Natural Language Processing HW 2

Jing Huang 933039277
Chao-Ting Wen 933311010
Yu-Hsiang Liu 933951866

1. Shannon Game and Entropy of English

1.1.

Jing Huang

```
The entropy for this experiment is 2.1708882
```

Chao-Ting Wen

```
The entropy for this experiment is 2.4364908
```

Yu-Hsiang Liu

```
The entropy for this experiment is 2.7866473
```

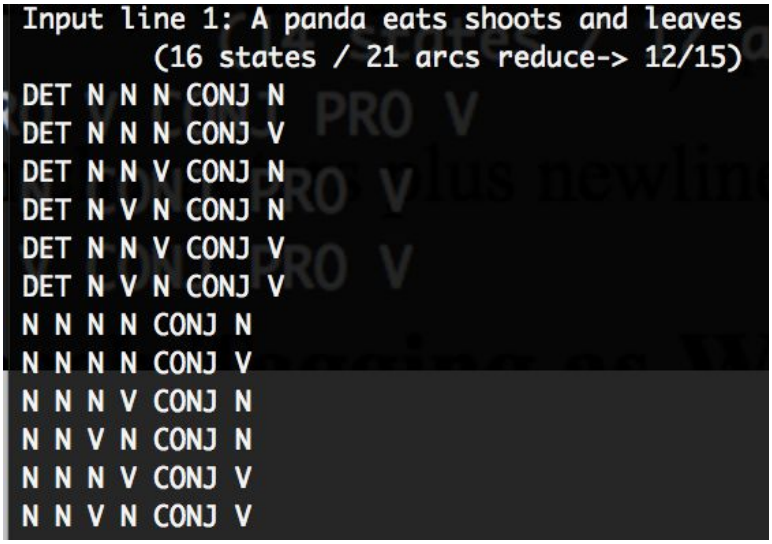
1.2. The entropy was calculated $\sum_{j=1}^{N} P(j)\, log(P(j))$, which N means the number of printable English characters plus newline.

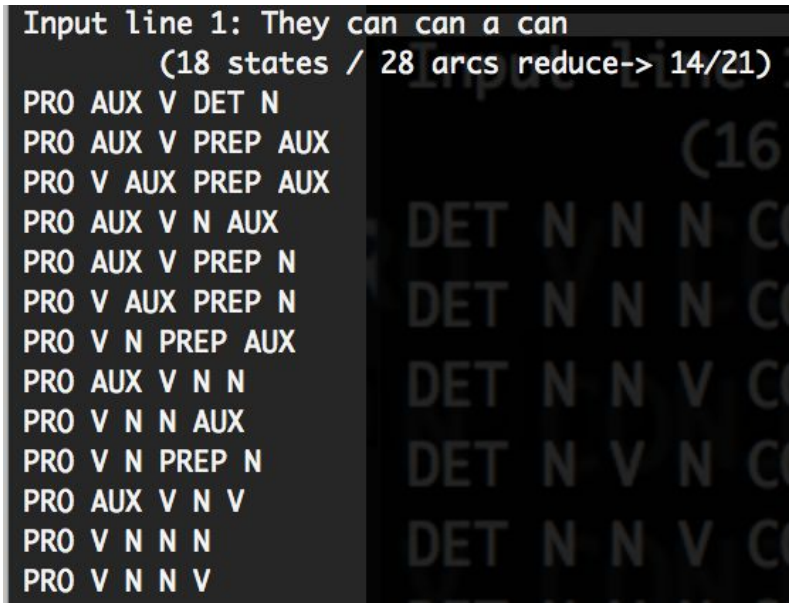
2. Part-of-Speech Tagging as WFST Composition

