Cross-Dialectal Transfer for Low-Resource Arabic: The Tunisian Arabic Dependency Treebank

Amal Aissaoui
CUNY Graduate Center
365 5th Ave, New York, NY 10016
aaissaoui@gradcenter.cuny.edu

Abstract

This paper presents a small-scale dependency treebank for Tunisian Arabic (TADT) developed within the Universal Dependencies framework, addressing the scarcity of linguistic resources for the Arabic varieties. The approach employs domain adaptation, leveraging a machine learning model (UDPipe 1.0) trained on Algerian Arabic data to annotate 100 Tunisian Arabic social media comments, followed by manual correction. This pilot study evaluates the feasibility of using machine learning-assisted annotation to scale resource development for spoken Arabic and identifies key challenges in cross-dialectal transfer for improving annotation quality and efficiency. This work contributes to more inclusive and fair representation of Arabic linguistic varieties in academic research and NLP applications.

Keywords: Dependency Treebank, Tunisian Arabic, Domain Adaptation

1. Introduction

Given the central role that linguistic data plays in academia and industry, data scarcity is the biggest obstacle that stands in the way of a fair and comprehensive representation of languages in linguistic research and language-driven applications. For a linguistic community to increase its presence and weight in language technology, linguistic data must be "accessible and processable" (Lynn, 2016, 1).

The first contributing factor to data scarcity is the absence of a standard orthography system. Without a standard orthography, a language cannot be easily documented, which limits the amount of textual data that can be generated and stored for that language. The second factor is the absence of government support. Since underresourced languages tend to be spoken in countries with a colonial history or within diglossic speech communities, they are often marginalized in educational and governmental institutions.

In Tunisia, the aforementioned factors played a major role in creating a severe data shortage problem. Tunisia exhibits the features of a true diglossic community where more than one language coexist within a single speech community, with each language being used for particular purposes and in different contexts. Considering Tunisia's smaller population compared to its neighboring states, its geopolitical and cultural ties to the member states of the Arab League, and its French colonial history, it is expected that the country faces social and economic pressure to adopt more widely spoken languages for better economic opportunities, further reducing the opportunities to document and record diverse and extensive data of Tunisian Arabic and keeping any standard writing system for the language from emerging.

In NLP, MSA is too formal to be used for NLP technologies used for daily tasks. For natural interactions between humans and voice assistants, spoken Arabic is preferable. Attempting to circumvent this issue by applying NLP tools trained on MSA to spoken Arabic will result in substantially poor performance (Al-Shargi and Rambow, 2015; Samih, 2017).

For spoken Arabic, even if the aspirations of policy makers to institute language reforms are present, they might take decades to implement, whereas the pressing urgency to advance in language technology calls for rapid actions with prompt results. This study adopts a proactive approach to bridging the existing gap by making the following contributions:

- Highlighting and assessing trends in resource development for the Tunisian Arabic variety.
- Presenting a small-scale Tunisian Arabic treebank developed within the Universal Dependencies framework.
- Evaluating the use of dialect-transfer as a strategy to accelerate the resource development process.

2. Related Work

2.1. Resource Development

Boujelbane et al. 2013 adopted a lexical transfer approach using the Penn Arabic Treebank

Proceedings of the Ninth Workshop on Universal Dependencies (UDW 2026), pages 258–267
May 16, 2026.

(Maamouri et al., 2005) and the MADA morphological analyzer (Habash et al., 2009) to build a bilingual lexicon of Tunisian Arabic and MSA. Zribi et al. 2013 created a bilingual dictionary consisting of roughly 30k words transcribed from radio and TV broadcasts in Tunisian Arabic and used it to propose a method to adapt an Arabic morphological analyzer designed for MSA to Tunisian Arabic.

Sadat et al. 2014 introduced a bilingual lexicon of 1.6k Tunisian Arabic words and their aligned MSA translation, and proposed a rule-based translation system to translate any social media text in TA into MSA. Bouamor et al. 2014 built a corpus that covers five varieties, namely Egyptian, Tunisian, Syrian, Jordanian, and Palestinian Arabic. The corpus includes 2k Egyptian sentences selected from the Egyptian-English corpus created by Zbib et al. 2012 then translated by native speakers into the four other varieties. The translators were instructed to avoid transliterating using the Roman script yet no orthographic guidelines for the Arabic spelling were provided. Meftouh et al. 2015 started by recording natural conversations and used the recordings of TV shows in Algerian Arabic which were translated into MSA then converted to Tunisian, Syrian, and Palestinian Arabic by native speakers of the respective languages. The parallel corpus of 32k sentences was then used to train machine translation models.

