Finite-State Morphological Analysis for Marathi

Vinit Ravishankar
Faculty of ICT
University of Malta
Msida MSD 2080, Malta
vinit.ravishankar@gmail.com

Francis M. Tyers
School of Linguistics
Higher School of Economics
Moscow, Russia
francis.tyers@uit.no

Abstract

This paper describes the development of free/open-source morphological descriptions for Marathi, an Indo-Aryan language spoken in the state of Maharashtra in India. We describe the conversion and usage of an existing Latin-based lexicon for our Devanagari-based analyser, taking into account the distinction between full vowels and diacritics, that is not adequately captured by the Latin. Marathi displays elements of both fusional and agglutinative morphology, which gives us different ways to potentially treat the morphology; philosophically, we approach our analyser by treating the morphology system as a three-layer affixing system. We use the *lttoolbox* lexicon formalism for describing the finite-state transducer, and attempt to work within a morphological framework that would allow for some consistency across Indo-Aryan languages, enabling machine translation across language pairs. An evaluation of our finite-state transducer shows that the coverage is adequate, over 80% on two corpora, and the precision is good (over 97%).

1 Introduction

This paper describes the development of free/open-source morphological descriptions of Marathi, an Indo-Aryan language spoken in the state of Maharashtra in India. Morphological descriptions are computational models of a language's morphology, and are used to output morphological analyses from word forms and vice versa.

In section 2, the paper gives an overview of Marathi morphology, and talks about some of the grammatical decisions we made during development of the analyser. Section 3 is a literature review of previous work done in the field. Section 4 describes the methodology we followed whilst working on the analyser, the formalisms we have used, and describes our lexicon. We continue with section 6, describing the evaluation metrics we have chosen, and how well our analyser performs on them. Section 7 describes potential future work we could do.

2 Marathi

Marathi is an Indo-Aryan language spoken primarily in the west Indian state of Maharashtra, and has approximately 62 million speakers, as of 2003 (Pandharipande, 2003). Despite being an Indo-European language, Marathi has borrowed several features - such as clusivity, and certain retroflex consonants (such as the retroflex lateral flap), either absent or relatively uncommon in other Indo-Aryan languages.

Whilst Marathi retains some fusional morphological aspects of its proto-language, Sanskrit, it displays morphological agglutination within many contexts. Our analysis broadly follows the perspective of Masica (1993). They consider the split morphological to be a form of morphological "layering"; with a primary layer, comprising mainly of inherited fusional elements (the "oblique" case), a secondary agglutinative layer, and a tertiary postpositional layer. These layers are, to a certain extent, universal amongst Indo-Aryan languages: they differ largely in the conditions under which they occur, and language-specific variations that may occur. A brief, specific definition of the layers in Marathi would, therefore, look like:

1. The "oblique" case; complex morphophonemic changes in the lemma. eg. मुलगा *mulagā* "boy" → मुला *mulā*

2. Agglutinative suffixes, similar to traditional cases that mark noun functions, like the nominative or the genitive.

3. Postpositions; morphologically and semantically complex elements. These can attach to an (optional) oblique genitive suffix in layer 2.

Certain particles, such as an emphasis particle -च -*c*, or particles like -ही -*hī* and -सुद्धा -*suddhā* "as well as", are quite common, and can attach as a suffix to most words, with the exception of conjunctions. Verbs, along with the optional negation particle, decline for

Proceedings of the 13th International Conference on Finite State Methods and Natural Language Processing, pages 50–55,
Umeå, Sweden, 4–6 September 2017. © 2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/W17-4006

tense, aspect and mood, and have adjectival and adverbial derivations. (1) is an example with two of the three case layers and two suffix particles.

(1) (to) ghar-ā-māge-hī
(he) house-OBL-behind.POST-too.PTCL
ge-l-ā-c
go-PFV-3MSG-FOC
"He definitely went behind the house too"

3 Prior work

There have been a number of efforts to develop morphological analysers for Marathi over the years. While morphological analysis for Marathi is fairly well studied, one downside of previous work is that the software and lexicon is not freely available. Dixit et al. (2005) present a spellchecker for the language based on a lexicon of 13,000 root words and morphological rules. They did an evaluation of spell-checking accuracy showing that out of 10,648 words classified as correctly spelt, only 0.45% were actually false positives. The morphological analyser of Bapat et al. (2010) is based on a word–paradigm approach modelled with a finite-state transducer and contains a lexicon of 24,035. They evaluate 21,096 unique word forms from a corpus and find that 97.18% receive all and only the correct morphological analyses; it is worth noting, however, is that their dictionary was created to specifically fit their evaluation corpus. Another analyser based on finite-state technology is described by Dabre et al. (2012), based on a gold standard of 1,341 words, achieved an accuracy of 72.18%. The size of the lexicon was not specified. Gawade et al. (2013) also use a finite-state transducer to model Marathi morphology, although their paper does not evaluate its effectiveness.

