# A Deep Learning Architecture for Corpus Creation for Telugu Language

Dhana L. Rao, Venkatesh R. Pala, Nic Herndon, and Venkat N. Gudivada(✉)

Cognitive Computing Lab, East Carolina University, Greenville, NC 27855, USA
gudivadav15@ecu.edu

**Abstract.** Many natural languages are on the decline due to the dominance of English as the language of the World Wide Web (WWW), globalized economy, socioeconomic, and political factors. Computational Linguistics offers unprecedented opportunities for preserving and promoting natural languages. However, availability of corpora is essential for leveraging the Computational Linguistics techniques. Only a handful of languages have corpora of diverse genre while most languages are *resource-poor* from the perspective of the availability of machine-readable corpora. Telugu is one such language, which is the official language of two southern states in India. In this paper, we provide an overview of techniques for assessing language vitality/endangerment, describe existing resources for developing corpora for the Telugu language, discuss our approach to developing corpora, and present preliminary results.

**Keywords:** Telugu · Corpus creation · Optical character recognition · Deep learning · Convolutional neural networks · Recurrent neural networks · Long short-term memory · Computational linguistics · Natural language processing



## 1 Introduction

As of 2019, there are 7,111 spoken languages in the world [1]. This number is dynamic as linguists discover new spoken languages which were hitherto undocumented. Moreover, the languages themselves are in a flux given the rapid advances in computing and communications, and social dynamic and mobility. About 40% of the world languages are *endangered* and these languages have less than 1,000 speakers [1]. A language gets endangered when its speakers begin to speak/teach another (dominant) language to their children. Sadly, about 96% of the languages of the world are spoken by only 3% of the world population. About 95% of the languages might be extinct or endangered by the end of the century. Especially, the indigenous languages are disappearing at an alarming rate. To bring awareness to this issue, the United Nations has declared 2019 as The International Year of Indigenous Languages.


B. Iyer et al. (eds.), *Applied Computer Vision and Image Processing*,
Advances in Intelligent Systems and Computing 1155,
https://doi.org/10.1007/978-981-15-4029-5_1

We will begin by defining some terminology. A person's *first language* is referred to as her *native language* or *mother tongue*, and this is the language the person is exposed to from birth through the first few years of life. These *critical years* encompass an extremely important window to acquire the language in a linguistically rich environment. It is believed that further language acquisition beyond the critical years becomes much more difficult and effortful [2,3]. The first language of a child enables reflection and learning of successful social patterns of acting and speaking. The first language is also an integral part of a child's personal, social, and cultural identity.

Each language has a set of *speech sounds*. Phonetics of the language is concerned with the description and classification of speech sounds including how these sounds are produced, transmitted, and received. A *phoneme* is the smallest unit in the sound system of a language. Ladefoged [4] describes all the known ways in which the sounds of the world's languages differ. Native speakers of a language learn the sounds of the language correctly. Therefore, when they speak the language, they speak natively—produce correctly accented sounds. In contrast, non-native speakers of a language speak it *non-natively*, which is evidenced by incorrectly accented sounds. We use the terms first language, $L1$, mother tongue, and arterial language synonymously.

A person is *bilingual* if she is equally proficient in two languages. However, correctly accented fluency is considered a requirement, which is difficult given that each language is associated with a set of characteristic sounds. The vocal chord development must be attuned to the distinct sounds of both the languages. Given the criticality of the mother tongue on a child's social and intellectual development, on November 17, 1999 the UNESCO designated 21 February as the *International Mother Language Day*.



There are many languages in the world which are widely spoken, but are on an accelerated path toward insignificance and eventual extinction [5]. This trend is manifesting through multiple indicators. First, the number of speakers who speak the language *natively* is declining. Second, even these native speakers lack fluency in the choice of correct words as their vocabulary is rather limited. Furthermore, they begin a sentence in their mother tongue and interject words of another language to compensate for their limited mother tongue vocabulary. Third, the dominance of English as the language of WWW is also a contributing factor. Though this is affecting most languages, there are exceptions. Languages such as Japanese, Spanish, Arabic, Turkish, German, Italian, and French have begun to flourish on the WWW. Speakers of these languages take pride in learning their mother tongue and government policies also nurture and promote these languages.

The fourth factor for the decline of languages is related to the global economy and associated mobility and migration. This is more pronounced in developing countries such as India. Fifth, the British colonialism actively promoted the use of English at the cost of native languages. Lastly, in developing countries such as India, not speaking one's mother tongue is viewed as elite social status. This is in sharp contrast with most European countries where people take pride in their mother tongue and native languages are used to study even disciplines such as medicine and engineering.

We believe that the recent and rapid advances in computing and communication technologies provide unprecedented opportunities for reversing the shift in the language use. More specifically, *computational linguistics* offers software tools and approaches to preserve and promote natural languages. However, these approaches require machine-readable corpora of diverse genre to reflect the entire language use. In this paper, we discuss the development of a machine learning-based approach for developing corpora for Telugu language. Telugu is the official language of two states in South India. There are many other spoken languages in these two states, which include Kolami, Koya, Gondi, Kuvi, Kui, Yerukala, Savara, Parji, and Kupia. This study does not address these dying languages, though the approach we develop can be used for all languages.

