DanTok: Domain Beats Language for Danish Social Media POS Tagging

Anonymous Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymouser Author
Affiliation / Address line 1
Affiliation / Address line 2
email@domain

Affiliation / Address line 3
Anonymousest Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

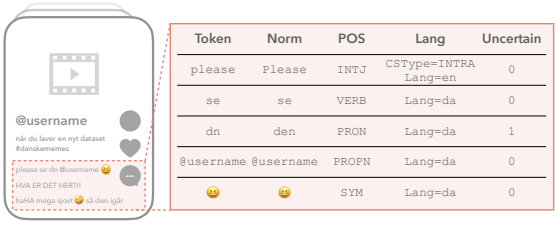
Abstract

Language from social media remains challenging to process automatically, especially for non-English languages. In this work, we introduce the first NLP dataset for TikTok comments and the first Danish social media dataset with part-of-speech annotation. We further supply annotations for normalization, code-switching, and annotator uncertainty. As transferring models to such a highly specialized domain is non-trivial, we conduct an extensive study into which source data and modeling decisions most impact the performance. Surprisingly, transferring from in-domain data, even from a different language, outperforms in-language, out-of-domain training. These benefits nonetheless rely on the underlying language models having been at least partially pre-trained on data from the target language. Using our additional annotation layers, we further analyze how normalization, code-switching, and human uncertainty affect the tagging accuracy.

1 Introduction

Language data from social media offer unique insights into how communities use language to communicate in a natural, spontaneous setting, using highly domain-specific vocabulary. This domain is, however, also subject to frequent changes, high noise, and variability, making it extremely difficult to process (Eisenstein, 2013) both for high (Gimpel et al., 2011; Derczynski et al., 2013) and especially for lower-resourced languages (Kaji and Kitsuregawa, 2014; Albogamy and Ramasy, 2015; Singh et al., 2018; Mæhlum et al., 2022). To understand better and improve how to target such

Code will be shared; data available upon request.



Token	Norm	POS	Lang	Uncertain
please	Please	INTJ	CSType=INTRA Lang=en	0
se	se	VERB	Lang=da	0
dn	den	PRON	Lang=da	1
@username	@username	PROPN	Lang=da	0
😆	😆	SYM	Lang=da	0

Figure 1: DanTok contains comments from Danish TikTok videos annotated for normalization, POS, code-switching, and annotator uncertainty. Fabricated example for illustrative purposes.

a highly specialized domain, we present the first linguistically annotated dataset of contemporary Danish social media from TikTok, a relatively new platform focused on short videos. The Danish Dependency Treebank (Kromann et al., 2003) from the PAROLE Corpus (Bilgram and Keson, 1998), and its Universal Dependencies (UD) conversion (Johannsen et al., 2015) is the primary part-of-speech (POS) resource for Danish NLP (Kirkedal et al., 2019) containing only well-edited text.

We present not only the first Danish social media dataset with POS annotation, but it is, to the best of our knowledge, the first linguistic exploration of TikTok language in any language. The dataset contains annotation for POS, normalization, code-switching as well as annotation uncertainty (Figure 1). It further prioritizes quality over quantity, covering +8k tokens, which have been manually verified to ensure relevance, correctness, and privacy.

It has been shown that training language models on the target domain can be used for cross-domain learning (Gururangan et al., 2020; Barbieri et al., 2022). In this work, we aim to tease apart the effect of the pre-training data of the language model and the source of the annotated training data for the target task. For this purpose, we contribute: ① DanTok: The first linguistically annotated dataset of TikTok comments and the first

Danish POS-annotated social media dataset; ② An extensive study of transfer learning targeting this highly specialized domain; ③ An in-depth analysis investigating unique features of this new domain, and how specific data/model properties help boost transfer performance.

2 DanTok

2.1 Collection

The social media platform TikTok shows a continuous, personalized feed of short videos on various topics. Users can express feedback (heart/thumbs-down) and add comments. In 2022, the number of Danish users of age $\geq$ 18 was estimated to be 1.14M. This is substantial, both in terms of the total population in Denmark of 5.8M, as well as compared to, e.g., Twitter with estimated 685k Danish users.[1]

We selected 15 hashtags with predominantly Danish comments. From these hashtags, we scraped all comments of five videos with at least 50 comments, leading to a total of 8,741 comments and 63,076 tokens. We then filtered these comments first based on % out-of-vocabulary tokens, followed by a deduplication procedure to maximize diversity. This deduplication procedure iteratively retains only one of the two most similar comments based on normalized Levenshtein distance until a certain threshold is reached. More details are reported in Appendix B.2.

