., 2023) because of the easy alignment between constructions from different sources.

Our work is closer to the quantitative syntax tradition (Bresnan et al. 2007; see also Thuilier 2012 for an example in a Romance language), where the main goal is to discover how some selected factors explain specific syntactic phenomena, such as the dative alternation or adjective order preferences. Such approaches have also been examined from a comparative perspective (e.g. Bresnan and Ford, 2010). In our case, we work with more than two languages and our approach consists of automatically identifying the predictive factors.

Descriptive Grammar Extraction

This paper builds on recent work on descriptive grammar extraction, where the main goal is to pro-

duce quantitative fine-grained grammar rules using traditional machine learning (ML) techniques. Chaudhary et al. (2020, 2022) formalise the task as a classification problem using decision trees to extract agreement, word order, and case marking rules across all UD treebanks. Herrera et al. (2024a,b) use a sparse linear classifier for agreement and word order for a few languages to obtain more expressive rules. They do this in single-corpus and one-to-one contrastive scenarios, but not for a collection of languages.

An advantage of (regularized) linear models is their ability to extract overlapping rules, which reflects how grammatical rules often behave. In contrast, decision trees create disjoint partitions of the data. In addition, their rule structure is highly sensitive to hyperparameters, such as tree depth. Tuning these hyperparameters can be challenging when the goal is descriptive insight rather than predictive accuracy. For instance, limiting a tree's depth to improve interpretability may result in uninformative residual nodes that use complex negative conditions to account for the remaining examples.

3 Task definition

Our main goal is to identify common and distinctive rules given a set of treebanks, focusing on object word order. Following Herrera et al.'s (2024a) formalisation, a quantitative rule posits a predictive relationship within the data, where the presence of a pattern P, identified within a starting sample S, increases the likelihood, by an α extent, of a phenomenon of interest Q.

$$S \implies (P \overset{\alpha\,\%}{\Longrightarrow} Q)$$

For example, to extract object word order rules for all objects in the sample, we look for linguistic patterns that favour the right or the left position with respect to its governor:

$$S : {}^{\text{OBJECT}}_{\text{RELATION}} \implies (P \overset{\alpha\,\%}{\Longrightarrow} Q : {}^{\text{OBJECT}}_{\text{POSITION}})$$

Given this formalisation, we consider a common grammatical rule to be evenly distributed across all languages. A distinctive rule, on the other hand, is unevenly distributed, though it may be shared equally by a subset of languages. If a corpus-based rule is quantitative in nature and captures gradient phenomena, a contrastive rule additionally posits whether the distribution of the predictive factor P is uniformly distributed across subsamples, in our case across languages.

To explore word order, we extract contrastive rules for the following four questions: (1) word order between two nodes connected by a dependency, (2) object word order, (3) pronominal object word order, and (4) nominal object word order. Analyzing object word order and its subtypes in two steps might be considered unnecessary, as the rules emerge when examining general object word order when the regularization parameter of the linear model is low enough. However, extracting rules for each type of object allows us to examine them in a more precise subspace. Furthermore, it is reasonable to explore pronouns in more detail as they are selected as good predictors of object word order. Overall, this multistep process yields fewer, more significant rules on average.

Before discussing the methods in detail, we first present the data, its preprocessing, and the sampling strategies employed.

4 Data

This study focuses on Romance languages for three main reasons. First, they are well represented in the UD treebank collection, providing us with a robust dataset. Second, we are experts and have native proficiency or possess a comprehensive understanding in all of the chosen languages. Finally, the use of these languages as a test case is advantageous due to their closeness. We hypothesize that the successful detection of differences and similarities between closely related languages will serve as an indication of the applicability of our method.

When multiple treebanks were available in the UD collection for a single language, the largest one in terms of tokens was selected, except in cases where annotation quality or methodology warranted a different choice. In Table 1 we provide an overview of each treebank, including its size, annotation process, and text sources where available.

