Finite-state morphological analysis for Marathi

Vinit Ravishankar [1] Francis M. Tyers [2]

[1] University of Malta, Malta

[2] Higher School of Economics, Moscow

Motivation

- No decent, wide coverage, open-source analyser for Marathi (that we know of)

Motivation

- No decent, wide coverage, open-source analyser for Marathi (that we know of)
- We wanted to enable machine translation for a language with traditionally rubbish MT

Motivation

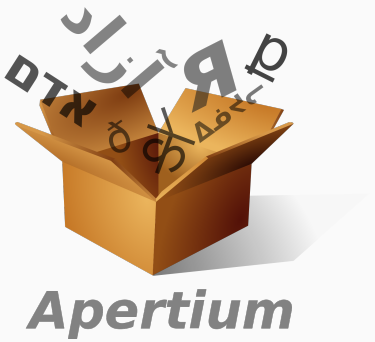
- No decent, wide coverage, open-source analyser for Marathi (that we know of)
- We wanted to enable machine translation for a language with traditionally rubbish MT
- Who doesn’t like morphological analysers?

Overview

1. Apertium
2. Marathi
3. Approach
4. Evaluation

Apertium

Apertium

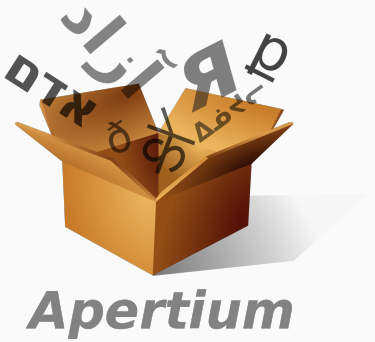


- Rule-based machine translation pipeline
- 46 translation pairs
- Chunking with XML rules that translate to inhumane `sed` chains
- Modular: you can use *hfst* instead of *lttoolbox* for morph analysis

---

`#apertium` on irc.freenode.net

Apertium

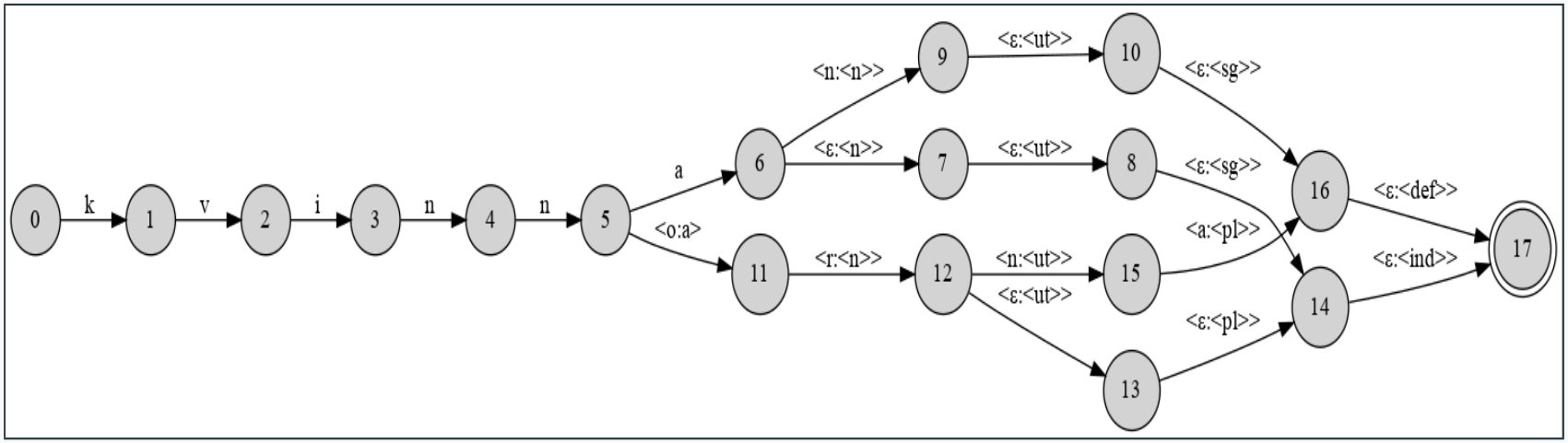


- Rule-based machine translation pipeline
- 46 translation pairs
- Chunking with XML rules that translate to inhumane `sed` chains
- Modular: you can use *hfst* instead of *lttoolbox* for morph analysis (but we didn't)

