TækTåk: Syntactic Analysis of Language Use on Danish TikTok

Thea K. R. Kristensen, Rob van der Goot
IT University of Copenhaagen
thkk@itu.dk, robv@itu.dk

Abstract

Language use is different across different language communities. Social media provides a rich source for studying how language varies, as it contains large data for a wide variety of sub-communities. In this paper, we study language usage on Danish TikTok. TikTok is a video-based platform, but most users are mainly active in the text-based comment sections. With the goal of analyzing language usage on this language variety, we contribute: 1) the first Danish social media treebank annotated for Universal Dependencies 2) evaluation of a variety of parsers using the new treebank, showing that cross-lingual in-domain data provides a valuable signal 3) a comparison of syntactic trends on standard Danish languages and TikTok language.

Keywords: Dependency parsing, Linguistic analysis, Domains, Social media, Danish

1. Introduction

Language use is known to differ across many dimensions (e.g. Biber and Conrad, 2009), where every sub-community has their own variety of language. These differences are relevant for Natural Language Processing practitioners, as they have shown to have a large impact on downstream performance (e.g. Eisenstein, 2013; Plank, 2016).

Social media provides an interesting source of data for NLP models, as it is rich in information and quantity, but provides challenges due its unique language varieties, including many sub-communities with their own language customs. Although there are linguistic conventions that are generally followed within sub-communities, social media is also characterized by its direct, spontaneous and fast nature, which leads to diversity in the language use. Much of the early NLP work on social media focused on Twitter data, because it was easy to collect. We focus on a more recently emerging platform, TikTok, which is a video-sharing platform with active text-based discussions under the video's.

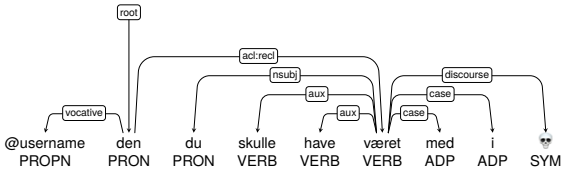
There has been much work on language specific social media dependency treebanks, including for Italian (Sanguinetti et al., 2018; Cignarella et al., 2019), English (e.g. van der Goot and van Noord, 2018; Liu et al., 2018a), Hindi-English (Bhat et al., 2018), French (e.g. Seddah et al., 2012; Kaljahi et al., 2015), German (Rehbein et al., 2019), Turkish (Pamay et al., 2015), and Arabizi (Seddah et al., 2020). Almost all of these are focusing on Twitter data, Kirstein Hansen et al. (2023a) is in that sense the most similar resource, which is a POS tagged Danish TikTok corpus. Most of aforementioned treebanks apply the Universal Dependencies (UD) guidelines, and add additional guidelines where necessary. Notably, Sanguinetti et al. (2020) provide unified annotation guidelines specifically tailored towards adapting UD for social media data. These treebanks have been mainly used to develop parsers for specific language varieties; we will go one step further, and use our resulting parser to analyze a larger amount of text for linguistic trends.

The previous work focused mainly on high-resource languages, whereas Danish can be considered a mid-resource language: it has less speakers, and was categorized by Joshi et al. (2020) into the "3. The Rising Stars", indicating high availability of plain text data, but not many annotated resources. Most other languages for which a social media treebank exists, are in categorized in the higher-resource "4. The Underdogs" and "5. The Winners" (Joshi et al., 2020). Danish, as a mid-resource language provides an interesting testcase, as there is quite some data available, but at the same time it is underrepresented in most multi-lingual studies and models. While there has been quite some NLP resources developed for Danish in recent years, the social media domain has not been studied extensively. The existing benchmarks and models consist of sentiment analysis (Pauli et al., 2021), named entity recognition (Plank et al., 2020; Enevoldsen et al., 2024), and POS tagging (Kirstein Hansen et al., 2023a).

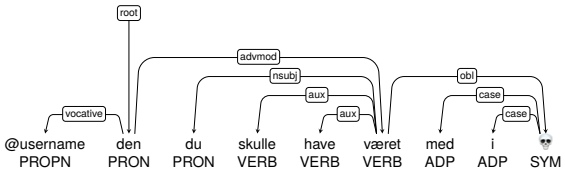
Our main contributions are:

- TækTåk, the first Danish social media dataset annotated with 583 comments annotated for dependency structures and part-of-speech tagging.
- An extensive evaluation of the effect of cross-domain, in-domain, cross-lingual, and in-language learning for automatic dependency parsing.
- An in-depth analysis of remaining errors for our best two parsers, including an analysis of errors that can be resolved with adding a small amount of in-domain, in-language data.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11524–11534
11–16 May 2026.



