LING 5801 – Introduction to Computational Linguistics (S20, Reese)

Assignment 2: Finite-state computational morphology with XFST

Due March 4, 2020

For this assignment you will create a morphological parser for Finnish noun inflection by writing XFST regular expressions in the file `finnish.xfst`. Do not change the name of this file. Submit your solutions via Canvas.

Remember that you execute the commands in an XFST script file using the `source` command at the XFST prompt:

```
xfst[0]: source <filename>
```

If you have any problems, get in touch with me right away. See the links on the Moodle Web page for additional information about XFST or contact me with XFST related questions. If any of the instructions or descriptions of the problem do not make sense to you, please get in touch with me right away. **Collaboration is encouraged, but everyone should submit their own solutions.**

Finnish Noun Inflection

This exercise is a simplified version of the Finnish Noun Inflection problem on the Beesley & Karttunen book web site.[1] The goal is to build a lexicon that contains Finnish nouns inflected for number and case. To keep the problem reasonably small, we consider only three cases: **Singular Nominative**, **Singular Partitive** and **Plural Partitive**. For the same reason, we consider only monosyllabic and bisyllabic noun stems. Even with these limitations, the problem is challenging because the shape of the different endings depends on the shape of the stem and stems are subject to several regular alternations.

The Facts

A Finnish noun begins with a stem. In all of the cases below, assume that the stem is identical with the nominative singular. A plural marker, if any, immediately follows the stem. After the stem and the possible plural marker comes one of several possible case endings. We consider only two cases: Nominative and Partitive. Singular nominative has no case marker. For the purpose of this exercise we, assume that the plural and the Partitive endings are:

- *I* – Plural marker for cases other than Nominative, realized as *i* or *j*.
- *Ta* – Partitive Marker marker, realized as *ta* or *a* in words containing only back vowels.

The list below illustrates some of the possible combinations with the noun *valo* 'light'.

- *valo* – Nominative Singular. No overt case ending
- *valoa* – Partitive Singular. The partitive marker *Ta* is realized here as *a*
- *valoja* – Partitive Plural. The plural marker *I* is realized as *j* here. The partitive marker *Ta* is realized here as *a*.

[1] The problem is adapted from Lauri Karttunen's LSA course; the course website can be found at `http://www.stanford.edu/~laurik/fsmbook/LSA-207/index.html`.

The realization of the *T* in the Partitive marker depends on the syllable structure of the word. After a monosyllabic stem such as *maa* 'earth', the *T* is realized as *t* as in *maata*. After a bisyllabic stem such as *valo* 'light', the *T* disappears as in *valoa*.

The plural marker *I* is realized as *j* between two vowels, otherwise it is realized as *i*.

Using this information you should be able to write the rules that correctly realize the plural marker and the partitive suffix in all environments.

The remaining problem is that the noun stem also undergoes alternations. We will use only back-vowel stems in order to avoid having to deal with vowel harmony. The other stem alternations, Vowel Rounding, Vowel Lowering, Vowel Dropping, and Vowel Shortening are illustrated in the table below.

Stem (Nom Sg)	Gloss	Part Sg	Part Pl
puu	tree	puuta	puita
maa	earth	maata	maita
suo	swamp	suota	soita
tikka	dart	tikkaa	tikkoja
pappi	priest	pappia	pappeja
kukka	flower	kukkaa	kukkia
tutti	pacifier	tuttia	tutteja
kauppa	shop	kauppaa	kauppoja
kuoppa	hole	kuoppaa	kuoppia
jalka	foot	jalkaa	jalkoja
linko	sling	linkoa	linkoja
sopu	harmony	sopua	sopuja
kampa	comb	kampaa	kampoja
piispa	bishop	piispaa	piispoja
vahti	guard	vahtia	vahteja
ilta	evening	iltaa	iltoja
sota	war	sotaa	sotia

Vowel Rounding

Short *a* is rounded to *o* in front of the plural marker *I*. Examples: *tikkaa* Partitive Singular, *tikkoja* Partitive Plural, *kauppaa* Partitive Singular, *kauppoja* Partitive Plural. This does not happen if the vowel nucleus of the preceding syllable starts with a rounded vowel (*o* or *u*). See the rule for Vowel Dropping.

Vowel Lowering

Short *i* is lowered to *e* in front of the plural marker *I*. Examples: *vahtia* Partitive Singular, *vahteja* Partitive Plural, *pappia* Partitive Singular, *pappeja* Partive Plural. See the rule for Vowel Dropping.

Vowel Dropping

A short *a* is deleted in front of the plural marker *I* if the nucleus of the preceding syllable consists of, or begins with, a rounded vowel (*u* or *o*). Note the different behavior of *kuoppa* where the *a* is dropped and *kauppa* where the *a* is rounded to *o* in the plural. Examples: *sotaa* Partitive Singular, *sotia* Partitive Plural, *kuoppaa* Partitive Singular, *kuoppia* Partitive Plural.

Vowel Shortening

In front of the plural marker *I*, the long vowels *aa*, *ee*, *ii*, *oo*, *uu*, are shortened to *a*, *e*, *i*, *o*, *u*, respectively. The diphthongs *uo* and *ie* shorten to *o* and *e*, respectively. Examples: *puuta* Partitive Singular, *puita* Partitive Plural, *suota* Partitive Singular, *soita* Partitive Plural.

The Task

Your task is to write an `xfst` script `finnish.xfst` that takes the stems mentioned above, assembles them into morphotactically correct underlying Finnish forms, and then handles the realization of surface forms according to the rules for the plural marker, the partitive endings, and stem-changing sketched above. You must combine the suffix realization rules with the stem alternation rules. Think about how to order the rules. It matters!

The noun stems you must account for are listed above (and also provided in the file `finnish_stems.txt`). Start by creating a FST for the noun stems (hint: you can use the file `finnish_stems.txt` to create an FST for just the noun stems using the `read text` command) concatenated with Number and Case morphemes. Use the tags `+Sg`, `+Pl`, `+Nom`, `+Part` for marking number and case on the lexical side. You should end up with lexical/intermediate pairs like the following:

```
kukka+Sg+Part      jalka+Pl+Part      suo+Pl+Part
kukkaTa            jalkaITa           suoITa
```

(**Tip**: do not forget about the distinction between concatenating single character symbols and a multicharacter symbol. This might be a good place to use the `print sigma` command to check for unintended multi-character symbols.)

Once you have the lexical transducer working properly, define a replacement rules that when composed in a series, lead to correct surface forms. Compose the lexical transducer and the rule transducer and leave the resulting FST on the `xfst` stack for testing. The final result should contain pairs such as:

```
kukka+Sg+Part      jalka+Pl+Part      suo+Pl+Part1
kukkaa             jalkoja            soita
```

Also, do not forget singular nominative forms, e.g. (`sota+Sg+Nom`, `sota`).

Verify that the lower side of the FST contains properly inflected surface forms by terminating the script with the command:

```
print lower-words
```

Make sure you get **all and only the forms given in the table above**.