Resources like BabelNet[1], generated by statistically machine translating WordNet ontologies, do not appear to be very useful - whilst BabelNet does contain some Marathi nouns, common verbs are all absent.

4 Development

We initially worked on the open word classes, many of which could be successfully scraped from the resources of the Language Technologies Research Centre (LTRC), at the International Institute of Information Technology, Hyderabad.[2] As the lexicon was in WX notation,[3] a transliteration script was used along, along with standard UNIX command line utilities, to extract and convert noun paradigms. Adjective paradigms were fairly trivial to convert; a significant number of adjectives do not inflect at all, and most others are very regular. Words were then *scraped* (extracted) from the lexicon and assigned to their respective paradigms.

Verbal declensions were stored with a different method; separate files existed, not for separate paradigms, but for separate word forms. Each file had a set of words, declined to match the particular form described by the file. The lexicon, however, was similar to the nominal lexicon, in that verbs were assigned a particular verb paradigm. Rather than merge these multiple files into a single paradigm, we created our own verbal paradigms, with Dhongade and Wali (2009) and Masica (1993) as references. The verb list, however, primarily consists of entries from the LTRC lexicon.

4.1 Formalisms

For the finite-state transducer we employ the *lttoolbox* formalism, an XML-based format used in the *Apertium* project (Forcada et al., 2011). This formalism is widely used for encoding language data, with Apertium having over 40 language pairs for machine translation. Although we could have used an FST toolkit like HFST (Lindén et al., 2011) or Foma (Hulden, 2009), with separate layers for processing morphonology and morphotactics, the lack of significant morphophonological processes relevant to Marathi orthography made *lttoolbox* a perfectly adequate choice.

4.2 Lexicon

The main source of lexical material for our analyser is from an existing morphological analyser published by the Language Technology Research Centre (LTRC) at IIIT Hyderabad. Unlike other work on Marathi, the lexicon is available under the free/open-source GPL licence. The source lexicon (see example in Figure 3) is composed of a dictionary table containing six columns. All text in Marathi is written in a Latin-based transliteration scheme.

The paradigms in the LTRC lexicon are essentially lists of different forms of a word; words are assigned paradigms based on their conformance to the inflection of the paradigm word. One of the biggest problems with this is the inefficient noun paradigm system; each paradigm lists forms that include bound postpositional morphemes (including adjectival postpositions); this is quite unnecessary, as postpositions (layer 3) are largely regular, and attach to the oblique case (layer 1) with an optional clitic (layer 2). This results in 968 forms per paradigm, where four would suffice - the singular and plural nominative and oblique. There were other minor problems, such as the inclusion of plural forms for uncountable nouns or abstract nouns.

4.3 Paradigms

The Apertium paradigm system essentially functions using finite-state transducers, defined in XML. Paradigms are expressed as an input side (within '<l></l>' tags), and a corresponding output side (within '<r></r>' tags); the transducer is made to re-

[1] http://babelnet.org/

[2] Available from the LTRC website at `http://ltrc.iiit.ac.in/showfile.php?filename=onlineServices/morph/index.htm`

[3] WX is an ASCII-based transliteration scheme for Indian languages; the name derives from the use of 'w' and 'x' for dental stops.

```
^ठेचा/ठेचा<n><m><sg><nom>$
^मिरची/मिरची<n><f><sg><nom>$
^,/,<cm>$
^चिंच/चिंच<n><f><sg><nom>$
^व/व<cnjcoo>$
^मीठापासून/मीठ<n><nt><sg><obl>+पासून<post><adv>$
^तयार केला/तयार करणे<vblex><perf><p3><m><sg>$
^जातो/जाणे<vblex><impf><p3><m><sg>$
^./.<sent>$
```