2.2 Annotation

We tokenized the data using the `Tweet-Tokenizer` from `NLTK` (Bird et al., 2009), with manual post-correction and replaced usernames with @*username*. Comments containing personal information were filtered out manually. After two rounds of inter-annotator agreement convergence with three native Danish annotators (more details in Appendix A), reaching a Fleiss' κ score of 94.28 (round 1) and 96.05 (round 2), we annotated each sentence once, with the following layers:

POS We followed the UD 2.11 guidelines and the social media-specific suggestions by Sanguinetti et al. (2020a). Where possible; we annotated the intended meaning of a token given its use in context, e.g., in "Would you like a 👍?", 👍 is a `NOUN`. The tokens *og*, *eller* and *men* ("and, or, but") are always `CCONJ`. Tokens that imitate pronunciation are `INTJ`. The possessive pronouns are always tagged as `DET`. For demonstrative pronouns that may also function as determiners, these were tagged as `DET` when followed by an adjective or number.

	COMMENTS	TOKENS	TYPES	TTR
Dev	429	4,000	1,520	0.38
Test	430	4,028	1,519	0.38
Total	859	8,028	2,512	0.31

Table 1: DanTok dataset statistics. TTR is type-token ratio.

Normalization We follow the format and guidelines of MultiLexNorm (van der Goot et al., 2021a): URL's and interjections are not normalized; splits and merges are included in the annotations (but insertions/deletions are not). Merging of tokens that have erroneously been written in multiple words is indicated in the normalization column of the last word, as compound words in Danish adopt the part-of-speech (POS) of the final word. For tokens that have erroneously been written as one word, an overall POS tag is given to this token alongside individual tags for the tokens resulting from the splitting.

Code-switching Code-switching is indicated via language and type. Type is either `INTER` (a full sentence unit), `INTRA` (individual tokens) or `MIXED` (a lemma from one language with inflection from another language) as described by Sanguinetti et al. (2020b). We used the official Danish dictionary[2] for cross-referencing, as many originally English words are considered Danish.

Certainty Following Bassignana and Plank (2022), the certainty of a POS tag was annotated with a binary label.

2.3 DanTok Statistics

Our final dataset consists of 8,028 tokens total and 2,512 unique types (Table 1). We observed 16.66% of tokens requiring normalization, 5.03% code-switching (all to English), and 5.12% of tokens with annotation uncertainty.

[1] https://datareportal.com/reports/digital-2022-denmark

[2] https://dsn.dk/ordboeger/retskrivningsordbogen/

	Langs	Train-sources	Type	Train-size	Architecture	#Params	%Unk.	Subw.-ratio	Reference
LLM									
DANISH-BERT-botxo	DA	web, wiki, subtitles		10gb	Bert-base	111M	0.15	1.28	https://github.com/certainlyio/nordic_bert
RØBÆRTA-base-danish	DA	web	-D+L	?	Roberta-base	125M	0.00	1.58	https://huggingface.co/DDSC/roberta-base-danish
ÆLÆCTRA-danish-small-cased	DA	legal, social, web, wiki, news		1,045M words	Electra-small	14M	0.04	1.39	https://github.com/MalteHB/-l-ctra
BERTWEET-Base	EN	social		850M tweets	Roberta-base	135M	0.10	1.65	Nguyen et al. (2020)
BERTWEET-Large	EN	social	+D-L	850M tweets	Roberta-large	355M	0.00	1.90	Nguyen et al. (2020)
TWITTER-ROBerta-base	EN	social		58M tweets	Roberta-base	125M	0.00	1.90	Barbieri et al. (2020)
TWITTER-XLM-roberta-base	30+	social		198M tweets	XLM-r base	278M	0.01	1.45	Barbieri et al. (2022)
BERNICE	66	social	+D-L	2.5B tweets	Roberta-base	278M	0.00	1.44	DeLucia et al. (2022)
TWHIN-bert-large	100+	social		7.5B tweets	new	561M	0.01	1.45	Zhang et al. (2022)
TREEBANK									
LINES	EN	fiction, nonfiction, spoken	-D-L	57,372 words					Ahrenberg (2015)
TWEEBANK2	EN	social	+D-L	24,753 words					Liu et al. (2018)
DDT	DA	fiction, nonfiction, spoken, news	-D+L	80,378 words					Johannsen et al. (2015)

