Estimating POS Annotation Consistency of Different Treebanks in a Language

Akshay Aggarwal
Univerzita Karlova
Matematicko-fyzikální fakulta
Malostranské náměstí 25
Prague, Czechia
akshayantimatter@gmail.com

Daniel Zeman
Univerzita Karlova
Matematicko-fyzikální fakulta
Malostranské náměstí 25
Prague, Czechia
zeman@ufal.mff.cuni.cz

Abstract

We introduce a new symmetric measure (called θ_{pos}) that utilises the non-symmetric KL_{cpos^3} measure (Rosa and Žabokrtský, 2015) to allow us to compare the annotation consistency between different treebanks of a given language, annotated under the same guidelines. We can set a threshold for this new measure so that a pair of treebanks can be considered harmonious in their annotation if θ_{pos} does not surpass the threshold. For the calculation of the threshold, we estimate the effects of (i) the size variation, and (ii) the genre variation in the considered pair of treebanks. The estimations are based on data from treebanks of distinct language families, making the threshold less dependent on the properties of individual languages. We demonstrate the utility of the proposed measure by listing the treebanks in Universal Dependencies version 2.5 (UDv2.5) (Zeman et al., 2019) data that are annotated consistently with other treebanks of the same language. However, the measure could be used to assess inter-treebank annotation consistency under other (non-UD) annotation guidelines as well.

1 Introduction

There exist a multitude of treebanks for different languages (Zeman et al., 2014). As noted by Kakkonen (2006), there exist a variety of formats and annotation schemes even for the treebanks for the same language. As an example, two well known POS tagging schemes for English language include the POS tagging scheme of the Penn Treebank[1] (Marcus et al., 1994) and the Universal POS tagset (Petrov et al., 2012).

The Universal Dependencies (UD) Project (Nivre et al., 2016b; Nivre et al., 2020) was introduced in 2014 as a means of unifying all the novel features of different annotation formats as a universal annotation scheme consistent across different languages. It has since become a standard reference to compare scores relating to parser performance (Che et al., 2018; Martínez Alonso et al., 2017), study of language-specific features (Alzetta et al., 2018), and for dependency parsing shared tasks on UD (Zeman et al., 2018).

UDv2.5 (Zeman et al., 2019) contains 157 treebanks in 90 languages, with multiple treebanks for some languages. Regardless of the differences in genre or the teams involved in building the treebanks, all treebanks of one language should be consistent with respect to the annotation guidelines, both intra and inter treebanks. However, this is often not the case, primarily because of the different sources of origin of the annotated data. The problem of determining the degree to which the different treebanks differ from each other has been studied in some detail over multiple years, but is not yet entirely solved.

The rest of the article is organised as follows. The literature relevant to the problem is discussed in Section 2, followed by a short introduction to the KL_{cpos^3} measure and a definition of the proposed measure in Section 3. Section 4 lists the constraints for choosing the dataset for the experiments as listed in Sections 5 and 6. The results of the experiments are summarised in Section 7. A discussion of the measure concludes the article in Section 8. The treebanks

[1] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

in UDv2.5 are marked for consistency or inconsistency of their POS annotation based on the proposed measure in Appendix A; Appendix B demonstrates the calculation of the measure for a concrete pair of treebanks.

2 Related Work

One of the most commonly used approaches to find inconsistencies in annotation is to train a high quality tagger or parser on the given training data, and evaluating the cases where the prediction from the trained model differs from the annotation of the test data. This approach can also be extended by bootstrapping different trained models, with the majority consensus being compared against the available annotation. Martínez Alonso and Zeman (2016) assessed the similarity of the Spanish treebanks in UDv1.3 (Nivre et al., 2016a) using dependency parsing. A high-efficiency parser was trained on one of the treebanks, and then tested on another. If a drop in parsing accuracy was more than what was intuitive, the treebanks were marked as not similar enough. The same technique was employed to evaluate the different Russian treebanks in UDv2.2 (Nivre et al., 2018) against each other by Droganova et al. (2018). It is worth stating here that the performance of the used tagger or parser may be a bottleneck, with the additional variables of the size and genre composition of the evaluated treebanks, among others. Furthermore, the acceptable variability in score in such cases depends on the architecture of the trained model, and is not comparable across different languages, or even when a different architecture is employed on the same data.

Dickinson and Meurers (2003a; 2003b) focus on finding an n-gram of tokens in the corpus that occurs in the same context (referred to as a *variation nucleus*) such that its different occurrences are annotated differently. Originally coined for continuous annotation,[2] the method was eventually adapted to look for inconsistencies in discontinuous annotation as well (Dickinson and Meurers, 2005).

Chun et al. (2018) compare the POS annotation consistency for several Korean treebanks by using the relative frequency of the individual POS tags, while also briefly mentioning the cause of the variation in their distribution. While such analysis is slightly helpful in terms of drawing a comparison, it does not consider the interaction of different POS tags with each other. To illustrate such interactions, an n-gram-based approach might be utilised.

3 KL_{cpos^3} and Measure Definition

In a delexicalised cross-language parser transfer scenario, Rosa and Žabokrtský (2015) show that the KL-Divergence score of POS trigrams, referred to as KL_{cpos^3}, can be effectively used for selection of the source language.

$$KL_{cpos^3}(tgt, src) = \sum_{\forall cpos^3 \in tgt} f_{tgt}(cpos^3) \log \frac{f_{tgt}(cpos^3)}{f_{src}(cpos^3)} \tag{1}$$

where $cpos^3$ is a coarse-grained[3] POS tag trigram, and

$$\begin{aligned} f(cpos^3) &= f(cpos_{i-1}, cpos_i, cpos_{i+1}) \\ &= \frac{count(cpos_{i-1}, cpos_i, cpos_{i+1})}{\sum_{\forall cpos_{a,b,c}} count(cpos_a, cpos_b, cpos_c)} \end{aligned}$$

with $count_{src}(cpos^3) = 1$ for each unseen trigram and a special value for $cpos_{i-1}$ or $cpos_{i+1}$ when $cpos_i$ lies on the sentence beginning or end.

[2]The annotation of the current token is based on the annotation of a contiguous token in word order. Discontinuous annotation implies the annotation of the current token is dependent on another token that might not be contiguous in the word order, as in the case of dependency parsing.