4.1 Data Processing and Sampling

In all our experiments, we use the listed treebanks and ensure cross-linguistic comparability by selecting the same number of relevant syntactic patterns for each language. For broad pattern types, e.g. all dependencies, we collect approximately 10% of the matches for the smallest treebank to ensure a balanced sample size across languages. For more

UD Treebank	Tokens \ Sentences	Genres	Annotation Process
Brazilian Portuguese (Porttinari) (Duran et al., 2023)	168k \ 8k	News	Morphology: automatically tagged (reviewed); Syntax: manual annotation
Catalan (AnCora) (Taulé et al., 2008)	553k\ 16k	News	Morphology: automatically tagged (reviewed); Syntax: manual annotation
French (GSD) (Guillaume et al., 2019)	300k \ 16k	News, blog, reviews, wiki	Morphology: automatically tagged (reviewed); Syntax: converted from non-UD (corrected)
Italian (ISDT) (Bosco et al., 2014)	298k \ 14k	News, legal, wiki	Originally manually annotated, converted from non-UD
Romanian (RRT) (Barbu Mititelu and Irimia, 2016)	218k \ 9k	News, legal, fiction, academic, etc.	Morphology: automatically tagged (reviewed); Syntax: converted from non-UD (corrected)
Spanish (AnCora) (Taulé et al., 2008)	568k \ 17k	News	Morphology: automatically tagged (reviewed); Syntax: manual annotation

Table 1: Overview of selected UD treebanks used in this study.

specific target patterns (for example all object dependencies), we set the match count to the minimum available across all languages (approximately 7260 matches in this case) to ensure that each treebank was equally represented.

To compile these datasets, we randomly sampled sentences from each treebank until the desired number of matches was reached. To improve consistency across samples, we applied an interquartile range (IQR) filter based on sentence length, where length was defined as the total number of nouns, verbs and adjectives in the sentence. Sentences whose length fell outside the IQR-adjusted bounds were excluded and replaced with others from the original treebank falling within those bounds.

In order to limit the amount of noise in our final results, we removed punctuation due to lack of consistency across UD, and any enhanced dependencies as not all treebanks contained them.

5 Methodology

Our method can be divided into three separate steps. First, we extract and rank the most overall salient patterns for a given linguistic phenomenon across the set of treebanks using a linear classifier. Secondly, in order to identify common and distinctive patterns, we assess if the distributions of selected patterns are statistically different from a proportionally expected distribution. Finally, for each distinctive pattern we cluster languages to find those that share the same behaviour.

5.1 Rule Extraction Method

In order to achieve our first objective of extracting a small set of important features, we employed Herrera et al.'s (2024a) method of automatically extracting and ranking fine-grained grammatical rules from the combined treebanks. We train a series of sparse logistic regression models on features of all nodes within a defined search space. The extraction task is framed as a classification problem, where the goal is to predict the likelihood of a linguistic phenomenon occurring based on its associated features. The linear model is trained using the **negative log-likelihood loss** and **L1-norm** regularization to force sparsity, which makes the model easier to interpret and the selected features less redundant.

For any given node, the search space includes the node itself, its parent and children. We consider only the universal features of UD, parts of speech (UPOS) and morpho-syntactic features (FEATS), in order to minimise noise in the decision process and make samples more comparable.

The core logic of this approach for the word order of objects is as follows: given all the object dependencies in the treebank the goal is to identify patterns P, like being a object pronoun, that better predict its position with respect to its governor. The scope S and the target question Q are manually defined, reflecting the linguistic phenomena of interest, while patterns P are selected by the ML model. The model outputs a binary indication for each feature, indicating whether it is a reliable indicator of Q (i.e., postverbal position) or $\neg Q$ (i.e., preverbal position).

Patterns selected by the classifier are not relevant to a single language, but rather to the general classification task. Importantly, the selected patterns are ranked via the regularization path determined by

the series of trained models. In each run, the regularization parameter is decreased to allow more features to be activated, providing a ranking of importance inherent to the model. For more details, refer to the cited paper.

Herrera et al. (2024a) proposed several descriptive measures to describe a corpus-based grammar rule. One of them is **precision**, defined as the probability of Q happening, given that P has already happened (e.g. out of all the objects (S), the number of objects placed after their governor (Q) or before it ($\neg Q$) that are pronouns (P)). For our purpose, we have applied this metric to all occurrences of a rule in each language as follows:

Rule	**Precision for Treebank** t
$S \implies (P \overset{\alpha\,\%}{\Longrightarrow} Q)$	$\frac{\#_t(S \wedge P \wedge Q)}{\#_t(S \wedge P)}$
$S \implies (P \overset{\alpha\,\%}{\Longrightarrow} \neg Q)$	$\frac{\#_t(S \wedge P \wedge \neg Q)}{\#_t(S \wedge P)}$

Since we are interested in how languages differ from each other, we compute the **Coefficient of Variation (CV)** to measure the dispersion over the precision scores of treebanks t:

$$\text{CV} = \frac{\text{standard deviation}(\{\text{prec}_t\})}{\text{mean}(\{\text{prec}_t\})} \quad \text{for } t \in T$$

The CV measures the spread of the sample standard deviation relative to the mean of the precision scores. Language subsets with higher CV values exhibit more diversity, whereas those with low CV values are more similar. The extracted rules can be explored and ranked not only by their precision or the predictive importance given by the linear model, but also by their dispersion.

