ISSN 2278-3091

Volume 10, No.3, May - June 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse371032021.pdf

https://doi.org/10.30534/ijatcse/2021/371032021

Universal Dependencies for Urdu Noisy Text

Amber Baig[1], Mutee U Rahman[2], Abdul Salam Shah[3], Suhni Abbasi[4]

[1]Department of Computer Science, Isra University, Hyderabad, Pakistan, amberbaig@gmail.com

[2]Department of Computer Science, Isra University, Hyderabad, Pakistan, muteeurahman@gmail.com

[3]Department of Computer Engineering, University of Kuala Lumpur (UniKl-MIIT), 50250 Kuala Lumpur, Malaysia, shahsalamss@gmail.com

[3]Information Technology Centre, Sindh Agriculture University, Tandojam, Hyderabad, Pakistan, suhni.abbasi@sau.edu.pk

ABSTRACT

In this paper, the process of creating a Dependency Treebank for tweets in Urdu, a morphologically rich and less-resourced language is described. The 500 Urdu tweets treebank is created by manually annotating the treebank with lemma, POS tags, morphological and syntactic relations using the Universal Dependencies annotation scheme, adopted to the peculiarities of Urdu social media text. annotation process is evaluated through Inter-annotator agreement for dependency relations and total agreement of 94.5% and resultant weighted Kappa κ = 0.876 was observed. The treebank is evaluated through 10-fold cross validation using Maltparser with various feature settings. Results show average UAS score of 74%, LAS score of 62.9% and LA score of 69.8%.

Key words: Dependency parsing, low-resourced languages, treebank, Universal Dependencies, Urdu tweets.

1. INTRODUCTION

Rise of Web 2.0 technologies such as Twitter have initiated the appearance of new forms of textual data online. As a result of which, internet users who were previously just information consumers became information producers. This leads to a mass insurgence of an interesting new source of information. According to a recent research by [11], the number of people using Twitter has reached up to 326 million till date. Every day, these users post 500 million tweets which means, every second 5,787 tweets are produced. These tweets share many properties of spoken language. These tweets are often usually informal and are not carefully edited, often lack punctuation, and can include ungrammatical structures. In addition, the data often comprises spelling errors and creative use of language, resulting in large number of unknown words [14]. Moreover, the limited number of letters for each tweet stimulates creativity and encourages an innovative and non-standard language usage together with Twitter-specific elements like emoticons, hashtags, retweet tokens and usernames [8] in the content. Accordingly, such data is referred to as noisy user generated text [14] to represent these deviations from the conventions used in corpora of well-edited news text.

Although, noisy and informal, such content may hold great value for applications such as language technologies, data analysis, sentiment analysis, event detection and opinion mining [16], [22], [34] to name a few. Nevertheless, prior studies e.g. [21], [10], [9] have shown that this noisiness prevents most natural language processing tools, particularly data-driven parsers to attain good performance on such data. These aspects highlight the need for parsers and parsing resources to be adapted to provide adequate coverage of such type of user-generated content. In response to this need, there has been growing number of research studies for parsing such noisy content for languages like English [9], [27], Italian [23], Arabic [2], and German [20] etc. However, Urdu language which is a widely spoken language in South Asia still lacks in this area of NLP.

Amongst 7105 languages of the world, Urdu ranks 19th and is used in countries like Pakistan, India, Afghanistan, Bangladesh and Iran by more than 50 million regular users [26]. However, from language technology perspective, Urdu is still considered as a low-resourced language [19]. Extending coverage of NLP tools and resources to social media data can shed significant light on various scientific questions, be it in theoretical or contrastive linguistics, in linguistic typology, or in NLP itself [15].

Universal Dependencies (UD) [17] is a treebank building project whose goal is to create syntactically annotated corpora in a range of languages using shared annotation principles. Such an approach has the advantage of producing comparable linguistic annotation across corpora, which in turn facilitates research in cross lingual parsing and machine translation, linguistic typology and contrastive linguistics [15]. Analyzing the syntax of noisy user generated text such as tweets can benefit from such adaptability.

