*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 7088–7095
Marseille, 20-25 June 2022


# GujMORPH - A Dataset for Creating Gujarati Morphological Analyzer

**Jatayu Baxi, Brijesh Bhatt**
Dharmsinh Desai University
Gujarat(India)
{jatayubaxi.ce, brij.ce}@ddu.ac.in

**Abstract**

Computational morphology deals with the processing of a language at the word level. A morphological analyzer is a key linguistic word level tool that returns all the constituent morphemes and their grammatical categories associated with a particular word form. For the highly inflectional and low resource languages, the creation of computational morphology-related tools is a challenging task due to unavailability of underlying key resources. In this paper, we discuss the creation of an annotated morphological dataset- GujMORPH for the Gujarati - an indo aryan language. For the creation of this dataset, we studied language grammar, word formation rules, suffix attachments in depth. This dataset contains 16527 unique inflected words along with their morphological segmentation and grammatical feature tagging information. It is a first-of-its-kind dataset for the Gujarati language and can be used to develop morphological analyzer and generator models. The dataset is annotated in the standard Unimorph schema and evaluated on the baseline system. We also describe the tool used to annotate the data in the standard format. The dataset is released publicly along with the library. Using this library, the data can be obtained in a format that can be directly used to train any machine learning model.

**Keywords:** Language resources, Morphology,Gujarati

## 1. Introduction

Morphological analysis is a key word level natural language processing task, used to analyze an inflected word. The morph analyzer outputs constituent morphemes and grammatical feature sets associated with an inflected word form. It is necessary to have a powerful morphological analyzer system for the downstream NLP tasks such as question answering, machine translation, sentiment analysis, etc. The morph analyzer usually requires knowledge in form of word formation rules, language grammar understanding, and extensive feature engineering in order to achieve decent accuracy (Goyal and Lehal, 2008; Sharma et al., 2021).

As the demand for web-based technologies is increasing day by day, language resources have a crucial role to play to prevent the extinction of the natural language. Nearly all European and East Asian languages are resource-rich languages but most of the south and southeast asian Languages including Gujarati can be still considered as low resourced.(Ali et al., 2020) Gujarati is an Indo-Aryan language, spoken mainly in the Gujarat state of India. It is the 26th most widely spoken language with approximately 55 million speakers across the world.(Magueresse et al., 2020; Suba et al., 2011)

In the literature, there are mainly two methods for building a morphological analyzer. The first is the classic rule-based method and the second is a machine learning-based methods(PJ and KP, 2012). For the rule-based methods, a substantial amount of knowledge is required in form of word formation rules, suffixes, and other language-specific characteristics. When the language is highly inflectional such as Gujarati, the task becomes more challenging due to a large number of possible suffixes and language-specific ambiguities in the word-formation rules. For the machine learning methods, most of the approaches are focused on the supervised learning methods. Supervised methods introduce two new problems: manual feature engineering and requirements of the training dataset. With the progress in the deep learning-based methods in the last few years, researchers have worked on deep neural network-based models for the morph analysis task. These models do not require any manual feature engineering but they require an enormous amount of training data for efficient results. This fact has introduced the need of creating a dataset for building a morphological analyzer. It is also desirable that the dataset should be annotated in some standard format so that it is widely accepted and remains inconsistent with the dataset of other languages. The typical format of such a dataset is each word should be mapped with its corresponding root word along with the possible set of grammatical features.(Kirov et al., 2018)

The existing work in the field of the morphological analyzer for different languages is largely done using the datasets available in either Universal Dependencies treebank(Nivre et al., 2016; Nivre et al., 2020) or Unimorph schema(Kirov et al., 2018). The Gujarati language is still not part of these

datasets. For the Gujarati language, not much efforts are made to build morphological analyzer. The reason for this is complex morphological structure of the language and the lack of a standard dataset for the validation of the system. We have studied Gujarati morphology in detail and created a unique dataset that can be used to create an efficient morph analyzer for the Gujarati language. We have also annotated our dataset in the standard Unimorph format to maintain consistency. The dataset is evaluated on the baseline system and overall decent results are obtained.
Our key contributions are as follows :

- We provide a detailed survey of Morphology for Gujarati Language
- We have created Gujarati Morphology dataset which provides details of morpheme segmentation and grammatical features.
- The dataset is annotated in standard Unimorph format.