5.2 Evaluating Distribution Proportionality

Selected rules are relevant to a given linguistic phenomenon. In practice, we capture general properties of our sample, such as the fact that the nominal object follows the verb or that being a prepositional phrase does not favour being an object. It is still unclear whether the selected rules are common or distinctive properties across languages. The aforementioned measures also do not account for the cross-lingual behaviour of each pattern. For instance, the CV indicates how the patterns P are spread across languages, but it does not reveal the significance of that spread or which languages are driving it. Inspired by Chaudhary et al. (2020), we perform a statistical test to evaluate whether this difference is significant enough to conclude that the rule applies differently across languages.

To asses whether a selected pattern P is common or distinctive, we evaluate if there is a statistically significant difference between the observed distribution and a uniform expected distribution across languages of occurrences of P (which satisfies condition Q). The expected distribution is based on the assumption that the probability of having pattern Q given P is the same across all languages (e.g., the probability of being a preverbal object when the object is a pronoun is the same for all languages). To test this, we formulate two hypotheses:

Null hypothesis: The occurrences of patterns P satisfying Q are uniformly distributed across languages, meaning each language has an equal probability of exhibiting the pattern P satisfying Q. The pattern is a **common pattern** across our sample.

Alternative hypothesis: The occurrences of patterns P satisfying Q are not uniformly distributed across languages, meaning certain languages show a higher or lower relative frequency of the pattern. The pattern is a **distinctive pattern** across our sample.

We employed a conservative significance level by applying the Bonferroni correction. We divide the base alpha level (p-value < 0.01) by the number of statistical tests performed, which is equivalent to the number of rules selected by the model.[1] We also report Cramér's V effect size. If the null hypothesis is rejected, we consider it as **distinctive rule**, otherwise we consider it a **common rule**. Rules with patterns that are not present in all languages are also considered distinctive patterns but no statistical test is computed. We test our hypothesis using χ^2 goodness-of-fit test between the expected and observed distribution, as follows:

$$\chi^2 = \sum_{t \in T} \frac{(O_t - E_t)^2}{E_t}$$

where O the observed counts and E the expected frequency under the null hypothesis. The expected values are computed as:

[1] A more exploratory approach should consider a lower significance level and using a weaker regularization parameter.

$$E_t = \#(P \wedge Q) \cdot \frac{\#_t P}{\#P}$$

This is equivalent to computing expected values that follow the same precision distribution.

Although the goodness-of-fit test separates common from distinctive patterns, it does not specify individual language behaviour relative to a selected pattern under the null hypothesis. Therefore, in order to provide a better description, we compute normalised residuals between observed and expected frequencies to identify which languages are driving the deviation and in which direction (more or less frequent than expected):

$$r_t = \frac{O_t - E_t}{\sqrt{E_t}}$$

Large residuals indicate a significant difference between the observed and expected counts. Residuals close to zero indicate that the observed counts are similar to the expected counts under the null hypothesis. More specifically, a $|r_t| > 2.58$ is highly significant.

It is important to note that statistically significant findings reflect substantial frequency differences between treebanks. These differences may arise from either genuine linguistic variation or systematic annotation discrepancies. Our conservative approach, while excluding variations of lower significance, might still capture major systematic differences, including potential annotation artifacts.

5.3 Language Clustering by Pattern

Distinctive rules apply differently across treebanks. To automatically regroup treebanks with similar behaviour, we cluster their precision scores for each rule. We employ a hierarchical and incremental approach using Euclidean distance and the Ward variance minimization algorithm to group languages that together have low variance. Early merges represent highly similar languages, while later merges, occurring at higher levels of the dendrogram, involve increasingly dissimilar groups that contribute more substantially to the overall variance.

6 Results

We present the raw results without postselection, even though some rules may be redundant. For the object order excluding the Catalan treebank (refer to Section 6.1 for the reason), we extracted 69 potential grammar rules or tendencies. Of these, 39 are common, 14 are distinctive, 4 have low-frequency occurrences, and 12 are present only in one treebank. For an overview of shared and distinctive rules after the clustering process refer to Appendix B for all languages.[2]

We evaluate the sparse linear models with the selected features on the treebank test sets, and they generalise well (refer to Appendix C). However, such an evaluation provides only limited insight into the extracted rules, as we are not interested in classification scores but in the selected features. Since it is not feasible to qualitatively evaluate all the extracted rules, we explore a few relevant rules to illustrate how to interpret them and what the limits are.