The remainder of the paper is organized as follows. We discuss tools and techniques for assessing language vitality and endangerment in Sect. 2. Section 3 summarizes the salient characteristics of the Telugu language. A brief overview of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks is presented in Sect. 4. Existing resources and approaches for generating corpora for the Telugu language are presented in Sect. 5. Our approach to corpus generation and preliminary results are discussed in Sect. 6 and Sect. 7 concludes the paper.



## 2 Assessing Language Vitality and Endangerment

The first step in preserving and promoting languages is to assess their current state using a nomenclature or metric. The Fishman's Graded Intergenerational Disruption Scale (GIDS) was the first effort in this direction [6]. There are eight stages on the GIDS scale—stage 1 through stage 8, and stage 8 refers to the highest level of language disruption and endangerment. The **stage 8** indicates a state where very few native speakers of the language exist, the speakers are geographically dispersed and socially isolated, and the language needs to be reassembled from their mouths and memories. The **stage 7** reflects the language situation that most users of the language are socially integrated and ethnolinguistically active population, but passed the child-bearing age. The **stage 6** refers to language situation where intergenerational informal oralcy exists, there is demographic concentration of native speakers with institutional reinforcement.

The **stage 5** reflects a situation where the language literacy exists at home, school, and community, but there is no extra-communal reinforcement of such literacy. The **stage 4** refers to the state where the language is used in primary and secondary education, and this is mandated by education laws. The state where the language is being used in *lower work sphere* beyond the native language's geographic region is referred to as **stage 3**. The **stage 2** reflects the state that the language is used in mass media, and lower governmental services. Lastly, the **stage 1** refers to the use of the language in the higher echelons of education, occupations, government, and media, but lacks political independence. The GIDS scale is quasi-implicational meaning that the higher scores imply all or nearly all of the lesser degrees of disruption as well.

The GIDS served as a seminal evaluative framework of language endangerment for over two decades. Recently, other evaluative frameworks have been

proposed. For example, UNESCO has developed a 6-level scale of endangerment [7]. Ethnologue uses yet another set of measures (five categories) to assess language vitality/endangerment [1]. Lewis et al. align the above three evaluative systems and proposed an evaluative scale of 13 levels [8], and is referred to as the E(xpanded) GIDS (EGIDS). Evaluating a language's vitality using the EGIDS involves answering five key questions about the language related to identity function, vehicularity, state of intergenerational language transmission, literacy acquisition status, and a societal profile of generational language use. To make this paper self-contained, the EGIDS scale is shown in Table 1.

**Table 1.** The EGIDS scale for assessing language vitality

| Level number | Level name | Level description |
|---|---|---|
| 0 | International | Used in international trade and policy, and knowledge exchange |
| 1 | National | Used in education, work, mass media, and government at the national level |
| 2 | Provincial | Used in education, work, mass media, and government within major administrative subdivisions of a country |
| 3 | Wider communication | Used in work and mass media without official language status across a region |
| 4 | Educational | Vigorously used, literature sustained, and an institutionally supported education system exists |
| 5 | Developing | Vigorously used, but the literature use is not widespread and sustainable |
| 6a | Vigorous | Used for spoken communication by all generations and the situation is sustainable |
| 6b | Threatened | Used for spoken communication by all generations, but is losing native speakers |
| 7 | Shifting | The child-bearing generation use the language among themselves, but not transmitted to children |
| 8a | Moribund | Only grandparent generation and older are the active users of the language |
| 8b | Nearly extinct | Only grandparent generation and older are the only active users of the language, and they have little opportunity to use the language |
| 9 | Dormant | No one has more than symbolic proficiency of the language |
| 10 | Extinct | No one uses the language |

Language Endangerment Index (LEI) is another metric for assessing the levels of language endangerment [9]. LEI is based on four factors: intergenerational transmission, absolute number of speakers, speaker number trends, and domains of use. Compared to other language endangerment assessments, LEI can be used even if limited information is available.

Dwyer [10] explores the uses and limits of various language vitality/endangerment tools through case examples of assessment, including successful language revitalization and maintenance efforts. The author also discusses the role of Non-Governmental Organizations (NGOs) in linguistic and cultural maintenance, especially in the Tibetan context.

Mihas et al. through a research monograph [11] address many complex and pressing issues of language endangerment. This volume specifically addresses language documentation, language revitalization, and training. The case studies of the volume provides detailed personal accounts of fieldworkers and language activists engaged in language documentation and revitalization work.

Lüpke [12] argues that the language vitality assessments used for African languages are rooted in Western language ideologies, and therefore, are inappropriate for the African context. The author proposes an alternative set of vitality parameters for African languages. Through a research monograph, Essegbey et al. [13] bring together a number of important perspectives on language documentation and endangerment in Africa.

A few resources are available for finding linguistic characteristics of languages and assessing their vitality/endangerment. The World Atlas of Language Structures (WALS) [14] is a free, online resource. It is a large database of the phonological, grammatical, lexical properties of the world languages. The WALS also features an interactive reference tool for exploring the database (https://www.eva.mpg.de/lingua/research/tool.php).