The remaining the paper is organized as follows. Section 2 describes the related work in the field of computational morphology. Section 3 describes the details of Gujarati morphology. Section 4 describes the dataset creation process. Section 5 describes the baseline system followed by the conclusion and future work.

## 2. Related Work

A significant development in the field of morphological analysis has been done for English and other European languages. For Indian languages also, research has been done in this field with the standard rule-based, statistical, machine learning-based and hybrid approaches. In this section, we present the related work initiated on the corpus development for morphological analysis along with the survey of some standard approaches for developing the morphological analyzer.
The analysis of words including morpheme structures and grammatical characteristics is a very important task for understanding the language and also for achieving effective results for the NLP tasks such as machine translation, semantic analysis, parsing etc.(Tkachenko and Sirts, 2018)
The classical ways to develop the morphological analyzer are rule-based approach and machine learning-based approach. In past few years, deep neural network based models are also extensively used to create morphological analyzer.
The initial efforts in the field of computational morphology can be traced back to 80s. (Koskenniemi, 1984) developed universal model for the analysis of morphologically rich language known as two level morphology. The enhancements of this model was done by(Beesley and Karttunen, 1992). Another popular methodology is finite state morphology. In this approach, FSA are used for the analysis and FST is used for the morphological generator.(Kenneth R. Beesley and Lauri Karttunen, 2003)(Beesley, 1998)(Beesley, 2003)(Megerdoomian, 2004). For many Indian languages, paradigm based approach has been extensively used for the morph analysis problem. In this approach, paradigms are defined based on the similarities observed in the word formation rules(Bharati et al., 2002). Other notable works in the same area for Indian languages are (Melinamath and Mallikarjunmath, 2011) (Sahoo, 2003)(Kumar et al., 2012).

Other than rule-based techniques, researchers have applied supervised and unsupervised machine learning-based methods for developing the morph analyzer. (Chakrabarty et al., 2016) developed Lemmatizer for Bengali language using neural architecture and observed its effectiveness for the word sense disambiguation task. (Heigold et al., 2016) explored various CNN and RNN-based neural networks for the morphological tagging task for the languages having rich morphology. The same work is extended by (Heigold et al., 2017) and results are observed for 14 different languages. (Chakrabarty et al., 2017) developed language independent and context sensitive Lemmatizer and evaluated it on two indic and two non indic languages. (Premjith et al., 2018) proposed RNN-based morpheme segmentation model for malayalam language. (Tkachenko and Sirts, 2018) proposed simple multiclass, multilabel multiclass and hierarchical models for the morphological tagging and evaluated it on 49 different languages. The encoder used for this work is same as the one used in (Lample et al., 2016). (Gupta et al., 2020) evaluated the performance of 4 different sequence labelling methods on Sanskrit, a morphologically rich, fusional Indian language.

The deep neural network and other supervised approaches require a dataset for training the system. The majorityy ofabove-citedcited works use the dataset available in universal dependency treebank(Nivre et al., 2016) or unimorph schema. The unimorph (Kirov et al., 2018) project has two modules, a language-independent schema for the annotation and language-specific datasets in which each inflected form is associated with a lemma, which typically carries its underlying lexical meaning, and a bundle of morphological features. Currently, the dataset of 142 languages is included in the Unimoprh schema. The other popular data source is the universal dependencies treebank. It is a framework for consistent annotation of grammatical features such as POS, Morph, syntactic dependencies etc. The annotation consists in a linguistically motivated word segmentation; a morpho-

logical layer comprising lemmas, universal part-of-speech tags, and standardized morphological features; and a syntactic layer focusing on syntactic relations between predicates, a,rguments and modifiers. Currently, around 200 treebanks in over 100 languages including Hindi are available in universal dependencies. (Bhat et al., 2016).