Medhaffar et al. 2017 presented TSAC, a Tunisian sentiment analysis corpus consisting of 17k manually annotated Facebook comments. Annotators were native speakers of Tunisian Arabic who labeled comments as positive or negative, accounting for code-switching and Arabic. Fourati et al. 2020 introduced TUNIZI[1], a Tunisian Arabizi sentiment analysis data set consisting of 9k manually annotated sentences of Arabizi from social media. Gugliotta and Dinarelli 2022 presented TArC[2], a multi-layer annotated corpus for user-generated Tunisian Arabic, containing approximately 45k annotated tokens. The corpus features multi-level linguistic annotation including language identification, transliteration, tokenization, POS tagging, and lemmatization, performed using a semi-automatic pipeline.

2.2. Trends and Limitations

While research on resource development for Tunisian Arabic is gradually emerging, three main trends can be identified. The first is the predominance of labeled resources lacking linguistic annotation, with Gugliotta and Dinarelli 2022 constituting, to date, the only resource annotated with linguistic information. Moreover, morphological and syntactic annotation have received no attention in the existing literature.

The second trend is the absence of shared guidelines. Once annotation projects are started, it is crucial to adopt common guidelines and best-practice recommendations for the development of language resources to ensure "interoperability" (Ide et al., 2017, 116). An abundance of resources that can only be used in isolation is as ineffective as having a shortage of resources.

The third trend is the reliance on native speakers to generate annotations. Native Arabic speakers possess language proficiency, but are not sufficiently qualified to assist in this mission since they lack linguistic expertise. As discussed in Section 1, under-resourced languages tend to lack a standardized writing system due to historical marginalization in educational and governmental contexts. While native speakers possess invaluable inherent intuition, the absence of formal instruction in their primary language often means they have not acquired the specific academic frameworks necessary for independent linguistic analysis. Consequently, while their contributions are essential, native speakers lack the technical training required to generate formalized linguistic judgments that can be fully validated without the methodological oversight of a trained linguist for annotation tasks, including part of speech tagging, morpho-syntactic analysis, semantic role labeling, co-reference resolution, or transcription for automatic speech recognition.

Given the small size the linguistic community and the limited metalinguistic awareness among native speakers, finding a team of expert annotators can be challenging. As a result, developing a gold-standard linguistic resource becomes an incredibly time-consuming process. This study follows a linguistically-informed approach that leverages machine learning, user-generated data, and an existing linguistic resource from a neighboring Arabic variety to reduce the burden of labor-intensive manual annotation and streamline the development process to expedite the creation of linguistic resources. It further seeks to orient the efforts of the Tunisian research community toward interoperability and community standards by promoting the Universal Dependencies framework. Building on this approach, the study presents the first attempt to develop a treebank for Tunisian Arabic in the CoNLL-U format and in accordance with the Universal Dependencies scheme.

[1] https://github.com/iCompass-ai/TUNIZI?tab=readme-ov-file

[2] https://github.com/eligugliotta/tarc

3. Resource Development and Description

3.1. Training Data

For the training of the machine learning model, the Algerian Arabic NArabizi[3] dataset in CoNLL-U format compiled by Seddah et al. (2020) was used as training input. The treebank consists of 1,500 fully annotated user-generated sentences in Algerian Arabic, that include morpho-syntactic annotations, along with translations at both the word and sentence levels. It is made freely available under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