The Catalogue of Endangered Languages (ELCat) is another online resource for information on the endangered languages of the world [15]. The ELCat project is a partnership between Google, Alliance for Linguistic Diversity, University of Hawai'i at Mānoa Linguists, and the LINGUIST List at Eastern Michigan University. This project is sponsored by a grant from the National Science Foundation.

The Open Language Archives Community (OLAC) is an international partnership of institutions and individuals whose goal is to create a worldwide virtual library of language resources [16]. The OLAC is a free online service. Linguistic Linked Open Data (LLOD) is another free, cloud service for linguistic data [17]. It logically integrates diverse license-free, linguistic data resources and provides a unified search feature. Its linguistic resources include corpora; lexicons and dictionaries; terminologies, thesauri, and knowledge bases; linguistic resource metadata; linguistic data categories; and typological databases. Ethnologue is a commercial resource for the world language data [1].

## 3 The Telugu Language

Telugu is the official language of two states in southern India—Andhra Pradesh and Telangana. Telugu is also spoken in the Yanam district of Puducherry (a union territory of India). It is also spoken by a significant number of linguistic minorities in other states of India including Odisha, Karnataka, Tamil Nadu, Kerala, and Maharashtra. Telugu is a member of the Dravidian language family and there are over 215 million speakers for this language family. Among the languages of the Dravidian family, Telugu is the most widely spoken language. Per BBC news article, Telugu is the fastest growing language in the United States [18].

Contrary to the popular myth and propaganda, India has no national language. The constitution of India recognizes 22 languages as **scheduled languages** and Telugu is one of them. Telugu is also one of the six languages to have the **classical language of India** designation, which is bestowed by the Government of India.

According to the 2001 census of India, Telugu has the third largest number of native speakers in India at 74 million. Furthermore, Telugu ranks 13th in the Ethnologue list of most-spoken languages in the world. However, these ranks have fallen after a decade. Per 2011 census of India [19], Telugu slipped to the fourth position in terms of the largest number of native speakers in India and to the 15th place in the Ethnologue list of most widely spoken languages worldwide.



The Telugu script is an **Abugida**, which is derived from the Brahmi script. Abudiga is a segmental writing system where most consonants are immediately followed by a vowel, which form syllables. Each consonant-vowel sequence is written as a unit. The Telugu *Varnamala*/alphabet consists of 57 symbols, of which 18 are *achulu*/vowels, 36 are *hallulu*/consonants, and 3 are vowel modifiers. Of the 18 vowels, 2 of them are not used now. Even among the consonants, two of them have fallen out of use. The Telugu Varnamala is also called *aksharamulu*. The three vowel modifiers are considered as belonging to both vowel and consonant groups, and are referred to as *ubayaksharamulu*. The number of syllables in the language is approximately equal to the product of the number of constants and the number of vowels (i.e., $36 \times 18 \simeq 648$). From an Optical Character Recognition (OCR) point of view, the number of classes is very high.

Unlike English, there is no distinction between upper- and lower-case letters in Telugu. The script is written from left to right and the basic units of writing are syllables. Telugu words are pronounced exactly the way they are spelled. The Telugu writing system won the second place in The World Alphabet Olympics held in 2012 (https://languagelog.ldc.upenn.edu/nll/?p=4253).

Some features of *metrical poetry* are unique to the Telugu language. The Telugu poetry (called *padyalu*) employs an elaborate set of rules called *chandhas* for defining structural features. The *Chandhas* also apply to prose and it generates rhythm to the literature and poetry. The assigned unicode code-points for the Telugu language are 0C00–0C7F (3072–3199).

The Telugu language has vast literature. It has unique literary traditions, for example, *Ashtavadhanam*. The latter is a public performance of an *Avadhani* (the

performer) whose goal is to demonstrate her sharpness of memory and retention, mastery of the language, literature, grammar, and linguistic knowledge. In *Ashtavadhanam*, eight peers take turns in posing questions to the *Avadhani* and also distract her with challenges. The *Avadhani* answers to peer's questions must be constructed in a way to adhere to certain grammatical constructions and other linguistic constraints. *Satavadhanam* and *Sahasravadhanam* are advanced versions of *Ashtavadhanam*, where 100 and 1000 peers, respectively, ask questions and distract the *Avadhani*. These events last over several weeks.

The main challenges that the Telugu language faces today are lack of new literature, fast declining language speakers who can speak the language natively, lack of institutional support, and societal apathy toward the language. These factors will reduce the language to a mere spoken language in the near-term and eventual extinction in the medium-term. Paradoxical it may sound, very few current generation speakers speak the language natively. Their native language vocabulary is dismal, ability to read classic texts is abysmal, and the number of qualified language teachers is astonishingly small and rapidly shrinking. Even the so-called professionals in the Telugu mass media have severe language performance issues.

The other challenges include the dominance of English on the World Wide Web (WWW), the colonial past of India, federal government policies, strong desire of people for migration to other countries, and changing cultural and societal values. The current Indian government equates nationalism with having one language (i.e., Hindi) under the motto "One Nation, One Language." The federal government aggressively enforces Hindi on non-Hindi speaking population with the eventual goal of replacing English with Hindi. The Telugu movie industry is also another endangerment for the language, where the actors and actresses lack language proficiency and set a bad example for the younger generation.