[3]For example, the coarse-grained POS associated with different nouns would be `NOUN` while the fine-grained POS would include `NN`, `NNP`, `NNS`, etc. We use UPOS tags for UD data, which are already coarse-grained in nature.

Considering that a treebank of the same language (despite the differences in the genres[4] covered) should be a better fit for POS transfer than a treebank from another language, we employ a symmetric variant of KL_{cpos^3}, called θ_{pos}, to assess the annotation consistency among the different treebanks of a language. θ_{pos} is a non-negative divergence measure. However, the measure scores cannot be compared directly across different languages. For a language-independent usage, there should be an empirical upper bound that needs to be placed on the θ_{pos} scores. As long as the θ_{pos} scores are lower than this empirical bound, the considered pair of treebanks can be considered harmonious in terms of their POS annotation. We denote this empirical upper bound by Θ_{pos}. The measures θ_{pos} and Θ_{pos} are linked together in the following definition:

Definition 1. Given two treebanks A and B, we say the treebanks are consistent in their POS annotation if the symmetric measure of their mutual divergence (given by θ_{pos}) is less than or equal to a threshold (given by Θ_{pos}). Formally, it can be represented as:

$$\theta_{pos}(A,B) = KL_{cpos^3}(A,B) + KL_{cpos^3}(B,A) \tag{2}$$

$$\leq \Theta_{pos}(A,B) \tag{3}$$

where $KL_{cpos^3}(P,Q)$ indicates the KL_{cpos^3} score of Q as an estimator for P.

Even though Θ_{pos} is an empirical bound on the θ_{pos} measure, the former is essentially a property of the latter. The empirical upper bound value would need to be estimated anew for a different set of annotation guidelines. In the remaining article, we estimate the empirical upper bound in a language-independent manner by looking at the influence of size of data, and the POS distribution in individual genres on θ_{pos} in different UDv2.5 treebanks (Zeman et al., 2019).

4 Assumptions while Working with UD Data

The UD website[5] provides a star ranking of individual treebanks within each language. The ranking is calculated heuristically[6], depending on multiple factors including the size of the treebank and the number of genres present in the data. The score also incorporates the output from the official UD validator[7] and from the search for known error types[8] in UDAPI (Popel et al., 2017). The treebank's compliance with the UD guidelines thus plays an important role in the score. While it is possible for a treebank to have a high score without being internally consistent, we assume that a treebank that adheres better to the guidelines also contains fewer inconsistencies. Therefore, we trust treebanks rated 3.5 stars or more (out of 5 stars).

Sometimes a whole treebank may not be sufficiently internally consistent because different genres have different distributions of POS n-grams. We may then require that the data belonging to one particular genre is annotated consistently.

5 Dataset Size and θ_{pos}

The value of θ_{pos} may depend on data size, as some POS trigrams may not be present in small datasets. We use k-fold cross validation to check the effect of presence or absence of POS trigrams in the data, based on the data size.

Experimental Setup

$KL_{cpos^3}(tgt, src)$ is defined on distributions of trigrams found in *tgt* and *src*. The calculated scores (and consequently θ_{pos} scores) are therefore affected by the presence or absence of the

[4]The usage of 'genre' in this context should also account for domain distinctions. In case such a distinction is available explicitly, data from each domain should be considered a separate 'genre'. To some extent this is actually the case with the 'genre' labels that are available in UD data and used in our experiments.

[5]`universaldependencies.org`

[6]For more details on the associated heuristics, refer to `https://github.com/UniversalDependencies/tools/blob/master/evaluate_treebank.pl`

[7]`https://github.com/UniversalDependencies/tools/blob/master/validate.py`

[8]`https://udapi.readthedocs.io/en/latest/_modules/udapi/block/ud/markbugs.html`

POS trigrams. In order to discount variability of θ_{pos} because of genre distribution, we use data from a single genre *(news)*. We take two UDv2.5 treebanks that have a large number of *news* sentences, high star ranking, and that belong to different language families: Czech-PDT (Indo-European, rated 4.5 stars) and Estonian-EDT (Uralic, rated 4 stars). For easier manipulation, we downsample the *news* data from either treebank as shown in Table 1.

Treebank	**Genre**	**Sentences**	**Downsampled to**
Czech-PDT	News	53,075	50,000
Estonian-EDT	News	13,557	12,000

Table 1: Sentence Counts in the *news* genre in Czech-PDT and Estonian-EDT.

To check the effect of data size on θ_{pos}, we run k-fold cross-validation on the downsampled data with different k-values. For each value of k, the downsampled data gets split to k folds, we select randomly one fold as test set and compute θ_{pos} of each of the remaining $k-1$ folds and the test set. This way we obtain $k-1$ values of θ_{pos}; their average is the θ_{pos} value we report for the given k in Table 2.

In addition to finding the values of θ_{pos}, we are also interested in finding its relationship with the count of unique trigrams common to the pair of distributions. We define coverage for a fold as the count of unique trigrams common to both training and test sets in the fold, expressed as a ratio of the count of all unique trigrams in the larger training set.

Experimental Scores and Inference

k **value**	θ_{pos} **Score**	**Coverage (in %)**
5	0.021 ± 0.001	83.872 ± 0.552
10	0.037 ± 0.001	75.447 ± 0.619
20	0.069 ± 0.002	66.131 ± 0.691
50	0.161 ± 0.005	52.768 ± 0.806
100	0.304 ± 0.011	42.373 ± 0.868
250	0.663 ± 0.028	29.345 ± 0.926
500	1.092 ± 0.053	20.784 ± 0.952

(a) *news* Data from UDv2.5 Czech-PDT, downsampled to 50,000 sentences

k **value**	θ_{pos} **Score**	**Coverage (in %)**
4	0.064 ± 0.002	76.139 ± 0.814
6	0.087 ± 0.003	69.742 ± 0.835
8	0.109 ± 0.004	65.177 ± 0.855
12	0.155 ± 0.005	58.72 ± 0.934
16	0.2 ± 0.007	54.142 ± 0.948
24	0.286 ± 0.011	47.727 ± 0.964
48	0.52 ± 0.022	37.094 ± 1.01
120	1.039 ± 0.052	24.474 ± 1.055