Table 2: An overview of the used language models and POS fine-tuning sets. %unk is the percentage of unknown subwords in our development data, sub.-ratio is the average amount of subwords per word. Capitalized name parts are handles.

3 Experiments

3.1 Setup

For a highly specialized dataset such as DanTok, transfer learning is key, as there is no training data matching the domain and language. Hence, we investigate training on 36 combinations of in/out-of-domain (+D: social media is the main source, -D: rest) as well as in/out-of-language (+L: trained on Danish, -L: rest) training data and large language models (LLMs). Further, all experiments are replicated on the normalized DanTok. The Danish LLMs are trained on web data, including some forum data, but none are explicitly optimized for social media. Table 2 lists the LLMs and datasets used in our experiments. All models consist of an LLM encoder plus a linear layer for POS labeling and are implemented in MaChAmp v0.4 (van der Goot et al., 2021b) using default hyperparameters. To avoid overfitting on DanTok, we use the transfer data's development set for model selection (Artetxe et al., 2020).

3.2 Results

Our main results are shown in Table 3. Unsurprisingly, the combination of in-domain, in-language (+D+L) training data and LLMs has the best overall performance. In general, having in-language data is more beneficial than in-domain data. However, when training on a single dataset, the in-domain English dataset (+D-L) leads to surprisingly high performance with the multilingual language models, even outperforming all scores obtained with the Danish training data. One reason for this could be the relatively high frequency of code-switching (5%). Interestingly, model size (Table 2) is not a good predictor of performance: Although the largest model, TWHIN, obtains the highest score overall, it requires large amounts of pre-training data and a specialized pre-training objective based on rich social engagements (Zhang et al., 2022). Meanwhile, ÆLÆCTRA's performance is very close while being 41 times smaller. Given these results, we conclude that the best strategy for obtaining a high-quality tagger would be to use domain-specific models when available (even if multilingual) and use in-domain fine-tuning data even if in another language (+ in-language if available).

Table 6 shows that using normalized data gives a consistent boost of 2–5 % points across all setups, with only a few exceptions. Furthermore, performance varies less.

	MODEL \ DATA	-D-L LINES	+D-L TWB	-D+L DDT	+D-L + -D+L TWB+DDT
-D+L	DANISH-BERT	44.02	49.60	77.98	84.08
	RØBÆRTA	58.43	60.82	70.17	78.72
	ÆLÆCTRA	49.50	63.30	74.20	84.95
+D-L	BERTWEET-B	27.80	38.00	67.90	79.47
	BERTWEET-L	25.92	36.55	67.40	81.50
	TWITTER-ROB	25.02	37.30	64.05	79.40
+D+L	TWITTER-XLM	67.58	77.15	72.15	83.28
	BERNICE	70.45	78.22	72.95	83.28
	TWHIN	69.30	81.38	72.65	**85.92**

Table 3: POS Tagging accuracy on the DanTok development set using combinations of in/out-of-domain (+D/-D) models and training data as well as in/out-of-language (+L/-L) models and training data.

LLM	-NORM	+NORM
TWHIN	86.05	88.18
ÆLÆCTRA	85.80	88.55

Table 4: Results on the test data of our two best models trained on TWEEBANK and DDT.

On test data TWHIN performs similarly as on the development data. After normalization, the smaller ÆLÆCTRA model outperforms it slightly (Table 4).

4 Analysis

4.1 Subword Analysis

The Subword ratio (Table 2) does not show a clear correlation with performance, so we qualitatively evaluate the subword segmentation of the two best-performing models ('TWHIN' and 'ÆLÆCTRA'). Surprisingly, we find that the multilingual model ('TWHIN') seems more capable of interpreting inflection suffixes than the Danish model. It correctly splits morphemes indicating definiteness, plurality, or adverbial status, which the Danish model sometimes does not split. There are also examples where the segmentation of the Danish model matches morphemes more closely, while both subword segmenters make a substantial amount of hard-to-motivate splits, suggesting that there is room for improvement.