The recent language policy introduced by the government of Andhra Pradesh fast-tracks the trajectory of Telugu language extinction. In November 2019, the government issued an executive order that all instruction in elementary and primary schools be delivered only in English, which was hitherto done in Telugu. The students will still learn Telugu as a language, but physical sciences, mathematics, and social sciences will all be taught in English. In private schools, teaching Telugu is optional. Replacing instruction in mother tongue (Telugu) with a foreign language (English) is a flawed policy and is a disaster in the making. This policy runs contrary to the research that suggests the critical need for learning in native language in primary and secondary schools [20–24].

## 4 CNNSs and RNNs

Computational Linguistics [25] plays a central role in preserving and promoting natural languages. However, it requires corpora of different genre in machine-readable format. Almost all significant and classical works in the Telugu language are copyright-free but are not in machine readable form. Though OCR is a solved problem for languages including English, Korean, Spanish, Mandarin, Turkish,

Arabic, and Japanese, it is an unsolved problem for many Indian languages including Telugu. Machine Learning (ML) in general and Deep Learning (DL) in particular offer unparalleled opportunities for solving the OCR problem for the *resource-poor* languages. Recently, ML and DL have been actively investigated in numerous domains and seminal works which include [26–31]. In the following, we provide a brief introduction to a major deep learning architecture used for OCR—Convolutional Neural Networks (CNNs).

The use of neural networks for classification problems dates back to the early days of machine learning, with the development of perceptron, one of the first algorithms for binary classification (Fig. 1). The perceptron models a neuron, by combining multiple inputs $(x_1, x_2, \ldots, x_n)$ to generate one output, similar to how a biological neuron takes its input from multiple other neurons connected to its dendrites, and produces one output through its axon. The perceptron assigns different weights to each input $(w_1, w_2, \ldots, w_n)$, in addition to its bias (i.e., $w_0$). If the value of the inputs multiplied by their weights, plus the bias, is larger than a threshold then its output is one, otherwise it is zero. This is the responsibility of the *activation function*.

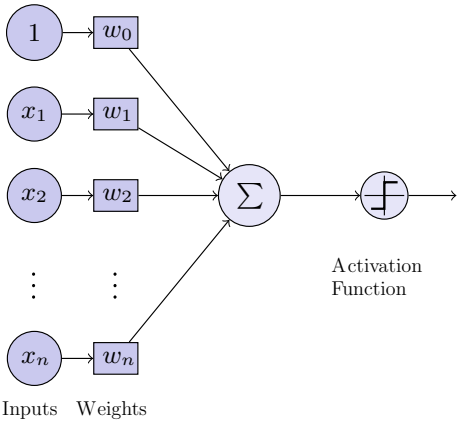


**Fig. 1.** A perceptron

Perceptrons can be extended to multiclass classification by using one perceptron for each class. This creates a basic neural network in which all inputs are connected to all outputs, i.e., a neural network with an input layer and an output layer. This model was improved by adding additional, hidden layers between the input and output layers, creating a multilayer, feedforward neural network, with the nodes from each layer connected to all the nodes in the subsequent layer. A feedforward neural network is shown in Fig. 2. This network has four layers—an input, an output, and two hidden layers. The first hidden layer has four neurons and the second hidden layer has three of them.

One of the first applications of neural networks was in computer vision, namely handwritten digit recognition. An image, represented by an array of

integers corresponding to the pixel values in the image, would be given as an input to the neural network, and the output would indicate the numerical value of the digit in the image. One main observation with this application is that the intensity of each pixel is not as significant as the differences between that pixel and adjacent pixels. This is because, due to the lighting conditions when each image is taken, a pixel might be lighter or darker. Thus, the emphasis should be on "local" and "differences."

Convolutional Neural Networks (CNNs) take this observation into consideration, and use intermediate layers that are not fully connected. Instead, in these layers, each node takes its inputs from a limited number of adjacent pixels—a patch of an image—and apply a filter to it. The role of each filter is to identify specific features in an image, such as a horizontal line, a vertical line, and so on. The output of these layers is passed to a fully connected neural network. For example, in an OCR task, if only the convolution filters for the horizontal line and for the diagonal line are activated, it might indicate that the handwritten digit is 7.

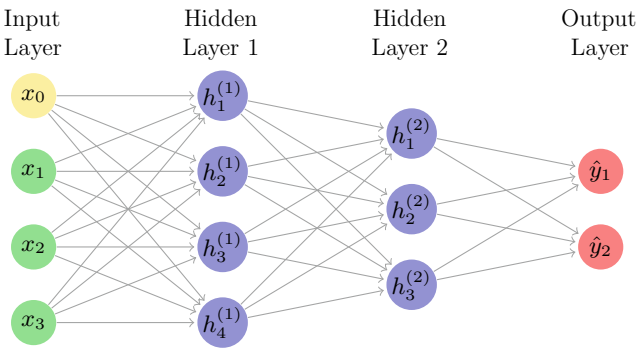


**Fig. 2.** A feedforward neural network

CNNs are specialized for processing grids of values. Another variation of multilayer neural networks are Recurrent Neural Networks (RNNs), which specialize in processing a sequence of values. One of the key characteristics of RNNs is that they allow cycles in the network, which enables them to create a memory of previous inputs that persists in the internal state of the network. Thus, they are appropriate for tasks in which previous inputs have an influence arbitraryly far into the future, such as in language-related tasks, where a current word influences the probability of seeing subsequent words, or a sequence of letters influence the combination of subsequent letters for creating a valid word.