(a) Interpretation from main annotator



(b) Interpretation from co-annotator

Figure 1: Different interpretations of the sentence *the one you should have been in* 💀

- A thorough comparative linguistic analysis of language use across four datasets: a Danish newspaper dataset, a Twitter dataset, TækTåk, and Tæktåk automatically annotated with the best-performing parser.

2. TækTåk

2.1. Data Collection

To examine how young people communicate with each other on social media and how it differs from more traditional media, we created a dataset by scraping the comment section of popular Danish TikTok content creators. The dataset, collected from the video IDs of the 20 most popular Danish content creators, initially consisted of approximately 1 million comments. The data was cleaned by removing comments that consisted solely of friend mentions, emoji strings, or single word reactions such as "haha" . These entries were excluded due to their lack of meaningful linguistic content.

Furthermore, we removed all non-Danish comments by applying a language detector (Shuyo, 2010). This approach introduces a bias, as some Danes could write comments that appear predominantly English due to the slang common on TikTok, while their intentions may not have been to write English. However, since we are mainly interested in the Danish language (there exist already many English social media treebanks, see Section 1), this filtering step was necessary to maintain the relevance of the dataset. We anonymized all usernames by replacing them with the placeholder *@username*. After filtering and anonymization, the resulting dataset comprised 199,569 comments—an 80 % reduction.

2.2. Annotation

From the large dataset, a subdataset was created with 2000 comments; Exactly 100 from every content creator. This was done to ensure variety in the final test set. From this dataset, we select 583 random comments, corresponding to 4,598 words[1], were manually annotated for Part-of-Speech (POS) tagging and dependency parsing, according to the guidelines provided by Universal Dependencies (Nivre et al.), with additional guidelines from an earlier in-house Danish annotation project, which will be released with the data.

Further, we decided to annotate words according to their intended meaning and not their literal meaning; this meant that spelling errors would be interpreted as the intended word. This was even if the misspelled word was a correct spelling of another word. In the example *Min bedste vend*, which directly translates into *my best turn*, however, the intended meaning was probably *my best friend* / *min bedste ven* and *vend* was tagged as `NOUN`, which was likely the intended meaning of the author, instead of `VERB`, which would have been the literal meaning. However, this approach may also introduce a higher disagreement, as it relies on annotators interpreting the intended meaning. The context may be misunderstood or influenced by the annotators' subjective perspective.

Our main annotator was a Danish 24 year old data science student, who is native speaker of Danish, a social media user, and has no prior experience with UD. Our second annotator was a 33 year old Danish language learner, with an A2 certificate in Danish and native speaker of Dutch, a closely related language. The second annotator had prior experience with UD (annotated two treebanks), so the annotators complemented each other.

2.3. Annotator Disagreements

Annotation was conducted in three phases. First, both annotators annotated the 100 shortest sentences of the data, after which they discussed annotation differences, doubts, and distinctions. After these 100, they both annotated another 50 sentences randomly sampled from the full data to obtain a more reliable inter-annotator agreement. Disagreements on these 50 sentences are then also inspected and resolved. Finally, the main annotator annotated the remainder of the data.

Some common misunderstandings were slang words, colloquial Danish language, and sentence interpretations. An example of this is seen in Figure 1, where the interpretation of the emoji changed the structure of the sentence. It could be given the label `discourse`, according to the standard often used for emojis. However, the author could have

[1] Following Sanguinetti et al. (2020), we do not perform any sentence segmentation

POS	UAS	LAS
86.5	87.0	77.8

Table 1: Inter annotator agreement in percentage accuracy over all 150 double annotated comments.

used the emoji as a word, as if something *they were in*, and in that case, `obl` would be more correct.

As the co-annotator is a non-native speaker with an A2 certificate in Danish, sometimes disagreements stemmed from misunderstandings of Danish words and phrasing, leading to a pessimistic inter-annotator agreement. One such example was the comment *luk dog køleren*, which roughly translates to *then shut the fridge*; however, the word *dog* translates poorly into English. In cases such as this, the judgment of the native speaker was typically accepted.

Looking at the results for the inter-annotator agreement in Table 1, we see that the agreements are much lower compared to what can be achieved for standard texts. It should be noted though that some disagreements stemmed from misunderstandings of the co-annotator related to the Danish language, so the results likely do not represent a accurate upperbound of performance. Discussions and conversations between annotators along the way, provided insights and improvement of the quality, which increased our certainty on how to proceed and made a good foundation for annotating the remaining comments.