4.2 Stratified Analysis

We explore the accuracy on different subsets of the development set according to our additional annotation layers. See Table 7 in the Appendix. Unsurprisingly, we observe that the models struggle more with tokens that got a normalized version during annotation and words that annotators were also uncertain of. For code-switched tokens, we observe a large performance drop for the in-language LLM (23.8) despite fine-tuning on English in-domain data. Surprisingly, the multilingual model, likewise fine-tuned on Danish and English in-domain data, also struggles with code-switched text (8.2 % point lower).

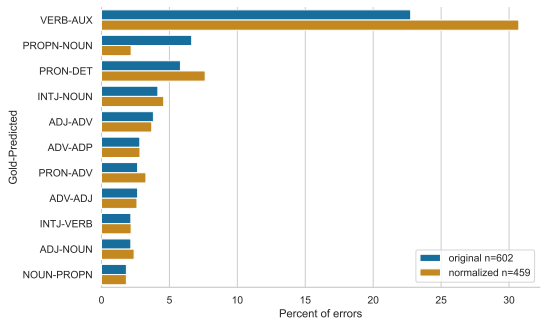


Figure 2: The 11 most frequent tag confusions for the ÆLÆCTRA model.

4.3 Qualitative Error Analysis

The most frequent tag confusion for the best ÆLÆCTRA model are given in Figure 2. TWHIN follows a similar pattern. Over half of the errors made by each tagger on the original data are shared with the other tagger. Some of the errors pertains to DanTok annotation. The most frequent error types can be categorized as follows:

VERB vs. AUX In DanTok, the present tense of the copula verb *være* ("to be") has been labeled VERB when it is the only verb in the sentence. However, the models prefer the tag AUX in 91.6% and 85.0% of the cases, respectively. This seems to be in line with the UPOS guidelines and is likely a result of the annotation of *er* in the DDT training set; here, 78.85% of *er* tokens have been tagged as AUX (the remaining being tagged as VERB).

Pronoun Confusion Tokens that can appear as pronouns and other POS confuse the taggers. The most frequent issue is PRON and DET confusion, which is arguably non-trivial in Danish[3]. PRON and ADV confusion is also prevalent; e.g., the token *der* can be either a relative PRON or the preliminary subject or an ADV of place. In the erroneous predictions, *der* is generally labeled ADV.

Proper Noun Inconsistencies Orthographic variations of social media language throw off the models; names written in lowercase are often labeled NOUN. On the normalized data, PROPN (gold) → NOUN errors decreased by 75% for ÆLÆCTRA and 62% for TWHIN. Likewise, when capitalized names are used in context, the models labeled them as PROPN, whereas we annotated the syntactical use of the token; e.g., *filming a TikTok*.

5 Conclusion

We presented DanTok, the first linguistically annotated TikTok dataset and the first Danish social media dataset with POS annotation. We conducted an extensive analysis of how to best transfer to a highly specialized domain in a mid-resource language. We demonstrated that LLMs benefit from common approaches such as normalization, while struggling with the same cases as the human annotators. Simultaneously, our results show that although in-language data/models form the basis for high performance, in-domain data, even from another language, should not be neglected in order to achieve state-of-the-art results.

[3]Consider, e.g., ***den*/PRON** *bog*/NOUN vs. ***den*/DET** *gamle*/ADJ *bog*/NOUN ("**that** book" vs. "**the** old book").

