A Universal Dependencies Treebank for Gujarati

Mayank Jobanputra[1*], **Maitrey Mehta**[2*], **Çağrı Çöltekin**[3]
[1]Department of Language Science and Technology, Saarland University
[2]Kahlert School of Computing, University of Utah
[3]Department of Linguistics, University of Tübingen
mayank@lst.uni-saarland.de, maitrey@cs.utah.edu, ccoltekin@sfs.uni-tuebingen.de

Abstract

The Universal Dependencies (UD) project has presented itself as a valuable platform to develop various resources for the languages of the world. We present and release a sample treebank for the Indo-Aryan language of Gujarati – a widely spoken language with little linguistic resources. This treebank is the first labeled dataset for dependency parsing in the language and the script (the Gujarati script). The treebank contains 187 part-of-speech and dependency annotated sentences from diverse genres. We discuss various idiosyncratic examples, annotation choices and present an elaborate corpus along with agreement statistics. We see this work as a valuable resource and a stepping stone for research in Gujarati Computational Linguistics.

Keywords: low-resource languages, universal dependencies, Gujarati

1. Introduction

The Universal Dependencies (UD) project (Nivre et al., 2016; de Marneffe et al., 2021) offers cross-linguistically consistent annotations for dependency treebanks, part-of-speech, and morphological features. The ever-expanding language base under the UD umbrella ensures that similar language patterns can be dealt with consistently when working with a new language. Further, language-specific features are brought to the fore for discussion. As a result, UD becomes the most fundamental of resources to be developed for a particular language.

Gujarati is an Indo-Aryan language originating from the western Indian state of Gujarat. The language is widely spoken by over 56 million speakers (Eberhard et al., 2022) and is one of the 22 languages with official status in India. Yet, the Gujarati Computational Linguistics community is still in its infancy. Joshi et al. (2020) classify Gujarati in the "Scraping-Bys" category (category 1) in their taxonomy indicating a scant availability of labeled datasets. Basic resources such as part-of-speech taggers, and named entity recognizers are not readily available. Hence, a dependency treebank in such a language can have a wide-reaching impact.

On the other hand, the UD community has already produced a handful of treebanks in various Indo-Aryan languages. As a result, we are equipped with resources in related languages like Marathi (Ravishankar, 2017), Hindi (Bhat et al., 2017; Zeman et al., 2017), and Punjabi (Arora, 2022). Such resources are of value while constructing a sample Gujarati treebank.

The benefits of building a sample Gujarati treebank are four-fold:

a) It presents as a valuable resource for the development of linguistic tools and resources in a low-resource language, i.e., Gujarati.

b) Gujarati uses a unique eponymous script that is not yet represented in the UD project. This can be especially valuable for future researchers interested in building resources for lesser-resourced languages such as Kutchi, and Bhili that also use the Gujarati script.[1]

c) It ensures annotation paradigms in similar contexts are adhered to and helps point out any discrepancies in existing treebanks.

d) We can point out some new idiosyncratic phenomena that might be Gujarati-specific, or missed by earlier works.

The above-mentioned reasons motivate us to propose a sample dependency treebank for Gujarati: *GujTB*.[2] In the subsequent sections, we explain the selected corpora, statistics and highlight some interesting discussion points encountered.

2. The Dataset

In this section, we provide details of the annotated corpora and the annotation process.

Corpora. We investigated available corpora that include Gujarati text such as IndicCorp (Kakwani et al., 2020) and Samanantar (Ramesh et al., 2022). However, we observe that these datasets majorly contain news and other formal

[*]Both authors contributed equally.

[1]https://www.omniglot.com/writing/languages.htm

[2]Code & Data available at: https://github.com/UniversalDependencies/UD_Gujarati-GujTB

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 56–62. May 25, 2024.

texts. Hence, we annotate a total of 187 sentences taken from diverse sources like Samanantar (news), UD Cairo (short),[3] Gujarati translations (from Mehta and Srikumar, 2023) of the French novella – *Le Petit Prince* (fiction) (The Little Prince, de Saint-Exupéry, 1943), and a Gujarati grammar book (grammar) (Raimond, 2004).

Annotation Process and Agreement. Two of the paper authors[4] annotated this dataset. The annotations were created separately, and followed by an initial correction phase to fix any obvious errors. A hundred-sentence subset of annotations was considered for the inter-annotator agreement (IAA) study.[5] The IAA for the part-of-speech (POS) tags is 99.87 (Cohen's κ). The head selection agreement is 99.44% and the relation agreement on the heads that matched is 99.88 (Cohen's κ). The head selection agreement is the proportion of dependents assigned the same head by both annotators (similar to the unlabeled attachment score).