2.4. Biases

We foresee at least three sources of bias in the data creation process. Starting with the data selection, we aim to capture Danish language on TikTok, but even within that, we have a bias towards popular content creators. The motivation was that it is easier to get a large amount of data, and that the popular videos will represent more common language. The second bias is in the further filtering of the data, by removing instances that are not considered Danish by a language classifier, we might have removed very non-standard sentences, and code-switched sentences. Finally, there is a bias in the annotation. Both of our annotators annotated according to their own interpretation.

3. Parsing experiments

3.1. Setup

Previous work has shown that in-domain treebanks in other languages can be highly beneficial when training parser (Stymne, 2020; Müller-Eberstein et al., 2021). Since no Danish in-domain treebank is available, we compare cross-lingual, in-domain training to in-language, cross-domain training based on data availability. More concretely, we train on the English social media treebank TweeBank (Liu et al., 2018b) (24,753 words), and on the Danish Dependency Treebank (Johannsen et al., 2015), which consist mainly of data from Danish news, fiction and non-fiction (80,378 words). We keep half of the TækTåk data separate for testing, and the remainder for training (2,299 words).[2] We remove double root labels from the Tweebank automatically by connecting all non-first roots to the first one, with the parataxis relation to follow the UD standard more closely and be able to use standard parsers.

Our language model selection is informed by the results of Kirstein Hansen et al. (2023b), who perform POS tagging on Danish TikTok data. We selected the two best performing Danish models (Ælæctra (Højmark-Bertelsen, 2021) and RøBÆRTa (Nielsen et al., 2021)) and the best performing Multi-lingual model (Twhin (Zhang et al., 2023)). We also include mBERT (Devlin et al., 2019) as a baseline. The latter two models have been trained on a mixture of languages including Danish.

We use the MaChAmp (van der Goot et al., 2021) framework for implementing our parser, which uses the Deep Biaffine parser as proposed by (Dozat and Manning, 2017). We used all default hyperparameters of MaChAmp v0.4.2, and train multi-task models that perform dependency parsing and POS tagging with shared weights of the language model.

3.2. Results

We use standard accuracy and Labeled Attachement Score (LAS) for evaluating the performance of respectively POS tagging and dependency parsing, as implemented by Zeman et al. (2018). We compare results when training on each source treebank, as well as on a combined set. We also mix in half of the TækTåk data to gauge the effect of in–domain, in-language training. Results (Table 2) show that performance in general is around 60-90% accuracy for POS tagging, and 50-70% LAS for dependency parsing, which is in a similar range as previous work on social media dependency parsing for other languages (Section 1) Furthermore, the results clearly show that for the Danish models, the Danish treebank (DDT) is a better training source, but for the multi-lingual models, the in-domain Tweebank leads to higher performance for both POS tagging and dependency parsing. Although it should be noted that the DDT is substantially larger, which could be considered a confounder, although we believe that training on more English data might not be

[2]Note that we have no dev split because of the limited size. Consequently, we do not perform any hyperparameter tuning.

Model	DDT POS / LAS	TB POS / LAS	TB/DDT POS / LAS	DDT/TB/TT POS / LAS	TB/TT POS / LAS
Ælæctra	75.58 / 52.10	59.11 / 41.64	85.43 / 60.94	86.74 / 66.11	80.40 / 64.90
RøBÆRTa	71.85 / 49.37	54.93 / 38.01	80.42 / 56.53	86.40 / 65.09	68.41 / 47.78
Twhin	74.89 / 51.25	84.67 / 61.46	85.07 / 60.49	88.44 / 67.86	**89.52 / 69.87**
mBERT	72.40 / 47.47	79.47 / 55.46	76.34 / 52.95	85.76 / 65.87	85.96 / 66.46

Table 2: Accuracy results for POS tagging and LAS for dependency parsing. Each model is trained on Danish Dependency Treebank (DDT), the TweeBank (TB), and/or TækTåk data (TT).

beneficial, as we hypothesize that the nuances that are learned at later stages will have less overlap cross-lingually, this is partially confirmed by convergence of performance during training. Perhaps unsurprisingly, the combined dataset performs best for all language models. Finally, we confirm findings by Blodgett et al. (2018), who show that even small samples of in-domain data can boost performance, we see a very similar trend when we add the TækTåk data (DDT/TB/TT column).

Based on this result, we hypothesized that training on the Tweebank and TækTåk would lead to better results, because the TækTåk seems to give quite a performance boost, but adding the DDT to Tweebank is not always beneficial. The results (TB/TT) confirm this hypothesis, as the highest performance is obtained by Twhin on the Tweebank TækTåk combination. Comparing this final result to the inter-annotator agreement scores (Table 1), we can conclude that the POS tagger is on a similar level as human agreement, but the parser is still 8 percentage points lower.

