Cross-Dataset Inconsistencies in Morphological Annotation: Evidence from Universal Dependencies

Vlasta Ohlídalová
Natural Language Processing Centre, Masaryk University, Brno, Czech Republic
Lexical Computing, Brno, Czech Republic
vlasta.ohlidalova@sketchengine.eu

Abstract

Ensuring annotation consistency is a challenging task in language dataset development. While difficulty is typically increasing at higher levels of linguistic complexity, we show that it is a critical issue even for fundamental linguistic tasks such as morphological annotation. Contrary to previous research that targeted intra-dataset inconsistencies, this study investigates inconsistencies across various pre-existing datasets for the same language. On the example of Universal Dependencies datasets, we examined what morphological categories exhibit the most disagreement. The analysis revealed that there are specific categories with low inconsistency score that indicates good agreement on these features (namely Case, Gender, Number and to a lesser extent Animacy). On the other hand, the Part-of-Speech (UPOS) tag stands out as a "red flag" due to high inconsistency score. Analysis of the most frequent inconsistencies suggest that they are dataset-specific artifacts rather than inherently language-specific phenomena.

Keywords: Morphological Analysis, POS Tagging, UD, Annotation Consistency

1. Introduction

The issue of annotation inconsistency, both in general corpus linguistics and specifically within the Universal Dependencies (UD) framework, has been continuously investigated in the past. Research efforts have been primarily directed toward the syntactic level, serving several key purposes:

- Most commonly, research focuses on developing methods to identify and flag inconsistencies within individual UD treebanks, providing a mechanism for data cleaning and error correction that serves the purpose of improving overall dataset quality.
- Exploiting the cross-lingual potential of UD datasets: UD datasets are created "with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective" (Nivre et al., 2020). As cross-lingual consistency is crucial for such tasks, the issue of annotation inconsistencies has arisen in works such as (Kanayama and Iwamoto, 2020).
- Enhancing Existing Treebanks: Efforts have been made to utilize UD principles to enhance the consistency and detail of dependencies within pre-existing, language-specific treebanks. (Nivre et al., 2018)

Compared to the syntactic level, the analysis of morphological annotation inconsistencies in the UD framework remains rather underexplored.

Prior work in this domain has been limited to either specific languages (e.g. analysing the annotation differences among various Hungarian datasets (Dömötör et al., 2025)) or to the topic of a universal linguistic typology system (Nivre and Croft, 2025; Brosa Rodríguez and López, 2025). On top of that, the works typically focus on the Part-of-Speech category alone and other morphosyntactic categories are completely omitted.

This is also true in, to the best of our knowledge, the most comprehensive investigation of morphological annotation inconsistencies that has been provided in (Aggarwal, 2020). This research introduced a metric designed to compare various treebanks of the same language – treebanks that differed in size and genre distributions, but adhered to the same annotation guidelines. As the goal was to provide an overall consistency score, the author did not provide a detailed breakdown of specific error types or the linguistic origins of the observed inconsistencies.

Our approach differs, as our primary objective is to perform a fine-grained assessment of inconsistencies within specific morphological categories[1]. We aim to answer two main questions:

- Are specific categories, across various datasets and languages, more prone to inconsistencies? Identifying such categories would then serve as a warning to annotators and corpus developers, highlighting areas requiring extra attention during the creation of new treebanks. Or conversely, are the

[1] For the purpose of this study, "morphological annotation" comprises of both the POS category (defined by UPOS in UD) and each individual morphological feature available in the Feats column.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11715–11723
11–16 May 2026.

observed inconsistencies predominantly language-specific, or even dataset-specific?

- On the other hand, are there specific categories with high consistency across all datasets? It was previously shown that simply combining two (or more) corpora of the same language not only failed to improve the results of a POS tagger, but even degraded them due to inconsistent annotations (Dömötör et al., 2025). However, if we found categories that are consistent across all datasets and used only these, perhaps it would be possible to overcome this problem.

2. Methodology

2.1. Tools

Universal Dependencies (de Marneffe et al., 2021) is a linguistic framework for consistently annotated datasets. Currently, it contains more than 200 treebanks in over 150 languages. As the number implies, many languages are covered more than once, with as many as 12 datasets for English.