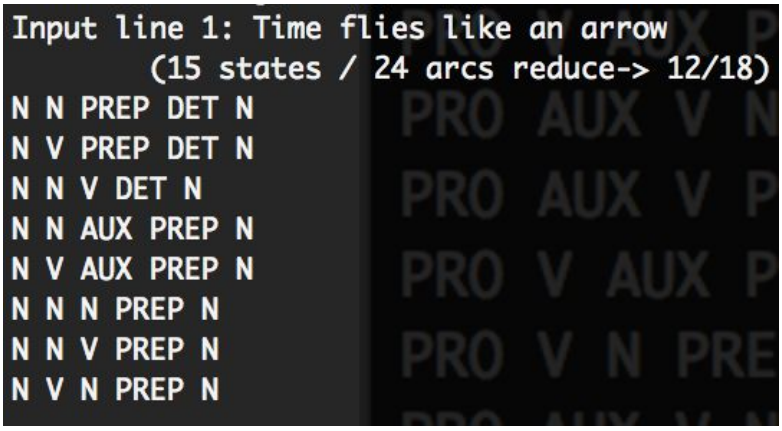
2.1.

```
Input line 1: I hope that this works
        (14 states / 17 arcs reduce-> 9/10)
PRO V CONJ PRO V
N N CONJ PRO V
N V CONJ PRO V
```

```
Input line 1: They can fish
        (9 states / 10 arcs reduce-> 7/7)
PRO AUX V
PRO V N
```







2.2. We set the probability by 1/n, where n is the number of states. For example, “Fish” can be a noun or verb, so the probability of both path is 0.5.

2.3. One of the funny things that we observed is the sentence “A panda eats shoots and leaves” can be tagged as “N N N N CONJ N”.
If we use trigram, the problem can be fixed such as the conjunction should connect the same type of words. For example, verb conjunction verb or noun conjunction noun.

3. Pronouncing and Spelling English

3.1.

```
Input line 1: PANDA
        (209214 states / 209213 arcs reduce-> 9/8)
P AE N D AH 1
0
0
0
0
```

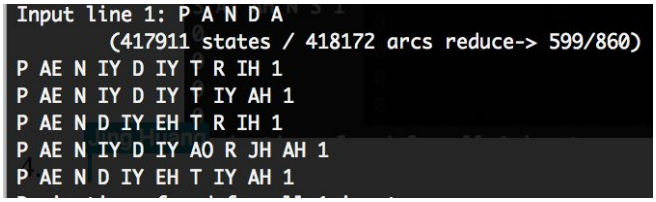
```
Input line 1: WORLD
        (209214 states / 209213 arcs reduce-> 8/7)
W ER L D 1
0
0
0
0
```

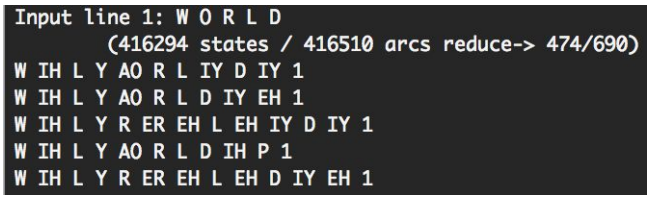
```
Input line 1: MILK
        (209214 states / 209213 arcs reduce-> 8/7)
M IH L K 1
0
0
0
0
```

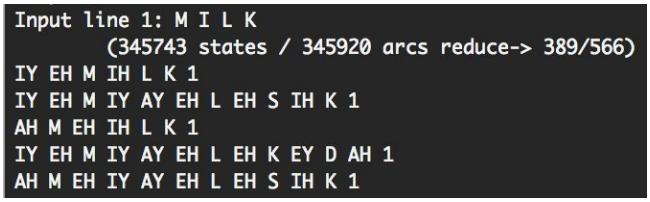
```
Input line 1: COMPUTER
        (209214 states / 209213 arcs reduce-> 12/11)
K AH M P Y UW T ER 1
0
0
0
0
```

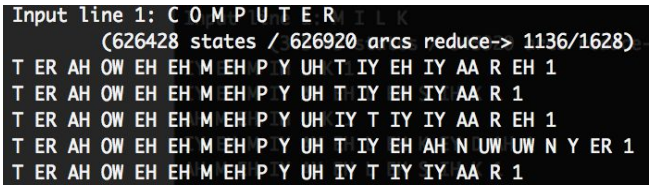
```
Input line 1: SCIENCE
        (209214 states / 209213 arcs reduce-> 9/8)
S AY AH N S 1
0
0
0
0
```

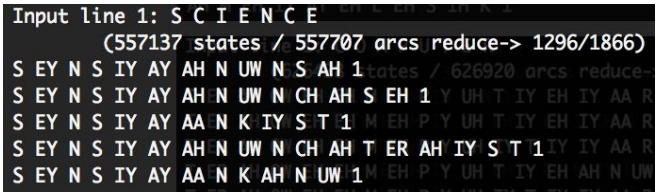
3.2.