4. Parser analysis

In this section, we will provide an analysis of the best performing multilingual model (Twhin TB/TT), and the best performing Danish model (Ælæctra DDT/TB/TT). We evaluate the effect of sentence length, and look at a diversity of quantitative metrics. Finally, we perform a qualitative analysis on both models, which we report separately because of the diversity in trends of errors.

4.1. Impact of sentence length

We evaluate the impact of sentence length by creating buckets of different lengths, and reporting performance within these buckets. Results (Table 3) show that longer sentences are harder to parse for both models, but Twhin is more affected. For POS tagging, the differences are smaller, but there is still a consistent downward trend. We hypothesize that the longer sentence lengths of DDT (see Table 5), the main training data of our Ælæctra model lead to the more robust performance of that model.

Length	Ælæctra POS / UD	Twhin POS / LAS
0–30	86.94 / 66.97	85.98 / 66.37
30–60	84.32 / 57.81	83.99 / 58.99
60–90	84.75 / 54.81	83.52 / 56.73
90–120	84.75 / 51.25	81.25 / 50.00

Table 3: POS accuracy and parsing LAS by sentence length in characters.

4.2. Quantitative analysis

To obtain a deeper understanding of where the main performance bottlenecks are, we report a variety of additional metrics for dependency parsing: UAS, which only considers the relations between the words, and not their labels. LAS: the default metric we used so far, which is the percentage of correct relations with their label. CLAS: including only content words, and MLAS, where head, relation, and POS have to match. We here perform an ablation evaluation of the two best models, where we also include the models trained without the in-domain, in-language TækTåk data to get a better insight into where the in-domain in-language data is beneficial.

Results (Table 4) show that the UAS is substantially higher compared to the LAS, indicating that there is still room for improvement in the labeling. The CLAS is only slightly lower than the the LAS for both parsers, so performance on content and function words is quite similar. Finally, the MLAS shows that the combination of tasks performs lower than the LAS, showing that errors across tasks are on different parts of the input.

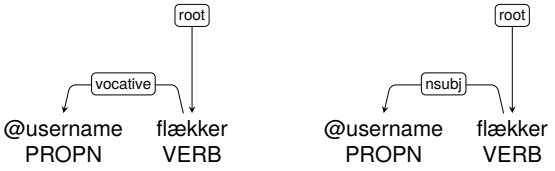
4.3. Qualitative analysis Ælæctra

We manualy analyzed the outputs of Ælæctra on the TækTåk test split, looking for trends in the errors compared to the gold standard. We again compare the model trained without and with the TækTåk training data to inspect which types of errors are resolved. Below, we will first discuss the errors made by both versions of the parser (with and without TækTåk in training), and then list the errors that were mostly resolved after adding TækTåk.

Metric	Ælæctra			Twhin		
	DDT/TB	+TT	↑	TB	+TT	↑
UAS	73.26	77.18	*5.35*	73.64	78.11	*6.07*
LAS	61.22	66.41	*8.48*	61.99	68.16	*9.95*
CLAS	60.69	66.97	*10.34*	61.41	68.97	*12.31*
MLAS	53.56	58.49	*9.06*	52.39	61.42	*17.23*

Table 4: F1 Score for different metrics, calculated for Ælæctra and Twhin. ↑ reports improvement in percentages after adding TikTok data.

Persistent errors An inconsistency between TækTåk and DDT annotations led to the parser consistently label possessive pronouns such as *mine* as `DET`, whereas they are annotated as `nmod:poss` in TækTåk. This is likely due to differences in annotation guidelines. After adding the other datasets, this error mostly remained, probably due to the high frequency of this phenomenon in DDT.

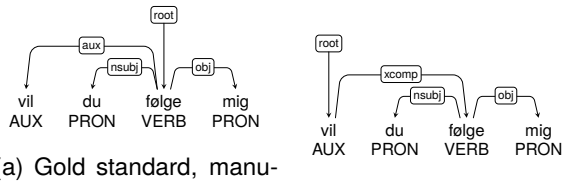


(a) Gold standard (b) DDT

Figure 2: Different dependency tags for the sentence "*@username cracking (up)*"

Resolved Errors One of the most frequent sources of error involved the dependency and POS tagging of emojis. In the gold standard, emojis are consistently labeled with the dependency relation `discourse` and assigned the POS tag `SYM`. The parser trained on DDT alone lacks exposure to emoji usage and consistently mislabels them, leading to inconsistent and seemingly arbitrary tagging across the test set. Since emojis are widely used across all social media platforms, including TikTok, this inconsistency significantly affects overall parsing accuracy. This issue was entirely resolved in the improved parser trained on in-domain data. Exposure to emoji-rich data in the TikTok and TweeBank datasets enabled the model to learn how to assign consistent and accurate labels.