Dataset Statistics. The dataset statistics by genre are given in Table 1. The distribution of POS tags in the corpus is given in Table 2. Furthermore, we provide the statistics regarding dependency relations in Table 3. Notably, our dataset is a representative set of all possible relations in Gujarati.

Genre	Sentences	Tokens
news	93	1159
short	20	178
fiction	40	331
grammar	34	217
Total	187	1885

Table 1: Data statistics by genre for GujTB.

3. Syntactic Relations

In this section, we discuss the many interesting dependency choices. While a large volume of dependency choices such as subjects, object, and light/serial verb constructions follow existing Indo-Aryan literature (Bhat et al., 2017; Ravishankar, 2017; Ojha and Zeman, 2020; Arora, 2022), our goal is to highlight the more subjective cases.

Interrogative/Question particles. The treatment of interrogative or question particles has largely varied in the UD literature.[6] We follow the preceding Indo-Aryan treebanks in assigning question particles with the respective dependency and POS tags as what would be assigned for a valid answer substitution. However, in cases where an obvious substitution is not viable (e.g., Yes/No questions) as shown in Example 1, we find that an aux relation fits the best.

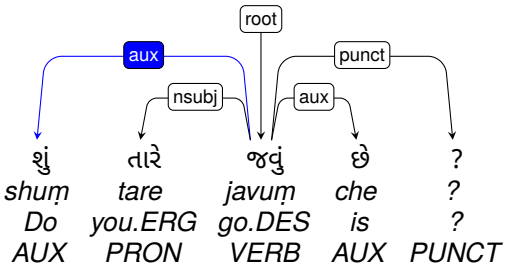
POS	Counts	POS	Counts
NOUN	425	CCONJ	50
PUNCT	250	PART	43
VERB	213	NUM	40
AUX	185	DET	23
ADP	152	INTJ	14
PROPN	145	SCONJ	13
ADJ	134	SYM	3
PRON	133	X	2
ADV	60	Total	1885

Table 2: Part-of-speech tag statistics.

Relation	Counts	Relation	Counts
punct	250	nummod	27
root	187	det	21
nsubj	174	acl	17
case	151	mark	14
aux	133	ccomp	13
nmod	129	appos	13
obl	110	parataxis	13
obj	99	iobj	11
amod	96	orphan	3
compound	70	dislocated	3
advmod	62	goeswith	3
conj	59	fixed	2
cc	52	xcomp	2
cop	51	vocative	1
discourse	44	reparandum	1
flat	36	Total	1885
advcl	35		

Table 3: Dependency relation statistics. All relation sub-types have been merged with their universal classes for representation.

(1)



'Do you want to go ?'

[3] https://github.com/UniversalDependencies/cairo

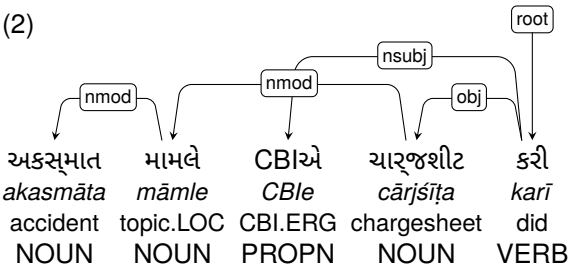
[4] Both are L1 speakers of Gujarati

[5] We release both the individual and adjudicated dataset as per Plank (2022)'s suggestion.

[6] https://github.com/UniversalDependencies/docs/issues/738

Non-projectivty. Bhat et al. (2017, pp.23) discuss non-projectivity in Hindi. Gujarati allows non-projective trees in a similar spirit. Partial free word order as shown in Example 2 can give rise to overlapping dependency edges.

(2)

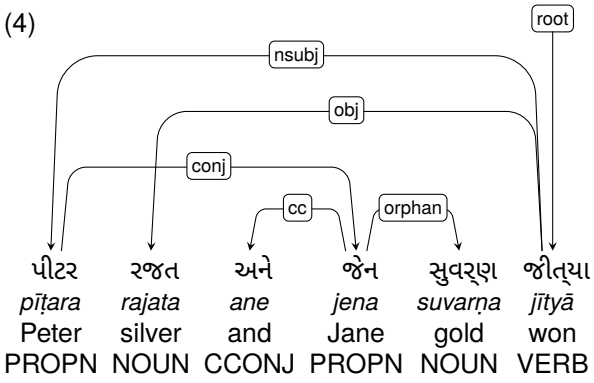


'CBI made a chargesheet about the accident'

Head-final conjunctions. UD guidelines necessitate that the head of a conjunctive phrase be the first conjunct. However, Gujarati carries case inflections and post-positional attachments on the final conjunct which mediate semantic relations between the governor and the conjunctive phrase (see Example 3). This may lead to unwarranted non-projectivity as shown in Example 4.