Figure 1: Example output from the analyser for the sentence ठेचा मिरची, चिंच व मीठापासून तयार केला जातो *ṭhēcā miracī, ciṃnca va mīṭhāpāsūna tayāra kelā jāto* “Pickles are prepared using chilis, tamarind and salt.” Note that the example has been manually disambiguated for brevity. The tag `cnjcoo` is coordinating conjunction, and `cm` is comma.

turn the lemma of a word and the corresponding tags. The Marathi dictionary had a few caveats regarding transliteration of the lexicon to Unicode; a paradigm with the invariant part of the word ending before a vowel would require additional entries in Unicode, depending on whether the final letter of the invariant part was a vowel sound or not. For instance, consider the pair *A/I* and *t/I*: the Unicode equivalents for this pair would be आ/ई and त/◌ी, with the vowel displayed as a diacritic in the second case. Both characters — the full vowel and the diacritic — are distinct Unicode code points. Whilst we can *infer*, from the WX transliteration equivalent, that the vowel ought to be a diacritic and not a full vowel, the distinction is explicit in the Unicode.

For several morphological contexts in which Marathi displays some form of agglutinativity, we have used the “join” operator, which essentially redirects the FST to another paradigm after it consumes the input for the first. This has resulted in a lot of ‘minor’ paradigms in the dictionary.

There are significant phonological differences between spoken and literary Marathi; these often reflect in informal written Marathi, which tends to modify spellings to match the spoken variant. Most paradigms include, therefore, multiple forms mapping onto the same analysis; there are, however, restrictions placed on the non-standard forms to prevent them from being generated during morphological generation. The most common example of this is neuter agreement - whilst literary Marathi uses the vowel /e/ ◌े *e* to mark the third-person neuter, informal Marathi uses a schwa, represented by a nasalisation diacritic ◌ं.

5 Grammar

During the development of this analyser, we made several linguistic decisions, some of which we shall attempt to describe and justify.

5.1 Light verbs

Marathi, like many other Indo-Iranian and Turkic languages, has frequent light verbs. These are, essentially, noun + verb constructs that represent a verbal predicate. These constructs have been fairly widely studied, particularly within the context of Persian (Karimi-Doostan, 2005). Whilst N + V combinations are, by far, the most common type of construct, there are several constructs where the first element cannot exist as an independent term (but are glossed as adverbs). (2) is an example of a sentence with a light verb construction, using the relatively uncommon verb मारणे *mārṇe* “to hit”.

```
^तयार करणे/तयार करणे<vblex><inf>$
^तयार/तयार<adv> करणे/करणे<vblex><inf>$
```

Figure 2: Example analyses for the light verb construct तयार करणे *tayār karṇe* “to prepare”. The first analysis is what we use; the second is an intended addition.

(2)	mī	(zamīn-ī-lā)	zhāḍū	mār-t-o
	I	(floor-OBL-DAT)	broom.N	hit.V-IPFV-1MSG

I sweep (the floor)

In our analyser, we attempt to create separate entries for every semantically valid light verb pair - i.e., an entry for each noun + verb combination, with a whitespace token separating the two. Whilst this approach does make things easier from the perspective of machine translation, it has two disadvantages - it is very laborious work, and it is not completely compatible with other linguistic resources, like the Universal Dependencies treebank project (Nivre, 2015), which requires that both the noun and the verb have separate analyses. We intend to eventually add support for both forms of analysis; Figure 2 shows the difference between the two analyses.

5.2 Verbal morphology

An issue we faced during development of the analyser was finding suitable names for all verb forms. Marathi’s relative verbal complexity, and the lack of consistency amongst our reference grammars along with the absence of descriptions of several verb forms, made this a fairly difficult task. We describe some of the forms:

- **Supine:** `<sup>`; name derived from the Latin supine, these forms indicate purpose for the action denoted by the verb; i.e. “in order to” carry out the action.
- **Transgressives:** `<trans>`; similar to Slavic transgressives, the forms indicate simultaneous (imperfective) and consecutive (perfective) actions.
- **Inceptives:** `<incp>`; inceptives typically form compounds with the verb लागणे *lāgṇe* “to attach”, and indicate the starting of the action denoted by the verb.
- **Predictive:** `<pred>`; these forms indicate an *intent* to carry out the action denoted by the verb.