In this work, we utilize the Stanford NLP Group's official Python NLP library Stanza (Qi et al., 2020). This is one of the most widely adopted and actively maintained Part-of-Speech (POS) taggers. While relying on a single tool's output could be debated, preliminary experiments conducted for Czech showed highly similar results when UDpipe was utilized instead of Stanza.

We use UD v2.15 treebanks, as that's the dataset version used for training the most recent Stanza models.

The analysis includes all languages for which Stanza provides at least two models, ensuring a minimum basis for inter-dataset comparison. The 29 languages analysed, along with the number of available models (datasets) for each, are listed below: Ancient Greek (2), Armenian (2), Czech (4), Dutch (2), English (6), Estonian (2), Finnish (2), French (5), Galician (2), German (2), Greek (2), Hebrew (2), Icelandic (3), Indonesian (2), Irish (2), Italian (7), Japanese (2), Korean (2), Latin (5), Lithuanian (2), Persian (2), Polish (2), Portuguese (3), Romanian (3), Russian (3), Slovenian (2), Spanish (2), Swedish (2), Turkish (7).

2.2. Methodology

For every language in our analysis, we utilized each model to evaluate all corresponding test sets – both the test set it was trained on and test sets from other datasets of the same language.

As expected, considerably higher scores were obtained when the test set belonged to the same dataset the model was trained on. Table 1 illustrates these score differences for the Czech language (using the UDpipe POS tagger). We assume that morphological categories exhibiting a larger percentage drop in score during cross-dataset testing are those with lower inter-dataset consistency. To ensure fair representation across categories with varying frequencies, we base our analysis on the proportional error rate of each category compared to other categories (as opposed to the raw error frequency, which would favor more frequent categories).

For a given category (C), the inconsistency score is defined as:

$$\text{IS}(C) = \text{ErrRate}_{test=train}(C) - \text{ErrRate}_{test!=train}(C)$$

A higher score therefore indicates lower consistency for the category.

We adopted two complementary approaches to analyse inter-dataset inconsistencies, with each method serving a different goal.

The first approach is a holistic approach that treats any change as an error, regardless of whether a category was completely omitted, added, or assigned a different value. The objective is to reveal the most important divergence between datasets, including both major structural/technical differences (e.g. whether a feature is annotated at all) and linguistic disagreements.

The second approach only analyses instances where a feature is present in both datasets but has a difference in its assigned value. This way, we can isolate and highlight linguistic phenomena that are inherently difficult to categorize consistently.

We utilized both options to thoroughly identify inconsistencies and will summarize the specific advantages and disadvantages of each approach in Section 4.3.

	PDT	CAC	CLLT	FicTree
PDT	98.31	81.21	74.13	68.82
CAC	80.42	97.77	81.16	73.77
CLLT	77.4	86.45	94.13	79.34
FicTree	76.24	80.16	87.47	96.89

Table 1: AllTags evaluation in cross-dataset settings on Czech using the UD pipe toool.(Ohlídalová et al., 2025) Colums refer to datasets, while rows refer to models.

3. Results: Overview

Figure 2 presents the final inconsistency scores for key morphological categories across all 29 languages included in the study. For clarity reasons, categories with negligible inter-dataset differences

or those used by only a few languages were omitted.

3.1. Morphological categories

3.1.1. High consistency categories

Overall, case emerges as the most consistent category across the entire language set. This result is primarily driven by high consistency (low negative scores) in languages such as German (-18.1), Greek (-12.0), Czech (-11.7), Turkish (-10.9), and Slovenian (-10.1). A notable exception to this trend is Polish (+1.1), which exhibits the highest inconsistency for case, surprisingly diverging from the low scores observed in other Slavic languages. The categories of Gender, Number, and Animacy also consistently demonstrate relatively high inter-dataset agreement.

3.1.2. Outliers and feature-specific issues

While most results are relatively unified, certain categories revealed significant outliers. For VerbForm, English (-7.3) acts as a consistent outlier, showing unusually high agreement, while Armenian (+8.5) presents a major negative outlier. This Armenian score is likely attributable to one dataset omitting the Gerundive value, which accounts for approximately 10% of VerbForm annotations in the other Armenian dataset.

3.1.3. The Part-of-Speech tag inconsistencies

The most critical finding is that the Part-of-Speech (UPOS) tag is the most problematic attribute in majority of languages. Although UPOS is often considered the most important piece of annotation, its consistency score frequently reaches up to 30 points for many languages. Only three languages – Armenian, Indonesian, and Italian – demonstrated relatively good consistency for the POS tag compared to their other attributes.