The selection of an Algerian Arabic treebank was motivated by two considerations. First, NArabizi is the only available UD treebank for a spoken Arabic variety and is provided in the user-generated transliteration of Arabic commonly referred to in the literature as the Arabizi script (Seddah et al., 2020), which is the same script in which Tunisian text data is available. Second, Algerian and Tunisian Arabic share a common linguistic foundation. Sociolinguistically, both varieties exist within a diglossic context, display frequent code-switching with French, and lack a standardized orthographic system, as discussed in Section 1. Lexically, they exhibit substantial overlap in their vocabularies, which derive primarily from shared Classical Arabic roots alongside borrowings from Berber, Italian, Turkish, and French. Structurally, both varieties show vowel reduction and consonant clustering, maintain closely parallel verbal conjugation systems, employ comparable strategies for pronoun attachment and negation morphology, and favor the SVO word order. Collectively, these factors contribute to the high degree of mutual intelligibility between speakers of the two varieties. For a detailed cross-varietal comparison involving Tunisian, Algerian, and other North African Arabic varieties, see Saadane and Habash 2015 and Harrat et al. 2015.

While the Algerian Treebank serves as a seminal resource for spoken Arabic as it represents the inaugural effort in this domain and provides the foundation for the current study, it presents certain methodological limitations. Specifically, inconsistencies originating from the NArabizi project framework affect the overall annotation reliability.

In Seddah et al., native speakers of different varieties of North African Arabic were trained to perform annotation tasks following the Universal Dependency guidelines. While the nature and length of the training was not specified by Seddah et al., a closer look at the segmentation of tokens casts doubts on the qualifications of the annotators. Native speakers of spoken Arabic have not received a formal education in their varieties or an academic linguistic training that equips them with the analytical skills to critically engage with Arabic varieties whose grammar was shaped by centuries of linguistic evolution and which remain underexplored by linguistic scholars. Without expert guidance, native speakers of Arabic are not sufficiently metalinguistically aware to conduct sound morphological analyses, which are essential to produce the systematic splitting of MWTs prescribed by the UD annotation guidelines. Table 1 illustrates MWT segmentation for a sentence from the NArabizi treebank, while Table 2 presents the corresponding expert segmentation performed by a linguist for the same instance.

ID	Form	Lemma	UPOS
1	allah	Dieu	PROPN
2	yahafdak	te_protège	VERB
3	w	et	CCONJ
4	ykhalika	te_laisse	VERB
5	lina	à_nous	PRON
6	ya	oh	INTJ
7	Mr	monsieur	NOUN
8	le	le	DET
9	président	président	NOUN
10	w	et	CCONJ
11	tahia	vivre	VERB
12	bouteflika	Bouteflika	PROPN
13	w	et	CCONJ
14	rana	nous_sommes	VERB
15	m3ak	avec_toi	PRON

Table 1: MWT segmentation and POS tagging in NArabizi.

Another issue with the training data is adopting French glosses instead of lemmas. Seddah et al. justified this choice on the grounds of the prevalence of non-Arabic lexical borrowings in Algerian Arabic that would have required an etymological analysis. This argument is valid only if lemma annotation equals extracting the consonantal root of a word in Arabic. Only under such assumption would etymological analysis of words in spoken Arabic be necessary. Since lemmatization is the process of stripping a word form from inflection such as number, tense, and case, tracing etymological origins is not necessary for the assignment of lemmas since lexical borrowings typically conform to the morphological and syntactic rules of the host language. The example *garjouma* used by Seddah et al. to support their reasoning may be originating from French *gorge* (throat) but its current use adheres to the Arabic broken plural pattern system that is still present in North African

[3] https://github.com/UniversalDependencies/UD_Maghrebi_Arabic_French-Arabizi/tree/master

ID	Form	Lemma	UPOS
1	allah	Dieu	PROPN
2-3	yahafdak	_	_
2	yahafd	protège	VERB
3	ak	te	PRON
4	w	et	CCONJ
5-6	ykhalika	_	_
5	ykhali	laisse	VERB
6	ka	te	PRON
7-8	lina	_	_
7	li	à	ADP
8	na	nous	PRON
9	ya	oh	INTJ
10	Mr	monsieur	NOUN
11	le	le	DET
12	président	président	NOUN
13	w	et	CCONJ
14	tahia	vivre	VERB
15	bouteflika	Bouteflika	PROPN
16	w	et	CCONJ
17	rana	nous_sommes	VERB
18-19	m3ak	_	_
18	m3a	avec	ADP
19	k	toi	PRON

Table 2: Expert MWT segmentation and annotation.