`#apertium` on irc.freenode.net

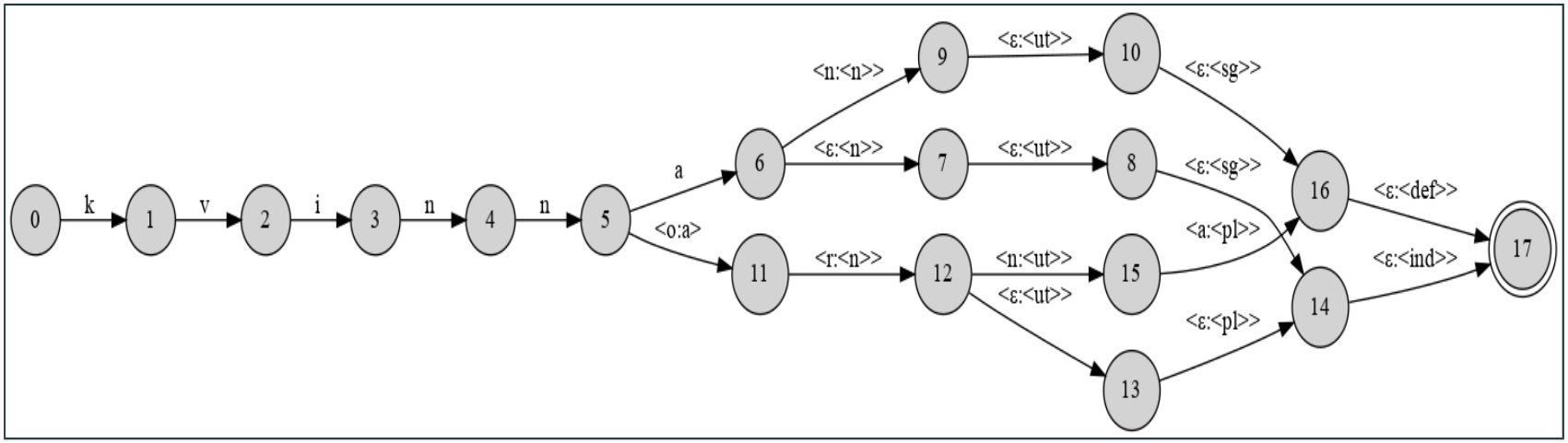
lttoolbox (in theory)

Apertium's pet morphological analyser



lttoolbox (in theory)

Apertium's pet morphological analyser



1. Define paradigms in XML
2. Add entries
3. Goto (1)

lttoolbox (realistically)

```
<pardef n="मुल/गा__n_m">
<e><p><l>गा</l><r>गा<s n="n"/><s n="m"/><s n="sg"/><s n="
    nom"/></r></p><par n="__emph"/></e>
<e><p><l>◌ा</l><r>गा<s n="n"/><s n="m"/><s n="sg"/><s n="
    obl"/><j/></r></p><par n="__clt"/></e>
<e><p><l>◌ा</l><r>गा<s n="n"/><s n="m"/><s n="sg"/></r></p
    ><par n="__sg_case"/></e>
<e><p><l>◌̇</l><r>गा<s n="n"/><s n="m"/><s n="pl"/><s n="
    nom"/></r></p><par n="__emph"/></e>
<e><p><l>◌ा</l><r>गा<s n="n"/><s n="m"/><s n="pl"/><s n="
    obl"/><j/></r></p><par n="__clt"/></e>
<e><p><l>◌ां</l><r>गा<s n="n"/><s n="m"/><s n="pl"/><s n="
    obl"/><j/></r></p><par n="__pl_case"/></e>
</pardef>
```

```
<pardef n="kvinn/a__n_ut">
<e><p><l>a</l> <r>a<s n="n"/><s n="ut"/><s n="sg"/><s n="
    ind"/></r></p></e>
<e><p><l>an</l> <r>a<s n="n"/><s n="ut"/><s n="sg"/><s n="
    ind"/></r></p></e>
<e><p><l>or</l> <r>a<s n="n"/><s n="ut"/><s n="pl"/><s n="
    def"/></r></p></e>
<e><p><l>orna</l> <r>a<s n="n"/><s n="ut"/><s n="pl"/><s n
    ="def"/></r></p></e>
</pardef>

<e lm="kvinna"> <i>kvinn</i><par n="kvinn/a__n_ut"/></e>
```

```
$ echo "kvinnorna" | lt-proc swe.automorf.bin
^kvinnorna/kvinna<n><ut><pl><def>/kvinna<n><ut><pl><def
    ><compound-R>$
```

Marathi

Marathi

- /məɾaʈʰi:/ ; Indo-European language
- Spoken mainly in Maharashtra (cities == Mumbai, Pune)
- ~71m native speakers, ~82% literate
- Script - Devanagari / देवनागरी + extra chars

Morphology

Partially fusional:

Morphology

Partially fusional:

- Paradigmatic stem alterations for non-nominative nouns:
  mulgā → *mulā* ‘boy’ ; *maitriṇ* → *maitriṇī* ‘girlfriend’
- Merged aspect and gender:
  basto`<vblex><impf><p3><m><sg>` ‘he sits’
  baslā`<vblex><perf><p3><m><sg>` ‘he sat’