For manyIndo-Aryann languages such as Hindi, the datasets in treebank and Unimorph is available and can be used for the experiments. However, for the low resource languages such as Gujarati, the dataset is not available and hence it creates a hurdle in further NLP development. In this work, we develop a morph dataset for the Gujarati language and also annotate it in standard Unimorph format. Overall, the dataset creation process is carried out in 3 steps: Corpus acquisition, creation of the dataset by feature identification and annotation of the dataset.

## 3. Gujarati Morphology

Gujarati is an Indo-Aryan language mainly spoken in the Gujarat state of India. It is a part of the greater Indo-European family. The Gujarati language is more than 1000 years. It is a modern Indo-Aryan language evolved from the Sanskrit language. As per the Central Intelligence Agency, 4.5% of the Indian populace (1.21 billion as per the 2011 registration) speaks Gujarati, which adds up to 54.6 million speakers in India. There are around 65.5 million speakers of the Gujarati around the world, making it the 26th-most-spoken local language on the planet.(Chauhan and Shah, 2021)

Outside the state of Gujarat, Gujarati is spoken in many other parts of south Asia such as Mumbai, Pakistan by the migrants. Outside Asia also Gujarati is one of the widely Indian spoken languages by the Gujarati diaspora in the Unites States and Canada. In UK's capital London, Gujarati is the fourth most commonly spoken language. Gujarati is also spoken in Southeast Africa, particularly in Kenya, Uganda, Tanzania, Zambia, and South Africa.

Gujarati is a verb-final language and has a relatively free word order. It is rich in morphology and highly inflectional language. A language is said to have rich morphology when there are more inflectional forms of the base word. Suffixes are added in series to the root word to form a particular word form. Various morphophonemic changes occur when suffixes are attached to the root word. In this section we describe the morphology of the Gujarati language and some unique observation about the grammatical structure of the language.

### 3.1. Noun

In Gujarati, nouns participate in three genders and two numbers. The genders are masculine, feminine and neuter and numbers are singular and plural. Gujarati nouns also inflect for various cases. Table 1 shows gender and number markers for Gujarati and Table 2 shows various cases with corresponding case markers.

| Noun Feature | Marker |
|---|---|
| Gender Male | ઓ (Ō) |
| Gender Female | ી(Ī) |
| Gender Neutral | ુ(U) |
| Number Singular | $\phi$ |
| Number Plural | ા (Ā) |

Table 1: Number and Gender Markers

| Case | Suffix |
|---|---|
| Nominative | $\phi$ |
| Genitive | નો,ની,નું,નાં ( Nō,nī,nuṁ,nāṁ) |
| Ergative | એ ( ē) |
| Objective/Dative | ને (nē) |
| Ablative | થી (thī) |
| Locative | માં (māṁ) |

Table 2: Case Markers for Gujarati Noun

### 3.2. Verb

Gujarati verbs inflect for gender, number, tense, aspect and mood features. Table 3 and Table 4 shows example of Gujarati verb with different mood and aspects respectively.

### 3.3. Adjective

Gujarati adjectives can be classified in two types based on their nature of inflections. One class of adjectives do not inflect while the other class inflect for gender and number. Table 5 shows example of each category.

## 4. The Dataset

In this section, we discuss the method used to create the GujMORPH dataset. This dataset can be used to train and evaluate the morphological analyzer model. The morph analyzer has two components, one which segments a word into constituent morphemes and another to morphologically tag an inflected word. For both of the above tasks, a dataset is required which consists of an inflected word, segmented morphemes, and a set of morphological tags associated with a given inflected word. We have developed a tool for the annotation of the data and also publicly release the dataset with the libraries that can be directly plugged in with any machine learning model.

| Mood | Example | Transliteration | English Translation |
|---|---|---|---|
| Indicative | ધન્વી ચોકલેટ ખાય છે. | *Dhanvī cōkalēṭa khāya chē.* | *Dhanvi is eating a chocolate.* |
| Imperative | સવારે વહેલો ઊઠજે. | *Savārē vahēlō ūṭhajē.* | *Get up early in the morning..* |
| Conditional | જો હું ત્યાં હોત, તો હું તમને મદદ કરી શક્યો હોત. | *Jō huṁ tyāṁ hōta, tō huṁ tamanē madada karī śakyō hōta.* | *Had I were there, I would have helped..* |
| subjunctive | એ અત્યારે દીપક ને ધેર હોવાનો. | *Ē atyārē dīpaka nē dhēra hōvānō.* | *He must be at Dipak's home right now.* |