(b) *news* Data from UDv2.5 Estonian-EDT, downsampled to 12,000 sentences

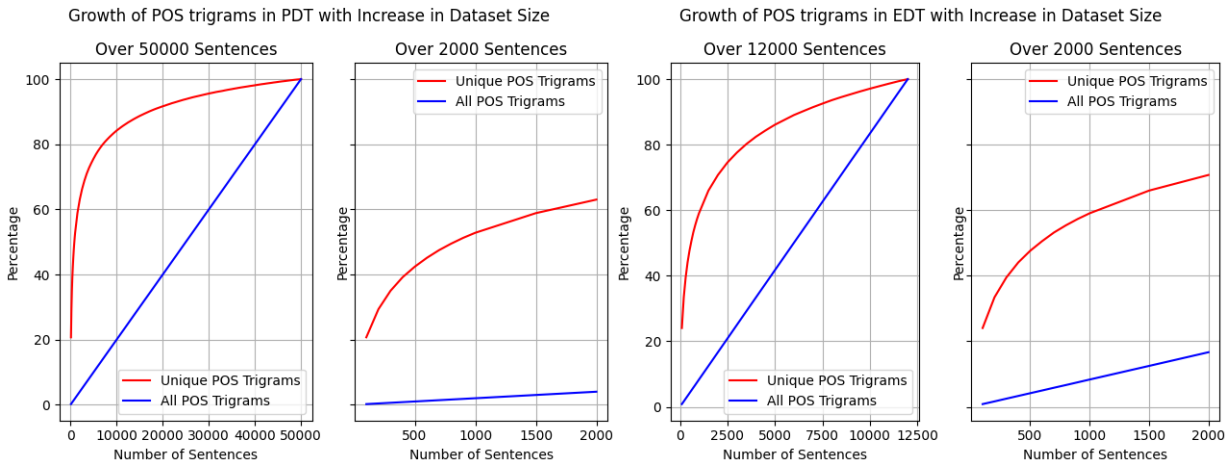
Table 2: θ_{pos} and coverage of POS trigram scores (± standard deviation) averaged over 100 different k-fold iterations. Each iteration results in a different downsample.

While there exists a strong negative correlation (Pearson correlation coefficient, r = -0.9075 and -0.9252 in Tables 2a, 2b respectively) between coverage of POS trigrams and the θ_{pos} scores, the coverage is, however, dependent on the size of the datasets being compared. Figures 1a and 1b show the variability in (i) number of distinct POS trigrams, and (ii) total number of POS trigrams, as the data size changes.

As evident from the figures, the growth pattern of counts is similar in both languages. The POS trigrams in a small part of the dataset obviously cannot be considered representative of those present in the entire dataset. Based on the observed coverage curve, we set 400 sentences[9] as the minimum size of a dataset whose consistency with another dataset is assessed.

However, difference in average sentence length is a factor that needs to be taken in account as well. If the two treebanks differ considerably in their average sentence length, then the size expressed in number of sentences does not reflect the number of tokens (and, consequently, the number of POS trigrams). For example, consider the Arabic treebanks in Table 3. If we take an equal number of sentences from Arabic-PUD and either of the other two treebanks, the total number of words will differ by a factor of almost 2.

[9]At about 400 sentences the percentage in Figure 1 crosses 40%.



(a) In *news* genre of Czech-PDT (b) In *news* genre of Estonian-EDT

Figure 1: Growth of POS trigrams with increase in dataset size

Counts	Arabic-NYUAD	Arabic-PADT	Arabic-PUD
Syntactic words	738,889	282,384	20,751
Sentences	19,738	7,664	1,000
Average length	37.434	36.845	20.751

Table 3: Average sentence lengths in Arabic treebanks. A syntactic word (node in the dependency tree) typically corresponds to a surface token but some tokens are split to multiple syntactic words.

Accommodating the dataset-size comparison, we can formally set the conditions such that the datasets can be compared amongst each other. Given two datasets A, B; the pair can be checked for annotation consistency if the following heuristic constraints are satisfied:

1. Individual dataset has at least 400 sentences, i.e. $\big(size(A) \geq 400 \;\;\&\;\; size(B) \geq 400\big)$; and
2. Dataset with smaller average sentence length has at least as many syntactic words as 400 sentences in the other dataset, i.e.

$$\boxed{\big(AvgSentLen(B) \leq AvgSentLen(A)\big) \implies \big(TotalSyntacticWords(B) \geq 400 \cdot AvgSentLen(A)\big)}$$

From Table 2, when the test split is composed of 500 sentences ($k = 100$ for Czech; $k = 24$ for Estonian), the θ_{pos} measure is ≈ 0.3. Considering that the larger values of k in either dataset do not satisfy heuristic constraint 1, we estimate the empirical upper bound of θ_{pos} based on $k = 100$ (Czech) and $k = 24$ (Estonian), respectively.

When estimating Θ_{pos}, we do not want to be too restrictive because the observed $\theta_{pos} \approx 0.3$ is based on internal consistency of a good treebank, which will be very hard to match for consistency between two different treebanks. We, therefore, round off the maximum observed θ_{pos} score from ≈ 0.3 to 0.5. Formally, if the datasets A, B contain data from the same genre, and the size of the datasets is comparable (as per heuristic constraints defined before), the upper limit on the θ_{pos} score can be specified in Equation 4.

$$\boxed{\theta_{pos}(A,B) \leq \Theta_{pos}(A,B) = 0.5} \tag{4}$$

6 Genre Distribution and θ_{pos}

In the previous experiments we assumed that the two compared datasets consist of the same language *and genre*. It is likely that the distribution of POS trigrams will differ when the two

datasets consist of different genres. We now proceed to investigate cross-genre variability inside a treebank that we believe is reasonably internally consistent. We are looking for Θ_{pos} thresholds that could be used to assess annotation similarity of two treebanks that differ in genre.

6.1 Inter-Genre Similarity

The Polish-LFG treebank in UDv2.5 (rated 4 stars) contains data from different genres,[10] the counts of which are shown in Table 4a. Table 4b shows the genres in UDv2.5 Finnish-TDT treebank (rated 3.5 stars). In this case, the data labeled *europarl* and *uni_articles* (university articles) is kept separate and not used in the estimation of variability of θ_{pos} across genres.