One of the limitations of RNNs is that the influence of a particular input on the hidden layer, and implicitly on the output layer, either increases or decreases exponentially due to the cycle in the network. This is known as the *vanishing gradient problem*. One method that addresses this problem is Long Short-Term Memory (LSTM) architecture. The LSTM introduces memory blocks—a

set of recurrently connected subnets. Each memory block has one or more self-connected memory cells as well as input, output, and *forget units* which enable write, read, or reset operations for the memory cells, respectively. This allows memory cells to save and retrieve information over long periods of time, making them suitable for applications such as learning context-free languages.

## 5 Existing Resources and Approaches for Generating Corpora for the Telugu Language

Google Books (https://books.google.com/) is an ambitious project of Google, Inc. Its goal is to scan books provided by publishers and authors through the Google Books Partner Program, or by Google's library partners through the Library Project. Google also partners with magazine publishers to digitize their archives. The scanned documents are converted into machine-readable text through Optical Character Recognition (OCR) technology. Google provides a full text search capability over these digitized archives. As of October 2019, the number of scanned books exceed over 40 million. However, most of these books are no longer in print or commercially available.

A criticism about Google Books is that the errors introduced into the scanned text by the OCR process remain uncorrected. Furthermore, the digitized documents are not organized in the form of corpora to enable computational linguistics research. Lastly, some critics dubbed Google Book project as *linguistic imperialism* enabler—the transfer of a dominant language to other people. Majority of the scanned books are in English and this entails disproportionate representation of natural languages. We noticed some classical Telugu works such as *Vemana Satakam* in Google Books.

*The Million Books Project* is a book digitization project led by Carnegie Mellon University from 2007 to 2008. This project partnered with government and research entities in India and China and scanned books in multiple languages. The project features over 1.5 million scanned books in 20 languages: 970,000 in Chinese; 360,000 in English; 50,000 in Telugu; and 40,000 in Arabic. *The Million Books Project* is now replaced by HathiTrust Digital Library (https://www.hathitrust.org/). Most of the the HathiTrust digitized collections are protected by copyright law and thus are not fully viewable.

Commercial APIs for OCR include IBM Watson Discovery API, Microsoft Azure Computer Vision API, Amazon Web Services Rekognition API, and Google Vision API. Our experimentation with all these services failed to produce results that are simply not useful (see Figs. 3, 4, 5, and 6). The Google Vision API performed slightly better than the others, still the results are not useful. We attribute the poor quality of OCR results from these systems to their grandiose goal of one system for hundreds of languages.

Tesseract (https://opensource.google/projects/tesseract) is an open-source OCR engine from Google. Google claims that Tesseract can recognize more than 100 languages out of the box. Tesseract can also be trained to recognize other languages.