Note that, in Example 4, the English translation fails to mark plurality on the verb "*won*" while in Gujarati "*jītyā*" has a plural inflection. As a result, the entire conjunctive phrase, not individual proper nouns (*Peter* or *Mary*), has to be the subject. At first sight, the non-projectivity in this example may seem avoidable by annotating promoted subject "*pīṭara*" as `root`, and attaching "*rajata*" to "*pīṭara*" as `orphan`, with the second clause attached as `conj` to the first clause. However, this would cause the plural verb to agree with a singular subject which is not the head of the coordinated structure. Similar issues also arise due to fixed head-initial coordination rule in UD for other head-final languages (Çöltekin, 2015; Kanayama et al., 2018; Tyers et al., 2017; Han et al., 2020). Hence, an argument can be made to mark the final conjunct as the head of the conjunctive phrase. However, we follow the UD guidelines and mark the first conjunct to be the head of the phrase.

(4)

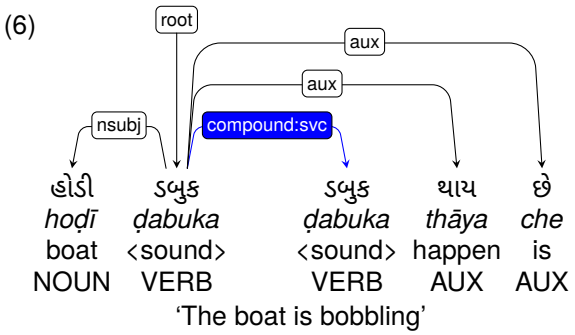


'Peter won silver and Jane gold'

Polarity/emphatic markers within serial verb constructions. Gujarati supports verb-verb constructions where the second verb is, usually, semantically bleached. Owing to the existence of partial free-word ordering discussed before, we observe that serial verb constructions are often separated by polarity or emphatic particles as seen in Example 5. To the best of our knowledge, this case is idiosyncratic to Gujarati. However, note that the treatment of these particles does not change.

Ideophonic verbs. In Gujarati, repetitions of a word can occur in two cases: discursive repetitions (બોલ બોલ ["tell tell"], જા જા ["go go"]) and onomatopoeias (ધમ ધમ ["dham dham"], the sound of Indian drums). Example 6 presents a case of onomatopoeias. Szubert et al. (2021) introduced `parataxis:repeat` for expressing adjectival repetitions in child-directed speech. Sulubacak et al. (2016) use `compund:redup` for reduplicated words. In our case, onomatopoeias are used to imitate different sounds that express actions and act as verbal repetitions. Hence, we suggest using `compound:svc`. To indicate the ideophonic nature of the verb, we mark the feature `VerbType=Ideo`.[7]

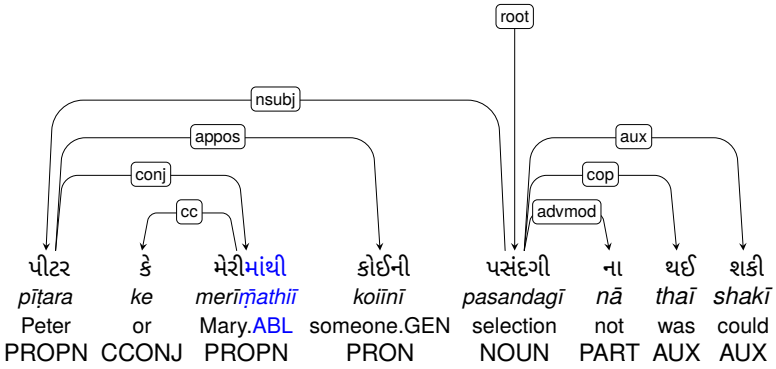
(6)



'The boat is bobbling'

