Evaluating the Quality of a Corpus Annotation Scheme Using Pretrained Language Models

Furkan Akkurt*, Onur Güngör*, Büşra Marşan†, Tunga Güngör*, Balkız Öztürk Başaran*, Arzucan Özgür*, Susan Üsküdarlı*

*Boğaziçi University
Istanbul, Turkey
{furkan.akkurt,gungort,balkiz.ozturk,arzucan.ozgur,
suzan.uskudarli}@bogazici.edu.tr,onurgu@pt.bogazici.edu.tr

†Stanford University
Stanford, CA, USA
busra@stanford.edu

Abstract

Pretrained language models and large language models are increasingly used to assist in a great variety of natural language tasks. In this work, we explore their use in evaluating the quality of alternative corpus annotation schemes. For this purpose, we analyze two alternative annotations of the Turkish BOUN treebank, versions 2.8 and 2.11, in the Universal Dependencies framework using large language models. Using a suitable prompt generated using treebank annotations, large language models are used to recover the surface forms of sentences. Based on the idea that the large language models capture the characteristics of the languages, we expect that the better annotation scheme would yield the sentences with higher success. The experiments conducted on a subset of the treebank show that the new annotation scheme (2.11) results in a successful recovery percentage of about 2 points higher. All the code developed for this work is available at `github.com/boun-tabi/eval-ud`.

Keywords: treebank annotation, large language models, Universal Dependencies, morphologically rich languages, Turkish

1. Introduction

The use of pretrained language models and large language models have led to a paradigm shift in solving several types of natural language processing (NLP) tasks. Pretrained language models like BERT (Devlin et al., 2019) and GPT (Brown et al., 2020) build a general model that encodes the characteristics of a language or multiple languages, and enable one to adapt this model to the task at hand. Large language models like ChatGPT (OpenAI, 2021) and LLaMA (Touvron et al., 2023a) go a step further. They provide models that can be used in different types of tasks in zero- or few-shot settings.

Universal Dependencies (UD) (De Marneffe et al., 2021) project is a framework that provides treebanks in a dependency grammar format (Bauer, 1979; Debusmann, 2000; De Marneffe and Nivre, 2019) in more than 100 languages (Nivre et al., 2020). It is commonly used for different purposes in the NLP community, including cross-lingual part-of-speech (POS) tagging (Parvez and Chang, 2021), semantic parsing (Reddy et al., 2017), and language identification (Toftrup et al., 2021). The UD project aims at unifying the annotations of the treebanks and arriving at consistent annotations by introducing a set of principles, universal tags and their language-specific subcategories related to morphosyntax. However, due to the varying characteristics of languages in different language families and the different theoretical frameworks followed by linguists, various approaches have emerged in annotating treebanks. Some linguistic phenomena in languages that the mechanisms in the framework cannot easily handle are attempted to be addressed in various ways by utilizing the MISC (miscellaneous) field in the CoNLL-U format. Even the treebanks of the same language may use different annotation strategies for the same linguistic phenomenon depending on the annotators' linguistic theory and assumptions. The annotation differences are typically in token splits, lemmas, part-of-speech tags and morphological features of tokens, and types of dependency relations between tokens. In addition to such conflicting annotations, the treebanks include noise at both morphological and syntactic levels.

There are different strategies to assess the quality of annotations in corpora, such as measuring the annotation agreement in a controlled setting or using the corpora in downstream tasks. The drawback of these approaches is that either they require a great deal of dedicated time and effort by area experts or suitable downstream tasks must be identified, which may require resources that are not easily attained for all languages.

In this work, we propose a novel method that

LREC-COLING 2024, pages 6504–6514
20-25 May, 2024.

makes use of large language models (LLM) in evaluating corpus annotation. As an application of the proposed approach, we compare the token-level annotations (specifically lemmas, part-of-speech tags, and morphological features) between two versions of the Turkish BOUN Treebank (Marşan et al., 2022) in the UD framework, the versions 2.8 (Türk et al., 2021) and 2.11 (Marşan et al., 2022). For each treebank, by feeding the annotations of the tokens in natural language to an LLM via a prompt, we expect the LLM to generate the tokens in order with their correct surface forms. We then compare the output sentences of both versions to determine which one is closer to the original sentence, which signals that the annotation of that sentence is of higher quality and expressivity than the other one. We show that, compared to other strategies, the proposed approach is highly efficient and applicable to any language due to the existence of multilingual LLMs.

The contributions in this work are as follows:

- A novel approach to assess the quality of corpus annotation, employing large language models;
- The application of the method to dependency treebanks to evaluate treebank annotations using a new strategy;
- A detailed analysis of the correctness of the annotations in two versions of a treebank based on several parameters, such as token recovery accuracy and morphological feature importance;
- The release of the code used for this work at `github.com/boun-tabi/eval-ud` to facilitate the use of the method for other languages and treebanks.