6.1 Object as a Word Order Cue

Before looking at the word order of objects, a good starting point is to check whether being an object is associated with word order in general. In other words, we examine whether and to what extent word order is predictable from being an object. To do this, we trained a classifier to select the most important features that predict word order for all pairs of nodes with a dependency relation. Selected patterns could be, as it was mentioned before, global properties of word order given our entire sample, or properties of sub-samples/treebanks.

Among the selected patterns, being an object is a salient pattern for general word order, but not, as expected, the most important one. Three rules concern the object order, all favouring the right position (see Table 2). The second pattern, involving nominal objects, is a common one in our sample and is part of a highly precise rule, where 99% of nominal objects are to the right of their governor (example 1 in Figure 1). The first and third patterns concern the objects in general and those governed by verbs. While both patterns are highly correlated, as UD objects should be governed by verbs, they show a significant difference in the distribution of precision scores per language.

A closer look reveals that the Catalan treebank is the big outlier, showing a much lower probability of objects to the right of their governors than other treebanks. This difference is an annotation artifact. Reflexive clitics are incorrectly labeled as objects in reflexive passive clauses, when they are dative oblique complements, and when they are part of a pronominal verb. This artificially multiplies the

[2]The results and code are available at https://github.com/s-herrera/contrastive-grex-syntaxfest-2025.

n rule	pattern P	rule precision	predicted order	λ	CV	type
32	X-[obj]->O	83%	XO	0.007	0.117	distinctive
33	X-[obj]->O; O[upos=NOUN]	99%	XO	0.007	0.004	common
129	X-[obj]->O; X[upos=VERB]	83%	XO	0.001	0.119	distinctive

Table 2: Word order rules for the object considering all pairs of nodes connected by a dependency. The pattern P is expressed in GREW (Guillaume, 2021) format. X corresponds to an undefined node, while O is the head of the object. λ is the value of the regularization parameter at the moment of the feature activation. For rule precision, CV, significance, see subsections 5.1 and 5.2. Results included the Catalan treebank.

number of objects, making comparisons based on pronominal object counts unrealistic.

As previously mentioned, a significant difference can reflect a real syntactic difference or a systematic difference in annotation. In this case, it is the latter. In the following, we exclude the Catalan treebank, ensuring that the rules involving pronominal objects analyzed are reliable.

In any case, this provides an overview of the order of objects, with the post-governor order being preferred. However, it is still unclear whether different languages have different strategies for ordering objects. To explore this question, we will focus on directly investigating which factors favor the order of objects.

6.2 Order of Pronominal Objects

The rule A.2 of Table 3 is the first common rule selected by the linear model (the second overall) for pronominal object order. It captures that, among all treebanks, 86% of pronominal objects of finite verbs are preverbal (example 2 in Figure 1). In other words, in all the considered Romance treebanks, the object pronouns are frequently placed to the left of finite verbs, although this is not the only order, and object position may not follow the same strategies in all languages. As stated, our significance threshold is highly conservative, and while we consider this a common pattern, the dispersion is not null. For example, Brazilian Portuguese has more postverbal object pronouns with finite verbs, because, among other reasons, finite verbs allow right object pronouns.

On the contrary, the most important rule (A.1) for our model shows a higher dispersion. It captures that 75% of object pronouns tend to be to the left of the verb and this is a good predictor of word order. However, CV is relatively high, and the observed deviation relative to the expected proportional distribution is statistically significant. It is important to note that this rule includes all types of pronouns. Rule A.10 restricts pronouns to

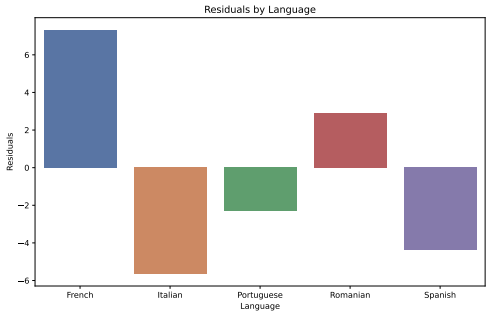

Figure 2: Residuals per language for rule A.10 of Table 3 on personal pronoun-object order (O[PronType=Prs]).

personal pronouns (*PronType=Prs*), a pattern that also favours the preverbal order. The difference in the expected distribution is also significant. If we examine the residuals for this rule in Figure 2, we see that there are more preverbal personal object pronouns in French, and Romanian, and much fewer in Italian, Brazilian Portuguese, and Spanish than expected under the null hypothesis. Being an object pronoun is then not a uniform predictor of word order in our sample, but it does discriminate well between two groups of languages where the behaviour with regards to this rule is different.