A similar issue is seen with the `vocative` dependency relation. While vocatives are not entirely absent in formal text, their frequency and function differ significantly on TikTok. On the platform, vocatives are commonly used when users tag friends in a comment, often followed by a personal reaction or emotional response to the video. This differs from traditional uses, where vocatives typically serve to directly address someone within a sentence. However, the lack of punctuation and often also the lack of subject –commonly seen in slang– makes it a case of ambiguity. As in the example in figure 2, it could mean that the *@username* is cracking up, and it would in standard Danish. However, in the context of TikTok, it is more likely that the person commenting is informing another person that they themselves are cracking up over a video, and would like them to see the video too.



(a) Gold standard, manually annotated (b) Twhin, trained on Twitter

Figure 3: Different dependency tags for the sentence "*Do you want to follow me / Will you follow me*"

4.4. Qualitative analysis Twhin

We perform a similar analysis as for the Ælæctra model, including a comparison of errors for the model trained with and without the TækTåk data.

Persistent Errors The Twhin model trained on the TweeBank had a high performance for POS tagging compared to the other models (84.67), but still room for improvement. Especially spelling errors and slang would often be categorized with a seemingly arbitrary POS-tag. Further, words with ambiguity posed a challenge for the model. Especially prepositions, such as *så*, which can be both an adverb and a preposition, and *ved*, which can be both a verb and a preposition. If the model predicted one of the ambiguous words incorrectly, the error would propagate, since the preposition changed the entire structure of the sentence. These POS errors show that the model has trouble interpreting Danish words, which also leads to error propogation in the dependency structures.

A common error for the parser trained only on TweeBank was its inability to correctly connect `AUX` to the verb it accompanies. This often led to the

root being different from the gold standard, and additionally, a much higher ratio of xcomp. The example shown in Figure 3 shows an illustrative case of this issue in a fairly common sentence structure. The problem was also noted during manual annotation rounds and highlights how Danish includes idiosyncratic features that make accurate predictions less likely for a parser trained solely on English data. A few cases of this phenomenon were corrected after training on TækTåk, but the majority remained, probably because of the small size of TækTåk.

Another common mistake seen for the Twhin parser trained only on TweeBank, was consistently labeling structures with compound, whereas the manual annotations used either flat or goeswith. This issue was most frequently observed with PROPN spans that covered more than one token. This divergence might stem from differences in how multi-token proper names are handled across languages. While compound for English is typically used for noun compounds (e.g., "ice cream"), the case is a bit different for Danish, as words most often are written as one word in Danish and therefore the use of compound is much rarer. The parser may overgeneralize the use of compound based on patterns learned from noun phrases, which it saw then training on the English corpus, which led to these systematic mismatches. This error was also partially, but not consistently, corrected with the improved parser. Often, compound was replaced by another tag that did not follow the gold standard, and sometimes the error was not corrected at all.

Resolved Errors The addition of TækTåk to the training data resolves several errors, particularly those likely caused by the lack of comparable content in the TweeBank dataset. A common example of this is the words *henne* [3] and *hedde*[4]. The parser trained solely on the TweeBank consistently assigned incorrect POS and dependency labels to these words. As neither *henne* nor *hedde* has a clear direct translation to English, the parser lacked relevant exposure and struggled to make accurate predictions. However, in the model trained with the inclusion of TækTåk, this error was resolved, and the parser aligned with the gold standard annotations.

Another challenging phenomena for the parser was questions. The word *hvor /where*, would often be tagged as root where the gold standard has it as an advmod of the main verb. The word *Hvad / what* was assigned a diversity of labels but in the gold standard, *hvad* was most commonly labeled as an obj (e.g. Figure 4a).

An exception to this pattern occurs when the main verb is assigned cop, in which case *hvad* is promoted to the head. In Figure 4b, similar sentences are automatically annotated by the parser trained solely on TweeBank. Here, the assigned head varies even among sentences that resemble each other structurally. This inconsistency is observed throughout the dataset and negatively affects parsing accuracy. The inconsistent treatment of *hvad* and *hvor* leads to lower LAS and UAS. In the gold standard, approximately 25% of the sentences contain a wh-word, meaning this mismatch has a substantial impact on overall performance. This type of error occurred far less frequently in the parser trained with the inclusion of TækTåk, suggesting that the model had learned to apply a more consistent annotation standard.

Conclusion Although our parser obtained better performance when training on the in-domain English TweeBank, we can clearly identify trends in the errors that show lack of understanding of Danish syntax. Adding even a small amount of TækTåk data, resolved some of these errors. This is in contrast to the Ælæctra model trained on the Danish data, which had different types of errors. We conclude that in-domain data is crucial for better performance, especially if it is available for a closely related language, although without language specific data, errors in common target language constructions are to be expected.

