

CS 655: Analyzing Sequences

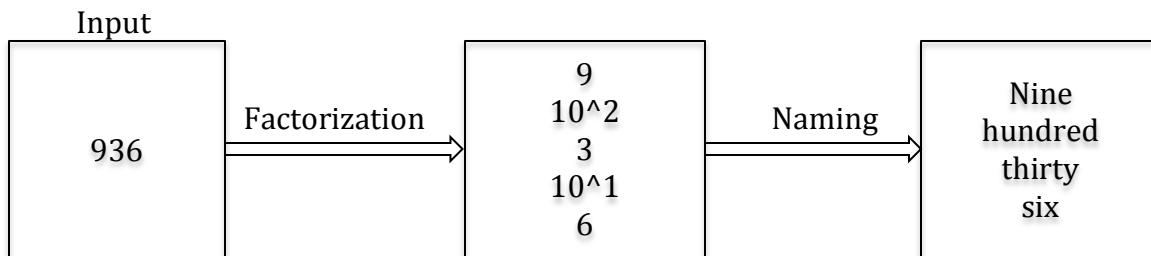
Homework 2

Part 1: Number-name generator

For this part I generated number names in three languages: English, Hindi and Telugu (South Indian language).

English:

I used OpenFST to generate number names upto 9,99,999 (nine lakh ninety nine thousand nine hundred ninety nine). In order to do this I first split the number to its powers of 10 and then converted the numbers and powers to their corresponding names. The final number name is generated by composing the input with the factorization transducer and then with the naming transducer.



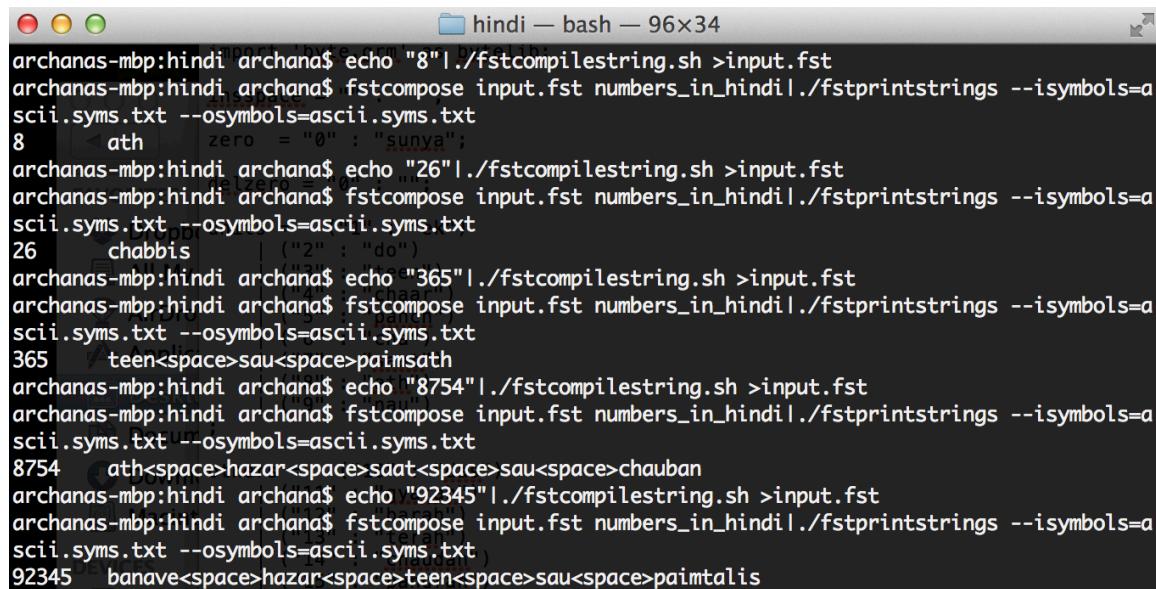
After implementing the number name generator, when I added spaces between the names, it worked fine when there are no zeros, but when there are zeros it resulted in some extra <space>s. Removing the extra spaces by tracking the nodes was becoming tedious. By this time I had implemented the other two parts in Thrax and realized that it is much simpler to do it in Thrax. So I shifted to Thrax for proper implementation with spaces.

```
Lakh_withspace — bash — 96x34
archanas-mbp:Lakh_withspace archana$ fstcompose input.fst fact.fst|fstcompose - name.fst|./fstprintstrings
957329 nine<space>lakh<space>fifty<space>seven<space>thousand<space>three<space>hundred<space>twenty<space>nine
archanas-mbp:Lakh_withspace archana$ fstcompile --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt --acceptor --keep_isymbols --keep_osymbols input.fst.txt >input.fst
archanas-mbp:Lakh_withspace archana$ fstcompose input.fst fact.fst|fstcompose - name.fst|./fstprintstrings
462 four<space>hundred<space>sixty<space>two
archanas-mbp:Lakh_withspace archana$
```