Genre (X)	*size*(*X*)	*AvgSentLen*(*X*)
fiction	7,252	7.124
news	6,744	8.401
nonfiction	1,273	7.719
social	526	6.977
spoken	1,253	6.047
academic	51	8.118
blog	136	7.772
legal	11	9.273

(a) In UDv2.5 Polish-LFG

Source (X)	*Size*(*X*)	*AvgSentLen*(*X*)
fiction	2,739	11.981
wiki	2,269	14.049
grammar	2,002	8.48
blog	1,781	12.533
legal	1,141	20.968
news	3,064	13.026
europarl	1,082	18.441
uni_articles	1,058	13.261

(b) In UDv2.5 Finnish-TDT

Table 4: Sources of genre data in UDv2.5 treebanks. Genres used in estimation of θ_{pos} scores are marked in bold.

As can be seen from Table 5, the different genres in Finnish-TDT are internally consistent in their annotation, as per the constraint in Equation 4. For each genre source, the dataset is downsampled to 900 sentences, and the results are presented on the individual folds resulting from 2-fold cross-validation on the downsampled data. The similar analysis for genres in Polish-LFG is omitted here because the *social* genre does not have enough data.

Genres	θ_{pos} ($\pm$ sd)	Θ_{pos}
fiction	0.316 $\pm$ 0.015	0.5
wiki	0.3 $\pm$ 0.017	0.5
grammar	0.427 $\pm$ 0.021	0.5
blog	0.332 $\pm$ 0.017	0.5
legal	0.216 $\pm$ 0.035	0.5
news	0.286 $\pm$ 0.015	0.5
europarl	0.233 $\pm$ 0.017	0.5
uni_articles	0.3 $\pm$ 0.014	0.5

Table 5: θ_{pos} ($\pm$ standard deviation) averaged over 100 runs for each genre in UDv2.5 Finnish-TDT. Each run results in a different downsample.

Experimental Setup

We compare different genres in the Polish-LFG and Finnish-TDT treebanks by presenting the θ_{pos} scores for each pair of genres (as per Table 4). Each genre is downsampled to the number of instances as listed in Table 6 such that the heuristic constraints for dataset comparison are satisfied.

Experimental Scores and Inference

Tables 7 and 8 list the θ_{pos} scores for data from Polish-LFG and Finnish-TDT, respectively. It is worth noting that for most genre pairs, the Θ_{pos} constraint as employed in Equation 4 is not enough, as θ_{pos} frequently surpasses the imposed limit of 0.5.

[10] `https://github.com/UniversalDependencies/UD_Polish-LFG#data-split-and-genres`

Genre (X)	Downsampled to	$\frac{TotalSyntacticWords(X)}{AvgSentLen(A)}$
fiction	500	424
news	500	500
nonfiction	500	459
social	500	415
spoken	600	432

(a) UDv2.5 Polish-LFG

Genre (X)	Downsampled to	$\frac{TotalSyntacticWords(X)}{AvgSentLen(A)}$
fiction	1,000	571
wiki	1,000	670
grammar	1,000	404
blog	1,000	598
legal	1,000	1,000
news	1,000	621

(b) UDv2.5 Finnish-TDT

Table 6: Counts of sentences for different genres in data downsampled from UDv2.5 treebanks. A in $Avg(A)$ in the third column refers to the genre with the highest number of average words per sentence in each language, marked in bold.

Genres	news	nonfiction	social	spoken
fiction	0.754 ± 0.047	0.556 ± 0.028	0.726 ± 0.032	1.059 ± 0.047
news	-	0.55 ± 0.032	0.906 ± 0.044	1.53 ± 0.071
nonfiction	-	-	0.624 ± 0.027	1.285 ± 0.046
social	-	-	-	1.178 ± 0.033

Table 7: θ_{pos} scores (± standard deviation) averaged over 100 runs for inter-genre analysis in downsampled UDv2.5 Polish-LFG data. Each run results in a different downsample.

Genres	blog	grammar	wiki	legal	news
fiction	0.356 ± 0.014	0.47 ± 0.019	1.552 ± 0.041	1.559 ± 0.04	1.323 ± 0.044
blog	-	0.504 ± 0.018	1.307 ± 0.042	1.328 ± 0.026	1.113 ± 0.043
grammar	-	-	1.166 ± 0.041	1.554 ± 0.036	0.888 ± 0.035
wiki	-	-	-	1.229 ± 0.032	0.473 ± 0.021
legal	-	-	-	-	1.078 ± 0.026

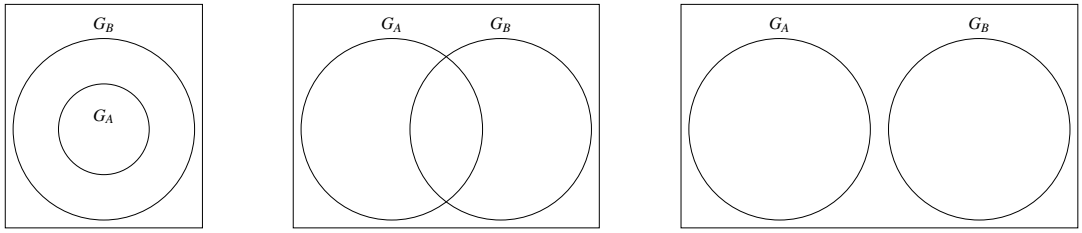
Table 8: θ_{pos} scores (± standard deviation) averaged over 100 runs for inter-genre analysis in downsampled UDv2.5 Finnish-TDT data. Each run results in a different downsample.

As expected, we need a higher threshold when comparing datasets whose genre does not match. While a threshold of 1.6 would accommodate data in Polish-LFG and Finnish-TDT, we again allow some room to reduce false alarms about inconsistent pairs of treebanks, and frame the empirical upper bound on θ_{pos} between genre x in dataset A (written as A_x) and genre y in dataset B (B_y) as in Equation 5, given below:

$$\boxed{\theta_{pos}(A_x, B_y) \leq \Theta_{pos}(A_x, B_y) = 2.0} \tag{5}$$

6.2 Combination of Genres

We denote the set of genres in treebank X as G_X. Given two treebanks with at least one different genre, the different genres in the two treebanks can interact in either of the three cases as shown in Figure 2. To see how the θ_{pos} scores are affected in either of the cases, we experiment with the data from UDv2.5 Polish-LFG.



(a) Case 1: $G_A \subseteq G_B$ (b) Case 2: $G_A \not\subseteq G_B$; $G_A \cap G_B \neq \phi$ (c) Case 3: $G_A \not\subseteq G_B$; $G_A \cap G_B = \phi$

Figure 2: Interaction of genres in treebanks A and B, such that $|G_A| \leq |G_B|$

Experimental Setup

We start by downsampling the data from the *fiction* and *news* genres to 2000 sentences each. Using 2-fold cross-validation, the downsampled data is then split into 2 halves, termed as *base* and *test* set for the genre. In addition, we downsample the data from the *spoken* genre to 1000 sentences and use it as a *test* set (without corresponding *base* set).