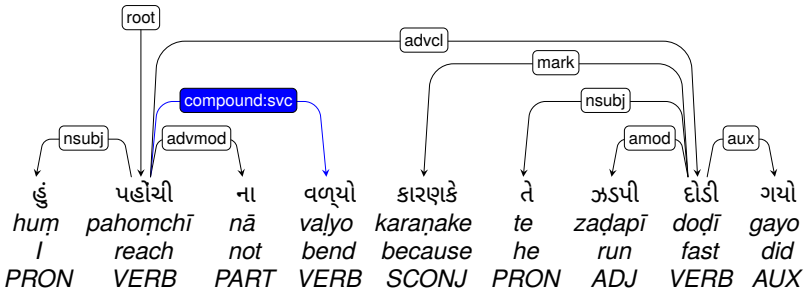
[7] As noted in https://github.com/UniversalDependencies/docs/issues/842

(3)



'No one from Peter or Mary got selected'

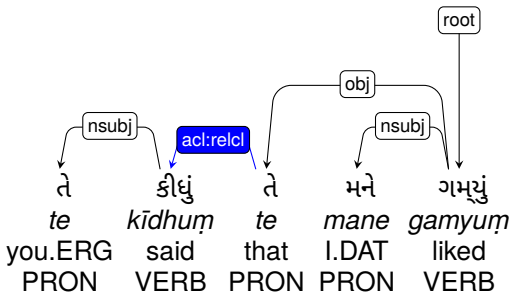
(5)



'I could not keep up because he ran fast'

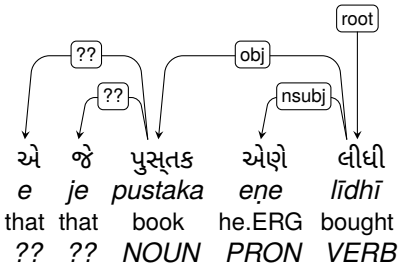
Absence of clausal subjects. We find that clausal subjects do not exist in Gujarati. We substantiate this argument using an English example, *"What she said is likable."*: i) A perfect translation of this sentence does not exist in Gujarati. A close translation is given in Example 7. Note that a co-referential pronominal તે [te, that] is added to construct a grammatically sound sentence. ii) Secondly, the presence of a dative nominal construction with experiencer semantics is permitted. Such constructions are considered grammatical subjects (Arora, 2022) which makes clausal subjects impossible. iii) Finally, the mandatory co-referential pronominal mediates the relation between the governor and the would-be subject clause.

(7)



'What you said is liked by me'

(8)



'That book which he took'

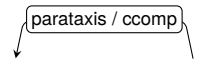
Challenging Construction. Example 8 depicts a case where arguments can be made for multiple possible annotations: i) Assigning `det:predet` to એ [e] and `det` જે [je] with પુસ્તક [pustaka] as their head ii) One may argue a change in order between "જે" and "પુસ્તક", where "જે" would act as a subordinating conjunction. However, we contend a semantic difference between this sentence and the one presented in Example 8. We lean towards the first annotation.

Quoter and Quotation. We encounter a screenplay dialog-style quotation that is yet to be resolved (see Example 9).[8] Recent guidelines recommend `ccomp` over `parataxis` for reported

[8]This is not a Gujarati-specific issue. Moreover, we have opened a discussion regarding this point:

speech.[9] We believe this to be a much more pervasive (and not a Gujarati-specific) issue; applicable, perhaps, when UD is extended to plays.

(9)



I play football : Mark
'I play football : Mark'

4. Tokenization and Part of Speech

Splitting Genitive Markers. Certain nominals (and, in some instances, verbs) in Gujarati are inflected for case. It is unclear if these suffixes should be separated from their heads. This is a known issue that has been raised in Ravishankar (2017). They choose to split genitive markers to be consistent with Hindi. We follow the same rule with the added incentive to separate out layer III postpositions that pair postpositions with preceding genitive markers (Masica, 1993).

The Case for Determiners. According to Gujarati grammars (Tisdall, 1892; Doctor, 2004), demonstrative pronouns like એ [e], તે [te], પેલું [peluṃ], etc. behave differently when attached to a nominal, versus when used independently. When occurring independently, we treat them as pronouns. Tisdall (1892) argues to treat them as adjectives when used with nominals (e.g., એ કૂતરો 'that dog'). Gujarati grammar does not discuss determiners as such. However, we see this usage closer to the UD definition of determiners and hence use the same.

