**LAHIRI — Digitisation**

<0 images>

**Digitisation and Automatic Classification of Indian Folktales through MAIF**

Kumar, R., Lahiri, B., Ojha, A. and Alok, D.

In this paper we discuss the development of a collaborative and crowd-based platform, the Multilingual Archive of Indian Folktales (MAIF). It is a comparative, multilingual, digital repository of the folk literature and their automatic classification in different languages and cultures of India. This repository is mainly targeted towards the lesser-known, endangered, and scarcely documented and described languages of the country. It is a platform for linguists, folklorists, performance researchers, sociologists, anthropologists, and anyone else interested in folk literature, culture, and language to deposit their own folktales as well as use the folktales already deposited in the repository for their own purposes.

It is expected that a large part of the contribution to the repository will be made by the field linguists who are actively involved in the documentation and description of various aspects of the endangered and lesser-known languages. This repository could be used for a large number of activities, including comparative study of the folktales and the language itself, development of teaching materials in the language as well as language technologies for these languages (we have developed some applications for Magahi and Bhojpuri using these data)—both of which could prove to be instrumental in revitalization of these languages.

**Motivation**

It has been estimated that out of around 6,000 languages spoken in the world, only 83 languages are spoken by around 80% of the population. On the other hand, just 20.4% of speakers use one of the 3,000 languages, and a meager 0.2% of the population use one of the other 3,586 languages across the world (Harrison, 2007). In addition, almost 14 years ago Nettle and Romaine (2000) had estimated that over 11% of the languages had fewer than 150 speakers each; it could be fairly assumed that quite a few of these languages would have ceased to be spoken by now. The situation in India is no less bleak. A large number of Indian languages, 196 as per UNESCO’s Atlas of Endangered Languages (Moseley, 2010), are considered endangered and in immediate need of revitalization and documentation.

One of the biggest reasons behind the endangerment of language (and consequently the indigenous knowledge system and culture of the speakers) is the negative attitude of the speakers towards the language and their unwillingness to pass on the language to the children. This negative attitude towards the language could be softened and transformed into a positive attitude, thereby initiating some kind of linguistic revitalization of the endangered languages, through various means. Ensuring a digital presence of the language could prove to be one of the motivations behind the revitalization of the language. Moreover, it will also ensure long-lasting documentation of the language. In addition to the other linguistic forms, folktales form a very significant part of the language documentation process since it not only provides linguistic data but also gives a peek into the sociocultural aspects as well as folk knowledge of the community. Thus, MAIF serves the twofold purpose of current documentation as well as serving as a tool for possible future revitalization of not just the language but also the culture of the community.

**The Frontend of MAIF**

This platform allows people with any kind of folk literature to upload it to the repository. Every folktale is transcribed using the International Phonetic Alphabets (IPA) along with the text in the native script, if available. The transcriptions is interlinearly glossed at the morphemic level. If morphemic-level information is not available or it is not possible to give information at that level, then a minimum word-to-word glossing is provided. Along with this, the text is also annotated with part-of-speech information. In addition, a free translation of every folktale in English, aligned at the sentence level, is provided. It not only helps in providing access to the folktales for a large number of people but also helps in the automatic classification of the folktales across different languages. The folktales could be either entered directly in the web interface of the archive or uploaded in the proper XML format. The folktales could also be accompanied with audio or audiovisual recording of the narration.

**The Backend of MAIF**

At the backend, MAIF is structured and maintained as an XML database. The annotation of each sentence at the morphological level and the part-of-speech level are kept at separate tiers of annotation. The XML schema for MAIF allows for as many additional levels of annotation as are required for a particular task. Throughout the data storage process we have followed the recommendations given by Text Encoding Initiative (http://www.tei-c.org/index.xml) to ensure that the documentation is long-lasting.

The part-of-speech annotation is carried out using the Bureau of Indian Standards (BIS) tagset for part-of-speech annotation. This is a generic tagset for annotating Indian languages. Out of this generic tagset, depending on the grammatical features of a language, a subset of tags is selected for annotating the data of a particular language. This subset serves as the tagset for that particular language. The generic tagset consists of 11 top-level categories: Noun, Pronoun, Demonstrative, Verb, Adjective, Adverb, Postposition, Conjunction, Particles, Quantifiers, and Residuals. It is expected that any subset of this generic tagset will consist of all these categories. Eight out of these 11 categories have subcategories at level 2 and level 3. Thus, this tagset provides a fine balance between being concise so that it is easier for machines to learn the tagset as well as descriptive enough to give a good idea of the lexical categories in the text.