We try to understand θ_{pos} variability in the scenarios depicted in Figure 2. The different genres combining together to form a dataset can be identified by the name of the concatenated dataset. The trailing *base* in the dataset name marks that it is composed of data from the *base* set of the genre(s). The datasets using *test* set of genre(s) can similarly be identified by trailing *test* in the dataset name.

Experimental Scores and Inference

We present the calculated scores for different cases in Table 9.

	news_base	*fiction_base*	*news_fiction_base*
news_test	0.257 ± 0.010	0.64 ± 0.034	0.3 ± 0.015
fiction_test	0.646 ± 0.034	0.278 ± 0.013	0.351 ± 0.021
spoken_test	1.503 ± 0.049	0.99 ± 0.036	1.144 ± 0.035
spoken_news_test	0.489 ± 0.022	0.499 ± 0.020	0.338 ± 0.014
spoken_fiction_test	0.854 ± 0.036	0.41 ± 0.018	0.498 ± 0.023
news_fiction_test	0.304 ± 0.016	0.348 ± 0.019	0.17 ± 0.007
all_genres	0.463 ± 0.022	0.351 ± 0.014	0.247 ± 0.011

	news_test	*fiction_test*	*news_fiction_test*
spoken	1.493 ± 0.048	0.987 ± 0.034	1.138 ± 0.03

Table 9: θ_{pos} (± standard deviation) scores averaged over 100 runs, reported for different genre combinations. Each run results in a different downsample. The scores marked in blue indicate that the genre sets overlap, while those in red indicate the genre sets are disjunct. The scores without color-code indicate that one genre set is a subset of the other.

It is noteworthy that the decomposition of a treebank into its constituent genres forms the first basis for the study of variance of θ_{pos} scores with a combination of the different genres. Upon a closer inspection, it was discovered that when there are multiple genres present in the treebank, the θ_{pos} measure score is dominated by the POS trigrams that are typical of the language, and the genre-specific POS trigrams become increasingly obscure.

Once the individual genres have been identified and checked for the inter-genre θ_{pos} scores, the overall measure score is less than the average of the measure scores calculated for individual pair of genres in the treebank(s). Formally, assuming treebanks A and B can be split into their constituent genres such that $G_A = \{A_1, A_2, ..., A_i\}$ and $G_B = \{B_1, B_2, ..., B_j\}$, the overall limit on the $\theta_{pos}(A,B)$ score can be specified as in Equation 6.

$$\boxed{\theta_{pos}(A,B) \leq \Theta_{pos}(A,B) \leq Average(\theta_{pos}(A_x, B_y))} \quad \forall[A_x \in G_A; B_y \in G_B] \tag{6}$$

6.3 Adulterant Genres

In our analysis so far, we have restricted ourselves to instances where the data in the different genres could be reliably compared. We define a genre in the dataset as *adulterant* if the number of sentences in the genre does not satisfy either or both the constraints pertaining to dataset comparison. In this subsection, we take a look at how the presence of adulterant genres affects the θ_{pos} scores.

Experimental Setup

To study the effect of adulterant genres, we first downsample data from the *fiction*, *news* and *spoken* genres in Polish-LFG to 500, 500 and 600 sentences respectively. For adulterant genres, we work with the data from the *academic*, *blog* and *legal* genres. The data from all the

adulterant genres is concatenated to form a dataset labeled *others*. Non-adulterant genres are then combined with adulterant genres to result in a dataset identified as *X*-*Y*, where *X* contains data from *news*, or *fiction*, a combination of the two genres. *Y* may be an individual adulterant genre, or a combination of all adulterant genres *(others)*. All the datasets created from the downsampled data are compared with the downsampled data from *spoken*.

Experimental Scores and Inference

The calculated θ_{pos} scores for each pair, averaged over 100 runs, are reported in Table 10.

	spoken
fiction	1.059 ± 0.047
fiction-academic	1.072 ± 0.046
fiction-blog	1.09 ± 0.044
fiction-legal	1.065 ± 0.047
fiction-others	2.413 ± 0.384

	spoken
news	1.53 ± 0.071
news-academic	1.552 ± 0.069
news-blog	1.54 ± 0.065
news-legal	1.547 ± 0.071
news-others	2.63 ± 0.334

	spoken
fiction_news	1.196 ± 0.048
fiction_news-academic	1.215 ± 0.048
fiction_news-blog	1.223 ± 0.046
fiction_news-legal	1.206 ± 0.048
all-genres	2.309 ± 0.358

Table 10: θ_{pos} Scores (± standard deviation) averaged over 100 different runs with adulterant genres present in Polish-LFG. Each run results in a different downsample.

From the table, we observe that a low number of adulterant genres in the data does not affect the θ_{pos} scores heavily. However, the presence of multiple adulterant genres pushes the θ_{pos} scores by almost 1.5 as compared to when there are no adulterants present. Taking into account also the standard deviation score, and the high annotation quality of the treebank, we can add a headroom of +2.0 if adulterant genres are present.

Formally, assuming treebanks A and B can be split into their constituent genres such that $G_A = \{A_1, A_2, ..., A_{n1}\}$ and $G_B = \{B_1, B_2, ..., B_{n2}\}$. Of all the constituent genres in $G_A \cup G_B$, the set of adulterant genres can be represented as $G_{adulterant}$. The overall limit on the $\theta_{pos}(A,B)$ score, as specified in Equation 6, can be updated as in Equation 7

$$\theta_{pos}(A,B) \leq \Theta_{pos}(A,B) \leq \begin{cases} Average(\theta_{pos}(A_x,B_y)) + 2.0 & \text{if } G_{adulterant} \neq \phi \\ Average(\theta_{pos}(A_x,B_y)) & \text{if } G_{adulterant} = \phi \end{cases} \quad (7)$$

$$\forall [A_x, B_y \in (G_A \cup G_B) - G_{adulterant}]$$

7 Framing the Overall θ_{pos} Limit

In a case when the data from individual genres in the data is not annotated consistently, the θ_{pos} score might be within the bounds of averaged scores for individual genres, therefore marking the pair as consistent. To avoid this, we calculate the idealistic Θ'_{pos} as the average of Θ_{pos} values for the genres.

$$\Theta'_{pos}(A,B) = Average(\Theta_{pos}(A_x,B_y)) \quad \forall [A_x, B_y \in (G_A \cup G_B)] \quad (8)$$

where $\Theta_{pos}(A_x,B_x) = 0.5$ and $\Theta_{pos}(A_x,B_y) = 2.0$ as per Equations 4 and 5, respectively.