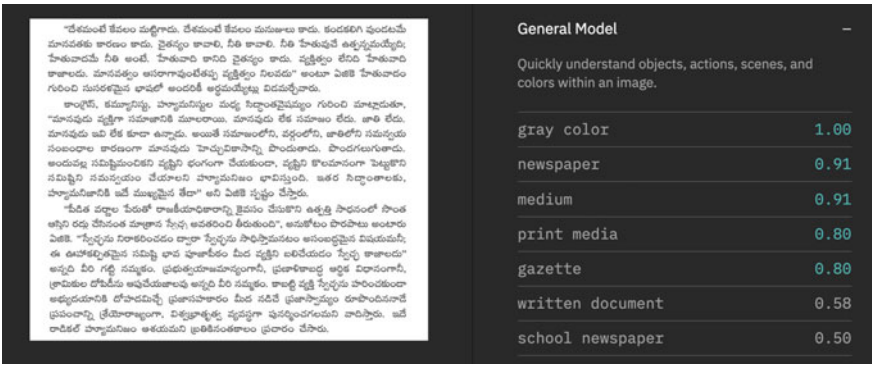


**Fig. 3.** OCR results from IBM Watson Discovery API

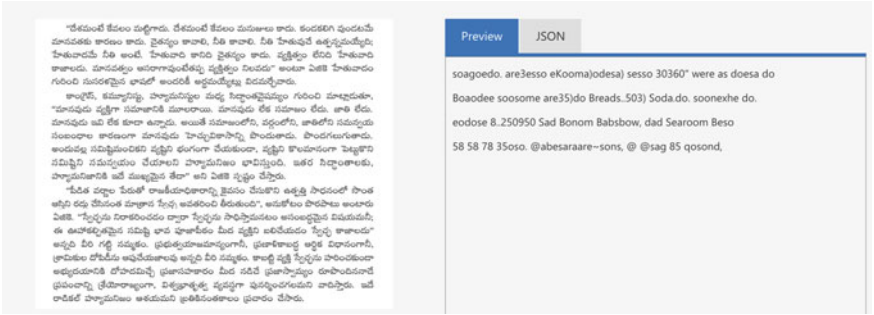


**Fig. 4.** OCR results from Microsoft Azure Computer Vision API

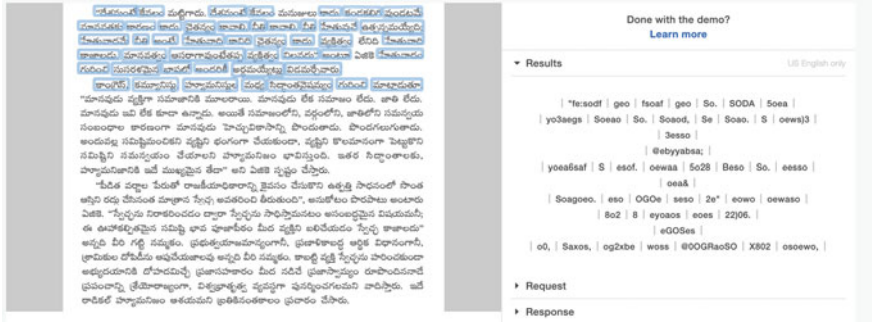


**Fig. 5.** OCR results from Amazon Web Services Rekognition API

https://github.com/TeluguOCR lists 8 code/corpora/font repositories for Telugu OCR. However, we were unable to make any of the code repositories compile and run. This is partly because of the code incompatibilities with the current versions of software libraries.

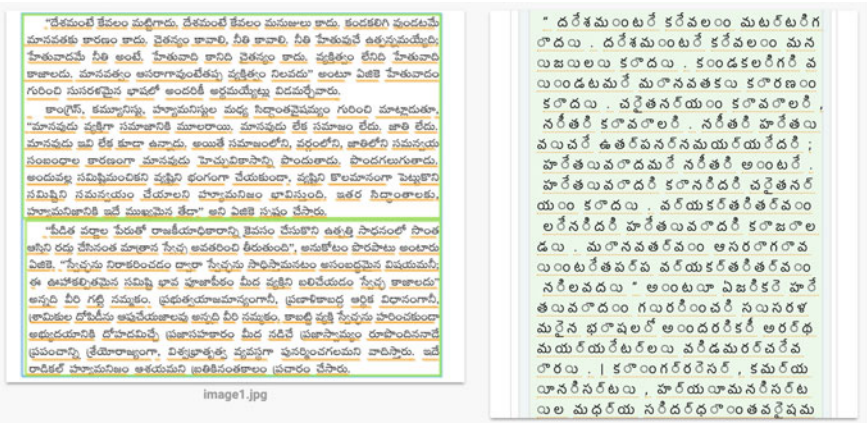


**Fig. 6.** OCR results from Google Vision API

## 6 Proposed Approach to Corpus Generation and Preliminary Results

The OCR for Telugu language poses several challenges due to its complex script and agglutinative grammar. Complex words in Telugu are derived by combining morphemes without changes to their spelling or phonetics. OCR is a two stage process—segmentation and recognition. The techniques used for segmenting Roman scripts are quite similar to each other. In such a script, the segmentation corresponds to identifying and demarcating units of written text. A unit corresponds to a contiguous region of text—a connected component. A robust segmentation algorithm should be able to effectively deal with noise from scanning, skew, erasure, and font variations. The recognition task is essentially a multiclass classification problem.

The proposed deep learning architecture for Telugu OCR employs a combination of CNNs and RNNs. Our model has three advantages. First, it requires no detailed annotations and learns them directly from the sequence labels. Second, the model learns informative feature representations directly from the segmented syllable images. This eliminates the handcrafting of features as well as some preprocessing steps such as the binarization component localization. Lastly, our model is lighter than a CNN model and requires less storage space.

The architecture of our model has three stacked components. The components from bottom to top layer are CNN, RNN, and a transcription layer. The CNN extracts the feature sequence, RNN predicts every frame of feature sequences passed from the CNN, and the top transcription layer translates the frame predictions into a label sequence. The entire network has a single loss-function.

We chose Amazon Web Services (AWS) infrastructure for implementing the proposed deep learning architecture. AWS infrastructure comes with pre-installed frameworks for deep learning. Amazon Web Services Rekognition API

provides excellent recognition accuracy for languages including English, Arabic, Chinese, Finnish, French, German, Hebrew, Indonesian, Italian, Japanese, and a few others. Obviously, Telugu is not one of these languages. We experimented with PyTorch and CUDA deep learning frameworks before deciding on TensorFlow.

We experimented with a few algorithms (binary, Gaussian, and MeanC) for converting a scanned text image into a binary thresholded image. We used Hough Line Transform (HLT) for skew detection. When HLT failed to detect skew, we used Minimum Bounding Rectangle (MBR) as a substitution. Figure 7 shows the word-level segmentation of our implementation on a scanned text image.