The remainder of this paper is organized as follows: Section 2 provides information about background information necessary to follow this work, Section 3 introduces related work, Section 4 introduces our proposed method, which is followed by experiments and results in Section 5. Finally, we conclude with Section 6.

2. Background

2.1. Universal Dependencies (UD) and BOUN Treebank

The UD Turkish BOUN treebank comprises 9,761 UD-style annotated sentences from various domains, such as biographical texts, newspaper articles, essays, instructional texts, and more. These sentences were randomly extracted from the Turkish National Corpus (Aksan et al., 2012), encompassing different registers. Importantly, the BOUN treebank data faithfully captures the unique characteristics of Turkish, which include its relatively free word order and licensing pro-drop in addition to object drop. For a comprehensive analysis of word order distribution within the BOUN treebank, refer to Table 1.

Order	Count	%
OV	5.744	37.21
SV	5.416	35.09
SOV	1.456	9.43
VS	1.116	7.23
VO	714	4.63
OVS	549	3.56
VSO	165	1.07
SVO	144	0.93
OSV	109	0.71
VOS	23	0.15

Table 1: Word order distribution in the BOUN treebank. *S: Subject, O: Object, V: Verb*

The older version of the UD Turkish BOUN treebank, v2.8 (Türk et al., 2022), had undergone a semi-automated annotation process in its inception. Initially, sentences were parsed and syntactically annotated using Kanerva et al.'s parsing pipeline, while morphological annotations were applied using Sak et al.'s morphological disambiguator. These annotations were then meticulously reviewed and improved by native Turkish speakers. It's worth noting that there were no newly introduced annotation practices or deviations from the Universal Dependencies (UD) standards in terms of annotation or lemmatization in this version.

This particular version suffered from notable limitations in terms of abstraction and expressiveness, primarily due to the disparities between the annotation framework of UD and the morphologically rich and highly syncretic nature of the Turkish language. Additionally, the prevalent use of morphophonologically null morphemes, such as the 0-morpheme copula, and highly productive derivational processes in Turkish contributed to these challenges.

UD, initially developed by speakers of morphologically sparse and highly isolating languages like Indo-European languages, adheres to a rigid and specific approach regarding word structure, derivation, lemma splitting, and null morphemes. This approach considers derivational processes and affixes to be opaque, based on the observation that in most languages derivational affixes typically precede inflectional ones. In contrast, languages such as Turkish exhibit a different pattern. For example, *-ki* which derives pronominals can follow genitive and locative case suffixes which are inflectional and further inflectional material such as the plural suffix

can be applied to the stem derived by *-ki* as shown in 2.1:

(1) ev -de -ki -ler
home LOCATIVE ki PLURAL
'[the ones] at home'

Additionally, the UD framework lacks a consistent and unified approach to handling morphophonologically null morphemes. As a result, languages featuring these morphemes have resorted to diverse strategies for annotation (Marton et al., 2013; Ravishankar, 2017; Dyer, 2022). These limitations within the UD framework have been recognized, prompting discussions and debates within the UD community (Gerdes and Kahane, 2016). Some topics like wordhood and word segmentation (Seyoum et al., 2018) remain subjects of ongoing debate in this context. In addition, researchers adopt different strategies to mitigate the shortcomings of UD in handling of the morphologically rich languages (More and Tsarfaty, 2016; Vincze et al., 2017; Seyoum et al., 2018), function words (Osborne and Gerdes, 2019), and various other language-specific phenomena like case-drop (Sundar Ram and Lalitha Devi, 2021).

Version 2.11 of the BOUN treebank addresses the challenges related to null morphemes, lemmatization, intertwining derivational and inflectional processes, and syncretism within the Turkish language, all while staying as faithful as possible to the UD framework. The primary objective behind creating v2.11 is to enhance the representational capabilities and both theoretical and practical accuracy of the BOUN treebank.

To achieve this goal, the entire treebank underwent a meticulous manual reannotation process, conducted by linguists who are native speakers of Turkish with domain expertise. Throughout this reannotation process, issues such as lemmatization errors, inconsistencies in dependency annotations, and the presence of missing or extraneous morphological features were systematically addressed by the annotators. Furthermore, novel annotation strategies were introduced (for more in-depth discussions of these strategies, please refer to Marşan et al. and Bedir et al.).

2.2. Large Language Models

The term "Large Language Model" refers to many aspects of a specific family of neural networks in a single phrase. The term "large" is relative, given that there is a larger one of every model. Moreover, we do not usually use earlier language models in the same way we use this type of language models.