In order to further examine this difference in behaviour across the treebanks, we focus on the word order of pronominal object. The most predictive rule (B.1) states that being governed by an infinitive verb (*VerbForm=Inf*) favours the postverbal position (example 3 in Figure 1). This contrasts with the aforementioned rule involving finite verbs, listed here in the third position, where the behaviour is fairly uniform across the treebanks. Figure 3 shows that, in the case of object pronouns governed by an infinitive verb, there are clearly two clusters of languages: Italian, Portuguese and Spanish where the object pronoun immediately follows the infinitive verb in the majority of cases, and the rest for which the opposite is true (example 4 in Figure 1)

The model also extracts salient rules for less fre-

	pattern P	rule precision	predicted order	λ	CV	significance
(A) OBJECT ORDER						
1	O[upos=PRON]	75%	OV	0.1	0.19	distinctive
2	O[upos=PRON]; V[VerbForm=Fin]	86%	OV	0.026	0.09	common
10	O[PronType=Prs]	77%	OV	0.006	0.23	distinctive
(B) ORDER OF PRONOMINAL OBJECTS						
1	V[VerbForm=Inf]	56%	VO	0.7	0.76	distinctive
4	V[VerbForm=Fin]; O[PronType=Prs]	95%	OV	0.06	0.03	common
6	O[PronType=Rel]	98%	OV	0.05	0.55	common*
48	V[VerbForm=Ger]	74%	VO	0.005	0.64	common*
61	V[Mood=Imp]	85%	VO	0.003	0.17	low freq
95	O[PronType=Int]	88%	OV	0.001	0.93	low freq*
(C) ORDER OF NOMINAL OBJECTS						
1	with {Vchild[PronType=Prs]}	2.6%	OV	0.002	0.86	distinctive
5	with {V-[expl]->Vchild}	21%	OV	0.001	2.1	low freq*

Table 3: Word order rules concerning (A) object dependencies, (B) pronominal objects and (C) nominal objects. Refer to Table 2 for columns descriptions. The *with* clause in rules C.1 and C.5 should be interpreted as indicating the existence of at least another dependent that satisfies the specified condition. χ^2 test is not calculated for patterns with low frequency. *Patterns are selected by the model but are not shared by all languages.

quent phenomena. Rule B.6 indicates that the order of relative object pronouns does not vary significantly across treebanks and are almost always in a preverbal position. Rule B.95 indicates that being a interrogative pronoun favors the preverbal position. Romanian is not taken into account in these two cases because it uses a different label. Rules B.48 and B.61 favour the post-verbal position when the verb is a gerund form or is in the imperative mood, respectively. The expected frequency of these rules is low, and therefore the assumptions for computing the χ^2 statistic are not met.

Overall, we identify the main patterns of clitic object variation. Some findings challenge established knowledge. Brazilian Portuguese, for example, has a higher frequency of enclitic pronouns with infinitives than proclitics, resembling Spanish and Italian (c.f., Roberts, 2016, p. 791). However, the rules' limited expressivity, including the absence of negative conditions, prevents capturing phenomena such as clitic climbing with modal and aspectual verbs, as well as person-case constraints (Roberts, 2016, p. 789).

6.3 Order of Nominal Objects

When focusing on nominal objects, fewer rules emerge compared to other scopes. Rules indicating a post verbal position of the object have very low precision, confirming the postverbal dominant position (99%). Most of them are less reliable and difficult to understand. However, they hide some regularities and syntactic tendencies. We focus on the first rule (C.1), labeled as distinctive, which concerns nominal objects whose verbal governor has at least one more personal pronoun as a dependent. In French, the rule captures preposed nominal objects in direct interrogatives, where the verb bears a clitic. In Romanian and in Spanish, it captures clitic doubling phenomena which is not obligatory but strongly preferred when the nominal object is preposed. In addition, in Spanish, we find several passive subjects annotated wrongly