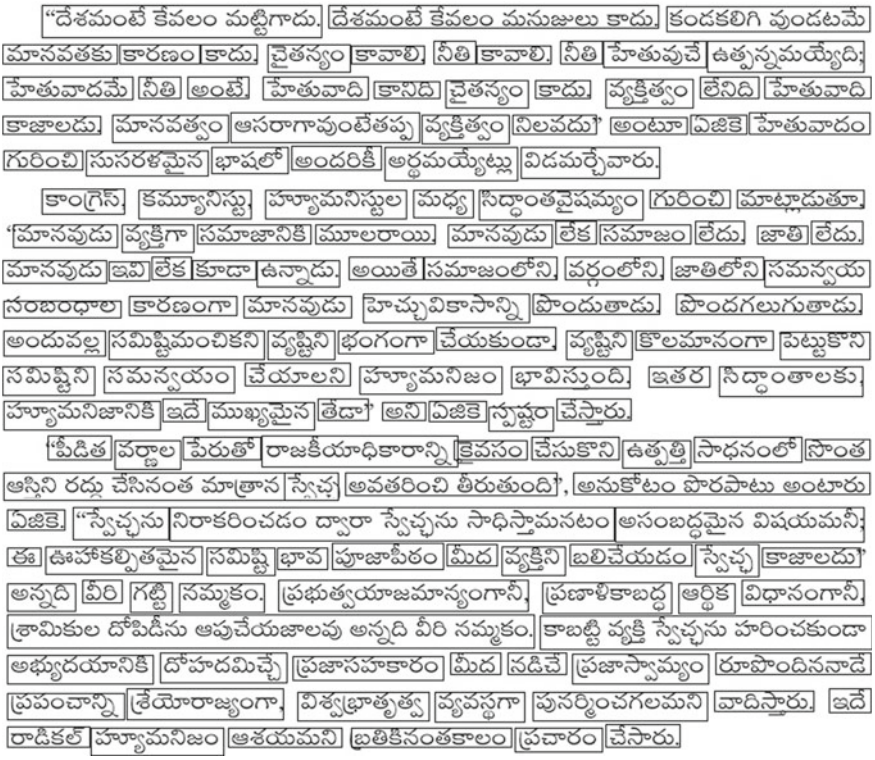

**Fig. 7.** Word-level segmentation performed by our system on a scanned text image

Our preliminary results indicate better performance for Telugu OCR compared to IBM Watson Discovery API, Microsoft Azure Computer Vision API, Amazon Web Services Rekognition API, and Google Vision API. This is expected given that these systems target hundreds of languages, whereas our system specifically focused on Telugu OCR. We plan to conduct additional evaluation of the system and further improve classification accuracy. Our end goal for this

research is to develop and open-source a robust Telugu OCR system, which in turn will enable creation of Telugu copora of various genre.

## 7 Conclusions

Language and culture are intricately intertwined. Language endangerment and cultural endangerment go hand in hand. The recent advances in computing, communications, and computational linguistics offer unparalleled opportunities for preserving and promoting linguistic diversity, and ensuring cultural autonomy and cultural pluralism. In a multi-cultural and linguistically diverse country such as India, cultural self-determination is crucial for India to succeed and excel in human development. Much to the chagrin of vested interests, the fact remains that India is much like the European Union and it should have been aptly named as the Unite States of India.

Though we have highlighted the case for the Telugu language, the issues are similar for other Indian languages such as Kannada, Marathi, Odiya, and Malayalam [32]. What is needed is strategies, tactics, and execution plans to preserve, promote, and celebrate linguistic diversity and the associated cultures. Toward this goal, the International Association for the Development of Cross-Cultural Communication issued a declaration at its 22nd seminar on Human Rights and Cultural Rights held in 1987 at Recife, Brazil. This declaration is referred to as the *Universal Declaration of Linguistic Rights* or *The Recife Declaration* and requires reformulation of national, regional, and international language policies.

Among other things, the Recife Declaration states that " ... Recognizing that the learning and use, maintenance and promotion of languages contribute significantly to the intellectual, educational, sociocultural, economic and political development of individuals, groups, and states. ... Asserting that linguistic rights should be acknowledged, promoted, and observed, nationally, regionally and internationally, so as to promote and assure the dignity and equity of all languages. ... Aware of the need for legislation to eliminate linguistic prejudice and discrimination and all forms of linguistic domination, injustice and oppression, in such contexts as services to the public, the place of work, the educational system, the courtroom, and the mass media. ... Stressing the need to sensitize individuals, groups, and states to linguistic rights, to promote positive societal attitudes toward plurilingualism and to change societal structures toward equality between users of different languages and varieties of languages. ... "

It is often said that the threatened languages are frequently surrounded by indifferent and unsympathetic insiders rather than the hostile outsiders (e.g., imposition of Hindi by the Government of India on non-Hindi-speaking populations). The need of the hour is not to wait for the governments to implement policies and provide resources to reverse the language shift and trajectory toward extinction. What is needed is grassroots movements to bring widespread awareness of the language and cultural endangerments, followed by strategy, tactics,

and clean execution. In the next stage, this awareness should be steered toward influencing electoral outcomes and effecting government policies and resource allocations.

The confluence of handheld computing devices, machine learning, computational linguistics, social media, and native language enthusiasts is an unstoppable force for language revitalization. Telugu youth with computing knowledge and skills can play an extraordinary role by creating open-sourced language learning apps including cross-word puzzles, grade-appropriate vocabulary lists, modern dictionaries, and short stories [33]. If the native speakers of the Telugu language do not lead this effort, who else will? The time is now and any further delay will cause irreversible damage.