In this paper, a new manually annotated treebank of Urdu noisy text (UNTDT) following UD guidelines is introduced together with a statistical parser model which is trained to parse such noisy text. The new treebank also addresses social media-specific challenges which are not addressed by UD guidelines.

The paper is organized as follows. In Section 2, a brief overview of previous research related to parsing of tweets is presented. Section 3 introduces UD framework for Urdu noisy text and the process of training the statistical parser and its evaluation is detailed in Section 4. Section 5 discusses the results of parser evaluation followed by conclusions in Section 6.

2. RELATED WORK

Most of the attempts found in literature largely concentrate on PoS-tagging of tweets, though some studies on syntactic annotation are also present. English is still the most studied language.

For English, [9] developed a 1,000 tweets and forum posts sentences dataset, with the purpose to investigate social media text parsing problems. Afterwards, other works endeavored to overcome these limitations by developing ad hoc training data resources for parsing. French Social Media Bank [25] followed this suite where a 1,700 user-generated sentences set was annotated with a modified French Treebank scheme [1], and TWEEBANK [12], constructed by manual addition of dependency parses to tweets extracted from [18]'s PoS-tagged Twitter corpus.

[2] used bootstrapping for developing a UD based Arabic tweets dependency treebank. A rule-based parser was used for creating a small treebank of 1,000 Arabic tweets. this treebank was used as seed training set by a data-driven parser for creating a bigger treebank. 6,738 Italian tweets were annotated within UD 2.0 framework by [22]. They evaluated the performance of two parsers on this dataset and found that their performance lagged significantly in comparison to their performance on the Italian UD Treebank. A new dataset of 500 tweets within the framework of UD 2.0 was developed and annotated by [5], out of which 250 tweets are in African American English. TWEEBANK V2 was developed by [13] by completely labelling TWEEBANK V1 according to UD 2.0 along with additionally sampled tweets, for a total of 3,550 tweets. They found it challenging to create coherent annotations because of frequent ambiguities in tweets interpretation. Nonetheless, a pipeline for tokenizing, tagging, and parsing the tweets was trained by them, and ensemble and distillation models were developed for parsing accuracy improvement. [20] developed and annotated TweeDe, the first German Twitter treebank, as a new training and test suite for UD parsing. TweeDe includes more than 12,000 tokens of informal private communication, annotated for PoS, morphology and UD syntactic dependencies. They also presented parsing baselines for their dataset, showing that combining small volume of in-domain Twitter data in combination with a larger out-of-domain data volume can yield parsing accuracies in the range of 83% (UAS) and 76% (LAS) on their new test suite.

As far as literature is concerned, the work presented in this paper is the first step towards developing a dependency treebank for Urdu tweets which can benefit a wide range of downstream NLP applications such as information extraction and machine translation.

3. UD FRAMEWORK FOR URDU NOISY TEXT

3.1 Corpus

This study borrows the corpus of 500 (12,723 tokens) Urdu tweets for hand annotation from [4]. This corpus is already preprocessed to be used for machine learning and the tokens of the corpus are already tokenized which is the basic requirement of UD.

3.2 Lemmatization

Lemmatization is a method of extracting a root or dictionary form (lemma) of a given word. Especially for languages with rich morphology it is important to be able to normalize words into their base forms to better support for example search engines and linguistic studies.

However, it is not possible to lemmatize token variations found in noisy text such as emoticons, email addresses and URLs; Twitter hashtags and Twitter at-mentions. Therefore, in such cases, the lemma remains the same as their word form

3.3 POS Tagging

17 POS tags are defined in UD specification with the requirement that all conforming treebanks use only these tags. New POS tags cannot be introduced in this universal POS tagset. However, language specific POS tags can also be used along with universal POS tags. UNTPOS, a new POS tagset for Urdu tweets by extending universal POS tagset proposed by [4] is used for language specific POS tagging. This tagset consists of 33 tags. The mapping between UD CPOS tagset and UNTPOS tagset is presented in Table 1.