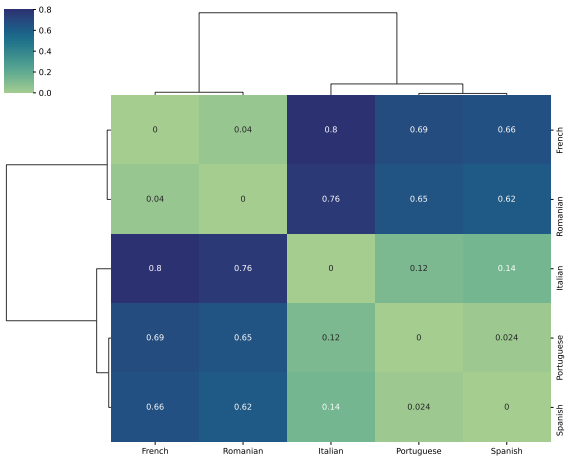

Figure 3: Clustermap of distances between precision distributions of each language for the rule B.1 of the Table 3 on pronominal object of infinitive verbs.

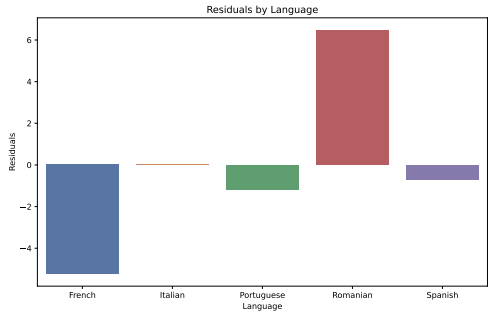

Figure 4: Residuals for rule C.1 of Table 3 on verbs having another pronoun dependant besides the object.

as objects. In Italian too, most occurrences are mediopassive constructions in *si* where the deep object promoted as a subject has been wrongly analyzed as an object (*il vero relax non si improvvisa* 'true relaxation cannot be improvised'). Residuals in Figure 4 show that Romanian in particular has more occurrences of the C.1 pattern than what is expected under the null hypothesis, while the opposite is the case for French. This is partly explained by the absence of the double clitic and the smaller number of preposed objects in French, as well as a seemingly productive strategy in Romanian. However, such low-frequency phenomena are probably influenced by the sample and its genres. This makes it impossible to draw a final conclusion about these patterns.

6.4 Annotation Inconsistencies and Error Detection

We mentioned that some results are affected by annotation inconsistencies as rules with extreme dispersion values reflect most of the time annotation inconsistencies. Our model uses these patterns in order to isolate anomalies in the sample, as they are extremely precise. This is the case, for example, for the feature *Emph=No*, which is used only in the French treebank to distinguish between emphatic and non-emphatic pronouns. It is also the case of the use of *PrepCase=Npr* in the Catalan and Spanish treebanks (both developed by the same team), to indicate that it is not a pronoun that changes form before a preposition. It should be noted that the other studied Romance languages also have this property, but do not use the feature. Sometimes the difference is not whether a label is used or not, but how it is used: the Italian treebank is the only one where the post-posted object clitics of infinitive verbs are annotated with the feature *Clitic=Yes*, even though this is a characteristic shared by all studied languages, except for French.

Cases like missing features in one or more treebanks are extreme, but annotation inconsistencies also arise from different annotation strategies. This is the case of rule C.5 of the table concerning clitic doubling. For instance, Romanian uses the relation `expl` (example 5 in Figure 1), while Catalan and Spanish use `obj`. In practice, the UD guidelines do not encourage the doubling of the *obj* dependency.

This illustrates how our method additionally captures regular error annotation, sharing thus conceptual ground with other error detection approaches. Related approaches use ML models trained on existing annotations to highlight inconsistencies between predicted and observed labels (Aquino et al., 2025) or compare annotations to predefined grammatical rules (Oepen et al., 2004). Hybrid systems combine both strategies (Agrawal et al., 2013; Ambati et al., 2011), while others identify consistent sequences in order to extract reliable patterns before extracting anomalies (Dickinson, 2015).

Our approach can be reframe as an error detection method for harmonizing annotations across corpora. First, we use ML techniques to extract salient syntactic patterns, treating those unique to one corpus as potential inconsistencies. Subsequently, statistical tests are used to analyze variations in shared patterns, which can signal either genuine linguistic differences or annotation discrepancies. Interpretation depends on corpus similarity: in closely related treebanks, variations more likely indicate annotation errors, while for distant corpora, linguistic knowledge is required to determine the cause.