Modal auxiliaries. There are several verbs that can be compounded with other verbs, nouns, or adjectives to form verb compounds. While most of these are semantically bleached, Gujarati identifies a fixed set of verbs to act as modal auxiliaries (Doctor, 2004). This fixed set includes verbs like 'જા [jā,go], આવ [āva,come], રહે [rahe,stay]' (temporal), 'કર [kara,do], લાગ [lāga,feel]' (compulsion), and 'પડ [pada,fell],જોઈ [joī,want]' (obligation). We mark these fixed set of verbs as auxiliaries while the rest are marked as regular verbs.

5. Conclusion and Future Work

We present the first dependency treebank in the Gujarati language and script. We provided detailed dataset statistics and discussed interesting examples and decisions. In a low-resourced language like Gujarati, we see this sample treebank as an enabler for future computational linguistics research. In the future, we aim to increase the size of the annotated corpora to help contribute a dependency parser. Furthermore, we also intend to provide annotations for the morphological features of Gujarati.

6. Ethics Statement

The dataset presented in this work is a voluntary annotation effort between the two authors of this paper. While the annotators speak different dialects of Gujarati, we are aware that our corpus might not contain diverse dialectical varieties.

Acknowledgements

We thank Dr. Atul Kr. Ojha and Aryaman Arora for their useful insights. We thank the anonymous reviewers for their valuable feedback. We highly appreciate the feedback received at the UniDive (CA21167) General Meeting 2023 where we presented an extended abstract version of this work. The format of the paper is inspired by Arora (2022).

7. Bibliographical References

Aryaman Arora. 2022. Universal Dependencies for Punjabi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5705–5711, Marseille, France. European Language Resources Association.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia. 2017. *The Hindi/Urdu Treebank Project*, pages 659–697. Springer Netherlands, Dordrecht.

Çağrı Çöltekin. 2015. A grammar-book treebank of Turkish. In *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 35–49.

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

de Saint-Exupéry, Antoine. 1943. *Le petit prince [The little prince]*. Reynal & Hitchkock (US), Gallimard (FR).

Doctor, Raimond. 2004. *A Grammar of Gujarati*. Lincom Europa.

https://github.com/UniversalDependencies/docs/issues/904

[9] https://universaldependencies.org/changes.html#reported-speech

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World. Twenty-fifth edition.* SIL International, Dallas, Texas.

Ji Yoon Han, Tae Hwan Oh, Lee Jin, and Hansaem Kim. 2020. Annotation issues in Universal Dependencies for Korean and Japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 99–108, Barcelona, Spain (Online). Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Kakwani, Divyanshu and Kunchukuttan, Anoop and Golla, Satish and N.C., Gokul and Bhattacharyya, Avik and Khapra, Mitesh M. and Kumar, Pratyush. 2020. *IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages*. Association for Computational Linguistics.

Hiroshi Kanayama, Na-Rae Han, Masayuki Asahara, Jena D. Hwang, Yusuke Miyao, Jinho D. Choi, and Yuji Matsumoto. 2018. Coordinate structures in Universal Dependencies for head-final languages. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 75–84, Brussels, Belgium. Association for Computational Linguistics.

Colin P Masica. 1993. *The indo-aryan languages*. Cambridge University Press.

Mehta, Maitrey and Srikumar, Vivek. 2023. *Verifying Annotation Agreement without Multiple Experts: A Case Study with Gujarati SNACS*. Association for Computational Linguistics.

Luís Morgado da Costa, Francis Bond, and Roger V. P. Winder. 2022. The Tembusu Treebank: An English Learner Treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4817–4826, Marseille, France. European Language Resources Association.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Atul Kr. Ojha and Daniel Zeman. 2020. Universal Dependency Treebanks for Low-Resource Indian Languages: The Case of Bhojpuri. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 33–38, Marseille, France. European Language Resources Association (ELRA).

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Doctor Raimond. 2004. A grammar of gujarati. *München: Lincom Europa. Search in*.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Vinit Ravishankar. 2017. A Universal Dependencies treebank for Marathi. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 190–200, Prague, Czech Republic.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan. The COLING 2016 Organizing Committee.

Ida Szubert, Omri Abend, Nathan Schneider, Samuel Gibbon, Sharon Goldwater, and Mark Steedman. 2021. Cross-linguistically Consistent Semantic and Syntactic Annotation of Child-directed Speech. *arXiv preprint arXiv:2109.10952*.

Tisdall, WS. 1892. *A Simplified Grammar of the Gujarati Language*. Sagwan Press.

Francis Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An assessment of Universal Dependency annotation guidelines for Turkic languages. In *5th International Conference on Turkic Language Processing (TURKLANG 2017)*, pages 356–377.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.