Table 1: CPOS Mapping with UNTPOS

Mapping of UD-CPOS with UNTPOS			
CPOS	**UNTPOS**	**CPOS**	**UNTPOS**
ADJ	ADJ	**ADV**	ADV
ADP	ADP		NEG
	ADPT	**AUX**	AUX
CCONJ	CCONJ	**DET**	DET
INTJ	INTJ	**NOUN**	NOUN

	INTJE	**PROPN**	PROPN
NUM	NUM	**PRON**	PRON
	NUMO		PRONR
	NUMQ		PROND
	NUMF		PRONP
	NUMY		PRONRD
X	ICO	**PART**	PART
	RET	**PUNCT**	PUNCT
	REP	**SCONJ**	SCONJ
	LINK	**SYM**	SYM
	HASH	**VERB**	VERB

UNTPOS adopts same UD CPOS tags for tagging adjectives (ADJ), adverbs (ADV), conjunctions (CCONJ and SCONJ), auxiliary verbs (AUX), determiners (DET), nouns (NOUN), proper nouns (PROPN), particles (PART), punctuations (PUNCT), symbols (SYM) and verbs (VERB).

As twitter usernames of form “@username” indicate a real user, UNTPOS considers twitter usernames (mentions) as PROPN. NUM category of numerals is expended by adding tags to represent Urdu cardinal (NUM), ordinal (NUMO), fraction (NUMF), quantitative (NUMQ) and multiplicative numerals (NUMY). Similarly, ADP is extended for representing Urdu prepositions (ADP) and postpositions (ADPT), PRON class expended to represent Urdu personal pronouns (PRON), possessive pronouns (PRONP), reflexive pronoun (PRONR), demonstrative pronoun (PROND) and relative pronoun (PRONRD).

UNTPOS introduced new tags NEG, RET, REP, LINK and HASH, for tagging negations and twitter specific elements like retweets, replies, email address or web addresses and hashtags respectively. Emoticons or emojis are classified in two categories. First category includes emoticons or emojis showing emotions (e.g. □, □, □ etc.) and second category contains general emoticons or emojis (e.g. , □). As interjections express emotions, volition and moods, INTJ category is expended by introducing a new tag INTJE for first category, and a new tag ICO is introduced for second category in UNTPOS.

3.4 Morphological Features

The UD annotation schema defines a set of 21 morphological features across languages. Features are divided into the categories of lexical features and inflectional features. Lexical features are the characteristics of the lemmas, whereas inflectional features are the characteristics of the word forms. In contrast to the POS tag, the language specification allows treebanks to introduce morphological features that are not included in this universal inventory. This suggests that morphological features can be drawn from the extended compilation of morphological features of other languages (Zeman, 2008). Table 2 summarizes the morphological features used in Urdu tweet treebank annotation.

Urdu nouns (common noun and proper noun) can have masculine or feminine gender, take singular or plural number, can be in first, second- or third-person form and have three cases: oblique, nominative and vocative. Adjectives (ADJ) have the gender, case and number features dependent on associated nouns.

Table 2: Morphological Features of Urdu Noisy Text

UNTPOS Tags	**Morphological Feature**
NOUN, PROPN	Case=Nom, Acc, Voc |Gender=Masc, Fem |Number=Sing, Plur|Person=1,2,3
ADJ	Gender=Masc, Fem |Number=Sing, Plur| Case=Nom, Acc
NUM, NUMO, NUMY, NUMF	NumType=Card, Ord, Mult, Frac
PRON, PROND, PRONR, PRONP, PRONRD	Case=Nom, Acc | Gender=Masc, Fem | Number=Sing, Plur| Person=1,2,3 | PronType=Rel, Dem, Prs
VERB, AUX	Mood=Ind, Sub |Number= Sing, Plur | Person=1,2,3 |Tense=Pres,Past,Fut | VerbForm=Fin,Inf,Part | Voice=Act, Pass | VerbType=Aux
ADV, NEG	AdvType=Deg, Man, Loc, Tim | Polarity=Neg | PronType=Neg
ADP, ADPT	AdpType=Prep,Pos
DET	Case=Nom, Acc | Number=Sing, Plur| Person=1,2,3 | Definite=Def, Ind
PART, SCONJ, CCONJ, INTJ, INTJE, X, RET, REP, LINK, HASH, ICO, PUNCT, SYM	-