Hindi:

The Hindi number-naming system is quite different from English for numbers from 21 to 99. In Hindi there is a separate name for each number from 21 to 99, with not much of an obvious pattern like English. I implemented the number-name generator for Hindi in two ways, one using the full name directly (hindi.grm) and the other by grouping the similar strings at the end of the word (hindi2.grm).

30	- Tis	40	- Chalis
31	- Ikatis	41	- Ikatalis
32	- Battis	42	- Bayalis
33	- Taimtis	43	- Taimtalis
34	- Chaumtis	44	- Chaumtalis
35	- Paimtis	45	- Paimtalis
36	- Chattis	46	- Chiyalis
37	- Saimtis	47	- Saimtalis
38	- Aratis	48	- Aratalis
39	- unchalis	49	- unachas

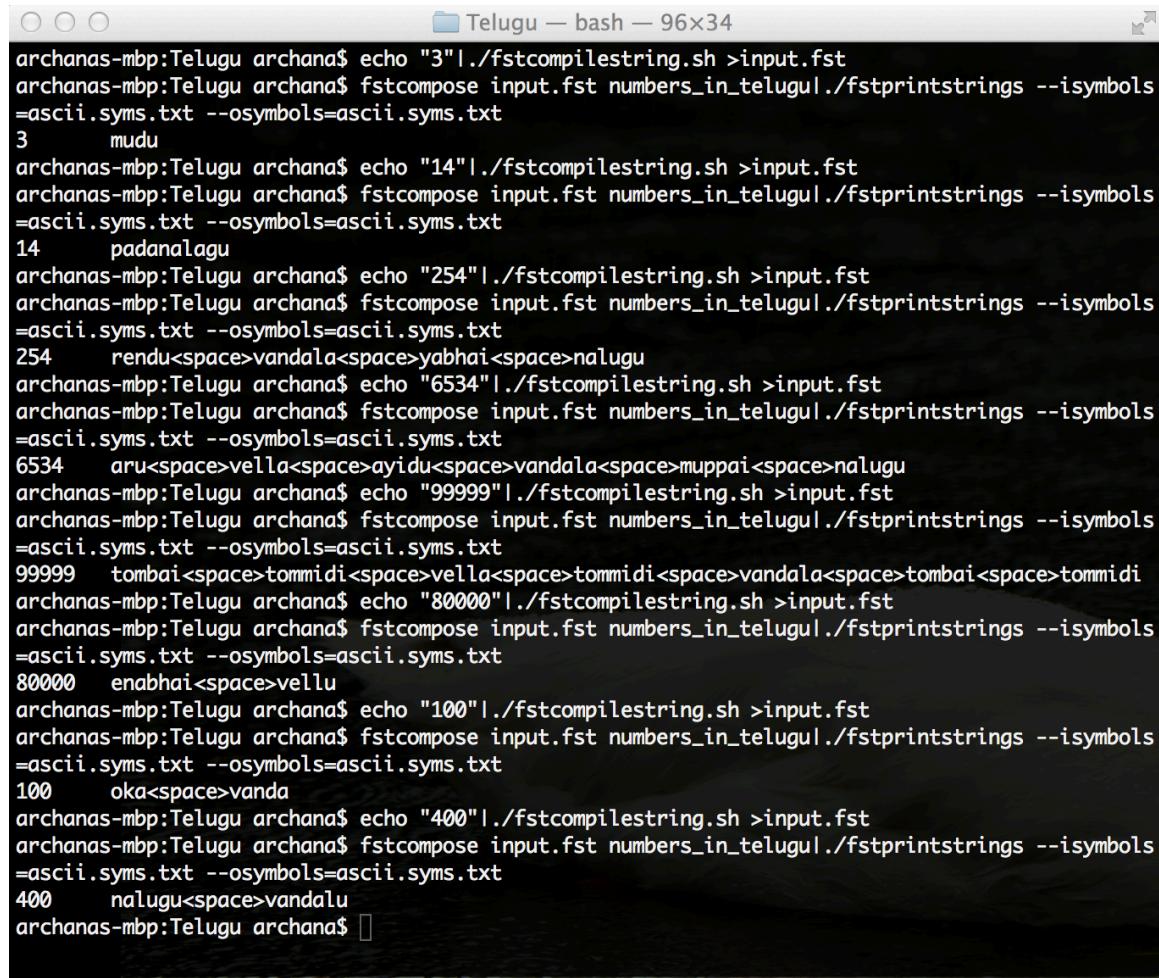


```
hindi — bash — 96x34
archanas-mbp:hindi archana$ echo "8" | ./fstcompilestring.sh >input.fst
archanas-mbp:hindi archana$ fstcompose input.fst numbers_in_hindi | ./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
8      ath      zero = "0" : "sunya";
archanas-mbp:hindi archana$ echo "26" | ./fstcompilestring.sh >input.fst
archanas-mbp:hindi archana$ fstcompose input.fst numbers_in_hindi | ./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
26      chabis   ("2" : "do")
archanas-mbp:hindi archana$ echo "365" | ./fstcompilestring.sh >input.fst
archanas-mbp:hindi archana$ fstcompose input.fst numbers_in_hindi | ./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
365     teen<space>sau<space>painsath
archanas-mbp:hindi archana$ echo "8754" | ./fstcompilestring.sh >input.fst
archanas-mbp:hindi archana$ fstcompose input.fst numbers_in_hindi | ./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
8754    ath<space>hazar<space>saat<space>sau<space>chauban
archanas-mbp:hindi archana$ echo "92345" | ./fstcompilestring.sh >input.fst
archanas-mbp:hindi archana$ fstcompose input.fst numbers_in_hindi | ./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
92345   banave<space>hazar<space>teen<space>sau<space>paintalis
```