For overall calculation of Θ_{pos} scores for treebanks with multiple genres, the overall computation can be given by:

$$\theta_{pos}(A,B) \leq \Theta_{pos}(A,B) = \begin{cases} Minimum(\Theta'_{pos}(A_x,B_y),\ Average(\theta_{pos}(A_x,B_y),\ 2.0) & \text{if } G_{adulterant} = \phi \\ Minimum(\Theta'_{pos}(A_x,B_y),\ Average(\theta_{pos}(A_x,B_y),\ 2.0) + 2.0 & \text{if } G_{adulterant} \neq \phi \end{cases} \quad (9)$$

$$\forall [A_x, B_y \in (G_A \cup G_B) - G_{adulterant}]$$

where $\theta_{pos}(A_x,B_y)$ refers to the θ_{pos} score calculated between genre x present in treebank A and genre y present in treebank B.

Regardless of the genre composition of the treebanks under consideration, the treebanks with $\theta_{pos} \leq 0.5$ are termed as consistent in their POS annotation. Similarly, the treebanks with $\theta_{pos} \geq 4.0$ are termed as inconsistent in their POS annotation. In case of multiple genres present in either treebank, Equation 9 can be employed if just the percentage composition of different genres in the treebanks is known, regardless of whether it is possible to split the treebank into the constituent genres. However, for a fine-tuned estimation, it is imperative to be able to split the treebank into its constituent genres.

For treebanks with adulterant genres, the higher Θ_{pos} limit on the θ_{pos} scores can be problematic. If possible, the adulterated genres should be isolated and the annotation consistency of the treebank should be checked without presence of any adulterant genre(s).

8 Discussion and Conclusion

8.1 Using θ_{pos} to Localise Inconsistency

While the θ_{pos} measure is primarily meant to identify whether two given treebanks are consistent in their POS annotation, the measure can also be employed to localise points of inconsistency, if required.

Consider the example of two Finnish treebanks in UDv2.5, FTB and TDT. While the data in the former is composed of a single genre, *grammar-examples*, the data in the latter consists of multiple genres, including *grammar-examples*. We can observe that

$$\theta_{pos}(\text{Finnish-TDT}_{grammar-examples}, \text{Finnish-FTB}_{grammar-examples}) = 0.707 > 0.5$$

which is a clear violation of the condition as specified in Equation 4. We believe that the inconsistency in the annotation can be localised to the *grammar-examples* part of Finnish-TDT. Consequently, concentrating simply on the instances from this genre should be enough to bring the overall θ_{pos} score between the two treebanks under the Θ_{pos} limit.

8.2 Split into Constituent Genres as a Requirement

The estimation of Θ_{pos} is primarily based on the requirement that the genre composition of treebanks is known. While the limit is best estimated when the genres can be isolated and the adulterant genres identified, it is possible to get a crude estimate of the limit. For example, one can estimate all the common genres with θ_{pos} scores of 0.5, and the different genres have a θ_{pos} score of 2.0. An average of these estimates should give a crude estimate on the Θ_{pos} limit without accounting for an adulterant genre. Data with multi-genre classification can also be handled in a similar manner.

8.3 Conclusion

We proposed a numeric measure based on the KL_{cpos^3} measure (Rosa and Žabokrtský, 2015) to attest the POS annotation consistency across treebanks that allegedly follow the same guidelines, for the same language. Through the use of the measure, we sought to answer how the different treebanks of a language, with variable size and genre distributions but following the same annotation guidelines, can be compared against each other. We also defined a reliable threshold on the proposed measure that would inform the annotators if the treebanks being compared are not consistent with each other. In addition, the measure can also be used intra-treebank to localize the genre(s) that cause the inconsistency with another treebank. We also evaluated different treebanks in UDv2.5 (Zeman et al., 2019) and identified the consistent and inconsistent treebank pairs based on the proposed measure. To the best of our knowledge, this is the first such measure that compares treebanks directly, without an added variable of tagger performance. At present, the measure does not allow checking for consistency in treebanks with syntactic annotation. Perhaps similar ideas might lead to a syntactic version of the measure in the future.

References

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2018. Universal Dependencies and Quantitative Typological Trends. A Case Study on Word Order. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, October. Association for Computational Linguistics.

Jayeol Chun, Na-Rae Han, Jena D. Hwang, and Jinho D. Choi. 2018. Building Universal Dependency Treebanks in Korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2194–2202, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Markus Dickinson and W. Detmar Meurers. 2003a. Detecting Errors in Part-of-speech Annotation. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 107–114, Stroudsburg, PA, USA. Association for Computational Linguistics.

Markus Dickinson and W. Detmar Meurers. 2003b. Detecting Inconsistencies in Treebanks. *IEEE Transactions on Learning Technologies - TLT*, 01.

Markus Dickinson and W. Detmar Meurers. 2005. Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 322–329, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, number 155, pages 53–66, Linköping, Sweden. Linköping University Electronic Press.

Tuomo Kakkonen. 2006. Dependency treebanks: methods, annotation schemes and tools. In *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, pages 94–104, Joensuu, Finland, May. University of Joensuu, Finland.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, page 114–119, USA. Association for Computational Linguistics.

Héctor Martínez Alonso and Daniel Zeman. 2016. Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, (57):91–98.

Héctor Martínez Alonso, Željko Agić, Barbara Plank, and Anders Søgaard. 2017. Parsing Universal Dependencies without training. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 230–240, Valencia, Spain, April. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Çağrı Çöltekin, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gokirmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Simon Krek, Veronika Laippala, Lucia Lam, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav

Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Loganathan Ramasamy, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jing Xian Wang, Jonathan North Washington, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2016a. Universal Dependencies 1.3. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016b. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Shoval Sadde, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2018. Universal Dependencies 2.2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*. European Language Resources Association (ELRA).

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096, Istanbul, Turkey, May. European Languages Resources Association (ELRA).

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden, May. Association for Computational Linguistics.