Verbs (VERB) have complicated inflection among Urdu word categories. Verbs may be in form of first, second or third person, may either be singular or plural, may have three moods: perfective, subjunctive and Imperfective with two voices: active and passive and three tenses: present, past and future, with different characteristics in different moods. Moreover, verbs are also distinguished as auxiliary verbs (AUX) using morphological feature VerbType. For adverbs (ADV), AdvType feature is used to classify adverbs of time, location, manner and degree. Negations (NEG) are defined as part of adverbs (ADV) using features polarity and PronType.

Pronouns are distinguished between personal (PRON), relative (PRONRD) and demonstrative (PROND) pronouns using PronType. Whereas possessive (PRONP) and reflexive (PRONR) pronouns are considered a subset of the personal pronouns. Pronouns also have case, person, gender and number features. For determiners and articles, the distinction between definite and indefinite articles and determiners is indicated by the Definite feature in the DET class.

Urdu cardinal, ordinal, fraction and multiplicative numerals have single morphological category: NumType. Whereas quantitative numerals have the same morphological categories as adverb. Feature AdpType is used to distinguish between two types of adpositions: prepositions (ADP) and postpositions (ADPT).

The other PoS tags for particles (PART), conjunctions (CCONJ, SSCONJ), interjections (INTJ, INTJE), punctuations (PUNCT, SYM), and the remainder class (X, ICO, REP, RET, LINK, HASH) do not have any features because they do not inflect.

3.5 Syntactic Relations

The Universal Dependencies V 2.0 grammatical relations listed in Table 3 are carefully followed for Urdu noisy text. however, there are some deviations from UD conventions due to noisiness in text which are explained below. No language-specific relations were included and csubj, csubjpass, expl, orphan and reparandum relations have not been implemented because they did not occur in our corpus but will be applied in future if found.

Table 3: UD V 2.0 Relations

Universal Dependencies V 2.0 Syntactic Relations				
acl	ccomp	discourse	mark	punct
advcl	clf	dislocated	nmod	reparandum
advmod	compound	expl	nsubj	root
amod	conj	fixed	nummod	vocative
aux	cop	flat	obj	xcomp
appos	csubj	goeswith	obl	
case	dep	iobj	orphan	
cc	det	list	parataxis	

For Urdu Noisy Text, hashtags, emoticons and emojis used at the end of sentence are marked with discourse relation. However, hashtags used in sentences are marked according to their role in a sentence. Example of emoticon annotation is shown in Figure 1, hashtag annotation is shown in Figure 2 and Figure 3 shows annotation of hashtags used between sentences.

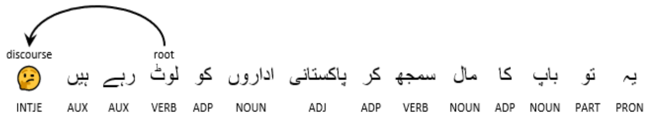
discourse
root
یہ تو باپ کا مال سمجھ کر پاکستانی اداروں کو لوٹ رہے ہیں
INTJE AUX AUX VERB ADP NOUN ADJ ADP VERB NOUN ADP NOUN PART PRON

Figure 1: discourse relation example 1

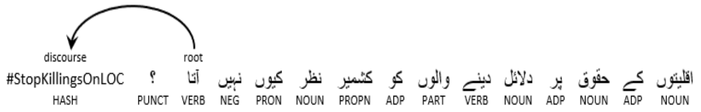
discourse
root
#StopKillingsOnLOC
اقلیتوں کے حقوق پر دلائل دینے والوں کو کشمیر نظر کیوں نہیں آتا ؟
HASH PUNCT VERB NEG PRON NOUN PROPN ADP PART VERB NOUN ADP NOUN ADP NOUN