Large language models are called "large" because they are large in the sense that they have tens of billions of parameters (e.g. 70B in Llama2 (Touvron et al., 2023b), 180B in Falcon (TII, 2023), 540B in PaLM (Chowdhery et al., 2022)), compare this to the figure of 213M of the biggest neural network based model for NLP in 2017 (Vaswani et al., 2017). Large language models are called "language models" because they are set up so that they can be used to predict the next word in a given sequence of words in addition to providing a fixed-length vector that represents the given sequence.

The models used in this paper are Transformer-based decoder-only architectures that were trained with corpora that usually contain trillions of words. *Transformer* is a model that proposes to represent each position in a given sequence as a function of the other words in the sequence (Vaswani et al., 2017). Moreover, this mechanism is implemented mostly through matrix multiplications which can be computed efficiently compared to previous approaches that usually involve sequential computation.

The literature in this area shows that these models are able to solve tasks that were not specifically included in the training phase, such as summarization, question answering, mathematical reasoning, even made up tasks like reversing the sequence of characters that make up a word (Brown et al., 2020).

The training of LLMs involve a high level of computational resources as the architectures usually include billions of trainable parameters. Research funding to conduct a training operation of this magnitude is usually hard to obtain. This causes researchers to rely on organizations that serve these models with mostly commercial interests (*OpenAI* and *Hugging Face* (OpenAI, 2023; Hugging Face, 2023)). This requirement restricts access to these models through the API services these organizations provide.

2.3. Evaluation of UD Resources

Evaluating language resources is a critical stage in ensuring quality, fine-tuning, and optimization. That is why significant efforts have been dedicated to evaluating and assessing UD-style treebanks (Nivre and Fang, 2017). These endeavors typically revolve around two primary aspects:

- Assessing the quality, consistency, and accuracy of annotations.
- Evaluating the performance of NLP systems trained on these resources in downstream tasks.

To evaluate annotation quality and accuracy, various metrics are employed, including head-attachment score (UAS, LAS), Kappa (McHugh,

2012), and various other inter-annotator agreement metrics. These metrics are designed to assess the consistency among different annotators who contribute to the same resource or to gauge the overall consistency across language resources in the same language that are annotated by different teams (Tyers et al., 2017; Grünewald and Friedrich, 2020).

The other aspect frequently involves utilizing a natural language processing application for a downstream task, such as parsing, POS tagging, named entity recognition, or semantic role labeling. With the emergence of more semantically-oriented approaches and foundation models, LLMs have been gaining prominence in UD evaluation tasks.

One important point to note here is that the majority, if not all, of UD evaluation tasks involving LLMs assess the abilities of these models in tasks like parsing (Al-Ghamdi et al., 2023; Kanerva et al., 2020), classification, learning and retaining syntax (Kulmizev et al., 2020; Limisiewicz et al., 2020), performing cross-linguistic tasks (Ahmad et al., 2019) or making generalizations about the structure of language (McCoy et al., 2019; Yang et al., 2019) using annotated UD data. The use of LLMs to evaluate the expressive capabilities and annotation accuracy of UD resources, or any language resources for that matter, remains relatively unexplored.

3. Related Work

Evaluating large language models is not as straightforward as evaluating other ML models. Evaluation of traditional NLP tasks might seem as simple as asking the LLM to give the name of the correct class among the possible classes in a text classification task. However, even this is complicated because the exact prompt that is used may affect the outcome. On the other hand, as LLMs are general-purpose AI tools that are expected to behave in many ways, a complete evaluation should include more than a number of accuracy results among several traditional NLP tasks. To alleviate this issue, it is suggested to assess the LLMs with scenarios that cover many dimensions like the domain, the intended users, the timeframe of the test data, and the language in addition to the task itself. In (Liang et al., 2022), it is also noted that the metrics like calibration, robustness, fairness, bias, toxicity, and efficiency should also be included.

GPT3 (Brown et al., 2020) was utilized in several different ways for a number of evaluation tasks. One of them involves using the log-probabilities reported by the system to arrive at a readability measure that correlates with the readability scores given by humans (Behre et al., 2022). A framework that uses large language models with Chain-of-Thought paradigm (Wei et al., 2022) to evaluate natural language generation outputs achieves a high Spearman correlation with human judgments for the summarization task (Liu et al., 2023). An explainable evaluation metric (Xu et al., 2023) is developed using LLaMA (Touvron et al., 2023a) and GPT4 (OpenAI, 2021) without using human feedback that shows better performance than other text generation metrics. GPT-4 was also used in evaluating automatically generated questions (Moore et al., 2023).