Arabic varieties as shown in Table 3.

Singular Form	Plural Form	Gloss
/ʕasˁfu:r/	/ʕsˁa:fir/	bird
/sˁandu:q/	/sˁna:diq/	box
/xanfu:sah/	/xna:fis/	bug
/garʒu:mah/	/gra:ʒim/	throat

Table 3: Plural forms of the CvCCv:C pattern.

3.2. Modeling

In preparing this treebank, the NArabizi dataset served as the training input for a machine learning model developed with UDPipe1.0 (Straka et al., 2016). Once trained, this model was applied to Tunisian Arabic text data (see Section 3.3) to generate annotations and facilitate the rapid production of the resource.

The selection of UDPipe for this project was motivated by its comprehensiveness as a tool. It offers all the essential tasks for linguistic annotation within a single framework. This integration eliminates the need to combine multiple tools for different tasks and increases methodological consistency. In addition, it is freely available as an open-source resource making it an attractive option for an academic project with limited funding and allows the study to be reproducible for other under-resourced Arabic varieties.

Although more recent versions have been released, UDPipe 1.0 was selected due to its architectural simplicity and ease of customization, which are particularly advantageous when working with a small-scale treebank, especially considering the scarcity of linguistic resources in the context of spoken Arabic.

3.3. Preprocessing and Annotation

The UDPipe model trained on the NArabizi corpus was applied to a subset of the CTAB[4] dataset (Amara et al., 2021). The CTAB dataset was collected between 2017 and 2021 through web scraping of public Facebook pages belonging to Tunisian media outlets, including Attessia TV and Ettounsiya TV. The data comprises 2,929 comments written in the Arabizi script, totaling 26,232 tokens and covering a diverse range of topics such as politics, soccer, and social issues.

A random sample of 1k comments was selected for the study. Given time constraints and the potential cognitive load associated with the annotation process, the length of selected comments was limited to a maximum of 20 tokens, defined by whitespace separation. When segmented according to UD guidelines, such comments could yield between 20 and 40 lines, depending on the original spacing and the morphological complexity of the tokens.

A sample of 100 comments was subsequently selected from the cleaned sample and processed by the model to generate linguistic annotations. The resulting CoNLL-U file was duplicated and one copy was manually corrected by the author, a native speaker of Tunisian Arabic with an advanced background in linguistics. As noted in Section 3.1, the model was trained on data with suboptimal tokenization. Therefore, the initial step in the manual annotation phase involved reviewing and correcting the tokenization of each generated annotation. This process included not only segmenting multi-token words but also making decision regarding the splitting of grammaticalized contractions derived from Classical Arabic such as *inchallah* and abbreviations commonly found in user generated text then documenting these decisions as part of a guideline to ensure consistency throughout the annotation process.

Once tokenization was completed for a given sentence, lemmas were added using Buckwalter transliteration for words in Tunisian Arabic, and the

[4] https://africarxiv.ubuntunet.net/items/c14b535f-ac17-49ef-91e5-31e75a46eec9/full

Roman alphabet in cases of code-switching. Subsequently, part-of-speech tags, morphological features, dependency relations, and other relevant annotations including the sentiment were applied.

3.4. Quality Control

To assess the internal consistency of the gold annotations, a series of automatic validation checks were applied using the script provided in the UD-tools repository. These checks verified structural integrity (Level 1) to ensure the file is a machine-readable CoNLL-U file. Additional checks verified for universal inventory (Level 2) to ensure the consistency of UPOS tags and DEPREL labels with the 17 official UD tags and 37 universal relations, respectively. The consistency of morphological features was verified to ensure the `Name=Value` format and alphabetical order. Furthermore, universal syntactic logic (Level 3) was checked to verify tree structure including single-root dependency trees and acyclic graphs.

The detected issues were manually reviewed and corrected, after which the treebank passed all Level 3 universal structural and logical checks. Since this is the first treebank for Tunisian Arabic, the validation script was only invoked to test validity up to Level 3. Validation checks for language-specific labels (Level 4) and language-specific guidelines (Level 5) could not be performed, as the language-specific documentation required by the UD validator has not yet been incorporated into the UD infrastructure. Given the long-term objective of officially releasing this treebank, future steps will include registering the language and defining its language-specific documentation, followed by an additional validation round to ensure compliance with Levels 4 and 5.