This widespread inconsistency highlights a significant challenge in the annotation process: the UPOS tag, being the only obligatory category for all tokens, is fully exposed to all annotation disagreements. Unlike optional features, where annotators might simply omit a feature they are unsure about (thereby masking inconsistency), any uncertainty in UPOS results in a concrete, potentially diverging, label.

3.2. Specific values

We conducted a deeper analysis by examining specific tag values. They will be discussed in this format: `Category=Correct Value->Wrong Value` (e.g., `POS=SYM->PROPN`, meaning the correct value `SYM` was replaced by the incorrect value `PROPN` in the Part-of-Speech category).

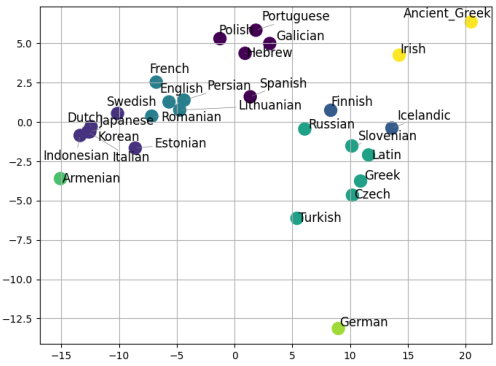

Figure 1: Language similarity based on the (in)consistency scores for each category.

A review of the most frequent errors suggests that mistakes generally cannot be easily generalized across the entire dataset. Most of them are highly language-specific or even dataset-specific. For instance, the error with highest inconsistency score, `POS=SYM->PROPN`, is unique to Italian, while the next one, `POS=PUNCT->PROPN` is not only specific to English, but namely to the Atis model. There are also inconsistencies typical for more than one language such as `POS=PART->ADV` that can be found in Czech, Russian and Latin.

The only specific error pattern that demonstrates a widespread issue across most languages is the misclassification `POS=NOUN->PROPN`. This is likely caused by the fact that the distinction between common nouns (`NOUN`) and proper nouns (`PROPN`) lies more in semantics (what the word refers to) than in morphosyntactic features (such as inflection or agreement), making it inherently difficult to annotate consistently across resources.

3.3. Conclusion on consistency and dataset specificity

One of our initial goals was to determine whether certain morphological categories are inherently more prone to annotation inconsistencies. Based on the analysis of consistency scores and the clustering results shown in Figure 1, we draw the following conclusions: some categories are typically more inconsistent – most notably the Part-of-Speech (UPOS) tag – as opposed to others – specifically Case, Gender, Number, and to a lesser degree Animacy. However, the specific causes for the inconsistencies within these categories are rather dataset-specific.

The visualization in Figure 1 plots languages

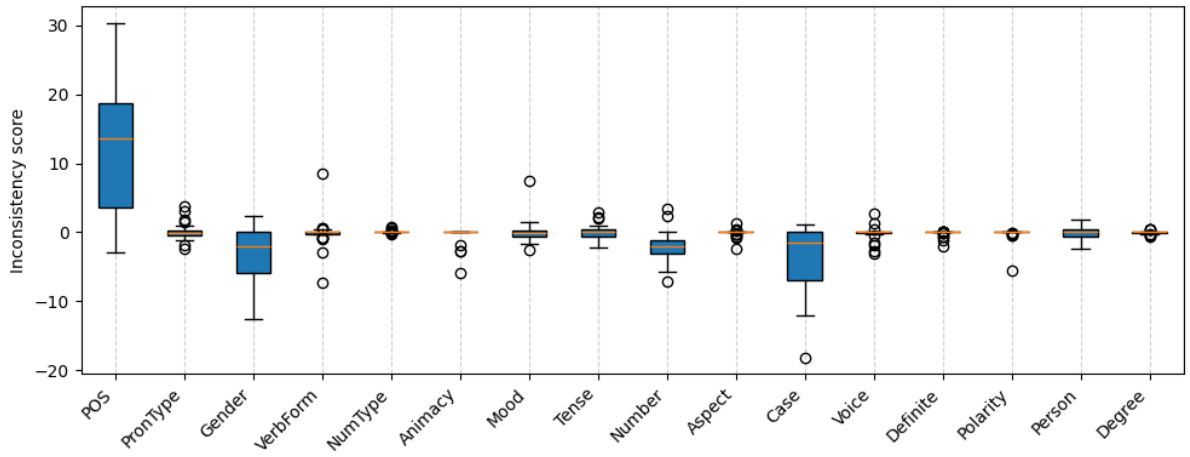

Figure 2: Inconsistency scores for morphological categories across all 29 languages in the study. Higher score means less consistent category.