Table 3: Moods of Gujarati Verb

| Mood | Example | Transliteration | English equivilant |
|---|---|---|---|
| Simple | રામ અમદાવાદમાં રહે છે. | *Rāma amadāvādamāṁ rahē chē.* | *Ram lives in Ahmedabad.* |
| Progressive | રામ અત્યારે પુસ્તક વાંચી રહ્યો છે. | *Rāma atyārē pustaka vāṁcī rahyō chē.* | *Ram is reading a book right now..* |
| Perfect | રામે પુસ્તક વાંચી લીધું. | *Rāmē pustaka vāṁcī līdhuṁ.* | *Ram has finished reading a book.* |
| Perfect Progressive | રામ સવારથી પૂજા કરી રહ્યો હતો. | *Rāma savārathī pūjā karī rahyō hatō.* | *Ram was doing pooja since morning.* |

Table 4: Gujarati Verb Aspects

| Type of Adjective | Example |
|---|---|
| Non-Inflected | ઉત્તમ (Uttama) |
| Inflected | સારો, સારી, સારું ,સારા (sārō,sārī,sāruṁ,sārā) |

Table 5: Gujarati adjective inflection

### 4.1. Corpus acquisition

For the creation of dataset, we did a survey of available corpus for the gujarati language. Table 6 shows various available corpus. The source for first 3 datasets is TDIL (Technology Development for Indian Language Programme, Govt. of India) and for the last corpus, the source is ELRA(European Language Resources Association) For the morphological analysis task, it is preferable to have the POS tagged data, hence we have selected Gujarati Monolingual Text Corpus ILCI-II corpus for the creation of the dataset. Under the Indian Languages, Corpora Initiative phase –II (ILCI Phase-II) project, initiated by the MeitY, Govt. of India, Jawaharlal Nehru University, New Delhi had collected this monolingual corpus in Gujarati. There are around 30k sentences in this corpus. The corpus contains sentences from various domains such as art and culture, entertainment, science, philosophy, and religion. A sample corpus of entertainment category is shown in the figure below. [2]

[2]TDIL :http://www.tdil-dc.in

| **Dataset Name** | **Description** |
|---|---|
| Gujarati Monolingual Text Corpus ILCI-II (Source : TDIL) | POS Tagged Data containing 30,000 Sentences |
| Gujarati News Corpus SRIMCA (Source : TDIL) | 517 News Articles from various Gujarati News Papers. |
| Gujarati Named Entity and Multi Word Expression List- CLIA (Source : TDIL) | List of Gujarati Named Entities(10,251) and Multi Word Expressions (2965) |
| The EMILLE-CIL Corpus (Source : ELRA) | Monolingual and Parallel Corpus containing 5,64,000 Words. |

Table 6: List of available corpus for gujarati

### 4.2. Dataset for Morpheme Segmentation

The word segmentation module segments an inflected word into its constituent morphemes. Because of rich morphology, it is often observed that more than one suffix is attached to a root form. For developing a morpheme segmentation module, it is required to have a dataset in which inflected words are mapped with their corresponding root

```
1 ID  Value
2 GJED3001    ગજવામાં\N_NN હાથ\N_NN નાખ્યું\V_VM
  છું\V_VAUX ટેરવે\N_NN કંઈ\DM_DMI અડકે\V_VM
  છે\V_VAUX .\RD_PUNC
3 GJED3002    ક્યારે\N_NST મેં\PR_PRP
  સેરવી\V_VAUX_VNP લીધું\V_VAUX ?\RD_PUNC
4 GJED3003    મને\PR_PRP ધ્રાસકો\N_NN પડે\V_VM
  છે\V_VAUX ત્યારે\N_NST હું\PR_PRP
  બબડતો\V_VAUX_VNP હોઉં\V_VAUX છું\V_VAUX
  :\RD_PUNC તમે\PR_PRP સાચવો\V_VM .\RD_PUNC
5 GJED3004    ખોવકીની\N_NN ઊધ\N_NN હું\PR_PRP
```

Figure 1: ILCI Gujarati Corpus - Source : TDIL[1]

word. We create such a dataset for the Gujarati language. Our dataset contains 20292 unique inflected words along with their root morpheme information.