Telugu:

The Telugu number-naming system is very similar to English, but differs in the way hundreds and thousands are named. They vary based on number present in the hundreds or thousands place and also on the numbers following them. This is illustrated below:

- 100 - oka **vanda**
- 200 - rendu **vandalu**
- 231 - rendu **vandala** muppai oka
- 1000 - oka **veyyi**
- 2000 - rendu **vellu**
- 2400 - rendu **vella** nalugu vandalu



```
archanas-mbp:Telugu archana$ echo "3" | ./fstcompilestring.sh >input.fst
archanas-mbp:Telugu archana$ fstcompose input.fst numbers_in_telugul./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
3      mudu
archanas-mbp:Telugu archana$ echo "14" | ./fstcompilestring.sh >input.fst
archanas-mbp:Telugu archana$ fstcompose input.fst numbers_in_telugul./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
14      padanalagu
archanas-mbp:Telugu archana$ echo "254" | ./fstcompilestring.sh >input.fst
archanas-mbp:Telugu archana$ fstcompose input.fst numbers_in_telugul./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
254     rendu<space>vandala<space>yabhai<space>nalugu
archanas-mbp:Telugu archana$ echo "6534" | ./fstcompilestring.sh >input.fst
archanas-mbp:Telugu archana$ fstcompose input.fst numbers_in_telugul./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
6534    aru<space>vella<space>ayidu<space>vandala<space>muppai<space>nalugu
archanas-mbp:Telugu archana$ echo "99999" | ./fstcompilestring.sh >input.fst
archanas-mbp:Telugu archana$ fstcompose input.fst numbers_in_telugul./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
99999   tombai<space>tommidi<space>vella<space>tommidi<space>vandala<space>tombai<space>tommidi
archanas-mbp:Telugu archana$ echo "80000" | ./fstcompilestring.sh >input.fst
archanas-mbp:Telugu archana$ fstcompose input.fst numbers_in_telugul./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
80000   enabhai<space>vellu
archanas-mbp:Telugu archana$ echo "100" | ./fstcompilestring.sh >input.fst
archanas-mbp:Telugu archana$ fstcompose input.fst numbers_in_telugul./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
100     oka<space>vanda
archanas-mbp:Telugu archana$ echo "400" | ./fstcompilestring.sh >input.fst
archanas-mbp:Telugu archana$ fstcompose input.fst numbers_in_telugul./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
400     nalugu<space>vandalu
archanas-mbp:Telugu archana$ ]
```

Part 2: ROT-13

The decoded sentence is

Her mother only scolded her for being nonsensical.