Morphology

Partially agglutinative:

Morphology

Partially agglutinative:

- Cases and postpositions agglutinate with the oblique:

  mulgā → *mulā* → *mulā-lā*$^{\text{DAT}}$

- Verbs can get longish:

  bas le lyā ṃ samor ūn ac
  sit PERF OBL PL in front of ABL PART

  ‘From ahead of the people that were sitting (and from nowhere else)’

‘Layered’ morphology

Three potential layers for nouns (and gerunds)

1. Base noun/gerund:

   mulgā$^{\text{NOM}}$ ‘boy’

2. (Optional) case suffixes on the oblique:

   mulā$^{\text{OBL}}$*-lā*$^{\text{DAT}}$ ‘to the boy’

3. Postpositions:

   mulā$^{\text{OBL}}$*-cyā*$^{\text{GEN}}$*-samor*$^{\text{POST}}$ ‘in front of the boy’

Cases and postpositions

No clear distinction between the two, but several ‘tests’:

	Cases	Postpositions
Attachment	Oblique noun	Oblique noun/genitive
Free morpheme	Impossible	Possible
Inflectable	Yes	No

Cases and postpositions

No clear distinction between the two, but several ‘tests’:

	Cases	Postpositions
Attachment	Oblique noun	Oblique noun/genitive
Free morpheme	Impossible	Possible
Inflectable	Yes	No
Tradition	Ugh	Ugh

Nouns

Quirks

- All tests have exceptions
- Pronouns are ugly/fusional
- Four layers can exist:
  mulā$^{\text{OBL}}$*-cyā*$^{\text{GEN}}$*-var*$^{\text{POST}}$*-ūn*$^{\text{ABL}}$ ‘from above the boy’
- Cases can/have to be pushed to layer 3 (sometimes):
  mājhyā$^{\text{OBL.GEN}}$*-hūn*$^{\text{ABL}}$ ‘from me’