5. Data analysis

Finally, we conduct a comparative analysis of how language use is different on TækTåk data compared to data from the TweeBank and the DDT. We use both our small manually annotated set, and the full set parsed by our best performing model (Twhin, trained on Tweebank and TækTåk).

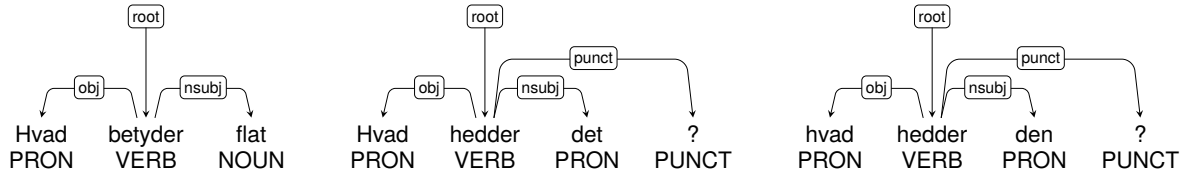
5.1. Label Distributions

Looking at Figure 5, there are some interesting syntactic differences between the datasets, which indicate how especially TweeBank, DDT, and TækTåk differ from each other and how different domains impact word use and structure.

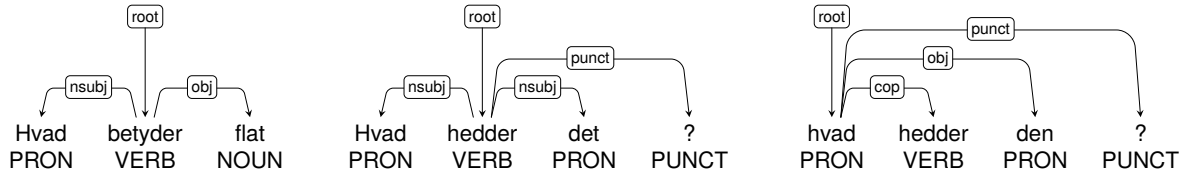
TækTåk uses PRON in a much higher degree than both DDT and TweeBank. A high ratio of pronouns is a typical sign of informal language (Szynalski, 2014). In DDT, which is based on formal Danish, we expect a low use of pronouns. While there are more pronouns in Twitter, indicating a more informal language than DDT, TikTok seems to be even more informal. Almost one in five words in the annotated

[3] Translations vary depending on context, *der henne* | *over there*, *jeg var henne i bageren* | *I was at the baker*, *det kan henne* | *it could be*

[4] *Hun hedder Pia* | *She is called Pia* or *She is named Pia*

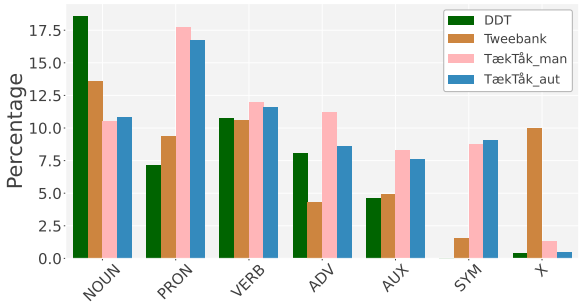


(a) Gold standard annotations for wh-questions

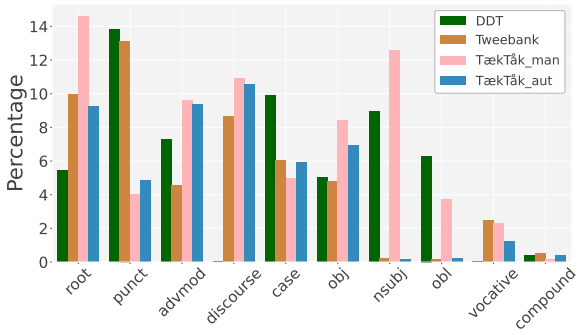


(b) Automatic annotations for wh-questions

Figure 4: Comparison between gold standard and automatically generated annotations for wh-questions



(a) Distribution of POS-tags for the four datasets



(b) Distribution of selected UD-tags for the four datasets

Figure 5: Distribution of POS and UD tags across the four datasets

TækTåk is a pronoun. This is reinforced when looking at the automatically annotated TækTåk, which is just as extreme. Even though TikTok is more similar to Twitter than to DDT, it still seems to be significantly more informal. The increased number of symbols for TækTåk and the lack of punctuation further confirm this suspicion. On TækTåk, commas and periods are rarely used. The most common punctuation found in the TikTok dataset was most often a series of either exclamation points or question marks, and in that use case, even indicating an informal language. TikTok users utilize emojis as one would traditionally use punctuation. Towards the end of a sentence, or sometimes in the middle: *"Den er mega god 😊 gå ind og se den" / "It is really good 😊 go in and see it"*. As in the example here, where an emoji is used where a comma would traditionally be used.