For the creation of this dataset, we first create separate wordlist files for each POS category. We then identify root words for noun, verb, and adjective categories. We also represent the data in such a way that it can be directly used to train any machine learning model. We represent the inflected word as a binary string and mark "1" in the position of the split character, the rest of the characters are marked as "0". Figure 2 shows morpheme splitting example.

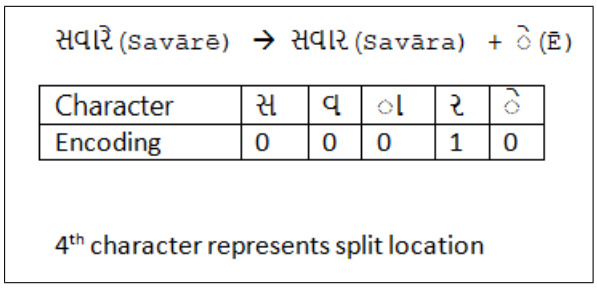



Figure 2: Morpheme Splitting Example

### 4.3. Dataset for Grammatical Feature Tagging

In this section, we describe the data related to the morphological feature tagging. Since each inflected word in the dataset has some morphological features associated with it, all words need to be annotated with corresponding morphological tags. Building such a dataset is more challenging as we need to first identify various morphological features for a particular part of the speech category, understand various inflections and then start tagging each word in the dataset. Our dataset for grammatical feature prediction contains 16527 unique words along with their feature tagging information as described in Table 7. Each feature for the inflected word has a number of labels associated with it. Table 8 shows labels associated with each feature. Figure 4 shows the distribution of various labels in the GujMORPH dataset.

| POS Category | Features | Number of Words |
|---|---|---|
| Noun | Gender, Number, Case | 6847 |
| Verb | Gender, Number, Tense, Aspect, Person | 6334 |
| Adjective | Gender, Number | 3346 |

Table 7: Details about Dataset

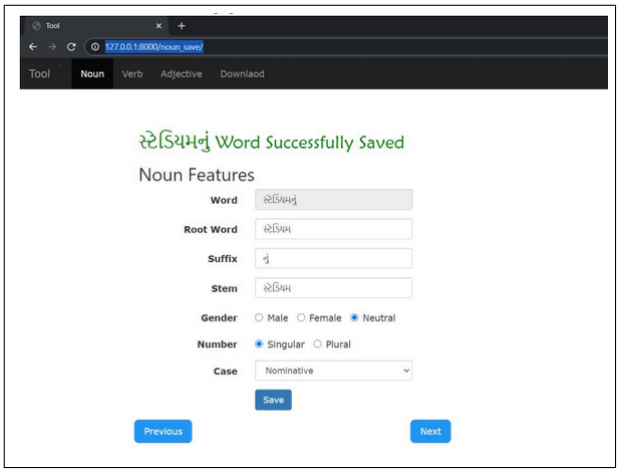
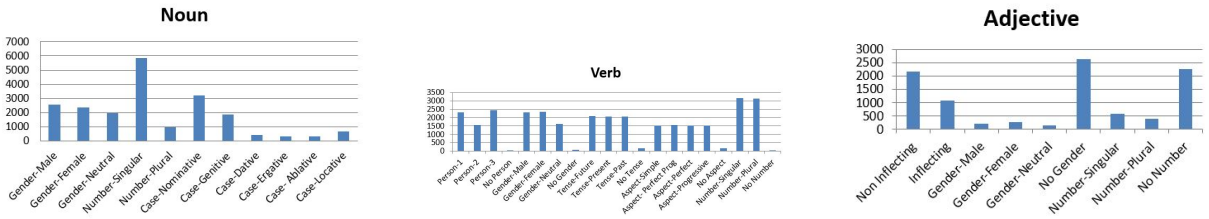



Figure 3: Graphical user interface of the web-based annotation platform used for the annotation of GuijMORPH dataset.