Rudolf Rosa and Zdeněk Žabokrtský. 2015. KLcpos3 - a Language Similarity Measure for Delexicalized Parser Transfer. 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 243–249, Beijing, China, July. Association for Computational Linguistics.

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2014. HamleDT: Harmonized Multi-Language Dependency Treebank. *Language Resources and Evaluation*, 48(4):601–637.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima

Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal Dependencies 2.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Appendix A: θ_{pos} Scores for UDv2.5 Treebanks, Annotated to Mark Consistent and Inconsistent Treebanks

This appendix lists the θ_{pos} scores in the UDv2.5 data (Zeman et al., 2019) with the annotations used as per Table 11. In the listing of scores, small treebanks where the total number of sentences is 1,000 or less are not included.

Color	Significance
Red	Inconsistent in POS Annotation
Green	Consistent in POS Annotation
Gray	Could Not Be Estimated

(a) Color Codes Used for Scores

Superscript	Significance
Asterisk (∗)	Cannot split into constituent genres
Dagger (†)	Adulterant Genre(s) Present

(b) Superscripts against Treebank Names

Table 11: Annotations Used in Table 13

Treebank1	Treebank2	θ_{pos}
Ancient_Greek-Perseus	Ancient_Greek-PROIEL	4.641
Arabic-NYUAD	Arabic-PADT	2.497
Dutch-Alpino	Dutch-LassySmall	0.664
Chinese-GSD	Chinese-HK	1.958
Estonian-EDT	Estonian-EWT	0.413
Finnish-FTB	Finnish-TDT	1.195
*Galician-CTG	Galician-TreeGal	0.714
Japanese-BCCWJ	*Japanese-GSD	0.951
*Korean-GSD	Korean-Kaist	2.56
Polish-LFG	*Polish-PDB	0.623
Portuguese-Bosque	*Portuguese-GSD	0.678
Romanian-Nonstandard	Romanian-RRT	1.233
†Slovenian-SSJ	Slovenian-SST	2.405
Spanish-AnCora	Spanish-GSD	0.352
Swedish-LinES	Swedish-Talbanken	0.443
Turkish-GB	Turkish-IMST	1.477

German	*GSD	*HDT
*HDT	0.49	-
LIT	1.383	1.1

Latin	ITTB	†Perseus
†Perseus	1.106	-
PROIEL	3.763	3.901

Norwegian	Bokmaal	Nynorsk
Nynorsk	0.095	-
NynorskLIA	2.291	2.375

Czech	CAC	CLTT	FicTree
CLTT	1.453	-	-
FicTree	1.138	2.657	-
PDT	0.373	1.935	1.006

Russian	*GSD	†Taiga
†Taiga	1.027	-
SynTagRus	0.567	0.631

English	EWT	GUM	LinES	ParTUT
GUM	0.26	-	-	-
LinES	0.407	0.455	-	-
ParTUT	0.62	0.432	0.581	-
ESL	0.592	0.799	0.564	0.823

French	†FQB	*GSD	†ParTUT	Sequoia	Spoken
*GSD	1.582	-	-	-	-
†ParTUT	1.942	0.683	-	-	-
Sequoia	1.693	0.248	0.524	-	-
Spoken	3.644	3.089	2.599	2.732	-
FTB	2.226	0.379	0.7	0.272	3.507

Italian	ISDT	ParTUT	*VIT	PoSTWITA
ParTUT	0.133	-	-	-
*VIT	0.121	0.194	-	-
PoSTWITA	1.67	1.478	1.764	-
TWITTIRO	1.501	1.376	1.594	0.347

Table 13: θ_{pos} Scores in UDv2.5 Marked for Consistency or Inconsistency in POS Annotation

Table 14 marks the Θ_{pos} limit for treebanks that were marked as inconsistent in the table above. We omit the Θ_{pos} limit for Ancient_Greek treebanks, since the reported θ_{pos} score for the treebanks in the language exceed the hard limit of 4.0.

Treebank Pair	θ_{pos}	Θ_{pos}	**Comments**
Arabic-NYUAD & Arabic-PADT	2.497	0.5	Same Genre Violation of Equation 4
Czech-CAC & Czech-CLTT	1.453	1.388	No Adulterant Genre Violation of Equations 4, 7
Czech-CLTT & Czech-FicTree	2.657	2.0	One Genre Each Violation of Equation 5
Czech-CLTT & Czech-PDT	1.935	1.688	No Adulterant Genre Violation of Equation 7
Finnish-FTB & Finnish-TDT	1.195	1.187	No Adulterant Genre Violation of Equations 4, 7
French-FTB & French-Spoken	3.507	2.0	One Genre Each Violation of Equation 5
French-Sequoia & French-Spoken	2.732	2.0	No Adulterant Genre Violation of Equations 5, 7
Latin-ITTB & Latin-PROIEL	3.763	1.25	No Adulterant Genre Violation of Equations 4, 5, 7
Latin-Perseus & Latin-PROIEL	3.901	3.625	Adulterant Genre Violation of Equations 4, 5, 7
Norwegian-Bokmaal & Norwegian-NynorskLIA	2.291	2.0	No Adulterant Genre Violation of Equations 5, 7
Norwegian-Nynorsk & Norwegian-NynorskLIA	2.375	2.0	No Adulterant Genre Violation of Equations 5, 7

Table 14: Comparison of θ_{pos} Score and Θ_{pos} Limit for Pairs of Treebanks Marked as Inconsistent in Table 13

There are a few important points that need to be specified here:

1. The affiliation of individual sentences in any given treebank is optional and not standardized. If the `README.md` file associated with a treebank in question does not specify how to split the treebank into the constituent genres, the information can be queried through the data providers of the treebank in question. Turkish-IMST could not be assessed for the annotation consistency with the other Turkish treebank as the information on their genre split could not be fetched through either source.

2. While the methods that we discussed can be applied for estimations across different guidelines, care must be taken while estimating the empirical upper bound for a new guideline. If the estimated value of Θ_{pos} is too large, we run the risk of saying the treebanks are harmonious even when they might not be. Also, if the value is too small, we could be overlooking at the effect of domain change and dataset size, to mistakenly announce the pair of treebanks as being non-harmonious to each other.

B Appendix B: Working Example to Mark Pair of Treebanks as Consistent or Inconsistent in POS Annotation

We demonstrate the calculation of Θ_{pos} in the case of the Latin-ITTB and Latin-PROIEL treebanks. Neither of them contains any adulterant genre. The sentence and word count statistics for the two treebanks can be seen in Table 15. The calculated θ_{pos} scores across genres in the two treebanks are shown in Table 16.