There are many similarities between the in-domain datasets. However, language also has an influence. The Danish datasets have a much higher ratio of adverbs and advmod relations. 8.05% of the words used in DDT are adverbs, almost twice as many as the TweeBank. TækTåk have the most, with 11.23%, and the automatically annotated a bit less. We see that the news data from DDT has a substantial higher amount of `case` relations, whereas the TækTåk datasets have more objects, suggesting that they are more focused towards actions. Overall, this shows that domain is not everything; Danish has some idiosyncratic features that make it stand out from English. From the results showing the accuracy of the parser seen in Table 2, it is clear that the in-domain parser works best, likely because the structure of the sentences is more similar. However, there is still some syntactic difference between English and Danish, which makes the parser lack in performance.

5.2. Informality degree

To understand differences in formality across the datasets, we analyzed length, along with trends typically associated with informal language. One such trend is the use of linguistic ellipsis—where the subject or verb is omitted and must be understood implicitly from context—which is particularly common in informal language. This often results from the high frequency of imperatives used in informal language, where the subject is left out and the recipient is directly instructed or ordered, as in

Statistic	DDT	TweeBank	TækTåk	TækTåk AA
% verbs without subjects	51.82	63.32	58.33	58.00
% sentences without verbs	8.99	15.25	13.38	8.72
% imperatives	1.40	5.16	9.97	4.25
Average length	18.34	15.10	6.85	10.88
Arc Length	2.96	2.74	1.98	2.53
Tree Height	4.91	3.68	2.71	3.37

Table 5: Summary statistics of informality measures across the datasets.

the example *Fortæl dem bror / tell them brother*. To identify imperatives, we used manually designed extraction heuristics.

Ellipsis and Imperatives As seen in Table 5, TækTåk and Tweebank have a higher degree of ellipsis and use of imperatives compared to DDT. Most striking is the use of imperatives in the TækTåk, where TækTåk uses imperatives seven times more often than DDT. Twitter also shows a high rate of imperatives, as well as sentences lacking verbs or missing subjects. These results suggests that data from Twitter is more informal than DDT, but TikTok data is even more informal.

The automatically annotated (AA) TækTåk and the gold standard also differ from each other. For sentences without a verb and imperative, there is a larger gap. This is likely an effect of the shorter sentence length and errors of the parser. For verbs without subjects, the differences are not as conspicuous.

Sentence Length Another characteristic of informal language is its length. For the DDT dataset, we measure sentence length directly, whereas for Twitter and TikTok, we assess the length of entire tweets or comments. While this approach isn't perfectly accurate, as a tweet or comment may contain multiple sentences, it still provides a useful indication of how length is distributed across datasets. One reservation is that tweets begin with the pattern *RT @username :*, which accounts for three tokens and artificially raises the average length of tweets. These results should therefore be seen as approximate indicators of informality rather than precise measurements.

Table 5 indicates that comments and tweets tend to be shorter than sentences found in newspapers. Further investigation is done by analyzing the tree height and arc length of the sentences. The DDT dataset has a higher mean tree height than both TikTok and Twitter. Indicating more complex structures. There does not seem to be as big a difference when it comes to average arc length. The TækTåk has a lower arc length than all the others. This might be explained by the sentence length being generally shorter.

6. Conclusion

We present TækTåk, a dataset scraped from the comment section of popular Danish TikTok content creators and manually annotated for dependency and POS. We evaluated model performance across both in-domain vs. out-of-domain and in-language vs. out-of-language datasets and a variety of LLMs. We found that in-domain datasets and models achieve the best performance, and beat both in-language datasets and models. Further, the impact of even a small, highly specific in-language, in-domain dataset has a positive effect on performance for all models.

Common errors found for out-of-language models included the parser's inability to predict correctly on Danish sentence structures and peculiarities. These errors were to some extent resolved by introducing a subsection of the TækTåk to the training data. The most common errors for the out-of-domain models are different, and mainly include the parser's inability to predict structures and out-of-vocabulary words. Many of these errors were resolved by introducing both the in-domain TweeBank and the TækTåk.

We found that generally, in-domain datasets are quite similar to each other despite their language difference. For the out-of-domain, in-language, they are structurally quite far from each other, except for a few language-specific similarities. Furthermore, we found that though TikTok data and Twitter are similar, TikTok still seems to have a higher degree of informality, perhaps due to its younger audience and the design of TikTok, which promotes a fast-paced mode of interaction.

