

Towards a Universal Dependencies Treebank for Gujarati

Maitrey Mehta^{1*} Mayank Jobanputra^{2*} Çağrı Çöltekin²

¹School of Computing, University of Utah

²Department of Linguistics, University of Tübingen

Relevant UniDive working groups: WG1, WG4

Abstract

The Universal Dependencies (UD) project has presented itself as a valuable platform to develop various resources for the languages of the world. In this proposal, we present our ongoing work around extending it to the Indo-Aryan language of Gujarati. We discuss the resources utilized and some cases of interest.

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 base of languages under the UD umbrella ensures that similar language patterns can be dealt with consistently and new language-specific features are brought to the foray for discussion when working with a new language. As a result, UD becomes one of the most fundamental 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). However, the Gujarati NLP community is still in its infancy. Basic resources like part-of-speech taggers, named entity recognizers, etc. 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 can be of value while constructing a Gujarati treebank.

The benefits of building such a treebank are three-fold: *a*) This presents as a valuable resource

for the development of NLP in a low-resource language, i.e., Gujarati. *b*) It ensures annotation paradigms in similar contexts are adhered to and helps point out any discrepancies in existing treebanks that need to be resolved. *c*) We can point out any new phenomena that might be Gujarati-specific or missed by earlier works. The above-mentioned reasons motivate us to propose a dependency treebank for Gujarati: *GujTB*. In the subsequent sections, we explain the selected corpora, and highlight some interesting discussion points we encountered during the course of our pilot annotation efforts.

2 Corpora

For our work, we investigated available corpora that include Gujarati text such as IndicCorp (Kakwani et al., 2020) and Samanantar (Ramesh et al., 2022). We observed that these datasets majorly contain news and other formal texts. For our pilot study, where we are doubly annotating a total of 300 sentences, we have taken sentences from Samanantar (news), UD Cairo (short),¹ and Gujarati translations of the French novella – *Le Petit Prince* (fiction) (The Little Prince, de Saint-Exupéry, 1943) for diversity purposes.

Genre	sentences	tokens
news	240	2666
short	20	173
fiction	50	658

Table 1: Pilot data statistics in the Gujarati UD corpus.

3 Discussion

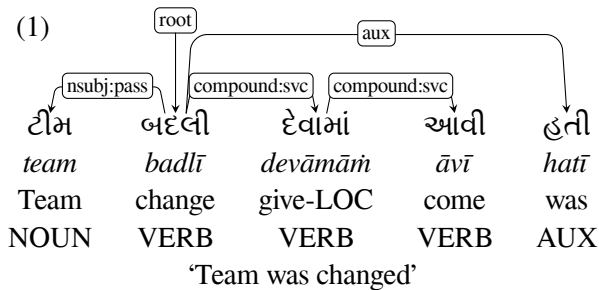
In this section, we discuss some interesting cases we have encountered in our pilot study.

Compound Verb Constructions with Locative Markers. A very common serial verb construction in Gujarati involves the usage of prototypical locative marker as shown in (1). Here,

¹<https://github.com/UniversalDependencies/cairo>

* Both authors contributed equally.

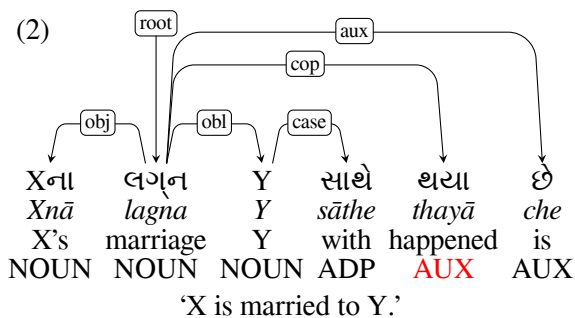
દેવું (*devuṃ*) and આવી are both semantically bleached with a locative marker attached to the former. The word આવી is an aspectual light verb. Arora (2022) argue to consider such tokens as a verb, and not an auxiliary. This serial verb construction is new. We propose (1) to handle such serial verb constructions. The annotations get tricky when બદલી + દેવામાં is replaced, while preserving meaning, by a ‘portmanteau’ બદલવામાં (*badalvāmām*). In such a case, a compound:svc relation exists from બદલવામાં to આવી. Does this mean that, in (1), the dependency relation should exist from બદલી to આવી instead? Moreover, the functionality of the locative marker and the morphological features it merits remains unclear.



Aspectual Verb vs. Independent Verb Another challenging annotation task involves deciding the role of a multifaceted verb *thayā* ‘happened’. It can occur in the text as an independent verb as well as a light verb.