into a 2D graph via PCA[2] based on their inconsistency profiles and clusters them into five groups using K-Means. Although some linguistically related languages cluster together as expected, the overall grouping is highly fragmented. An example is the Slavic language family, represented by Czech, Russian, Slovenian, and Polish. While Czech, Russian, and Slovenian are clustered into a single group, Polish is positioned far away and grouped with languages like Portuguese, Galician, Hebrew, and Spanish. This divergence mirrors our previous finding that Polish exhibited the lowest consistency score for the case category, contradicting the high consistency observed in the other Slavic languages.

The findings indicate that categorical inconsistencies are not primarily attributable to the intrinsic linguistic properties of specific languages or language families. On the contrary, they appear to be dataset-specific. A factor that is difficult to isolate and likely influenced the results is the automated conversion of datasets from original formats that may have followed diverse annotation guidelines.

4. Inconsistencies between datasets: A case study

In the next section, we will describe the most notable inconsistencies between datasets of two languages: Czech and Polish.

4.1. Polish

For Polish, these two datasets were compared: PDB (Wróblewska, 2018) with 350 thousand tokens and LFG (Patejuk and Przepiórkowski, 2018) with 130 thousand tokens. Inconsistency scores in Table 2 are presented both in the version including the omitted/added categories (Score1) and only counting value transitions (Score2).

4.1.1. Gender

The findings, as presented in Table 2, reveal substantial discrepancies within the Animacy and SubGender categories. This divergence is largely due to the established classification of masculine genders in the Polish language, as documented on the UD website. Specifically, Polish distinguishes between human masculine, animate masculine, and inanimate masculine genders. In the PDB treebank, this classification is encoded via the Animacy attribute. Conversely, the LFG treebank incorporates a language-specific SubGender feature to manage this subclassification. Given that gender is a feature present across numerous Part-of-Speech categories (including nouns, adjectives, and verbs), this difference accounts for a significant portion of the overall attribute divergence.

4.1.2. NumType and NumForm

Discrepancies related to the NumType category are tied to the value "sets", which is utilized in the PDB dataset for both paired body organs and the lemma "dziecko" (children). This specific tag is consequently assigned to both the numeral and the associated noun within the phrase.

Furthermore, the PDB dataset employs the NumForm attribute to distinguish between numerals written as words and those expressed as digits. Numerals written as words do not receive the NumType attribute and the corresponding cardinal value.

[2]The first two principal components explained 84.9% of the total variance in the dataset with PC1 = 73,4% and PC2 = 11.4%.

Category	Score1	Score2
Gender	-16.14	-5.13
POS	-8.86	**12.11**
Case	-8.17	**1.13**
Aspect	-6.55	-0.22
Polarity	-3.28	-0.09
VerbForm	-3.24	-0.11
Number	-2.95	**2.4**
Degree	-2.49	0.06
Voice	-1.72	-0.09
PronType	-1.19	-0.22
Tense	-1.15	-0.3
Number[psor]	-0.43	-0.24
Person	-0.29	0.04
Mood	0.09	0.25
PunctType	0.44	**3.12**
PunctSide	**0.88**	**0.9**
NumForm	**1.35**	_
NumType	**1.52**	0.04
SubGender	**25.39**	_
Animacy	**26.74**	_

Table 2: The most and least consistent categories in Polish datasets.

One more distinction in the Number category is the value "Ptan" used within PDB dataset to mark pluralia tantum. These are included in the common "plural" category in the LFG datase.

4.1.3. Punctuation

The PunctSide category is designed to classify punctuation marks as either being in the initializing or finishing position of a phrase or sentence. While the LFG dataset applies this distinction consistently, the PDB dataset only differentiates this category for brackets, failing to do so for quotation marks. This difference is likely an artifact of implementation complexity: marking brackets based on their form is straightforward, whereas the automatic classification of quotation marks poses a more substantial challenge.