References

Lars Ahrenberg. 2015. Converting an English-Swedish parallel treebank to Universal Dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 10–19, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Fahad Albogamy and Allan Ramasy. 2015. Towards POS tagging for Arabic tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 167–171, Beijing, China. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

GNU Aspell. 2019. GNU Aspell. `http://aspell.net/`. Version da-1.6.36-11-0.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Elisa Bassignana and Barbara Plank. 2022. CrossRE: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Thomas Bilgram and Britt Keson. 1998. The construction of a tagged Danish corpus. In *Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998)*, pages 129–139, Copenhagen, Denmark. Center for Sprogteknologi, University of Copenhagen, Denmark.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Mark Dredze, and Philip Resnik. 2022. Bernice: A multilingual pre-trained encoder for twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

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.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021a. Multilexnorm: A shared task on multilingual lexical normalization. In *Proceedings of the 2021 EMNLP Workshop W-NUT: The Seventh Workshop on Noisy User-generated Text*, pages 493–509. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021b. 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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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

Nobuhiro Kaji and Masaru Kitsuregawa. 2014. Accurate word segmentation and POS tagging for Japanese microblogs: Corpus annotation and joint modeling with lexical normalization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 99–109, Doha, Qatar. Association for Computational Linguistics.

Andreas Kirkedal, Barbara Plank, Leon Derczynski, and Natalie Schluter. 2019. The lacunae of Danish natural language processing. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 356–362, Turku, Finland. Linköping University Electronic Press.

Matthias T Kromann, Line Mikkelsen, and Stine Kern Lynge. 2003. Danish Dependency Treebank. In *Proc. TLT*, pages 217–220. Citeseer.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. 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.

Petter Mæhlum, Andre Kåsen, Samia Touileb, and Jeremy Barnes. 2022. Annotating Norwegian language varieties on Twitter for part-of-speech. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 64–69, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. 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. 2020a. 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, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020b. Treebanking user-generated content: A proposal for a unified representation in universal dependencies. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 5240–5250, Marseille. European Language Resources Association (ELRA).

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. A Twitter corpus for Hindi-English code mixed POS tagging. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 12–17, Melbourne, Australia. Association for Computational Linguistics.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.

Appendix

A Data Statement

The following dataset characteristics are outlined following Bender and Friedman (2018):

A. CURATION RATIONALE This dataset aims to provide high-quality, linguistically annotated data from contemporary Danish social media, in order to allow for analyses of how language use is evolving in these specialized domains, and how NLP methods can best be adapted to these changes.

B. LANGUAGE VARIETY The data consists of comments from TikTok videos collected in January 2023. The language covered is manually verified Danish (da-DK) with code-switching to English (en), and orthographic variations specific to the social media domain.

C. SPEAKER DEMOGRAPHIC Nothing specific is known about speaker demographics, as the data was scraped from 75 videos spanning different topics.

D. ANNOTATOR DEMOGRAPHIC Three Master's students, all native Danish speakers, one with previous experience in dataset creation for POS tagging. The annotators were paid for their efforts.

E. SPEECH SITUATION Comments under TikTok videos represent informal, written language produced largely spontaneously with the intent to address the video creator or express an opinion to other viewers.

F. TEXT CHARACTERISTICS The text contains domain-specific terms and abbreviations, some degree of typographical and orthographic errors as well as occasional ellipsis of sentence subject. Code-switching to English makes up 5% of tokens in the full dataset (development + test), though the dataset contains several additional tokens that exist with the same meaning in both English and Danish, e.g., *shit* and *like*.

G. RECORDING QUALITY N/A

H. OTHER N/A

I. PROVENANCE APPENDIX N/A

B Data Collection Details

B.1 Hashtags

Videos from the following 15 hashtags were scraped during data collection.

- #børn ("children")
- #danskememes ("Danish memes")
- #danskhumor ("Danish humor")
- #glædeligjul ("merry Christmas")
- #godtnytår ("happy new year")
- #gørdetselv ("do it yourself")
- #landsholdet ("the national team")
- #madlavning ("cooking")
- #mitarbejde ("my job")
- #morgenrutine ("morning routine")
- #parforhold ("relationships")
- #selvtak ("you're welcome")
- #sommerprojekt ("summer project")
- #tobiasrahim ("Tobias Rahim")
- #træning ("workout")

B.2 Collection

It is challenging to gather data from TikTok in a reproducible way as user feeds are populated algorithmically, differing based on user characteristics (e.g., Danish IP-address). Simultaneously, there is no way to retrieve all videos or filter by a specific language. To increase reproducibility, we, therefore, make use of the fact that it is possible to browse videos directly by hashtag in a non-random order while logged out. We selected 15 Danish hashtags (Appendix B.1) through an explorative approach, where videos were browsed to determine whether the content, including comments, was predominantly in Danish.