Finally, the gold annotations were systematically compared to those predicted by the model. Evaluation is conducted across multiple annotation layers using the evaluation script[5] by Zeman et al. 2018 from the CoNLL 2018 shared task.

3.5. Corpus Description

The Tunisian Arabic Dependency Treebank comprises 100 sentences in the Arabizi script extracted from social media comments, containing a total of 1,466 tokens. Sentence length varies from 5 to 46 syntactic words, with a mean of 14.66 and a median of 13. The vocabulary consists of 281 multiword tokens, roughly 23.7%, reflecting moderate morphological complexity typical of a clitic-heavy language such as Tunisian Arabic.

[5] https://github.com/ufal/conll2018/blob/master/evaluation_script/conll18_ud_eval.py

3.5.1. Parts-of-Speech

The tree includes 16 out of the 17 official UD POS tags defined in the UD v2 scheme. Table 4 shows that the distribution of Universal parts of speech (UPOS) tags is dominated by verbs, nouns , then pronouns, which together constitute nearly half of the corpus. XPOS tags were not added in this study since the focus was on the annotations layers actually used by the UD community for cross-linguistic research.

UPOS Tag	%
VERB	17.5
NOUN	16.3
PRON	15.6
ADP	10.0
DET	7.3
PART	5.7
PROPN	4.6
ADJ	4.6
CCONJ	4.1
Other (INTJ, ADV, etc.)	14.2
Total	**100**

Table 4: Distribution of universal part-of-speech (UPOS) tags ($N = 1,466$).

3.5.2. Morphological Analysis

As summarized in Table 5, the treebank is heavily marked for nominal features, with singular number (42.6%) and masculine gender (25.2%) being the most common. The high frequency of 3rd and 2nd person markings along with the active voice and the finite verb form align with the verbal-heavy nature of the distribution showed in Table 4.

Feature	%*
Number=Sing	42.6
Gender=Masc	25.2
Gender=Fem	16.2
Voice=Act	15.8
Person=3	14.9
VerbForm=Fin	14.9
PronType=Prs	13.5
Mood=Ind	12.6
Definite=Ind	11.7
Number=Plur	10.0

**Percentages calculated against total syntactic tokens.*

Table 5: Distribution of key morphological features ($N = 1,466$).

3.5.3. Dependency Relations

Syntactically, the treebank is dominated by core arguments and modifiers, including direct objects, case markers, nominal modifiers, and nominal subjects as shown in Table 6. This distribution indicates that the treebank captures the richness and complexity of argument structure, providing a useful foundation for future parsing tasks and research on the syntactic structure of the language. In the current version, enhanced dependencies are not included, as the focus is on establishing a solid basic UD foundation.

Relation	%
obj	9.1
case	8.9
nmod	7.6
det	7.0
root	6.8
nsubj	6.8
parataxis	6.8
obl	6.1
discourse	5.3
Other	35.6
Total	**100**

Table 6: Most frequent dependency relations in the corpus.

3.5.4. Misc

The corpus exhibits a typical multilingual profile of the North African region, with Tunisian Arabic serving as the main language, accounting for 94.07% of the total syntactic tokens. As detailed in Table 7, code-switching is observed primarily with MSA (2.86%) and French (2.73%). A negligible portion of the data (0.34%) consists of English. This distribution confirms that the treebank successfully captures the code-switching patterns characteristic of contemporary Tunisian speech.

Language	%
Tunisian Arabic (`aeb`)	94.07
Modern Standard Arabic (`arb`)	2.86
French (`fr`)	2.73
English (`en`)	0.34
Total	**100.0**

Table 7: Distribution of languages and code-switching ($N = 1,466$).

3.5.5. Lemmas

As noted in Section 3.1, this study questions the decision of Seddah et al. to use French glosses as lemmas and rejects the claim that lemmatization necessarily requires an etymological analysis of the language. Table 8 summarizes the lemmatization conventions by part of speech and language applied in TADT. In cases of code-switching, the lemma field is annotated in accordance with the existing UD guidelines for the language of the token. However, when a token originates from a foreign language but conforms in its usage to the morphological and syntactic system of Tunisian Arabic, such as the noun *garjouma* in Table 3, the lemma and all other layers of annotations follow the Tunisian Arabic conventions.