Corpus	Tokens	Cov. (%)	Mean ambig.
Wikipedia	4.0M	80.2	1.7
Bible	751K	80.7	1.9
Average	–	80.45	1.8

Table 1: Corpora used for naïve coverage tests

> The same forms are also used for the desiderative, to imply a desire to do something - we, however, chose to use `<pred>`, similar to Dhongade and Wali (2009).

It is worth noting that gerunds[4], like nouns, can take affixes functionally similar to the second and third layer in nominal affixing. The gerund itself is assumed to be the oblique, and therefore does not undergo any further modification before it takes case or postpositional suffixes. (3) is an example of a gerund with a postposition.

(3) tu-jhyā basṇ-yā-nantar mī ge-l-o
you-GEN sit-GER-after.POST I go-PFV-1MSG
I went after you sat *(after your sitting)*

6 Evaluation

We have evaluated the morphological analyser in two ways. The first was by calculating the naïve coverage and mean ambiguity on freely available corpora. Naïve coverage refers to the percentage of surface forms in a given corpora that receive at least one morphological analysis. Forms counted by this measure may have other analyses which are not delivered by the transducer. The mean ambiguity measure was calculated as the average number of analyses returned per token in the corpus.

Our analyser was also used by Ravishankar (2017) to evaluate coverage on generated back-transliterated corpora; their coverage was comparable to our evaluation, at 72.65% for their best system.

6.1 Corpora

We used two freely-available corpora for the evaluation. The first is the Marathi Wikipedia,[5] and the second is the Bible in Marathi.[6]

6.2 Precision and recall

Precision and recall are measures of the average accuracy of analyses provided by a morphological transducer. Precision represents the number of the analyses given for a form that are correct. Recall is the percentage of analyses that are deemed correct for a form (by comparing against a gold standard) that are provided by the transducer. To calculate precision and recall, it was necessary to create a hand-verified list of surface forms and their analyses. We extracted 1,000 unique surface forms at random from a Wikipedia corpus, and checked that they were valid words in the languages and correctly spelled. Where a word was incorrectly spelled or deemed not to be a form used in the language, it was discarded.

	Precision	Recall
Known tokens	0.97	0.97
All tokens	0.97	0.71

Table 2: Precision and recall over all tokens and only known tokens. Out of 699 tokens which were checked, the stems of 347 were not in the lexicon.

This list of surface forms was then analysed with the most recent version of the analyser, and each analysis was checked. Where an analysis was erroneous, it was removed; where an analysis was missing, it was added.[7] This process gave us a 'gold standard' morphologically analysed word list of 699 forms. The list is publicly available for each language in Apertium's SVN repository.

We then took the same list of surface forms and ran them through the morphological analyser once more. Precision was calculated as the number of analyses which were found in both the output from the morphological analyser and the gold standard, divided by the total number of analyses output by the morphological analyser.

Recall was calculated as the total number of analyses found in both the output from the morphological analyser and the gold standard, divided by the number of analyses found in the morphological analyser plus the number of analyses found in the gold standard but not in the morphological analyser.[8]

The results for precision and recall are presented in table 2.

6.3 Qualitative

After performing the manual evaluation, we found that a majority (59.6%) of the missing analyses were nouns or proper names. For the former class, the missing analyses were often English loanwords; amongst these, we rejected ones that would not conventionally occur in

[4] Dhongade and Wali (2009) do not treat these forms as gerunds; we disagree.

[5] The dump file used was `mrwiki-20150901-articles.xml.bz2` from `http://dumps.wikimedia.org`.

[6] Downloaded from `https://www.wordproject.org/bibles/mar/`. Corpus statistics and coverage are presented in table 1.

[7] By this we mean that analyses which the morphological analyser produced which were erroneous were removed from the gold standard in order to be able to determine how frequently the analyser produces erroneous analyses.

[8] For example, for the surface form *wound* in English, if the gold standard has `{wound<n><sg>, wind<vblex><pp>, wind<vblex><past>, wound<vblex><inf>, wound<vblex><pres>}` and the output of the morphological analyser is `{wound<n><sg>, wound<vblex><inf>, wound<vblex><pres>}` then the recall will be $\frac{3}{3+2} = \frac{3}{5} = 0.6$.