Treebank (A)	**Genre (x)**	$size(A_x)$	$TotalSyntacticWords(A_x)$	$AvgSentLen(A_x)$
Latin-ITTB	*nonfiction*	21,011	353,035	16.802
Latin-PROIEL	*nonfiction*	6,626	90,600	13.673
Latin-PROIEL	*bible*	11,785	109,563	9.297

Table 15: Statistics of constituent genres in Latin-ITTB and Latin-PROIEL

$TreebankA_{GenreA}$	$TreebankB_{GenreB}$	$\theta_{pos}(TreebankA_{GenreA}, TreebankB_{GenreB})$
Latin-ITTB$_{nonfiction}$	Latin-PROIEL$_{bible}$	3.702
Latin-ITTB$_{nonfiction}$	Latin-PROIEL$_{nonfiction}$	3.558
Latin-ITTB$_{nonfiction}$	Latin-PROIEL$_{nonfiction,bible}$	3.763

Table 16: Calculated θ_{pos} for different genres in Latin-ITTB and Latin-PROIEL

From Table 16, we notice

1. θ_{pos}(Latin-ITTB$_{nonfiction}$, Latin-PROIEL$_{nonfiction}$) $= 3.558 > 0.5$, which is a violation of Equation 4
2. θ_{pos}(Latin-ITTB$_{nonfiction}$, Latin-PROIEL$_{bible}$) $= 3.702 > 2.0$, which is a violation of Equation 5

Given the θ_{pos} score calculations, we can estimate the Θ_{pos} threshold in accordance with Equation 6 as follows:

$$\begin{aligned}
\theta_{pos}(\text{Latin-ITTB}_{nonfiction}, \text{Latin-PROIEL}_{bible}) &= 3.702 \\
\theta_{pos}(\text{Latin-ITTB}_{nonfiction}, \text{Latin-PROIEL}_{nonfiction}) &= 3.558 \\
Average(\theta_{pos}) &= \frac{3.558 + 3.702}{2} \\
&= 3.63 \\
\Theta'_{pos}(\text{Latin-ITTB}_{nonfiction}, \text{Latin-PROIEL}_{nonfiction,bible}) &= \frac{0.5 + 2.0}{2} \\
&= 1.25
\end{aligned}$$

$$\boxed{\Theta_{pos}(\text{Latin-ITTB}_{nonfiction}, \text{Latin-PROIEL}_{nonfiction,bible}) = Minimum(Average(\theta_{pos}),\ \Theta'_{pos},\ 2.0) = 1.25}$$

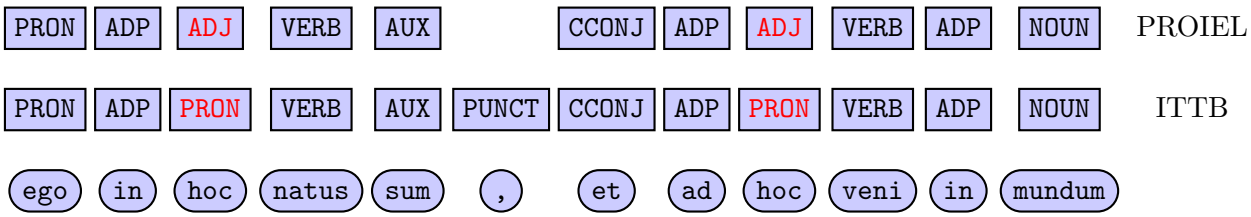
We observe that the calculated θ_{pos} score exceeds the estimated Θ_{pos} threshold, thereby judging the pair of treebanks as **inconsistent** in their POS annotation.

To further validate that the two treebanks are not consistent in their POS annotation, consider the following sentence present in either treebank.[11] The difference in annotation is shown beneath the example.

[11] Latin-IITB contains the sentence as such, without any modifications, while the sentence in Latin-PROIEL is without punctuation marks.

(1) *ego in hoc natus sum , et ad hoc veni in mundum , ut testimonium perhibeam*
I in this born am , and to this I came in world , that testimony I bestow
veritati .
to truth .

'*I was born, and for this I came into the world, to testify to the truth.*'



The Latin-PROIEL treebank's lack of punctuation marks is also well reflected in its trigram distribution. Table 17 shows the 10 most frequent POS trigrams in different Latin treebanks listed in order of their frequency in the corresponding treebank.

Latin-PROIEL		**Latin-Perseus**		**Latin-ITTB**	
POS Trigram	**Freq (%)**	**POS Trigram**	**Freq (%)**	**POS Trigram**	**Freq (%)**
NOUN VERB #	1.06	VERB PUNCT #	4.65	NOUN PUNCT #	2.086
VERB ADP NOUN	1.005	NOUN VERB PUNCT	3.604	VERB PUNCT #	1.976
NOUN CCONJ NOUN	0.843	NOUN PUNCT #	2.114	NOUN VERB PUNCT	1.702
NOUN NOUN VERB	0.787	NOUN NOUN VERB	1.541	VERB ADP NOUN	1.374
ADP NOUN VERB	0.77	VERB NOUN PUNCT	1.469	ADP NOUN PUNCT	1.104
# CCONJ VERB	0.735	ADJ NOUN VERB	1.174	NOUN NOUN PUNCT	0.993
NOUN ADP NOUN	0.726	ADJ VERB PUNCT	1.095	NOUN ADP NOUN	0.844
ADP NOUN NOUN	0.692	VERB VERB PUNCT	1.081	NOUN ADJ PUNCT	0.836
ADJ NOUN VERB	0.615	VERB NOUN NOUN	0.988	ADP NOUN NOUN	0.811
ADP NOUN ADJ	0.606	NOUN VERB NOUN	0.982	ADJ NOUN PUNCT	0.772

Table 17: Most Frequent POS Trigrams in Different Latin Treebanks with Frequency Percentage
Note: # denotes the POS of sentence boundary token

From the table, the reason of Latin-PROIEL treebank being inconsistent in annotation with the other two is clear. While the POS tag associated with punctuation (`PUNCT`) contributes to at least 6 of the top 10 trigrams in Latin-Perseus and Latin-ITTB, the POS tag (and therefore the trigrams) is missing in Latin-PROIEL.