In line with UD standards, nouns are lemmatized to their singular nominative forms where applicable, or to the masculine singular nominative form in the case of nouns with masculine–feminine alternations. For verbs, a cross-linguistic convention is followed: the infinitive is used for French and English, while for Arabic the third-person masculine singular perfective is adopted, following the guidelines of the Prague Arabic Dependency Treebank, which serves as the reference standard for Arabic UD projects.

For English and French word forms, lemmas are encoded in the Roman alphabet. For Arabic, Buckwalter transliteration is used, as it is widely adopted in NLP due to its reversibility and its avoidance of non-ASCII characters and complex diacritics that are both present in other romanization systems.

POS	Lang	Lemma
VERB	aeb, arb	3SG.M.PV (perfective)
	fr, en	Infinitive
NOUN	aeb, arb, fr	singular/masculine singular*
	en	singular
PRON	aeb, arb, en, fr	nominative
DET	aeb, arb, en	identical word form
	fr	masculine singular (*le/un*)
ADJ	aeb, arb, fr	masculine singular
	en	identical word form
OTHER	aeb, arb, en, fr	identical word form

**For nouns with gendered counterparts (e.g., professions).*

Table 8: Lemmatization conventions by part-of-speech and language.

The treebank exhibits a a high degree of morphological and functional diversity, with 495 unique lemmas identified as shown in Table 9.

The distribution is dominated by functional elements, particularly the definite marker Al (5.87%) and the conjunction w (4.37%). As the CTAB dataset is extracted from social media, the high frequency of pronominal forms, such as <inoti (2.87%), hiy~a (2.73%), and |naA (2.66%) as well as the frequent occurence of the vocative particle yaA (1.57%) reflect the interactive nature of the social media comments.

Lemma	POS	%
Al (the)	DET	5.87
w (and)	CONJ	4.37
<inoti (you)	PRON	2.87
hiy~a (she)	PRON	2.73
fiy (in)	ADP	2.73
|naA (we/us)	PRON	2.66
maA (not)	PART	2.59
huw~a (he)	PRON	1.91
$ (not)	PART	1.77
l (for/to)	ADP	1.64

Table 9: Top 10 most frequent lemmas ($N = 495$).

4. Results

Table 10 presents the complete set of results generated using the evaluation script by Zeman et al.. Performance is reported using precision, recall, and F1 score for all metrics as well as aligned accuracy (AlignedAcc) for annotation layers that require word level alignment. The findings suggest that the model achieves a strong performance at the tokenization level with an F1 score of approximately 90%. The sentence segmentation is perfect across all metrics.

In the UD framework and the CoNLL-U format, words refer to syntactic units which can occur separately as a separate token or attach to other syntactic words to form a MWT. Performance on words, which reflects the segmentation of syntactic words including the expansion of MWTs is substantially lower than for tokens. The model attains an F1 score of 60.76%. The significant drop in recall indicates that approximately 46% of syntactic words were not detected by the model.

Universal part-of-speech (UPOS) tagging performance is low across all metrics. The model identifies basic parts of speech only about 29%. The aligned accuracy metric indicates that 47% of the words the model was capable of aligning correctly with the gold annotation received the correct UPOS tag. The markedly low performance on language-specific-POS tags (XPOS) does not reflect any failure on the part of the model since the gold annotations did not include any.

The annotation of morphological features (UFeats) also exhibits low performance with F1 score of 18% and the aligned accuracy further confirms that even alignment is not achieved, approximately 70% of the features are predicted incorrectly. When POS tags and morphological features are evaluated jointly (AllTags), performance deteriorates sharply. This drop is expected since this metric requires all the tags of the three layers above to be simultaneously correct.

For dependency parsing, the model achieves an F1 score of 18.34 for the unlabeled attachment score (UAS). These results indicate that even when ignoring dependency labels, only 18% of the words are pointing to the correct parent word. For words that were correctly aligned, 30.18% have the correct head. Performance decreases further when dependency labels are taken into account, with the labeled attachment score (LAS) yielding a score of 11.42. Only 11% have both the correct head word and have the correct dependency relation label. For words that were correctly aligned, only 18.80% have the correct head and label.