4. Method

In this work, we compare the annotation schemes in two versions of the UD Turkish BOUN treebank, versions 2.8 and 2.11. We ask a large language model to recover the surface form (i.e., original text) of a sentence given the lemmas, parts of speech, and morphological features in its annotation. Since models like ChatGPT (OpenAI, 2021) are able to handle natural language and associate linguistic feature signals with linguistic material, we opted to use as natural a language as possible in the prompts. In this respect, we converted the annotations in the treebank into natural language sentences. For instance, the `POS` tag of a word being `NOUN` is represented with the phrase *it is a noun* in the prompt.

The important point in guiding a large language model is deciding on a prompt that enables the model to generate the desired output. We tested alternative prompts and arrived at the one that is formed of three parts. The prompt first gives the description of the task and what follows in the following paragraphs. In this description, we also included explanations to the abstraction of copula form to ensure that LLM understands the prompts better.

After the description, we provide an example question and answer to guide the model in a one-shot setting. The example includes providing the token count of the sentence, and each token's lemma, part-of-speech tag and morphological features, if any, in separate lines. Morphological feature annotations are ordered as they would surface in a token, as applies to Turkish morphology. As the annotations of the tokens are fed to the model through these natural language sentences, we inform the LLM that its task is recovering the *surface form* of the given sentence and we give the answer for the example sentence. In the last part of the prompt, we provide the annotations of a new sentence in the same manner and ask the LLM to output the original sentence.

For each sentence, its annotations in the treebank versions 2.8 and 2.11 are queried as described above. The sentence output by the LLM

for each version is compared with the original sentence. This process is repeated for all the sentences in the set and accuracy scores for both treebanks are computed. The comparison is done by the sequence matching algorithm of Python (Foundation, 2022). We use the `SequenceMatcher` module of the `difflib` library to get a ratio of matching sequences in the two texts, the original sentence and the output sentence. After comparing each sentence in this way, we calculate an average score of accuracy for the entire set.

Table 2 shows a prompt to reconstruct the Turkish sentence “Tepelere sisler indi." (“Fogs descended on the hills."). The preamble includes a one-shot example (the sentence “Meşrutiyetin ilanından önceki siyasi faaliyetlere katıldı.” - “He/she participated in the activities prior to the constitutional monarchy.”) to illustrate the expected behavior. The CoNLL-U annotation of the example sentence is given in Table 3. The prompt only provides the annotations of the tokens to the large language model for the model to generate the surface forms of the sentence. The prompts are automatically generated by a script using a template.

In the preliminary experiments we observed that the models may output some English explanations about the recovered sentence and some filtering steps are needed to remove these explanations. For these, we used heuristic filtering functions.

5. Experiments and Results

Several experiments using different large language models were conducted to compare alternative annotations of the same treebank. Specifically, the GPT-3.5, GPT-4 (OpenAI, 2023), Claude Instant, and Claude 2 (Anthropic, 2023) models were prompted using the Poe API (Poe, 2023).

The proposed method was implemented with 500 randomly selected sentences that are consistent with the distribution of the number of tokens per sentence in the treebanks. A prompt was generated for each annotated version of these sentences as explained in Section 4. API calls were made to the above-mentioned models with these prompts and their responses were processed. Table 4 shows the accuracies by sequence matching for both versions and several models. Across all experiments, we observe a consistent increase of approximately 1.5% in accuracy for UD Turkish BOUN v2.11. This suggests that its annotations better capture the features of the sentences.

We also experimented with a treebank of another language, the UD English EWT treebank (Silveira et al., 2023), to validate the method. As GPT-4 produced the best results in the UD Turkish BOUN treebank experiments, we used only this model for this experiment. The table shows that the model yields better recovery than all the models used for Turkish. This may be regarded as an expected result since the GPT-4 model better captures the English language, and English sentences are somewhat easier to recover as they do not possess complex morphological features such as morphologically rich languages like Turkish do.

For the experiments with the Turkish treebank, we have also experimented with smaller and open-source models such as Llama 2 70B (Touvron et al., 2023a) and Mixtral 8x7B Chat (Jiang et al., 2024). The results from these models show that, they produce significantly lower accuracies compared to the GPT-4 model. This is expected since they have not intentionally been trained on Turkish, which is considerably different from English in terms of morphology and syntax. Since these models are insufficient for understanding and generating based on Turkish morphology, they were not applicable to the study at hand; thus, we don’t include their results in this paper.

