CSCI 4342 Natural Language Processing
Semester 1 2021/22

Assignment 1: Finite State Transducer for Swahili numbers
(Group Submission)
Due: Sunday, 14/11/2021 (11.59 pm)

Transliteration from Swahili Numbers to English

Transliteration is the process of converting from one text/script to another by swapping or replacing the letters of the source with the target text/script. For example, the Arabic script كتابي (i.e., written) can be converted to the Latin script ***kitabi*** for English or, the Greek word ***Ελευθερία*** when transliterated into English becomes ***Eleutheria***. Other uses of transliteration include mapping between **loanwords** from the language of origin to another. A **loanword** is a word adopted from one language (origin) that is incorporated into another language without being translated such as the word ***aeo-con*** in Korean which is a loanword from the English word **air-cond.**

Transliteration can also refer to the process of spelling out a script within a language to its own in a longer form, just like how they are pronounced. For example, the numeral ***24*** is spelled out as ***ishirini-na-nne*** in Swahili/Kiswahili which means ***twenty-and-four*** in English/Latin. Swahili/Kiswahili is a native language of at least three (3) African countries; Uganda, Kenya and Tanzania. **Table 1** shows examples of transliterations of Swahili numbers to English/Latin

Numeral	Swahili	English		Numeral	Swahili	English
0	sifuri	zero		55	hamsini na tano	fifty and five
1	moja	one		60	sitini	sixty
2	mbili	two		70	sabini	seventy
3	tatu	three		80	thamanini	eighty
4	nne	four		90	tisini	ninety
5	tano	five		100	mia moja	hundred one
6	sita	six		300	mia tatu	hundred three
7	saba	seven		136	mia moja thalathini na sita	hundred one thirty and six
8	nane	eight		999	mia tisa tisini na tisa	hundred nine ninety and nine
9	tisa	nine		1000	elfu moja	thousand one
10	kumi	ten		1997	elfu moja mia tisa tisini na saba	thousand one hundred nine ninety and seven
11	kumi na moja	ten and one		2000	elfu mbili	thousand two
12	kumi na mbili	ten and two		5498	elfu tano mia nne tisini na nane	thousand five hundred four ninety and eight
17	kumi na saba	ten and seven		10000	elfu kumi	thousand ten
20	ishrini	twenty		100000	elfu mia moja/laki	thousand hundred one
25	ishrini na tano	twenty and five		1/2	nusu	half
30	thalathini	thirty		2 1/2	mbili na nusu	two and half
40	arubaini	forty		1/4	robo	quarter
50	hamsini	fifty		47 3/4	arubaini na saba na robo tatu	forty and seven and quarter three

Table 1 Transliterations example for Swahili-English numbers

Based on **Table 1** given above, build an **FST** that receives any given **number in Swahili as input** and **outputs the equivalent transliteration of the number in English**. For example, for the number ***elfu mbili (i.e., 2000)***, *your FST* will output ***thousand two*** in in **English** with ***elfu*** mapping to **thousand** and **mbili** maps to **two**. You may test your program with the following inputs: (your program will also be tested using random inputs).

Input	Output	Numeral	FST status
elfu-mbili	*thousand-two*	*2000*	*accept*
thalathini na saba	*thirty-and-eight*	*38*	*accept*
embili-elfu	*thousand-two*	*2000*	*reject*
mia-sita-sabini-na-tisa	*hundred-six-seventy-and-nine*	*679*	*accept*
mia-tano-hamsini-na-mbili	*hundred-five-forty-and-two*	*542*	*reject*
elfu-nne-mia-moja-ishrini-na-tatu	*thousand-four-hundred-one-forty-and-three*	*4123*	*accept*

Print all ACCEPTED input-output mappings into an output file named **Swahili-trans.dat** in the following format:

elfu-mbili −−> thousand-two
thalathini na saba −−> thirty-and-eight
mia-sita-sabini-na-tisa −−> hundred-six-seventy-and-nine
elfu-nne-mia-moja-ishrini-na-tatu −−> thousand-four-hundred-one-forty-and-three

Figure 1: Example of mappings of Swahili numbers to English/Latin

Note: To use string as input or output symbols, enclose the word in square brackets (not in parentheses). Example: to add an arc that takes the string ***elfu*** as input and returns ***thousand*** when going from state 1 to 2, you should use:

f.add_arc('1','2',['elfu'],['*thousand*'])

Submit the following for your assignment:

i) a modified python FST program (based on the file *recognize-sol2.py* in italeem) for the *Swahili-English* transliteration

ii) an output file that prints the mappings of the transliterations as shown in Figure 1

iii) an FST construction generated by the Python program – *Tkinter* (attach image or paste image on word document)