4.1.4. POS level

In this subsection, a few examples of inconsistencies within the category of Part-of-Speech (UPOS) are described:

- więc (so) is labelled as a `CCONJ` in the LFG dataset, but as a `SCONJ` in the PDB dataset.
- Words representing large numerical units (million, thousand, ...) often fall between the categories of noun and numeral. The PDB dataset resolves this ambiguity by labelling them as noun, whereas the LFG dataset tags them as numerals.
- Jak (how): This is a notorious case of a token that is difficult to categorize. The PDB dataset predominantly tags it as `SCONJ` while the LFG dataset more often classifies it as an `ADV`.
- There is an interesting discrepancy involving multi-word expressions that function semantically as prepositions but are structurally composed of an adposition and noun. In some cases, LFG dataset chooses the semantic representation, tagging the entire phrases as adposition.

4.2. Czech

We selected Czech for our second language analysis, which is represented in Universal Dependencies by four datasets: PDT (Hajič et al., 2018) with 1.9 million tokens, CAC (Hladka et al., 2008) with 494 thousand tokens, FicTree (Jelínek, 2017) with 167K tokens, and CLTT (Kříž and Hladká, 2018) with 36K tokens.

During our preliminary analysis, we found that the CLTT dataset, despite its small size, introduced a surprisingly high number of inconsistencies due to severe under-representation of basic linguistic features. Because CLTT consists of highly specialized texts, it lacked training examples for elemental phenomena – for instance, containing no tokens tagged with `Person=1` or `Person=2`. Consequently, models trained on CLTT performed poorly when evaluated against more generic test sets.

This issue highlighted a flaw in the comparative method when applied to highly genre-specific data, as the resulting "inconsistency score" primarily reflected a lack of training coverage rather than true annotation disagreement. Therefore, in this section, where our main goal is to identify annotation differences across the datasets rather than genre differences, we excluded the CLTT dataset. That's only the case for this part, in the overall results, the dataset is present.

4.2.1. Numerals

There are two main areas of inconsistencies between nouns and numerals: denominators of fractions and words for large quantities.

Official guidelines align with traditional Czech grammar, classifying denominators of fractions (polovina, třetina, čtvrtina, pětina, ...) as a distinct class of cardinal numerals. However, this rule is applied inconsistently across the datasets: CAC dataset adheres to the guidelines, tagging these words as `NUM`, PDT categorically tags all fraction denominators as `NOUN` and lastly, in the FicTree dataset, the word "polovina" (half) is considered a noun, while the other fragments are tagged as numerals.

For large quantity words (tisíc, milión, miliarda, ...), guidelines suggest that they should be considered nouns if they inflect like nouns, although the boundaries are acknowledged to be fuzzy.

In practice, they are always annotated as numerals in the PDT dataset, while the other two datasets use both POS categories, likely trying to distinguish their usage according to the guidelines.

A final, more technical, inconsistency involves the NumForm feature. The PDT dataset uniquely recognizes the Roman value for this feature (e.g., I, V, X), whereas the CAC and FicTree datasets use the more general Digit value for such forms.

4.2.2. Adverbs and function words

In Czech, the boundaries between several Part-of-Speech categories are rather unclear, leading to significant annotation difficulties. This is either due to words legitimately belonging to multiple categories depending on their context, or their systemic position being undecided.

The most common areas of overlap are found between adverbs and participles (jen, nejen, především, třeba, již, zřejme, dokonce, ...), adverbs and subordinating conjunctions (dokud, jakoby, jak, ...) and finally between coordinating conjunctions and participles (tedy, ovšem, ani, i, ...).

4.2.3. Various missing attributes

Beyond POS disagreements, various Czech datasets exhibit inconsistencies through the omission of specific attribute values or features.

An example of this would be the NameType attribute used for proper nouns, as the PDT dataset does not use the surname value. All personal names are uniformly annotated as given names, introducing false negatives when models trained on other datasets are applied to PDT.

For verbs, a special group of verbal nouns is annotated within both the PDT and CAC datasets, but it is omitted in the FicTree dataset. The FicTree dataset also disregards the aspect category if a perfective verb appears in its finite form.

On the other hand, FicTree is the only dataset that annotates animacy for personal pronouns, as long as the distinction is linguistically possible.

4.3. Findings

The differences observed across datasets for the same language fall into two distinct categories, reflecting fundamentally different types of annotation inconsistencies.