Figure 2: discourse relation example 2

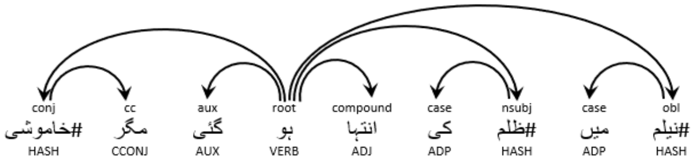
conj
cc
aux
root
compound
case
nsubj
case
obl
#نیلم میں #ظلم کی انتہا ہو گئی مگر #خاموشی
HASH CCONJ AUX VERB ADJ ADP HASH ADP HASH

Figure 3: discourse relation example 3

In UNTDT, retweets are treated with dislocated relation. Example is shown in Figure 4.

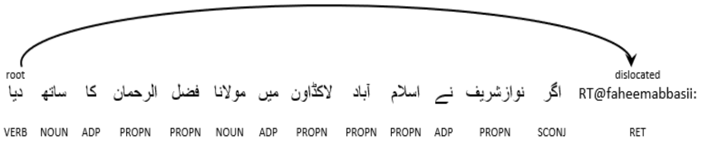
root
dislocated
RT@faheemabbasii:
اگر نوازشریف نے اسلام آباد لاکڈاون میں مولانا فضل الرحمان کا ساتھ دیا
VERB NOUN ADP PROPN PROPN NOUN ADP PROPN PROPN PROPN ADP PROPN SCONJ RET

Figure 4: dislocated relation example

URLs and email address in UNTDT are marked as list. Example is shown in Figure 5.

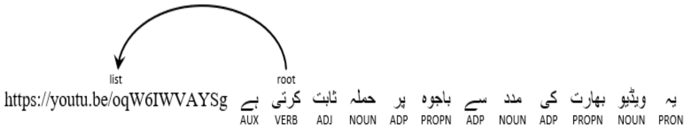
list
root
https://youtu.be/oqW6IWVAYSg
یہ ویڈیو بھارت کی مدد سے باجوہ پر حملہ ثابت کرتی ہے
AUX VERB ADJ NOUN ADP PROPN ADP NOUN ADP PROPN NOUN PRON

Figure 5: list relation Example

Since vocative relation is used to mark entity being addressed directly in the dialog, therefore, tweet mentions, and tweet replies are also treated as vocatives in UNTDT. Example is shown in Figure 6.

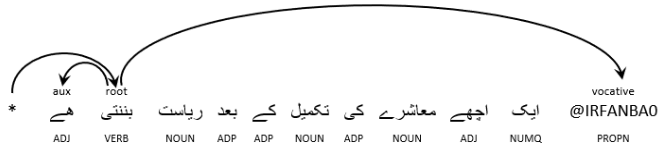
aux
root
vocative
@IRFANBA0
ایک اچھے معاشرے کی تکمیل کے بعد ریاست بننتی ہے *
ADJ VERB NOUN ADP ADP NOUN ADP NOUN ADJ NUMQ PROPN

Figure 6: vocative relation Example

In UNTDT, tweet sentences in which emoticons/emojis are used in between words of a sentence; those emoticons/emojis are marked as xcomp as shown in Figure 7.

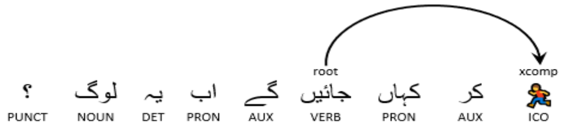
root
xcomp
کر کہاں جائیں گے اب یہ لوگ ؟
PUNCT NOUN DET PRON AUX VERB PRON AUX ICO

Figure 7: xcomp relation example

3.6 Manual Annotation and Correctness Evaluation

Since the corpus of 500 tweets (12,723 tokens) was borrowed from [4], it was already tagged with language specific UNTPOS tags. Manual annotation of the corpus with lemma, CPOS, morphological features and syntactic relations was done by two annotators. Both the annotators are native Urdu speakers having knowledge and command of the annotation process. The manual annotation was then revised and corrected by an expert, also a native Urdu speaker and has years of research and development experience in the field of NLP. WebAnno [28] is the primary software used for manual annotation, review, and correction process.