Figure 1 illustrates the impact of adding and removing features during the re-annotation phase. The most substantial improvement, attributed to the unique morphological feature key and value, was achieved by removing the third-person annotation (i.e., Person=3) from a single token in the treebank during the transition from v2.8 to v2.11. This change aided in producing the correct token 288 times, while it also led to the model producing an incorrect token 189 times. The fact that the same feature key and value pair (the third-person annotation) resulted in both the most significant improvement and worsening of the performance underscores its significance in token recovery. In Turkish, the third person is represented by two exponents: a null morpheme for singular and *-lAr* for plural. The fact that, as discussed in the previous sections, addressing a significant issue in Turkish that the UD framework has overlooked has had a notable impact on token recovery underscores the necessity to reconsider certain preconceptions and assumptions made by the UD framework regarding null morphemes.

While the removal of case and number can reduce accuracy in certain instances, it generally enhances overall performance. This suggests that the v2.8 version of the treebank may have had some level of clutter. Furthermore, the consistent improvement in annotation precision is observed when new features are added. It is essential to note that the addition of new features should be undertaken judiciously to prevent clutter and the annotation of non-existing layers of meaning.

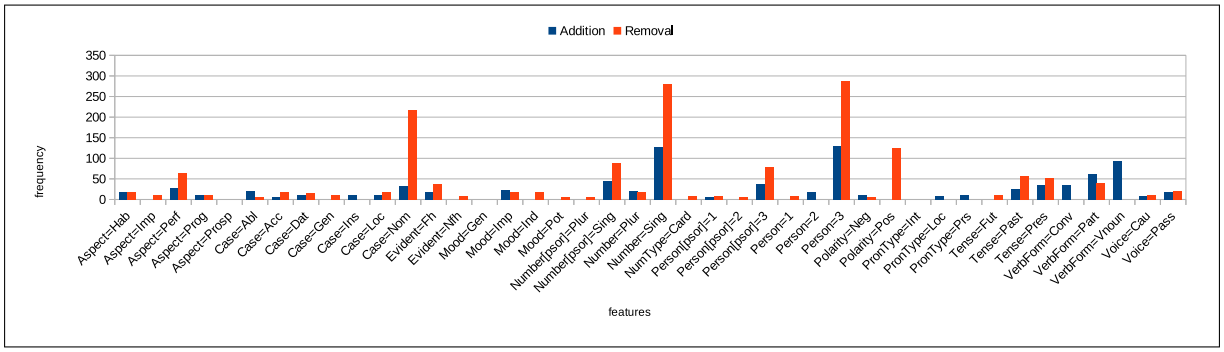
To gain insight into the impact of annotations on recovering the surface forms of the sentences, various failures have been examined. The reported accuracy scores are based on sequence matching

```
The following sentences detail linguistic features of a Turkish sentence with
lemmas, parts of speech and morphological features given for each token.  Lemma
"y" represents the overt copula in Turkish and surfaces as "i".

The sentence has 7 tokens.

1st token's lemma is "meşrutiyet", its part of speech is proper noun, its case is
genitive, its number is singular number, and its person is third person.
2nd token's lemma is "ilan", its part of speech is noun, its person is third
person, its number is singular number, its possessor's person is third person,
its possessor's number is singular number, and its case is ablative.
3rd token's lemma is "önceki", and its part of speech is adjective.
4th token's lemma is "siyasi", and its part of speech is adjective.
5th token's lemma is "faaliyet", its part of speech is noun, its person is third
person, its number is plural number, and its case is dative.
6th token's lemma is "kat", its part of speech is verb, its voice is reflexive
voice, its polarity is positive, its tense is past tense, its aspect is perfect
aspect, its person is third person, its number is singular number, and its
evidentiality is first hand.
7th token's lemma is ".", and its part of speech is punctuation.

Your task is to find the surface form of the sentence.  For example, your answer
for the previous parse should be:

"Meşrutiyetin ilanından önceki siyasi faaliyetlere katıldı."

Now, analyze the following test example and try to find the surface form of the
sentence.  It has 4 tokens.  Please include all the tokens in your answer in order.
Output only the surface form without any explanations or sentences in English.

1st token's lemma is "tepe", its part of speech is noun, its person is third
person, its number is plural number, and its case is dative.
2nd token's lemma is "sis", its part of speech is noun, its person is third person,
its number is plural number, and its case is nominative.
3rd token's lemma is "in", its part of speech is verb, its polarity is positive,
its tense is past tense, its aspect is perfect aspect, its person is third person,
its number is singular number, and its evidentiality is first hand.
4th token's lemma is "...", and its part of speech is punctuation.
```