7 Takeaways

We present a comprehensive framework to extract contrastive grammar rules and tendencies from syntactic treebanks. It allows us to induce a concise set of grammar rules that reflect statistical differences between closely related languages. A more exploratory and less conservative approach is possible by adjusting a few hyperparameters. Indirectly, the method can be used to find annotation inconsistencies across treebanks. Experiments also show the limitations of doing automatic grammar extraction and linguistic analysis with universal collections. For this reason, we encourage UD contributors maintaining related language treebanks to work together to harmonise annotation choices.

Limitations

Our sample presents three potential limitations. Firstly, our Romance sample does not cover all the language family diversity. Additionally, we only focus on object clitics, leaving out locative or genitive clitics. More critically, the heterogeneity of genres present within the employed treebanks introduces a confounding variable. Weak statistical trends may be attributable to variations or properties inherent to specific genres, rather than solely reflecting inherent linguistic characteristics. Third, the sample suffers from annotation inconsistencies and errors, introducing some noise in our results. Finally, concerning our methodology, it is important to emphasize that extracted grammar rules should be interpreted as having a predictive or directional nature, and not as causal factors.

Acknowledgments

This work is supported by the Université Paris Nanterre and the French National Research Project Autogramm (ANR-21-CE38-0017). It has benefited from discussions within CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). Finally, we would like to thank Guillaume Bonfante for his valuable insights and the three anonymous reviewers for their constructive feedback.

References

Bhasha Agrawal, Rahul Agarwal, Samar Husain, and Dipti M Sharma. 2013. An automatic approach to treebank error detection using a dependency parser. In *Computational Linguistics and Intelligent Text Processing*, Lecture notes in computer science, pages 294–303. Springer Berlin Heidelberg, Berlin, Heidelberg.

Diego Alves. 2025. Information theory and linguistic variation: A study of Brazilian and European Portuguese. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 9–19, Abu Dhabi, UAE. Association for Computational Linguistics.

Diego Alves, Božo Bekavac, Daniel Zeman, and Marko Tadić. 2023. Analysis of corpus-based word-order typological methods. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 36–46, Washington, D.C. Association for Computational Linguistics.

Bharat Ram Ambati, Rahul Agarwal, Mridul Gupta, Samar Husain, and Dipti Misra Sharma. 2011. Error detection for treebank validation. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 23–30, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Angelina A. Aquino, Lester James V. Miranda, and Elsie Marie T. Or. 2025. The ud-newscrawl treebank: Reflections and challenges from a large-scale tagalog syntactic annotation project. *Preprint*, arXiv:2505.20428.

Verginica Barbu Mititelu and Elena Irimia. 2016. Linguistic data retrievable from a treebank. In *Proceedings of the Second International Conference on Computational Linguistics in Bulgaria (CLIB 2016)*, pages 19–27, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Q. Bertrand, Q. Klopfenstein, P.-A. Bannier, G. Gidel, and M. Massias. 2022. Beyond l1: Faster and better sparse models with skglm. In *NeurIPS*.

Cristina Bosco, Felice Dell'Orletta, and Simonetta Montemagni. 2014. The evalita 2014 dependency parsing task. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa*. pisa university press.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Krämer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Arts and Sciences, Amsterdam.

Joan Bresnan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language*, 86(1):168–213.

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.

Aditi Chaudhary, Zaid Sheikh, David R Mortensen, Antonios Anastasopoulos, and Graham Neubig. 2022. Autolex: An automatic framework for linguistic exploration. *Preprint*, arXiv:2203.13901.

Hee-Soo Choi, Bruno Guillaume, Karën Fort, and Guy Perrier. 2021. Investigating dominant word order on Universal Dependencies with graph rewriting. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 281–290, Held Online. INCOMA Ltd.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Markus Dickinson. 2015. Detection of annotation errors in corpora. *Language and Linguistics Compass*, 9(3):119–138.

Magali Duran, Lucelene Lopes, Maria das Graças Nunes, and Thiago Pardo. 2023. The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics (2021-...)*, 6(1):17.

Stefan Th. Gries, Marlies Jansegers, and Viola G. Miglio. 2020. Quantitative methods for corpus-based contrastive linguistics. In Renata Enghels, Bart Defrancq, and Marlies Jansegers, editors, *New Approaches to Contrastive Linguistics*, pages 53–84. De Gruyter.

Bruno Guillaume. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Kiev/Online, Ukraine.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL : traitement automatique des langues*, 60(2):71–95.

Hilde Hasselgård. 2020. Corpus-based contrastive studies: Beginnings, developments and directions. *Languages in Contrast*, 20(2):184–208.