In order to decode the sentence, the FST of the input sentence was first composed with the ROT-13 FST, which deciphered the ROT-13-encoded message. It was then composed with the scrambling FST to produce sentences with all possible combination of vowels. This FST of sentences is in turn composed with `pride_and_prejudice_sentences.fst` to find a match to produce the final decoded sentence.

```
archanas-mbp:thr archana$ echo "Ule zhguve nayl fpryqbq ube she ovrat ahafvafnpry." | ./fstcompilestring.sh >input.fst
archanas-mbp:thr archana$ fstcompose input.fst rot13|fstcompose - vowel_scramble|fstcompose - pride_and_prejudice_sentences.fst |./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
Ule<space>zhguve<space>nayl<space>fpryqbq<space>ube<space>she<space>ovrat<space>ahafvafnpry.
Her<space>mother<space>only<space>scolded<space>her<space>for<space>being<space>nonsensical.
archanas-mbp:thr archana$
archanas-mbp:thr archana$ echo "My name is Archana" | ./fstcompilestring.sh >input.fst
archanas-mbp:thr archana$ fstcompose input.fst rot13|./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
My<space>name<space>is<space>Archana    Zl<space>anrz<space>vf<space>Nepunan
archanas-mbp:thr archana$ fstcompose input.fst rot13>output.fst
archanas-mbp:thr archana$ fstcompose output.fst rot13|./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
My<space>name<space>is<space>Archana    My<space>name<space>is<space>Archana
archanas-mbp:thr archana$
archanas-mbp:thr archana$ echo "is" | ./fstcompilestring.sh >input.fst
archanas-mbp:thr archana$ fstcompose input.fst vowel_scramble|./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
is      as
is      es
is      os
is      us
is      ys
archanas-mbp:thr archana$ fstcompose input.fst vowel_scramble|fstcompose - rot13|./fstprintstrings --isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
is      nf
is      rf
is      bf
is      hf
is      lf
archanas-mbp:thr archana$ 
```

Part 3: T9 Text Entry

A transducer is defined to replace a number with its corresponding letters as per the telephone keypad code ("2" with "a", "b" or "c"; "3" with "d", "e" or "f", and so on). For a given group of numbers, composing with the above-defined FST gives all possible combinations of the letters that the numbers represent. Composing these combinations of letters with a wordlist containing valid English words, yields all possible English-language words those digits could represent.

The predictor can be made more accurate by decreasing the weight of the words that occur more commonly, and increasing the weight of words that occur rarely. For example when the user types '843', the possible outputs are *the*, *tid*, *tie* and *vie* as shown below. But most of the time the user would end up using it for the word *the*. Decreasing the weight for path *the* will increase the chance of predicting '843' as *the*.

```
○ ○ ○ thrax1 — bash — 99x33
archanas-mbp:thrax1 archana$ echo "843" | ./fstcompilestring.sh > input.fst
archanas-mbp:thrax1 archana$ fstcompose input.fst output | fstcompose - wordlist | ./fstprintstrings -isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
843    the
843    tid
843    tie
843    vie
archanas-mbp:thrax1 archana$ echo "25277" | ./fstcompilestring.sh > input.fst
archanas-mbp:thrax1 archana$ fstcompose input.fst output | fstcompose - wordlist | ./fstprintstrings -isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
25277  clasp
25277  class
archanas-mbp:thrax1 archana$ echo "47" | ./fstcompilestring.sh > input.fst
archanas-mbp:thrax1 archana$ fstcompose input.fst output | fstcompose - wordlist | ./fstprintstrings -isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
47    is
archanas-mbp:thrax1 archana$ echo "8379" | ./fstcompilestring.sh > input.fst
archanas-mbp:thrax1 archana$ fstcompose input.fst output | fstcompose - wordlist | ./fstprintstrings -isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
8379   very
archanas-mbp:thrax1 archana$ echo "46837378464" | ./fstcompilestring.sh > input.fst
archanas-mbp:thrax1 archana$ fstcompose input.fst output | fstcompose - wordlist | ./fstprintstrings -isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
46837378464   interesting
archanas-mbp:thrax1 archana$ echo "2774466368" | ./fstcompilestring.sh > input.fst
archanas-mbp:thrax1 archana$ fstcompose input.fst output | fstcompose - wordlist | ./fstprintstrings -isymbols=ascii.syms.txt --osymbols=ascii.syms.txt
2774466368   assignment
```