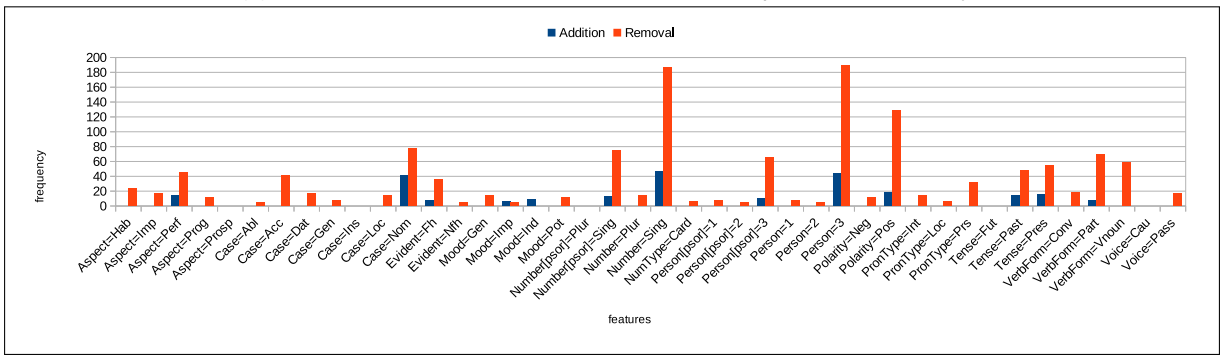
Table 2: A prompt for LLMs to reconstruct the surface form of a Turkish sentence given the annotations of its lemmas.

ID	Form	Lemma	POS	Feats
1	Meşrutiyetin	meşrutiyet	PROPN	Case=Gen | Number=Sing| Person=3
2	ilanından	ilan	NOUN	Case=Abl | Number=Sing | Number[psor]=Sing | Person=3 | Person[psor]=3
3	önceki	önceki	ADJ	—
4	siyasi	siyasi	ADJ	—
5	faaliyetlere	faaliyet	NOUN	Case=Dat | Number=Plur | Person=3
6	katıldı	kat	VERB	Aspect=Perf | Evident=Fh |Number=Sing |Person=3 | Polarity=Pos | Tense=Past | Voice=Rfl
7	.	.	PUNCT	—

Table 3: The CoNLL-U annotation, with only relevant columns, of the example sentence in the prompt.



(a) The addition or removal of the features that improved the recovery.



(b) The addition or removal of the features that reduced the recovery.

Figure 1: The impact of the addition and removal of features on the accuracy.

Model	v2.8	v2.11
GPT-3.5-Turbo	83.58	84.52
Claude-instant-100k	85.95	87.36
GPT-4	89.97	91.28
Claude-2-100k	89.00	90.53
	UD_English-EWT	
GPT-4	94.00	

Table 4: The accuracies by sequence matching of the experiments with various large language models provided for UD Turkish BOUN v2.8 and UD Turkish BOUN v2.11. The result for UD English EWT is provided for reference.

over the entire sentence. To gain a better understanding of the parts of the sentences that were not recovered accurately, the words within the sentences are examined along with their annotations. By doing so, we aimed to understand the impact of the annotations in this task. Since Turkish is a morphologically rich agglutinative language, the recoveries of the suffixes and their ordering are highly significant.

The percentage of produced tokens that matched the tokens and their order in the original sentences via prompting the GPT-4 model with the annotations from UD Turkish BOUN v2.8 and v2.11 are 73.8% and 76.9% respectively (total number of tokens is 6.247). The frequencies of the top five features for the words that were not accurately recovered are shown in Table 5.

Except for polarity, these features belong to the nominal domain of Turkish, which is characterized by a high degree of syncretism. For example, the third-person possessive marker, *-ı*, shares the same form as the accusative marker, *-ı*. Consequently, this is an area where language models or parsers not engaged in syntax may struggle, as seen in v2.8. However, this confusion appears to have been minimized in v2.11.

Additionally, the impact of modifying the values of features was examined. Table 6 shows the frequencies of improvement and worsening recovery for the most frequently used feature keys.

6. Conclusions

In this work, we examine the utility of large language models in evaluating the quality of corpora. The proposed approach was employed to examine the re-annotation of a large Turkish treebank in the Universal Dependencies framework with a focus on linguistic improvements. For this purpose, we compared the annotations in the two different versions of the treebank.

Several LLM models were examined with prompts generated from the annotations of the two

Feature	v2.8		v2.11	
	#	%	#	%
Person=3	1.185	37	1.167	35
Number=Sing	1.072	36	1.054	35
Polarity=Pos	459	49	570	51
Case=Nom	590	36	455	31
Person[psor]=3	335	43	361	44

Table 5: The top 5 features of the words that were not accurately recovered for UD Turkish BOUN v2.8 and UD Turkish BOUN v2.11, using GPT-4.