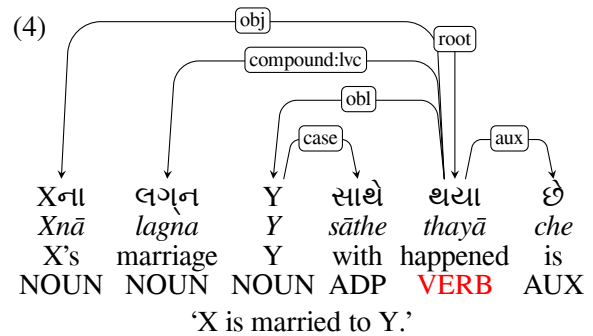
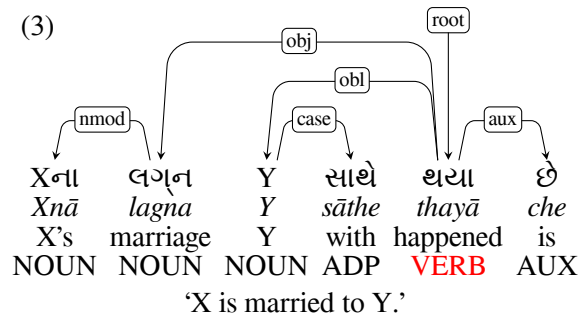
We take an example from our development set and discuss the annotation possibilities. We want to investigate two major questions as follows:

- Do we consider *thayā* as a VERB or an AUX (2)?
- If we consider *thayā* to be a VERB, should we annotate it as an independent verb (3) or as an aspectual light verb (4)?



In example 2, we consider *thayā* as an AUX. In this case, the sentence becomes a copula construction,

where the noun *marriage* becomes the root, and the dependency relation from the root to *thayā* should be *cop*.



On the contrary, in examples 3 and 4, we consider *thayā* as a VERB. In these cases, *thayā* becomes the root, but the choice of dependency relation relies on the type of the verb. If we consider *thayā* an independent verb then it becomes root and obj relation between *thayā* and *lagna*.

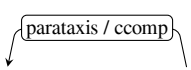
Another possibility is to consider *thayā* as a light aspectual verb compounding with *lagna* as shown in (4). We are leaning towards considering this option as here the meaning of the *thayā* being bleached by the noun *lagna* which provides a strong indication for the use of *compound:lvc* dependency relation.

Genuine disagreements. During the annotation process, we came across an example where both annotators disagreed on the root selection. This disagreement arises due to focus ambiguity in the sentence (Refer to (6)). Both annotators agree that both variations are possible and to remove this ambiguity more context is required. While earlier work from Da Costa et al. (2022) suggests mitigating all the disagreements before adjudication, we agree with Plank (2022) that such disagreements should be preserved, and we keep both annotations in adjudicated treebank. We also plan to release individual treebanks from all the annotators along with the adjudicated treebank.

Splitting Genitive Markers. Certain nominals (and, in some instances, verbs) in Gujarati are inflicted 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 where certain Gujarati postpositions often pair with preceding genitive markers to form layer III postpositions (Masica, 1993).

The Case for Determiners. According to Gujarati grammars (Tisdall, 1892; Doctor, 2004), demonstrative pronouns like એ, તે, તેણે 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 ફૂતલે ‘that dog’). Gujarati grammars do not discuss determiners as such. However, we see this usage closer to the UD definition of determiners and hence use the same.

Quoter and Quotation. We encounter a screenplay dialog-style quotation that has not been resolved yet. Example 5 shows such a case. Early UD literature suggests solutions for cases with speech verbs. Recent guidelines recommend ccomp over parataxis for reported speech.² We believe this to be a much more pervasive (and not a Gujarati-specific) issue; applicable, perhaps, when UD is extended to plays.³

(5) 
 I play football : Mark
 ‘I play football : Mark’

4 Conclusions

In this work, we present our on-going annotation effort on developing Gujarati UD. We describe the corpus we use, and a few interesting cases we have encountered so far in our pilot study. We plan to extend this work to cover a larger corpus, and bring forth any novel phenomena and highlight existing inconsistencies going forward.

²<https://universaldependencies.org/changes.html#reported-speech>

³This is an ongoing discussion: <https://github.com/UniversalDependencies/docs/issues/904>

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, et al. 2017. The Hindi/Urdu treebank project. In *Handbook of Linguistic Annotation*. Springer Press.
- Luis Morgado Da Costa, Francis Bond, and Roger VP Winder. 2022. The Tembusu treebank: An English learner treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4817–4826.
- Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.
- Antoine de Saint-Exupéry. 1943. *Le petit prince [The little prince]*. Reynal & Hitchcock (US), Gallimard (FR).
- Raimond Doctor. 2004. *A Grammar of Gujarati*, volume 28. Lincom Europa.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World. Twenty-fifth edition*. SIL International, Dallas, Texas.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- Colin P Masica. 1993. *The indo-aryan languages*. Cambridge University Press.
- 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).
- Barbara Plank. 2022. The ‘problem’ of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

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.

WS Tisdall. 1892. *A Simplified Grammar of the Gujarati Language*, volume 22. Sagwan Press.

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 Drohanova, 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.

A Disagreement Example