Marathi. Compound nouns were also quite common; many of these were loanwords from Sanskrit, a language with significant compounding, or from Perso-Arabic. For proper nouns, the missing analyses are easily explainable by the fact that many foreign names are not part of the lexicon, and by the fact that we have not yet split proper names into multiple paradigms: we assume similar inflectional paradigms for all of them.

Amongst the incorrect analyses, a few were due to our duplication of the -त *-t* suffix; we treated it as both the locative case, and as a postpositional suffix. Both have exactly the same semantic meaning, and treating it as a case suffix is largely due to convention and dialectal differences. Most other errors were ambiguities between verbs and nouns. There were also several errors with distinctions between verbal adverbs and postpositional adverbs; this distinction was removed in subsequent revisions of the analyser.

7 Future work

Most of our future work will involve expanding the size of the lexicon to improve coverage. Two specific domains, however, would make for interesting future expansions:

7.1 Apertium

By modifying the morphology standards to fit the entire Indo-Aryan language family, creating rule-based machine translation systems across Indian languages would be an interesting future project. The syntactic differences between Indo-Aryan languages are relatively more minor than the morphological differences, which make them very suitable to Apertium's chunking-based transfer system.

7.2 Universal Dependencies

The Universal Dependencies project (Nivre, 2015) is a collection of dependency parsed treebanks in several languages. The ConLL-U format used by UD contains, along with dependency labelling, fields with morphological analyses of each word. Using Apertium's morphological analyser, along with a script to automatically convert Apertium-style tags to UD-style tags, would simplify the process of creating a Marathi treebank.

8 Conclusions

We have presented, to our knowledge, the first free/open-source morphological descriptions for Marathi. The analyser has reasonable coverage over two available test corpora, and the precision is high (over 0.97).

Acknowledgements

We would like to thank the anonymous reviewers for their insight and helpful comments. We would also like to thank Memduh Gökırmak for inspiring discussion. The first author of this paper is funded by a stipend from the Erasmus Mundus Language and Communication Technology programme.

Column					
1	2	3	4	5	6
`boWatapaNe`		`AVY`	`avy`		
`wAra`		`basa`	`v`		
`pusakata`		`hAwa`	`adj`		
`tAMgava`		`basava`	`v`		
`anuBavI`		`AVY`	`avy`		
`meNekarI`		`BikArI`	`nm`		

Figure 3: A random sample of entries in WX notation from the LTRC Marathi lexicon. The first column is the stem, the third column is the inflection paradigm name, and the fourth column is the part of speech (`avy` *uninflected*, `v` *verb*, `adj` *adjective*, `nm` *masculine noun*). The second, fifth and sixth columns are not relevant to our work.

References

Bapat, M., Gune, H., and Bhattacharyya, P. (2010). A paradigm-based finite state morphological analyzer for Marathi. In *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP)*, pages 26–34.

Dabre, R., Amberkar, A., and Bhattacharyya, P. (2012). Morphological analyzer for affix stacking languages: A case study of Marathi. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, pages 225–234.

Dhongade, R. and Wali, K. (2009). *Marathi*. London Oriental and African language library. John Benjamins Publishing Company.

Dixit, V., Dethe, S., and Joshi, R. K. (2005). Design and implementation of a morphology-based spellchecker for Marathi, an Indian language. *Archives of Control Sciences*, 15:301–308.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Gawade, P., Madhavi, D., Gaikwad, J., Jadhav, S., and Ambekar, R. (2013). Morphological analyzer for Marathi using NLP. *International Journal of Engineering Research and Applications*, 3(2).

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 29–32. Association for Computational Linguistics.

Karimi-Doostan, G. (2005). Light verbs and structural case. *Lingua*, 115(12):1737–1756.

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T. A., and Silfverberg, M. (2011). Hfst—framework for compiling and applying morphologies. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 67–85. Springer.

Masica, C. (1993). *The Indo-Aryan Languages*. Cambridge Language Surveys. Cambridge University Press.

Nivre, J. (2015). Towards a universal grammar for natural language processing. *Computational Linguistics and Intelligent Text Processing*, 9014:3–16.

Pandharipande, R. (2003). Marathi. In Cardona, G. and Jain, D., editors, *The Indo-Aryan Languages*, pages 698–728. Routledge, Abingdon.

Ravishankar, V. (2017). Finite-State Back-Transliteration for Marathi. *The Prague Bulletin of Mathematical Linguistics*, 108(1).