Feature	Old Value	New Value	+	-
Aspect	Hab	Perf	–	12
Aspect	Imp	Prog	–	61
Aspect	Imp	Prosp	17	–
Aspect	Prog	Imp	33	–
Aspect	Prosp	Imp	–	11
Case	Acc	Nom	–	66
Case	Dat	Nom	11	9
Case	Gen	Nom	5	–
Case	Loc	Nom	–	6
Case	Nom	Acc	58	–
Case	Nom	Dat	6	–
Case	Nom	Loc	5	–
Number	Plur	Sing	16	12
Number	Sing	Plur	12	12
Person	3	1	–	5
Person	3	2	–	6

Table 6: The impact of modifying the value of an existing feature. The columns '+' and '-' show the frequencies for the increase and decrease in recovery.

treebank versions to reconstruct the original sentences' surface forms. We observed, as expected, that LLMs that aim to capture linguistic characteristics provide useful information about the annotated treebanks and the annotation scheme. All models suggest an improvement regarding the re-annotation, with GPT-4 demonstrating the best performance. These results provide insights regarding the treebanks and the annotation scheme. Similarly, an error analysis of LLM outputs can provide valuable insights and actionable items for improving the annotation quality.

As future work, we plan to apply the proposed LLM-based evaluation scheme to other types of NLP resources. We believe that LLMs will be valuable in evaluating and gaining insight into language resources. This approach will contribute to creating higher-quality resources.

7. Ethical Statement

We have gathered results for 2 different languages, namely Turkish and English. The data we used for Turkish and English are publicly available, licensed under Creative Commons licenses with attribution, more specifically CC BY-SA 4.0 and CC BY-NC-SA 4.0. We have not used any personal data in our experiments. We do not foresee any ethical issues arising from our methodology or results.

8. Acknowledgements

This work was supported by Boğaziçi University Research Fund Grant Number 16909.

9. Bibliographical References

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019. Cross-lingual dependency parsing with unlabeled auxiliary languages. *arXiv preprint arXiv:1909.09265*.

Yesim Aksan, Mustafa Aksan, Ahmet Koltuksuz, Taner Sezer, Ümit Mersinli, Umut Ufuk Demirhan, Hakan Yilmazer, Gülsüm Atasoy, Seda Öz, Ipek Yildiz, et al. 2012. Construction of the Turkish national corpus (TNC). In *LREC*, pages 3223–3227.

Sharefah Al-Ghamdi, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2023. Fine-tuning bert-based pre-trained models for arabic dependency parsing. *Applied Sciences*, 13(7):4225.

Anthropic. 2023. Product Anthropic. https://www.anthropic.com/product. [Online; accessed 19-October-2023].

Laurie Bauer. 1979. Some thoughts on dependency grammar.

Talha Bedir, Karahan Şahin, Onur Güngör, Suzan Uskudarli, Arzucan Özgür, Tunga Güngör, and Balkız Öztürk Başaran. 2021. Overcoming the challenges in morphological annotation of Turkish in universal dependencies framework. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 112–122.

Piyush Behre, S.S. Tan, Amy Shah, Harini Kesavamoorthy, Shuangyu Chang, Fei Zuo, Chris Basoglu, and Sayan D. Pathak. 2022. Trscore: A novel gpt-based readability scorer for asr segmentation and punctuation model evaluation and selection. *ArXiv*, abs/2210.15104.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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

Marie-Catherine De Marneffe and Joakim Nivre. 2019. Dependency grammar. *Annual Review of Linguistics*, 5:197–218.

Ralph Debusmann. 2000. An introduction to dependency grammar. *Hausarbeit fur das Hauptseminar Dependenzgrammatik SoSe*, 99(1):16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bill Dyer. 2022. New syntactic insights for automated wolof universal dependency parsing. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 5–12.

Python Software Foundation. 2022. Python language reference. Accessed: 2023-10-19.

Kim Gerdes and Sylvain Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. In *LAW X (2016) The 10th Linguistic Annotation Workshop: 131*.

Stefan Grünewald and Annemarie Friedrich. 2020. Unifying the treatment of preposition-determiner contractions in german universal dependencies treebanks. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 94–98.

Hugging Face. 2023. Hugging Face. https://huggingface.co/. [Online; accessed 21-October-2023].

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 133–142.

Jenna Kanerva, Filip Ginter, and Sampo Pyysalo. 2020. Dependency parsing of biomedical text with bert. *BMC bioinformatics*, 21:1–12.

Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do neural language models show preferences for syntactic formalisms? *arXiv preprint arXiv:2004.14096*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Tomasz Limisiewicz, Rudolf Rosa, and David Mareček. 2020. Universal dependencies according to bert: both more specific and more general. *arXiv preprint arXiv:2004.14620*.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *ArXiv*, abs/2303.16634.