To measure the reliability of syntactic relations, inter-annotator agreement (IAA) on the annotated tweets is calculated. Cohen's Kappa [7] that measures the degree of agreement between the assigned labels, correcting for agreement by chance is used for computing IAA. The observed total agreement is 94.5% and resultant weighted Kappa reached κ = 0.876, which indicates that the annotations are reliable [6].

4. PARSER TRAINING AND EVALUATION

After manual annotation and revision of 500 gold standard tweets (12,723 tokens), a statistical parser model was trained, and a baseline parsing score intended to be used to compare future parsing models was established. MaltParser [29], a multilingual transition-based parser which provides numerous effective deterministic parsing algorithms capable of producing a dependency tree in linear or quadratic time was trained on this 500 gold standard tweets and its performance was evaluated through 10-fold cross validation technique.

For parser performance evaluation, MaltEval [30] is used and Labelled Attachment Score (LAS), Unlabelled Attachment Score (UAS) and Label Accuracy (LA) are used as evaluation metrics. These three measures are basically token-level accuracies, that accounts for all test data tokens, giving equal weightage to each token in the evaluation. Formula for calculating LAS, UAS and LA are shown in equations (1), (2) and (3).

$$LAS = \frac{number\ of\ correct\ head\ \&\ dependency\ labels}{total\ tokens} \quad (1)$$

$$UAS = \frac{number\ of\ correct\ head\ labels}{total\ tokens} \quad (2)$$

$$LA = \frac{number\ of\ correct\ labels}{total\ tokens} \quad (3)$$

To test the impact of lemma, POS tags and morphological features on parsing performance, three different feature settings were used to train MaltParser models for experiments. The first feature setting includes word forms along with POS tags (CPOS + FPOS). The second feature setting includes the word form, lemma and the POS tags. The final feature setting includes the word form, lemma, part-of-speech and morphological features. The purpose of all these feature settings was to establish best performing feature set combination and a baseline score which can then be used to compare future parsing models. For training, Covington non-projective and LIBLINEAR learner algorithms of MaltParser were used.

5. RESULTS

The average results of 10-fold cross validation experiments are shown in table 4. Although, there was just a minor difference between average scores of the three parser feature model settings, the model with best average UAS score of 74%, LAS score of 62.9% and LA score of 69.8% was the one which utilized information from the word form, lemma, both CPOS and FPOS tags and morphological features.

Table 4: Average 10-fold Cross Validation Result

Model	LA	UAS	LAS
Form+POS	68.7	73.1	61.8
Form+Lemma+POS	69.26	73.85	62.03
Form+Lemma+POS+Feats	69.8	74	62.9

While comparing baseline scores with other smaller size treebanks; the Irish treebank [31] with 300 sentences had an accuracy of 63.3% for LAS and 73.3% for UAS. Tamil UD treebank with 600 sentences resulted in an accuracy of 64.8% for UAS and 56.3% for LAS, Buryat treebank (418 sentences) yielded an accuracy of 65.44% for UAS and 43.29% for LAS and Romanian treebank (633 sentences) reported 68.4% for UAS and 56.4% for LAS [32]. Yorùbá Treebank [33] with 100 sentences had an accuracy of 63.12% UAS and 53.07% LAS. All these treebanks are of standard well-edited text languages whereas UNTDT comprised of non-standard texts i-e. tweets. Therefore, the results achieved on baseline can be deemed as promising on this small dataset. At the time of this study, there were no similar works available for Urdu noisy text to compare our work with. However, in terms of other language's noisy treebanks, the largest one is German tweets treebank tweeDe, with more than 12,000 tweets. tweeDe has an accuracy of 80.65% UAS and 72.69% LAS [20]. PoSTWITA-UD, an Italian tweet treebank is second largest Tweet treebank with 6,700 tweets and has an accuracy of 86.95% UAS and 81.5% LAS [22]. In comparison to these treebanks, UNTDT is small in volume. However, it is expected that with increase in training data, parsing accuracy will significantly increase.