CLAS, which restricts evaluation to content words, drops further to an F1 score of 6.74 and an aligned accuracy of 11.82%. As the strictest dependency metrics, MLAS and BLEX require the correct prediction of dependency attachment in addition to the accurate morphology or lemma prediction and are near zero with F1 values of 0.52 and an aligned accuracies below 1%, thereby revealing the compounding impact of errors introduced at the early stages of annotation pipeline.

5. Discussion

Findings from Table 10 reveal a clear decline in performance across annotation layers. While sentence segmentation is handled perfectly, this metric does not reflect anything on the ability of the model to identify sentence boundaries since the data was already segmented into sentences in the pre-processing stage. Surface tokenization also achieves a strong performance indicating that the model generally identifies the physical boundaries of strings with high accuracy, but the gap between precision and recall suggests that the model is leaning towards undertokenizing.

However, performance starts to deteriorate moving from token to word-level segmentation. The drop in recall highlights the weakness of the model in handling MWTs. Since standard CoNLL-U evaluation assumes a one-to-one correspondence between gold and predicted lines, the model's failure to split a MWT results in systematic misalignment between gold and predicted annotations, thereby preventing reliable evaluation and comparison of morphosyntactic annotations. This misalignment is reflected in the sharp decline observed for mor-

Metric	Precision	Recall	F1 Score	AlignedAcc
Tokens	93.00	87.15	89.98	–
Sentences	100.00	100.00	100.00	–
Words	70.58	53.34	60.76	–
UPOS	33.48	25.31	28.83	47.44
XPOS	0.54	0.41	0.47	0.77
UFeats	21.84	16.51	18.80	30.95
AllTags	0.09	0.07	0.08	0.13
Lemmas	4.24	3.21	3.65	6.01
UAS	21.30	16.10	18.34	30.18
LAS	13.27	10.03	11.42	18.80
CLAS	7.51	6.11	6.74	11.82
MLAS	0.58	0.47	0.52	0.91
BLEX	0.58	0.47	0.52	0.91

Table 10: Evaluation results comparing gold and predicted CoNLL-U annotations.

phological and lexical annotations.

While the model struggles with POS tagging and the capture of fine-grained morphological distinctions, the slight improvement observed in the AlignedAcc scores suggests that its overall performance might be dragged down by its weakness in handling the morphosyntactic complexity of MWTs, given that it was trained with poorly segmented data.

Although the small size of the treebank can also affect the reliability of results and the overall poor performance of the model is discouraging, the boost in performance observed in the Aligned Acc scores for UPOS, UFeats, and UAS may be viewed as an encouraging indication for further investigation using refining techniques. These findings further suggest that a cross-dialectal transfer approach in the context of spoken Arabic might be effective only under two conditions. First, a strict adherence to shared guidelines governing the segmentation and tagging of Arabic varieties. Second, the delegation of annotation to trained linguists rather than to native speakers who, despite their language proficiency, lack the analytical skills required for sound linguistic judgment, potentially introducing errors in annotation that lead to cascading consequences.

Regarding the model-assisted workflow employed in this study, a comparative analysis between manual and semi-automatic annotation efficiency was precluded by resource constraints. Specifically, as the author served as the sole annotator, the simultaneous development of a manual baseline from the Tunisian dataset was methodologically unfeasible. The model generated a first-pass CoNLL-U file with reliable tokenization, thereby eliminating the initial and most repetitive stage of manual annotation.

Due to the model's word-level segmentation performance, the manual effort was sufficiently restricted to 39.3% of instances—a considerablyly less taxing requirement than defining all syntactic words from a blank string. Furthermore, despite lower performance in categories such as UPOS, the model successfully automated the assignment of approximately one-third of the tags, reducing the cognitive burden of manual labeling. Consequently, the model's output reallocated the annotator's effort toward higher-level linguistic tasks that required extensive manual correction, such as morphological analysis and dependency parsing.