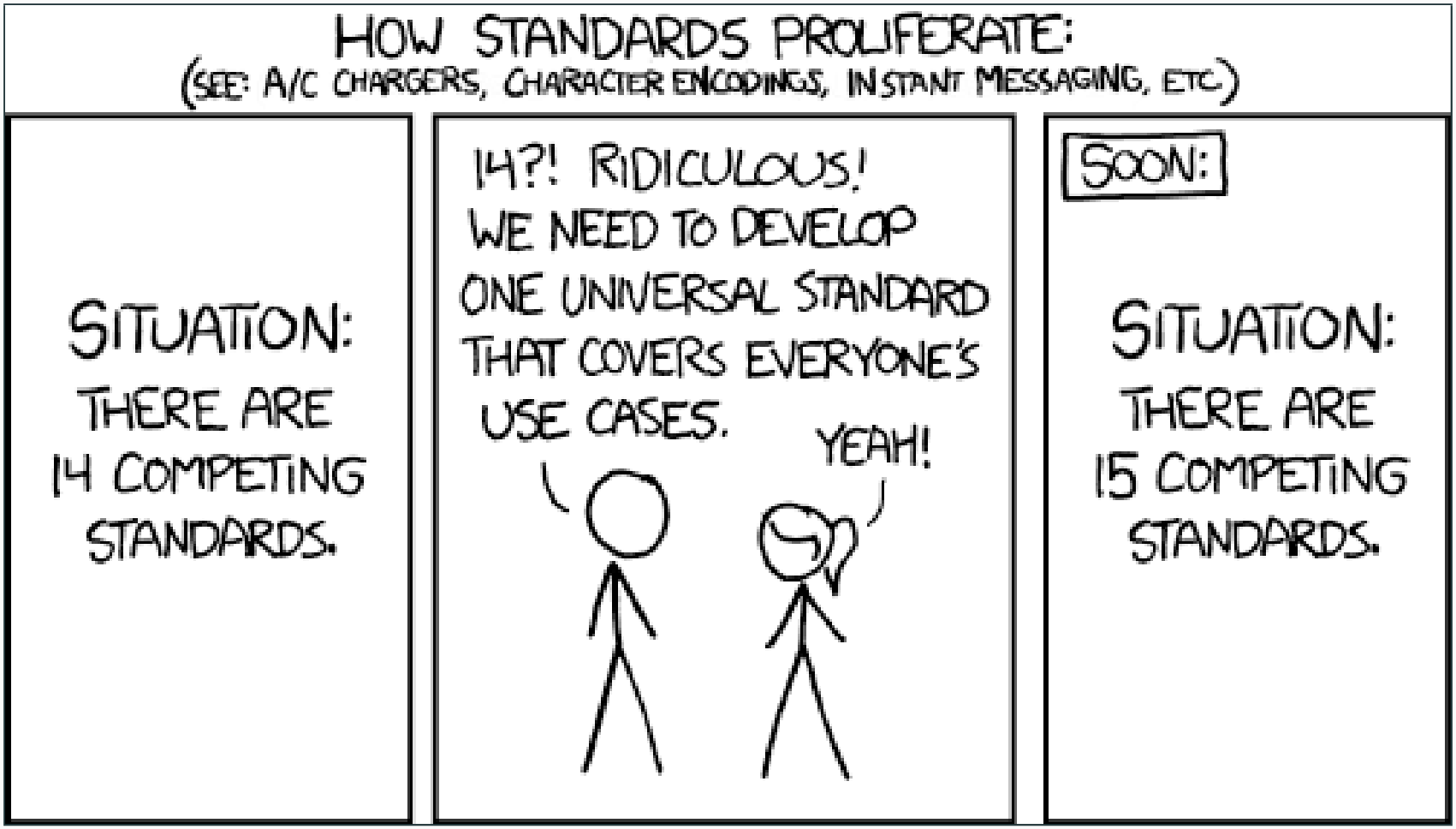
Verbs

- Multiple ‘participles’ with very specific uses (-at, -atā, -tānā, -tānṃā, -ṇe, -ū, -ūn, ...)
- Multiple negative verbs
- Nobody agrees on names:

  -ṇār(ā):

  `<pros>` (Wali) = `<ptcp><fut>` (Navalkar) = `<agt>` (Masica) = ☹ (us)
- Can’t rip better-resourced languages like Hindi off
- Dire need for standards

Verbs


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Approach

Lexicon

- (Partially) scraped from an analyser created at the International Institute of Information Technology, Hyderabad
- Modified to fit a consistent paradigm system
- Verbal paradigms created, nominals significantly extended - hard-coded postpositions replaced with joins
- Manually adding entries to improve coverage

Lexicon

- (Partially) scraped from an analyser created at the International Institute of Information Technology, Hyderabad
- Modified to fit a consistent paradigm system
- Verbal paradigms created, nominals significantly extended - hard-coded postpositions replaced with joins
- Manually adding entries to improve coverage
- Caffeine

Executive decisions

Grey areas!

- Non-genitive case information is on the root morpheme:
  ^मुलाला/मुलगा<n><m><sg><dat>$ ‘to the boy’
- Genitive cases and postpositions are treated as separate morphemes:
  ^मुलासमोर/मुलगा<n><m><sg><obl>+समोर<post><adv>$ ‘in front of the boy’
- We pretend pronouns are more agglutinative than they are:
  ^माझा/मी<prn><p1><mf><sg>+चा<gen><m><sg>$ ‘my (single mularculine thing)’

Executive decisions

- Treat causativity as non-productive (uh oh!)
- Ignore a tonne of derivational morphology; upset people
- Rely on intuition to separate loanwords from foreign words
- Light verbs - fake words are now real words, eg. *māhit asṇe* ‘to know’

Evaluation

Precision

Gold standard:

^wound/wound<n><sg>/wind<vblex><pp>/wind<vblex><past>
/wound<vblex><inf>/wound<vblex><pres>$

Analyser:

^wound/wound<n><sg>/wound<vblex><inf>/wound<vblex><pres>
/wound<adj>$

$$P = \frac{3}{3+1} = 0.75$$

Recall

Gold standard:

^wound/wound<n><sg>/wind<vblex><pp>/wind<vblex><past>
/wound<vblex><inf>/wound<vblex><pres>$

Analyser:

^wound/wound<n><sg>/wound<vblex><inf>/wound<vblex><pres>
/wound<adj>$

$$R = \frac{3}{3+2} = 0.6$$

Results

	Tokens	Coverage	Mean ambig.
Wikipedia	4.0M	80.2	1.7
Bible	751K	80.7	1.9

Results

	Tokens	Coverage	Mean ambig.
Wikipedia	4.0M	80.2	1.7
Bible	751K	80.7	1.9

	Precision	Recall
Known tokens	0.97	0.97
All tokens	0.97	0.71

Future

- Bring coverage up to the 90s
- Write extensive documentation and guidelines that can apply to other Indic languages
- Use the analyser on a Universal Dependencies treebank; evaluate
- Create a proof-of-concept translation pair with moderate lexical coverage

Future

- Bring coverage up to the 90s
- Write extensive documentation and guidelines that can apply to other Indic languages
- Use the analyser on a Universal Dependencies treebank; evaluate
- Create a proof-of-concept translation pair with moderate lexical coverage (Marathi - Crimean Tatar???)

Future

- Bring coverage up to the 90s
- Write extensive documentation and guidelines that can apply to other Indic languages
- Use the analyser on a Universal Dependencies treebank; evaluate
- Create a proof-of-concept translation pair with moderate lexical coverage ~~(Marathi - Crimean Tatar???)~~ (Marathi - Gujarati)

Thanks!