Category	Score
POS=PROPN->NOUN	1.26
POS=ADV->PART	1.31
POS=NOUN->PROPN	1.53
POS=PART->ADV	2.14
POS=NUM->NOUN	2.32
NameType=Giv->Sur	2.45

Table 3: The least consistent substitutions for Czech when additions/omission are not included.

Category	Score
POS=NUM->NOUN	1.21
VerbForm=Vnoun->_	2.23
Aspect=Perf->_	3.5
Animacy=Anim->_	3.79
Aspect=_->Perf	4.96
Aspect=Imp->_	5.37
Aspect=_->Imp	6.05

Table 4: The least consistent substitutions for Czech when additions/omissions are included.

4.3.1. Inconsistency of scope: omission or addition

The most notable category of disagreement is not *how* to annotate a feature but *whether* to annotate it at all.

In Czech, this is also evident in the confusion between common nouns/proper nouns and the category X, which denotes foreign words. For example, this UPOS tag is used for some cities (Los Angeles, Las Vegas) in the PDT dataset, while similar entities are annotated as proper nouns (e.g. Washington).

While these variations might be potentially problematic for training robust cross-lingual models, they generally reflect deliberate structural decisions by the dataset creators rather than a linguistic phenomenon upon which human annotators do not agree.

4.3.2. Inconsistency of value: substitution

The second, more subtle category of inconsistencies involves the substitution of one feature value for another. Analysing these errors is more complex, as it requires in-depth linguistic knowledge of the language, but they are crucial because they pinpoint genuine, unresolved linguistic phenomena that cannot be easily categorized. These errors are almost exclusively found within the Part-of-Speech category. They highlight real areas of ambiguity where a word's position in the linguistic system is fuzzy. These are not modelling flaws but rather core challenges that researchers should be aware of when utilizing or comparing language data.

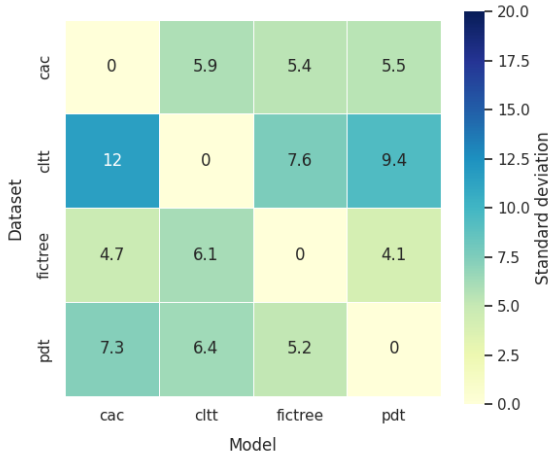

Figure 3: Dataset similarity expressed as standard deviation of inconsistency scores across morphological features for all Czech datasets.

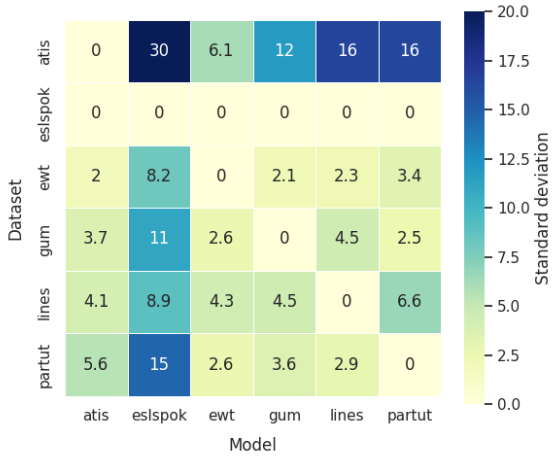

Figure 4: Dataset similarity expressed as standard deviation of inconsistency scores across morphological features for all English datasets.

5. Combining datasets based on their similarity

Figures 3 and 4 illustrate a comparative analysis of Czech and English datasets, respectively. In these figures, dissimilarity is quantified as the standard deviation of inconsistency scores across all morphological categories, calculated according to the formula in 2.2. For instance, the value of 2.1 between the EWT dataset and the GUM model represents the standard deviation of the differences in error rates observed between the intra-dataset setting EWT_{model} / EWT_{data} and the cross-dataset setting GUM_{model} / EWT_{data}.

5.1. Dataset-selective training

Regarding the English corpora, we conducted two experiments focusing on multi-dataset training. Initially, we combined all datasets while excluding those identified as significant outliers (Atis and ESLSpok). However, this configuration yielded suboptimal results across all datasets (decreases of up to 1.5% for UPOS and more than 6% for UFeats), suggesting that the degree of inter-dataset similarity was insufficient for a broad merger.