Büşra Marşan, Salih Furkan Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Betül Özateş, Suzan Üsküdarlı, Arzucan Özgür, Tunga Güngör, and Balkız Öztürk. 2022. Enhancements to the BOUN treebank reflecting the agglutinative nature of Turkish. In *Proceedings of the ALTNLP The International Conference and workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing*, pages 71–860.

Yuval Marton, Nizar Habash, and Owen Rambow. 2013. Dependency parsing of modern standard

arabic with lexical and inflectional features. *Computational Linguistics*, 39(1):161–194.

R Thomas McCoy, Junghyun Min, and Tal Linzen. 2019. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Steven Moore, Huy A Nguyen, Tianying Chen, and John Stamper. 2023. Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. In *European Conference on Technology Enhanced Learning*, pages 229–245. Springer.

Amir More and Reut Tsarfaty. 2016. Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 337–348.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95.

OpenAI. 2021. Gpt-3.5 (or chatgpt) language model. https://www.openai.com/chatgpt. Accessed on 19-October-2023.

OpenAI. 2023. Models - OpenAI API. https://platform.openai.com/docs/models. [Online; accessed 19-October-2023].

Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of universal dependencies (ud). *Glossa: a journal of general linguistics (2016-2021)*.

Md Rizwan Parvez and Kai-Wei Chang. 2021. Evaluating the values of sources in transfer learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5084–5116, Online. Association for Computational Linguistics.

Sergio Pelaez, Gaurav Verma, Barbara Ribeiro, and Philip Shapira. 2023. Large-scale text analysis using generative language models: A case study in discovering public value expressions in ai patents. *ArXiv*, abs/2305.10383.

Poe. 2023. Welcome to Poe for Developers - Documentation. https://developer.poe.com. [Online; accessed 19-October-2023].

Vinit Ravishankar. 2017. A universal dependencies treebank for marathi. In *Proceedings of the 16th international workshop on treebanks and linguistic theories*, pages 190–200.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark. Association for Computational Linguistics.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2011. Resources for Turkish morphological processing. *Language resources and evaluation*, 45:249–261.

Binyam Ephrem Seyoum, Yusuke Miyao, and Baye Yimam Mekonnen. 2018. Universal Dependencies for Amharic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vijay Sundar Ram and Sobha Lalitha Devi. 2021. Dependency parsing in a morphological rich language, Tamil. In *Proceedings of the First Workshop on Parsing and its Applications for Indian Languages*, pages 20–26, NIT Silchar, India. NLP Association of India (NLPAI).

TII. 2023. Falcon LLM. https://falconllm.tii.ae/. [Online; accessed 21-October-2023].

Mads Toftrup, Søren Asger Sørensen, Manuel R. Ciosici, and Ira Assent. 2021. A reproduction of apple's bi-directional LSTM models for language identification in short strings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 36–42, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and

Arzucan Özgür. 2022. Resources for Turkish dependency parsing: Introducing the BOUN treebank and the BoAT annotation tool. *Language Resources and Evaluation*, pages 1–49.

Francis Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An assessment of universal dependency annotation guidelines for turkic languages.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Veronika Vincze, Katalin Ilona Simkó, Zsolt Szántó, and Richárd Farkas. 2017. Universal dependencies and morphology for hungarian-and on the price of universality. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback. *ArXiv*, abs/2305.14282.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5284–5294.

10. Language Resource References

Marşan, Büşra and Akkurt, Salih Furkan and Türk, Utku and Atmaca, Furkan and Özateş, Şaziye Betül and Berk, Gözde and Bedir, Seyyit Talha and Köksal, Abdullatif and Başaran, Balkız Öztürk and Güngör, Tunga and Özgür, Arzucan. 2022. *Turkish BOUN Treebank of Universal Dependencies 2.11*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-4923. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Nivre, Joakim and de Marneffe, Marie-Catherine and Ginter, Filip and Hajič, Jan and Manning, Christopher D. and Pyysalo, Sampo and Schuster, Sebastian and Tyers, Francis and Zeman, Daniel. 2020. *Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection*. European Language Resources Association. PID http://hdl.handle.net/11234/1-5287.

Silveira, Natalia and Dozat, Timothy and Schuster, Sebastian and Connor, Miriam and de Marneffe, Marie-Catherine and Schneider, Nathan and Chi, Ethan and Bowman, Samuel and Manning, Christopher and Zhu, Hanzhi and Galbraith, Daniel and Bauer, John. 2023. *English EWT Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Türk, Utku and Atmaca, Furkan and Özateş, Şaziye Betül and Berk, Gözde and Bedir, Seyyit Talha and Köksal, Abdullatif and Başaran, Balkız Öztürk and Güngör, Tunga and Özgür, Arzucan. 2021. *Turkish BOUN Treebank of Universal Dependencies 2.8*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-3683. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.