Santiago Herrera, Caio Corro, and Sylvain Kahane. 2024a. Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15114–15125, Torino, Italia. ELRA and ICCL.

Santiago Herrera, Ioana-Madalina Silai, Bruno Guillaume, and Sylvain Kahane. 2024b. Building quantitative contrastive grammars from syntactic treebanks. *Langues & Langages à la croisée des Disciplines (LLcD)*.

Natalia Levshina. 2022. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, 26(1):129–160.

Badr Moufad, Pierre-Antoine Bannier, Quentin Bertrand, Quentin Klopfenstein, and Mathurin Massias. 2023. skglm: improving scikit-learn for regularized generalized linear models.

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1159–1176, Online. Association for Computational Linguistics.

Stephan Oepen, Dan Flickinger, and Francis Bond. 2004. Towards holistic grammar engineering and testing. *Beyond shallow analyses–formalisms and statistical modelling for deep analysis (workshop at the first international joint conference on natural language processing (IJCNLP-04)*.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.

Ian Roberts. 2016. Object clitics. In *The Oxford Guide to the Romance Languages*. Oxford University Press.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Juliette Thuilier. 2012. *Contraintes préférentielles et ordre des mots en français*. Ph.D. thesis, Université Paris-Diderot - Paris VII, Paris, France.

A Sparse Logistic Regression Hyperparameters

We use skglm (Bertrand et al., 2022; Moufad et al., 2023) Sparse Logistic Regression implementation. We use default hyperparameters, except for the regularization parameter, which we vary from 0.1 to 0.001 in 100 steps.

B Distinctive Rule Distributions

To analyze the interrelationships between languages according to their distinctive rules, we present two series of matrices. The first series consists of co-occurrence matrices (Figure 5), which quantify how often languages fall into the same cluster for a given rule. The second series, composed of difference matrices (Figure 6), shows the inverse: how frequently languages are separated into different clusters. The matrices reveal the nuanced relationships across treebanks. For example, in the case of pronominal object word order (28 rules in total), Spanish and French only co-cluster in eight rules and are separated in 20, primarily due to their differing behavior in infinitive and gerund constructions. Conversely, Romanian and French are often grouped together. This is not due to a strong resemblance between the two languages, but rather because they are both systematically different from the more homogeneous Italo-Iberian group.

(a) Object WO (14)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	2	-			
Portuguese	0	11	-		
Romanian	11	4	2	-	
Spanish	10	6	3	10	-

(b) Pronominal Object WO (28)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	3	-			
Portuguese	0	24	-		
Romanian	25	3	0	-	
Spanish	8	22	20	8	-

(c) Nominal Object WO (2)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	2	-			
Portuguese	2	2	-		
Romanian	0	0	0	-	
Spanish	2	2	2	0	-

Figure 5: Co-occurrence matrices of shared rules for object, pronominal object, and nominal object word order across treebanks. Languages are ordered alphabetically.

(a) Object WO (14)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	12	-			
Portuguese	13	3	-		
Romanian	3	10	12	-	
Spanish	4	8	11	4	-

(b) Pronominal Object WO (28)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	24	-			
Portuguese	28	4	-		
Romanian	3	25	28	-	
Spanish	20	6	8	20	-

(c) Nominal Object WO (2)

	French	Italian	Portuguese	Romanian	Spanish
French	-				
Italian	0	-			
Portuguese	0	0	-		
Romanian	2	2	2	-	
Spanish	0	0	0	2	-

Figure 6: Difference matrices of distinctive rules for object, pronominal object, and nominal object word order across treebanks. Languages are ordered alphabetically.

C Model Evaluation

Table 4, on the next page, shows the evaluation scores for the selected rules on the training and test sets. Sparse linear models generalize well across all test sets. Performance scores for the nominal object order model were excluded due to extreme class imbalance. Macro-averaged measures are reported. The simplicity of the task makes the evaluation scores relatively uninformative.

Dataset	Accuracy	Precision	Recall	F1-score	Majority-class Baseline
GENERAL WORD ORDER					0.57
Train	0.93	0.93	0.94	0.93	
Test	0.93	0.93	0.93	0.93	
OBJECT ORDER					0.87
Train	0.98	0.97	0.95	0.96	
Test	0.98	0.97	0.95	0.96	
ORDER OF PRONOMINAL OBJECTS					0.72
Train	0.98	0.97	0.97	0.97	
Test	0.97	0.97	0.96	0.96	

Table 4: Scores on train and test (25%), with the selected features, excluding the Catalan treebank.