The interlinear glossing is carried out using the Leipzig Glossing Conventions (http://www.eva.mpg.de/lingua/resources/glossing-rules.php). These conventions consist of 10 rules for the syntax and semantics of interlinear glossing. These rules specify the labels for category as well as different conventions for specifying these categories through the labels. The glossing helps in automatic morphological analysis as well as linguistic analysis of the data.

We are also planning to add semantic annotations to the texts in the corpus using Proppian fairy tale Markup Language, PftML (Malec, 2001). It will serve as a resource for understanding the segments of the folktales across languages and understanding the structure and semantics of the tales in different languages. This markup language is based on the work of Russian folklorist Vladimir Propp, who divided the plot of folktales into a set of basic plot components known as ‘character functions’ (Propp, 1968). This is a functional classification method based on roles that actors in a story perform—how they interact with each other such that this sequence of actions builds up the complete storyline. It has been noted in some of the previous studies that PftML is a coarse-grained way of annotating the tales, and supplementing these annotations with deeper linguistic information would be very helpful in proper categorisation and understanding of the tales (Lendvai et al., 2010). MAIF incorporates this kind of linguistic information in its basic structure itself and could thus prove to be a very rich resource.

**Automatic Interlinear Glossing in MAIF**

At the backend of MAIF, we have also developed a system for automatic training of the glossing system using supervised machine learning techniques. The data for this supervised training are provided by the interlinearly glossed data of the repository itself.

This system works at all three levels of annotation—morphological, lexical (the lexical categories), and semantic. The system currently runs support vector machines (SVM) for training and automatically annotating new data with this information. This, in effect, implies that the interlinear glossing of any new folktale entered into the system is an almost completely automated task. The algorithm and the set of features used for the automatic glossing is based on several experiments conducted on the data entered in the archive (Kumar et al., 2012). The best algorithm and feature set combination for each language has been used for the automatic glossing process of that particular language. This process requires that a substantial amount of data is entered and manually glossed before the system could be trained to automatically gloss new data. Once sufficient data are available and the automatic glossing system begins to work efficiently, then new data could be processed very quickly and with great precision. And as more data are created, it could be further used to make the classification system even more robust.

**Progress till Now**

In its current form, the repository has folktales from three languages of India: Magahi, Bhojpuri, and Paite. Magahi and Bhojpuri are Eastern Indo-Aryan languages, spoken in the Magadh and Bhojpur regions, respectively. Paite is a Tibeto-Burman language spoken in the state of Manipur. There are around 100 folktales in Magahi and Bhojpuri each, while there are around 10 folktales in Paite. The folktales in Magahi are annotated at the morphological and lexical level, and there is an automatic part-of-speech annotation system and morphological analyser running for Magahi. There is also an automatic part-of-speech annotation system for Bhojpuri running at the backend. This ensures that any new Magahi or Bhojpuri folktale that is submitted on the system need not be annotated manually for this information.

**References**

**Harrison, K. D.** (2007). *When Languages Die: The Extinction of the World’s Languages and the Erosion of Human Knowledge*. Oxford University Press, Oxford.

**Kumar, R., Lahiri, B. and Alok, D.** (2012). Developing a POS Tagger for Magahi: A Comparative Study. In *Proceedings of the 10th Workshop on Asian Language Resources*. Mumbai: COLING 2012 Organizing Committee, pp. 105–14.

**Lendvai, P., Declerck, T., Darányi, S., Gervás, P., Hervás, R., Malec, S. and Peinado, F.** (2010). Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC–2010)*, Valletta, Malta, 19–21 May 2010, pp. 1996–2001.

**Malec, S.** 2001. Proppian Structural Analysis and XML Modeling. In *Proceedings of the CLiP (Computers, Literature and Philology) Conference*, University Gerhard-Mercator, Duisburg, 6–9 December 2001.

**Moseley, C.** (ed.). (2010). *Atlas of the World’s Languages in Danger*. 3rd ed. Paris: UNESCO Publishing, http://www.unesco.org/culture/en/endangeredlanguages/atlas.

**Nettle, D. and Romaine, S.** 2000. *Vanishing Voices: The Extinction of the World’s Languages*. Oxford University Press, Oxford.

**Propp, V.** (1968). *Morphology of the Folktale*. Scott, L. and Wagner, L. A. (transl.). University of Texas Press, Austin.