Consequently, the follow-up experiment was restricted to the most closely related datasets: EWT and GUM (exhibiting standard deviations of 2.1 and 2.6). As detailed in Table 5, this targeted combination resulted in marginal improvements for both datasets in UPOS tagging and in case of EWT, also in the UFeats metric. Conversely, the GUM dataset experienced a significant performance degradation in UFeats, decreasing by approximately 4.5%.

5.2. Feature-selective training

In alignment with prior research (Dömötör et al., 2025), our baseline results suggest that a naive concatenation of Czech datasets is ineffective. A possible reason for this could the pronounced dissimilarity scores observed between the corpora. However, these discrepancies can be mitigated by adopting a more selective approach – prioritizing a subset of the most consistent morphological attributes rather than incorporating the entire feature set.

Consequently, for the follow-up experiment, we restricted multi-dataset training to these features: Case, Number, Animacy, Gender, and Person. A major limitation of this strategy lies in the role of the Universal Part-of-Speech tag. While our data reveals that UPOS tagging is highly inconsistent across datasets, omitting it is practically unfeasible, as it typically dictates which morphological features are applicable to a given token. Consequently, these inconsistencies will inevitably affect the accuracy of the UPOS evaluation itself.

Table 5 compares this multi-dataset configuration against an intra-dataset baseline, where models were trained on individual corpora using the same restricted feature set.

The CLTT corpus was the only dataset to benefit from this selective feature approach. CLTT is a small (36K tokens), domain-specific collection of legal texts that underperformed in default settings (achieving only 97.3% for UPOS and 88.22% for UFeats). The dataset also exhibits low compatibility with other Czech datasets (see Figure 3), having previously suffered a decline of over 10 percentage points in unconstrained multi-dataset environments.

	All attributes		Consistent only	
	Intra-dataset	Inter-dataset	Intra-dataset	Inter-dataset
CAC	98.52 / 93.64	-0.35 / -6.14	98.9 / 95.39	-0.84 / -0.83
CLTT	97.38 / 88.22	-0.34 / -10.45	97.82 / 89.83	+0.55 / +3.44
FicTree	98.45 / 96	-1.27 / -11.74	98.5 / 96.55	-1.18 / -5.09
PDT	98.88 / 95.84	-0.36 / -2.34	98.81 / 96.58	-0.17 / -0.9
EWT	96.69 / 96.79	+0,49 / +0,4	–	–
GUM	97.87 / 93.05	+0,28 / -4,47	–	–

Table 5: Performance of the Stanza POS tagger on Czech and English datasets (expressed as UPOS / UFeats accuracy). Inter-dataset results are reported as the relative change compared to the intra-dataset baseline (where training and testing were performed on the same dataset).

5.3. Conclusion on combining datasets

Unsurprisingly, our experiments confirm that a simple merging of diverse datasets generally fails to yield performance improvements in morphological tagging. Improvements appear to be achievable only in highly specific scenarios:

- When datasets exhibit very low dissimilarity scores (e.g., the EWT and GUM corpora), targeted merging can provide marginal gains.
- When dealing with very small, domain-specific datasets (like CLTT), gains are possible by restricting training to a consistent subset of morphological features, effectively mitigating the noise introduced by conflicting annotations in larger corpora.

Nevertheless, further investigation would be needed to validate this trend across other domains and languages.

Another question that would require further research is whether the performance degradation we have witnessed in a simple dataset merging is due to inter-dataset inconsistencies or rather a result of varying data genres.

6. Conclusion

The consistency of morphological annotation is often an overlooked area. This study undertook a fine-grained, quantitative assessment of morphological annotation consistency across 29 languages in the Universal Dependencies framework, moving beyond the traditional focus on intra-dataset or syntactic consistency. The metric used in this study does not provide a holistic score of how consistent an entire dataset is; rather, it quantifies the inter-dataset consistency of specific morphological categories within those datasets. Our primary objective was to determine which of these individual categories is most prone to inconsistencies.

We demonstrated that consistency varies significantly by attribute. Features such as Case, Gender, and Number typically reach low inconsistency scores, indicating a high degree of stability and agreement across different datasets. Conversely, the Part-of-Speech (UPOS) tag proved to be the most problematic attribute, showing the highest inconsistency scores across the majority of languages. This finding highlights a fundamental challenge: since UPOS is the only obligatory category for all tokens, it is fully exposed to every instance of annotation uncertainty, unlike optional features where disagreement can be masked by omission.