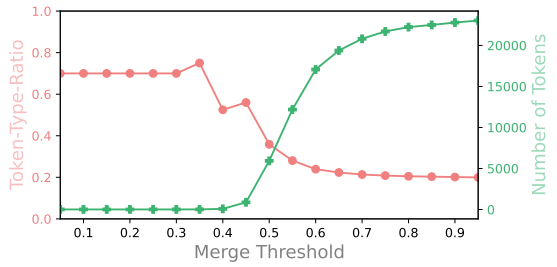


Figure 3: Deduplication using varying merge thresholds, and their effects on the number of tokens and their token-type-ratio.

Filtering Starting from all comments, we next apply automatic filtering to improve data diversity while reducing annotation workload. First, the out-of-vocabulary filter (OOV) attempts to remove non-Danish comments while retaining interesting diversions from orthographic norms. Using the Aspell dictionary for Danish (Aspell, 2019), comments with $< 10\%$ or $> 60\%$ OOV tokens are removed, leaving 2,887 comments with 23,244 tokens which are almost exclusively Danish.

Deduplication Next, as comments often have a high degree of repetitiveness (e.g., "wow", "woow!!"), we deduplicate them, in order to maximize the diversity of the data being annotated. To do so, we iteratively merge the two most similar comments, keeping only one, until there are no more pairs with a similarity greater than a threshold t. Similarity is measured using the normalized Levenshtein distance, and we retain the comment with a higher total similarity to all other comments (i.e., more prototypical samples). Determining the best threshold is non-trivial, so we conduct a sweep over $t \in (0.0; 1.0)$ in 0.05 increments (Figure 3), and choose $t = 0.60$ which retains 1,844 comments with 17,067 tokens and a token-type-ratio (TTR) of 0.24. This set retains relatively high diversity while accounting for the downstream, manual filtering of irrelevant comments down to our target of annotating 8k tokens.

B.3 Statistics

Table 5 lists the distribution of POS labels in DanTok, in addition to the statistics on the other annotation layers reported in Section 2.1.

Label	%
VERB	14.03
PRON	13.02
NOUN	11.70
ADV	11.32
SYM	9.62
PUNCT	6.73
PROPN	6.60
ADJ	6.00
ADP	5.53
INTJ	4.31
DET	3.46
AUX	2.86
CCONJ	1.88
NUM	1.16
SCONJ	1.13
PART	0.61
X	0.04

Table 5: POS label distribution in DanTok

	Model \ Data	-D-L LINES	+D-L TWB	-D+L DDT	+D-L + -D+L TWB+DDT
-D+L	DANISH-BERT	42.69	48.12	80.19	85.75
	RØBÆRTA	60.79	63.48	73.80	82.36
	ÆLÆCTRA	50.87	61.69	78.48	88.45
+D-L	BERTWEET-B	28.24	38.48	70.88	82.83
	BERTWEET-L	27.46	37.48	70.98	85.35
	TWITTER-ROB	25.75	38.13	68.76	84.19
+D+L	TWITTER-XLM	69.85	80.12	75.16	86.26
	BERNICE	72.01	80.27	75.61	85.15
	TWHIN	70.95	83.09	75.33	**88.80**

Table 6: POS Tagging accuracy on the normalized DanTok development set using combinations of in/out-of-domain (+D/-D) models and training data as well as in/out-of-language (+L/-L) models and training data, plus a concatenation covering +D and +L.

LLM	Certain			Canonical			In fine-tune vocab			Code-switched		
	n	-	+	n	-	+	n	-	+	n	-	+
TWHIN	203	62.1	87.2	662	70.8	88.9	1234	83.3	87.1	3808	86.3	78.1
ÆLÆCTRA		58.6	86.4		66.8	88.6		83.1	85.8		86.1	62.5

Table 7: Stratified accuracy on the 4000-token development set of the two best models trained on TWEEBANK and DDT. n is the number of tokens in the - category, e.g., 662 tokens were non-canonical.

C Additional Results

C.1 Normalization

Table 6 shows that using normalized data boosts the results by 2-5 points across all setups except a few. Furthermore, performance differences across models are less pronounced.