These findings underscore the importance of domain alignment in NLP, showing that even small, targeted datasets can significantly enhance model performance in informal and niche language environments like TikTok.

Ethical Statement

We created the TækTåk by scraping the comment section of public content creators' comments. This is social data, where users have voluntarily made the data by posting on social media (Olteanu et al., 2019), but have never imagined that their comment

could have a researcher as their audience, thus breaking their expectation of where their data ends. Social data is beneficial because it can capture a raw glimpse of how humans, and in this case, especially youth, write to and with each other. However, one must keep the ethical considerations in mind.[5] While the comments were posted for anyone to see, their intention was never that they could be used for research anywhere. As it is also pointed out by Boyd and Crawdford: *"Just because content is publicly accessible does not mean that it was meant to be consumed by just anyone"* (Boyd and Crawford, 2012).

While it is true that users did not expect their comments to be read or analyzed by researchers, care has been taken to anonymize the data. The analysis is conducted at an aggregated level and does not focus on individual users or specific comments. Given this, the risk of harm to the individual user is low. While the ethical concern of users, especially due to their age, is valid, the actual impact is likely minimal.

7. Bibliographical References

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 987–998, New Orleans, Louisiana. Association for Computational Linguistics.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*, 1 edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15:662–679.

[5]We will not publicly release the annotated treebank, interested researchers should contact the last author.

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France. Association for Computational Linguistics.

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.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Kenneth Enevoldsen, Emil Trenckner Jessen, and Rebekah Baglini. 2024. DANSK: Domain generalization of Danish named entity recognition. *Northern European Journal of Language Technology*, 10:14–29.

Malte Højmark-Bertelsen. 2021. Ælæctra - a step towards more efficient danish natural language processing.

Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2015. Universal dependencies for danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 157.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Rasoul Kaljahi, Jennifer Foster, Johann Roturier, Corentin Ribeyre, Teresa Lynn, and Joseph Le Roux. 2015. Foreebank: Syntactic analysis of customer support forums. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1341–1347,

Lisbon, Portugal. Association for Computational Linguistics.

Kia Kirstein Hansen, Maria Barrett, Max Müller-Eberstein, Cathrine Damgaard, Trine Eriksen, and Rob van der Goot. 2023a. DanTok: Domain beats language for Danish social media POS tagging. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 271–279, Tórshavn, Faroe Islands. University of Tartu Library.

Kia Kirstein Hansen, Maria Barrett, Max Müller-Eberstein, Cathrine Damgaard, Trine Eriksen, and Rob van der Goot. 2023b. DanTok: Domain beats language for Danish social media POS tagging. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 271–279, Tórshavn, Faroe Islands. University of Tartu Library.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018a. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018b. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dan Saattrup Nielsen, Malte Højmark-Bertelsen, Morten Kloster Pedersen, Kasper Junge, Per Egil Kummervold, and Birger Moëll. 2021. RøbÆrta - danish roberta base.

Joakim Nivre, Dan Zeman, Marie de Marneffe, Chris Manning, Lori Levin, Nathan Schneider, Francis Tyers, and Amir Zeldes. Universal Dependency Relations — universaldependencies.org. https://universaldependencies.org/u/dep/. [Accessed 13-05-2025].

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Tuğba Pamay, Umut Sulubacak, Dilara Torunoğlu-Selamet, and Gülşen Eryiğit. 2015. The annotation process of the ITU web treebank. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 95–101, Denver, Colorado, USA. Association for Computational Linguistics.

Amalie Brogaard Pauli, Maria Barrett, Ophélie Lacroix, and Rasmus Hvingelby. 2021. DaNLP: An open-source toolkit for Danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 460–466, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ines Rehbein, Josef Ruppenhofer, and Bich-Ngoc Do. 2019. tweeDe – a Universal Dependencies treebank for German tweets. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 100–108, Paris, France. Association for Computational Linguistics.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5240–5250, Marseille, France. European Language Resources Association.

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, Online. Association for Computational Linguistics.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a treebank of noisy user generated content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India. The COLING 2012 Organizing Committee.

Nakatani Shuyo. 2010. Language detection library for java.

Sara Stymne. 2020. Cross-lingual domain adaptation for dependency parsing. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 62–69, Düsseldorf, Germany. Association for Computational Linguistics.

Tomasz P. Szynalski. 2014. Formal and informal english.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Rob van der Goot and Gertjan van Noord. 2018. Modeling input uncertainty in neural network dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4984–4991, Brussels, Belgium. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5597–5607.