For the creation of the dataset, we have developed a tool through which annotation can be easily done. We upload wordlist files as an input and the tool displays one word at a time along with the annotation options including root word informa-

| Feature | Labels |
|---|---|
| Gender | Male, Female, Neutral, No Gender |
| Number | Singular, Plural, No Number |
| Case | Nominative, Dative, Ergative, Genitive, Ablative, Locative |
| Tense | Future, Present, Past, No tense |
| Aspect | Simple, Perfect, Progressive, Perfect Progressive, No Aspect |
| Person | 1st, 2nd, 3rd, No person |
| Type(Adjective) | Inflective, Non-Inflective |

Table 8: Features with corresponding Labels




(a) Noun (b) Verb (c) Adjective

Figure 4: The label distribution in GujMORPH Dataset

| Unimorph Style Features | Verb Form | Root_Form |
|---|---|---|
| V;1;M;PST;LGSPEC1;SG | રમ્યો | રમવું |
| V;1;F;PST;LGSPEC1;SG | રમી | રમવું |
| V;1;N;PST;LGSPEC1;SG | રમ્યું | રમવું |
| V;2;M;PST;LGSPEC1;SG | રમ્યો | રમવું |
| V;2;F;PST;LGSPEC1;SG | રમી | રમવું |
| V;3;M;PST;LGSPEC1;SG | રમ્યો | રમવું |
| V;3;F;PST;LGSPEC1;SG | રમી | રમવું |
| V;3;N;PST;LGSPEC1;SG | રમ્યું | રમવું |
| V;1;M;PRS;LGSPEC1;SG | રમું છું | રમવું |
| V;1;F;PRS;LGSPEC1;SG | રમું છું | રમવું |
| V;1;N;PRS;LGSPEC1;SG | રમું છું | રમવું |
| V;2;M;PRS;LGSPEC1;SG | રમે છે | રમવું |
| V;2;F;PRS;LGSPEC1;SG | રમે છે | રમવું |

Figure 5: Annotation in Unimorph format

tion and a grammatical feature list. The annotations are saved in standard Unimoprh format. The dataset along with the Python library and detailed documentation is publicly available at `https://github.com/jhbaxi/gujmorphdataset`. The python library available along with the dataset has following features:

- Load and describe the dataset.
- Character based tokenization.
- Performing binary encoding for the morpheme segmentation task.
- Converting the data into the training and testing list for the grammatical feature prediction task.

## 5. Baseline System

The baseline system for the Gujarati morphological analyzer using this proposed dataset was implemented using Bi-LSTM-based model in our previous work.(Baxi and Bhatt, 2021) In this work, the dataset is not publicly released but it is used to train and evaluate the system. For the morpheme boundary detection task, the system gives 89.05% accuracy and for the grammatical feature prediction task, the F1 scores are 0.68, 0.12 and 0.68 for the noun, verb and adjective POS categories.

## 6. Conclusion and future work

The activities like obtaining corpus, prepossessing, and annotation of low resource language data like Gujarati is a challenging task. We use the POS tagged corpus and create a unique dataset for the evaluation of Gujarati morph analyzer because it is observed that morphological features are POS category-specific. The identification of grammatical feature set and understanding minute details about the morphology of the language is a labor-intensive task. The annotation is done in standard Unimorph format and the dataset will be added to the Unimorph 4.0 release. We manually validate the GujMORPH dataset for the verification of proper annotation which is an expensive but essential activity to maintain the gold standard of the dataset. The proposed dataset is publicly released and also tested on the baseline system. In the future, we aim to expand the dataset by adding more examples and adding the remaining part of the speech categories.

## 7. Bibliographical References

Ali, W., Lu, J., and Xu, Z. (2020). SiNER: A large dataset for Sindhi named entity recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2953–2961, Marseille, France, May. European Language Resources Association.

Baxi, J. and Bhatt, B. (2021). Morpheme boundary detection grammatical feature prediction for gujarati : Dataset model. In *Proceedings of the 18th International Conference on Natural Language Processing*, NIT, Silchar, December.