6. CONCLUSION

With the aim of developing a gold standard corpus for training and testing of statistical dependency parser on social media text, this paper presents a new dependency tree bank for Urdu noisy text, UNTDT. The treebank is manually annotated at morphological and syntactic level by adopting UD framework to the particularities of social media text by two annotators. The consistency and correctness of this treebank is checked by a linguistic expert while inter-annotator agreement for dependency relations was also calculated with observed total agreement of 94.5% and resultant weighted Kappa κ = 0.876. Currently, the treebank has 500 gold standard Urdu tweets with 12,723 tokens. A 10-fold cross validation of this treebank using Maltparser with various feature settings was also performed with best average UAS score of 74%, LAS score of 62.9% and LA score of 69.8%.

Future work includes further expansion of the treebank developed in this study and the development of parsers based on this dataset to aid further research in Urdu NLP where the lack of training data has remained an obstacle. The treebank developed in this study is publicly available at: https://github.com/amberbaig/Urdu-Noisy-Text.

ACKNOWLEDGEMENT

This work is partially supported by Isra University Hyderabad, Sindh, Pakistan. We also acknowledge the efforts of our annotators in producing the first version of the syntactic annotations, and in facilitating the process of UD adoption for Urdu tweets.

REFERENCES

1. A. Abeillé, L. Clément, and F. Toussenel. **Building a treebank for French**, in *Treebanks*, ed: Springer, 2003, pp. 165-187.
2. F. Albogamy, A. Ramsay, and H. Ahmed. **Arabic tweets treebanking and parsing: A bootstrapping approach**, in *Proceedings of the Third Arabic Natural Language Processing Workshop*, 2017, pp. 94-99.
3. W. Ammar, G. Mulcaire, M. Ballesteros, C. Dyer, and N. A. Smith. **Many languages, one parser**, *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 431-444, 2016.
4. A. Baig, M. U. Rahman, H. Kazi, and A. Baloch. **Developing a POS Tagged Corpus of Urdu Tweets**, *Computers*, vol. 9, p. 90, 2020.
5. S. L. Blodgett, J. Wei, and B. O'Connor. **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)*, 2018, pp. 1415-1425.
6. J. Carletta. **Assessing agreement on classification tasks: the kappa statistic**, *Computational Linguistics*, vol. 22, pp. 249–254, 1996.
7. J. Cohen. **A coefficient of agreement for nominal scales**, *Educational and psychological measurement*, vol. 20, pp. 37-46, 1960.
8. J. Eisenstein. **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*, 2013, pp. 359-369.
9. J. Foster, O. Cetinoglu, J. Wagner, J. Le Roux, S. Hogan, J. Nivre, *et al.* **# hardtoparse: POS Tagging and Parsing the Twitterverse**, 2011. Available online: https://hal.archives-ouvertes.fr/hal-00702445/file/aaai_mt_2011.pdf (accessed on 5 November 2020).
10. K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, *et al.* **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: Short Papers-Volume 2*, Portland, OR, USA, 19–24 June 2011.
11. P. Cooper. **Twitter stats all marketers need to know in 2020**, *Published October*, vol. 30, p. 2019.
12. L. Kong, N. Schneider, S. Swayamdipta, A. Bhatia, C. Dyer, and N. A. Smith, **A dependency parser for tweets**, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1001-1012.
13. Y. Liu, Y. Zhu, W. Che, B. Qin, N. Schneider, and N. A. Smith. **Parsing tweets into universal dependencies**, *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. Vol. 1 (Long Papers), pp. 965–975, 2018.
14. T. Lynn, K. Scannell, and E. Maguire. **Minority language twitter: Part-of-speech tagging and analysis of Irish tweets**, *In Proceedings of the ACL 2015Workshop on Noisy User-generated Text*, Bejing, China, 31 July 2015; pp. 1–8.
15. A. Miletic, M. Bras, M. Vergez-Couret, L. Esher, C. Poujade, and J. Sibille. **Building a Universal Dependencies Treebank for Occitan**," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 2932-2939.
16. R. U. Mustafa, M. S. Nawaz, M. I. U. Lali, T. Zia, and W. Mehmood. **Predicting the cricket match outcome using crowd opinions on social networks: A comparative study of machine learning methods**, *Malaysian Journal of Computer Science*, vol. 30, pp. 63-76, 2017.
17. J. Nivre, M.-C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, *et al.* **Universal dependencies v1: A multilingual treebank collection**, in *Proceedings of the Tenth International Conference on Language*

Resources and Evaluation (LREC'16), 2016, pp. 1659-1666.