This work is supported in part by the National Science Foundation IUSE/PFE:RED award #1730568.

## References

1. Eberhard, D.M., Simons, G.F., Fennig, C.D. (eds.): Ethnologue: Languages of the World, 22nd edn, SIL International, Dallas, Texas (2019). http://www.ethnologue.com
2. Lenneberg, E.H.: Biological Foundations of Language. Wiley, New York, NY (1967)
3. Miozzo, M., Rapp, B. (eds.): Biological Foundations of Language Production. Psychology Press, Special Issues of Language and Cognitive Processes (2011)
4. Ladefoged, P., Maddieson, I.: The Sounds of the World's languages. Wiley-Blackwell, New York, NY (2009)
5. Hale, K.: Endangered languages **68**, 1–42 (1992). https://doi.org/10.2307/416368
6. Fishma, J.A.: Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages. Multilingual Matters (1991)
7. United Nations Educational: Scientific and Cultural Organization: UNESCO's language vitality and endangerment methodological guideline: Review of application and feedback since **2003**, (2011)
8. Lewis, M.P., Simons, G.F.: Assessing endangerment: expanding fishman's GIDS. Revue roumaine de linguistique **LV**(2), 103–110 (2010). https://www.lingv.ro/RRL-2010.html
9. Lee, N.H., Way, J.V.: Assessing levels of endangerment in the Catalogue of Endangered Languages (ELCat) using the Language Endangerment Index (LEI). Lang. Soc. **5**, 271–292 (2016). https://doi.org/10.1017/S0047404515000962
10. Dwyer, A.M.: Tools and techniques for endangered-language assessment and revitalization. In: Vitality and Viability of Minority Languages, Trace Foundation Lecture Series Proceedings. New York, NY (2011). http://www.trace.org/about
11. Mihas, E., Perley, B., Rei-Doval, G., Wheatley, K. (eds.): Responses to Language Endangerment: In Honor of Mickey Noonan. Studies in Language Companion, John Benjamins (2013)
12. Lüpke, F.: Ideologies and typologies of language endangerment in Africa. In: Essegbey, J.A., Henderson, B., Mc Laughlin, F. (eds.) Language Documentation and Endangerment in Africa, Handbook of Statistics, Chap. 3, pp. 59–105. John Benjamins (2015). https://doi.org/10.1075/clu.17.03lup
13. Essegbey, J., Henderson, B., Laughlin, F.M. (eds.): Language Documentation and Endangerment in Africa. Culture and Language Use, John Benjamins (2015)

14. The world atlas of language structures (wals) (2013). http://wals.info/
15. Catalogue of Endangered Languages (ELCat) (2019). http://www.endangeredlanguages.com/
16. OLAC: The open language archives community (2019). http://olac.ldc.upenn.edu/
17. LLOD: linguistic linked open data (2019). http://linguistic-lod.org/
18. BBC News: Do you speak telugu? welcome to America (2018). https://www.bbc.com/news/world-45902204
19. of India, G.: Data on language and mother tongue (2011). http://censusindia.gov.in/2011Census/Language_MTs.html
20. What is mother tongue education? (2019). https://www.rutufoundation.org/what-is-mother-tongue-education/
21. Jain, T.: Common tongue: the impact of language on educational outcomes. J. Econ. Hist. **77**(2), 473–510 (2017). https://doi.org/10.1017/S0022050717000481
22. Noormohamadi, R.: Mother tongue, a necessary step to intellectual development. Pan-Pacific Association of Applied Linguistics **12**(2), 25–36 (2008). https://files.eric.ed.gov/fulltext/EJ921016.pdf
23. Savage, C.: The importance of mother tongue in education (2019). https://ie-today.co.uk/Blog/the-importance-of-mother-tongue-in-education/
24. Seid, Y.: Does learning in mother tongue matter? evidence from a natural experiment in Ethiopia. Econ. Educ. Rev. **55**, 21–38 (2016). https://doi.org/10.1016/j.econedurev.2016.08.006
25. Clark, A., Fox, C., Lappin, S.: The Handbook of Computational Linguistics and Natural Language Processing. Wiley-Blackwell, New York, NY (2012)
26. Barber, D.: Bayesian Reasoning and Machine Learning. Cambridge University Press, New York, NY (2012)
27. Faul, A.: A Concise Introduction to Machine Learning. Chapman and Hall/CRC, Boca Raton, Florida (2019)
28. Goldberg, Y.: Neural Network Methods in Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, California (2017)
29. Ji, Q.: Probabilistic Graphical Models for Computer Vision. Academic Press, Cambridge, Massachusetts (2019)
30. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. The MIT Press, Cambridge, Massachusetts (2009)
31. Murphy, K.: Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts (2012)
32. Economist, T.: Language identity in India: one state, many worlds, now what? (2013). https://www.economist.com/johnson/2013/06/25/one-state-many-worlds-now-what
33. Open-source software can revitalize indigenous languages (2019). https://en.unesco.org/news/open-source-software-can-revitalize-indigenous-languages-0