Beesley, K. R. and Karttunen, L. (1992). Two-Level Rule Compiler. *Technical Report. Xerox Palo Alto Research Center. Palo Alto, California.*

Beesley, K. R. (1998). Arabic morphology using only finite-state operations. *Proceedings of the Workshop on Computational Approaches to Semitic languages. Association for Computational Linguistics*, page 50.

Beesley, K. (2003). Finite-State Morphological Analysis and Generation for Aymara: Project Report. *Proceedings of the Workshop of Finite-State Methods in Natural Language Processing, 10th Conference of the European Chapter of the*

*Association for Computational Linguistics.*, 5:2–5.

Bharati, A., Chaitanya, V., Sangal, R., and Gillon, B. (2002). *Natural Language Processing: A Paninian Perspective*. prentice hall of india.

Bhat, R. A., Bhatt, R., Farudi, A., Klassen, P., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., Vaidya, A., Vishnu, S. R., et al. (2016). The hindi/urdu treebank project. In *Handbook of Linguistic Annotation*. Springer Press.

Chakrabarty, A., Chaturvedi, A., and Garain, U. (2016). A neural lemmatizer for Bengali. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2558–2561, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Chakrabarty, A., Pandit, O. A., and Garain, U. (2017). Context sensitive lemmatization using two successive bidirectional gated recurrent networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1481–1491, Vancouver, Canada, July. Association for Computational Linguistics.

Chauhan, U. and Shah, A. (2021). Improving semantic coherence of gujarati text topic model using inflectional forms reduction and single-letter words removal. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(1), mar.

Goyal, V. and Lehal, G. S. (2008). Hindi morphological analyzer and generator. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 1156–1159. IEEE.

Gupta, A., Krishna, A., Goyal, P., and Hellwig, O. (2020). Evaluating neural morphological taggers for Sanskrit. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 198–203, Online, July. Association for Computational Linguistics.

Heigold, G., Neumann, G., and van Genabith, J. (2016). Neural Morphological Tagging from Characters for Morphologically Rich Languages. *arXiv e-prints*, page arXiv:1606.06640, June.

Heigold, G., Neumann, G., and van Genabith, J. (2017). An extensive empirical evaluation of character-based morphological tagging for 14 languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 505–513, Valencia, Spain, April. Association for Computational Linguistics.

Kenneth R. Beesley and Lauri Karttunen. (2003). Finite-State Morphology. (January).

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S. J., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2018). UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. (July):178–181.

Kumar, D., Singh, M., and Shukla, S. (2012). FST Based Morphological Analyzer for Hindi Language. *arXiv preprint arXiv:1207.5409*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for Computational Linguistics.

Magueresse, A., Carles, V., and Heetderks, E. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. *arXiv e-prints*, page arXiv:2006.07264, June.

Megerdoomian, K. (2004). Finite-state morphological analysis of Persian. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. Association for Computational Linguistics*, page 35.

Melinamath, B. C. and Mallikarjunmath, A. G. (2011). A morphological generator for Kannada based on finite state transducers. *ICECT 2011 - 2011 3rd International Conference on Electronics Computer Technology*, 1:312–316.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (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.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

PJ, A. and KP, S. (2012). Computational Morphology and Natural Language Parsing for Indian Languages : A Literature Survey. *International Journal of Scientific & Engineering Research*, 3(3):136–146.

Premjith, B., Soman, K. P., and Kumar, M. A. (2018). A deep learning approach for Malayalam

morphological analysis at character level. *Procedia Computer Science*, 132:47–54.

Sahoo, K. (2003). Oriya nominal forms: A finite state processing. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 3:730–734.

Sharma, D., Sahai, S., Chaudhari, N., and Bruguier, A. (2021). Improved pronunciation prediction accuracy using morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 222–228.

Suba, K., Jiandani, D., and Bhattacharyya, P. (2011). Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati. *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP)*, pages 1–8.

Tkachenko, A. and Sirts, K. (2018). Modeling composite labels for neural morphological tagging. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 368–379, Brussels, Belgium, October. Association for Computational Linguistics.