18. O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. **Improved part-of-speech tagging for online conversational text with word clusters**, in *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2013, pp. 380-390.
19. A. A. Raza, A. Habib, J. Ashraf, and M. Javed. **A review on Urdu language parsing**, *Int. J. Adv. Comput. Sci. Appl*, vol. 8, pp. 93-97, 2017.
20. I. Rehbein, J. Ruppenhofer, and B.-N. Do. **tweeDe–A Universal Dependencies treebank for German tweets**," in *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, 2019, pp. 100-108.
21. A. Ritter, S. Clark, and O. Etzioni. **Named entity recognition in tweets: an experimental study**, in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 1524-1534.
22. M. Sanguinetti, C. Bosco, A. Lavelli, A. Mazzei, O. Antonelli, and F. Tamburini. **PoSTWITA-UD: An Italian Twitter treebank in universal dependencies**, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
23. M. Sanguinetti, C. Bosco, A. Mazzei, A. Lavelli, and F. Tamburini. **Annotating Italian social media texts in universal dependencies**, in *Fourth International Conference on Dependency Linguistics (Depling 2017)*, 2017, pp. 229-239.
24. M. Sanguinetti, C. Bosco, A. Mazzei, A. Lavelli, and F. Tamburini. **Annotating Italian social media texts in universal dependencies**, in *Fourth International Conference on Dependency Linguistics (Depling 2017)*, 2017, pp. 229-239.
25. D. Seddah, B. Sagot, M. Candito, V. Mouilleron, and V. Combet. **The French Social Media Bank: a treebank of noisy user generated content**, in *COLING 2012, 24th International Conference on Computational Linguistics*, Mumbai, India, 2012, pp. 2441– 2458.
26. G. F. Simons and C. D. Fennig, *Ethnologue: languages of Asia*: sil International Dallas, 2017.
27. H. Wang, Y. Zhang, G. L. Chan, J. Yang, and H. L. Chieu. **Universal dependencies parsing for colloquial singaporean english**," *arXiv preprint arXiv:1705.06463,* 2017.
28. S. M. Yimam, I. Gurevych, R. E. de Castilho, and C. Biemann. **Webanno: A flexible, web-based and visually supported system for distributed annotations**, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2013, pp. 1-6.
29. J. Nivre, J. Hall, and J. Nilsson. **Maltparser: A data-driven parser-generator for dependency parsing**, in *LREC,* 2006, pp. 2216-2219.
30. J. Nilsson and J. Nivre, "MaltEval: an Evaluation and Visualization Tool for Dependency Parsing," in *LREC*, 2008.
31. T. Lynn, O. Cetinoglu, J. Foster, E. Uí Dhonnchadha, M. Dras, and J. van Genabith, **Irish treebanking and parsing: A preliminary evaluation**, in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12*) 2012, pp. 1939-1946
32. E. Badmaeva and F. M. Tyers. **A Dependency Treebank for Buryat**, in *TLT*, 2017, pp. 1-12.,
33. O. Ishola. **Universal Dependencies for Yorùbá**, Master's thesis, Eberhard Karls Universität Tübingen & Charles University, 2019.
34. N. F. Abd Yusof, C. Lin, X. Han, and M. H. Barawi. **Split Over-Training for Unsupervised Purchase Intention Identification**, *International Journal of Advanced Trends in Computer Science and Engineering,* Vol. 9, No. 3, 2020, pp. 3921–3928, 2020. https://doi.org/10.30534/ijatcse/2020/214932020