Our analysis revealed that annotation inconsistency is predominantly a dataset-specific artifact rather than being inherent to a language typology.

The suggested metric allows for detecting the most notable inconsistencies between datasets of the same languages. These fall into two groups: inconsistencies of scope (deliberate decisions *whether* to tag specific attribute rather than *how* to tag it) and inconsistency of value (pinpointing unresolved linguistic ambiguities that require focused attention from annotators and corpus developers). Cases of both of these were demonstrated on Czech and Polish datasets.

By identifying categories that maintain high consistency (Case, Gender, Number), we offer a basis for future work aiming to overcome the consistency problem by potentially prioritizing these stable features in cross-lingual learning or corpus combinations. Conversely, the persistent issues surrounding the UPOS tag serve as a clear warning to corpus developers that this foundational category requires a lot of attention during the creation of new treebanks.

7. Bibliographical References

Akshay Aggarwal. 2020. *Consistency of Linguistic Annotation*. Ph.D. thesis, Charles University.

Antoni Brosa Rodríguez and M. López. 2025. Beyond the data: The impact of annotation inconsistencies in ud treebanks on typological universals and complexity assessment. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 43–51.

Andrea Dömötör, Balázs Indig, and Dávid Márk Nemeskey. 2025. Variety delights (sometimes) - annotation differences in morphologically annotated corpora. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 270–278, Vienna, Austria. Association for Computational Linguistics.

Hiroshi Kanayama and Ran Iwamoto. 2020. How universal are Universal Dependencies? exploiting syntax for multilingual clause-level sentiment detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4063–4073, Marseille, France. European Language Resources Association.

Joakim Nivre and William Croft. 2025. Reference and modification in Universal Dependencies. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 1–10, Ljubljana, Slovenia. Association for Computational Linguistics.

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 Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Joakim Nivre, Paola Marongiu, Filip Ginter, Jenna Kanerva, Simonetta Montemagni, Sebastian Schuster, and Maria Simi. 2018. Enhancing Universal Dependency treebanks: A case study. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 102–107, Brussels, Belgium. Association for Computational Linguistics.

Vlasta Ohlídalová, Pavel Rychlý, and Miloš Jakubíček. 2025. Are we there yet? a thorough evaluation of pos tagging on czech. In *Text, Speech, and Dialogue*, pages 263–274, Cham. Springer Nature Switzerland.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

8. Language Resource References

de Marneffe, Marie-Catherine and Manning, Christopher D. and Nivre, Joakim and Zeman, Daniel. 2021. *Universal Dependencies*. MIT Press.

Hajič, Jan and Bejček, Eduard and Bémová, Alevtina and Buráňová, Eva and Hajičová, Eva and Havelka, Jiří and Homola, Petr and Kárník, Jiří and Kettnerová, Václava and Klyueva, Natalia and Kolářová, Veronika and Kučová, Lucie and Lopatková, Markéta and Mikulová, Marie and Mírovský, Jiří and Nedoluzhko, Anna and Pajas, Petr and Panevová, Jarmila and Poláková, Lucie and Rysová, Magdaléna and Sgall, Petr and Spoustová, Johanka and Straňák, Pavel and Synková, Pavlína and Ševčíková, Magda and Štěpánek, Jan and Urešová, Zdeňka and Vidová Hladká, Barbora and Zeman, Daniel and Zikánová, Šárka and Žabokrtský, Zdeněk. 2018. *Prague Dependency Treebank 3.5*. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University.

Hladka, Barbora and Hajič, Jan and Hana, Jirka and Hlaváčová, Jaroslava and Mírovský, Jiří and Raab, Jan. 2008. *The Czech Academic Corpus 2.0 Guide*.

Jelínek, Tomáš. 2017. *FicTree: a Manually Annotated Treebank of Czech Fiction*.

Vincent Kříž and Barbora Hladká. 2018. *Czech Legal Text Treebank 2.0*. European Language Resources Association (ELRA).

Agnieszka Patejuk and Adam Przepiórkowski. 2018. *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Institute of Computer Science, Polish Academy of Sciences.

Wróblewska, Alina. 2018. *Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format*. Association for Computational Linguistics.