6. Conclusion

This paper presented the first Tunisian Arabic Treebank focused on user-generated content, capturing the code-switching patterns characteristic of North African Arabic varieties. Annotated according to Universal Dependencies standards, this resource promotes interoperability among Arabic language resources. Despite its limited scale, the treebank serves as a critical asset for both linguistic research and NLP development; it is currently being finalized for official release within the UD framework.

Methodologically, this study evaluated the feasibility of employing machine learning to assist linguists in generating labeled corpora, aiming to mitigate existing resource gaps. While initial results suggest limited immediate efficacy, observed performance trends provide encouraging evidence for the approach's potential. Future work will focus on expanding the treebank and implementing an iterative fine-tuning process to enhance model performance through optimized data quality. The ultimate objective remains to achieve a level of predictive reliability that significantly accelerates the development of gold-standard resources.

7. Bibliographical References

F Al-Shargi and O Rambow. 2015. DIWAN: A Dialectal Word Annotation Tool for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 49–58.

H Bouamor, N Habash, and K Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245.

R Boujelbane, E Khemekhem, M, S Ben Ayed, and H Belguith, L. 2013. Building builingual lexicon to create Dialect Tunisian corpora and adapt language model. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 88–93.

C Fourati, A Messaoudi, and H Haddad. 2020. TUNIZI: a Tunisian Arabizi sentiment analysis Dataset. In *1st AfricaNLP Workshop Proceeding, AfricaNLP@ICLR 2020*, pages 1–3.

E Gugliotta and M Dinarelli. 2022. TArC: Tunisian Arabish Corpus, First complete release. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1125–1136.

N Habash, O Rambow, and R Roth. 2009. MADA + TOKAN : A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, pages 102–109.

S. Harrat, K. Meftouh, M. Abbas, S. Jamoussi, M. Saad, and K. Smaili. 2015. Cross-Dialectal Arabic Processing. In *Computational Linguistics and Intelligent Text Processing*, pages 620 –632. Springer.

N Ide, N Calzolari, J Eckle-Kohler, D Gibbon, S Hellman, K Lee, J Nivre, and L Romary. 2017. Community Standards for Linguistically-Annotated Resources. In N Ide and J Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 113–165. Springer.

Teresa Lynn. 2016. *Irish Dependency Treebanking and Parsing*. Ph.D. thesis, Dublin City University and Macquarie University.

S Medhaffar, F Bougares, Y Esteve, and L Hadrich-Belguith. 2017. Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop (WANLP)*, pages 55–61.

K Meftouh, S Harrat, S Jamoussi, M Abbas, and K Smaili. 2015. Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34.

H Saadane and N Habash. 2015. A Conventional Orthography for Algerian Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 69–79.

F Sadat, F Mallek, R Sellami, M Boudabous, M, and A Farzindar. 2014. Collaboratively Constructed Linguistic Resources for Language Variants and their Exploitation in NLP Application – the case of Tunisian Arabic and the Social Media. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 102–110.

Y Samih. 2017. *Dialectal Arabic Processing Using Deep Learning*. Ph.D. thesis, Düsseldorf, Heinrich-Heine-Universität.

D Seddah, F Essaidi, A Fethi, M Futeral, B Muller, J Ortiz Suárez, P, B Sagot, and A 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.

M Straka, J Hajič, and J Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

R Zbib, E Malchiodi, J Devlin, D Stallard, S Matsoukas, R Schwartz, J Makhoul, F Zaidan, O, and C Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59.

D Zeman, J Hajič, M Popel, M Potthast, M Straka, F Ginter, J Nivre, and S 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.

I Zribi, M Khemakhem, and H Belguith, L. 2013. Morphological Analysis of Tunisian Dialect. In *International Joint Conference on Natural Language Processing*, pages 992–996.

8. Language Resource References

A Amara, and H Turki, and M A Hadj Taieb, and M Ben Aouicha, and K Ellouze. 2021. *CTAB: Corpus of Tunisian Arabizi [Dataset]*. Zenodo. PID https://doi.org/10.5281/zenodo.4781769.

M Maamouri, and A Bies, and T Buckwalter, and H Jin, and W Mekki. 2005. *Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic textscAnalysis) LDC2005T20*. Linguistic Data Consortium. PID https://catalog.ldc.upenn.edu/LDC2005T20.