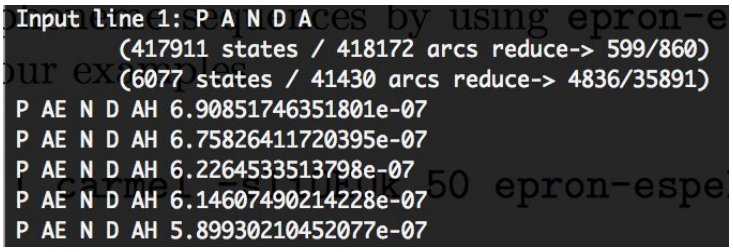




Yes. Because each state will have many different paths, which have different pronunciation.

For example, ‘M’ could be ‘AH M EH’ when we pronounce it as a character, but when it stands before ‘i’, it would be pronounced ‘M’, and ‘IH’ for ‘i’.

3.3.



```
Input line 1: M I L K
        (345743 states / 345920 arcs reduce-> 389/566)
        (3915 states / 26824 arcs reduce-> 3285/23247)
M IH L K 1.64154193781496e-06
M IH L K 6.08434055372412e-07
M IH L K 4.14917220380547e-07
M IH L K 4.59328209664788e-08
M IH K 2.01791355460004e-08
```

```
Input line 1: W O R L D
        (416294 states / 416510 arcs reduce-> 474/690)
        (5234 states / 42652 arcs reduce-> 4536/37844)
W ER L D 1.6408073137713e-06
W ER L D 6.97408038615515e-07
W ER L D 1.38542963076205e-07
W ER L D 5.8886241749423e-08
W UH L D 3.34985621362871e-08
```

```
Input line 1: C O M P U T E R
        (626428 states / 626920 arcs reduce-> 1136/1628)
        (12806 states / 123671 arcs reduce-> 10863/111802)
K AH M P Y UW T ER 9.11358666061392e-09
K AA M P Y UW T ER 7.93379601997658e-09
K AA M P Y UW T ER 6.88402153838876e-09
K AH M P Y UW T ER 5.1812584937802e-09
K AH M P Y UW T ER 4.74499968935245e-09
```

```
Input line 1: S C I E N C E
        (557137 states / 557707 arcs reduce-> 1296/1866)
        (16040 states / 174843 arcs reduce-> 13406/159826)
SH AH N S 1.49179092654124e-08
S AY AH N S 1.24310622879286e-08
S IY EH N S 1.22744838288307e-08
S IY AH N S 8.86299154125618e-09
S CH AH N S 4.32226680451323e-09
```

3.4.

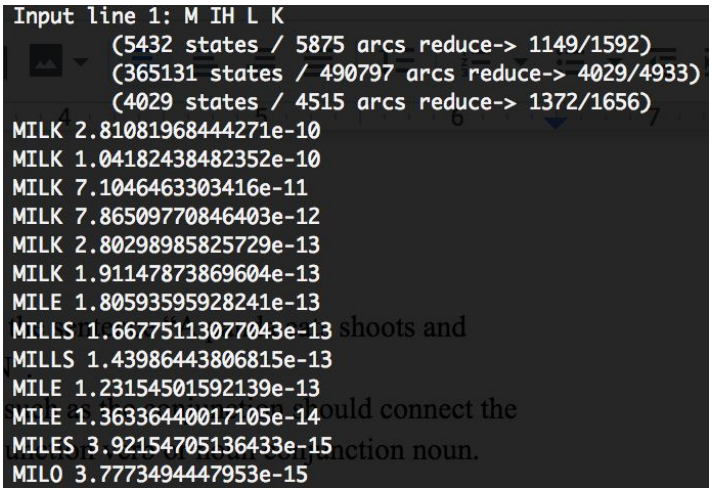
```
Input line 1: P AE N D AH
        (5572 states / 6139 arcs reduce-> 1517/2084)
P A N D U 0.90578738074
P A N D U 0.869418816126019
P A N D E 0.739449170139
P A N D U 0.645467
P A N D U 0.611552775972357
P A N D U 0.586998120960891
P A N D A 0.524832
P A N D A 0.513417428832
P A N D U 0.47729003750211
P A N D A 0.473016386304
P A N D A 0.466910129427181
P A N D A 0.44816308831377
P A N D A 0.448163068503102
P A N D A 0.430168745953176
P A N D A 0.349771574314641
P A N D O 0.238524006577
P A N D E 0.162335
```

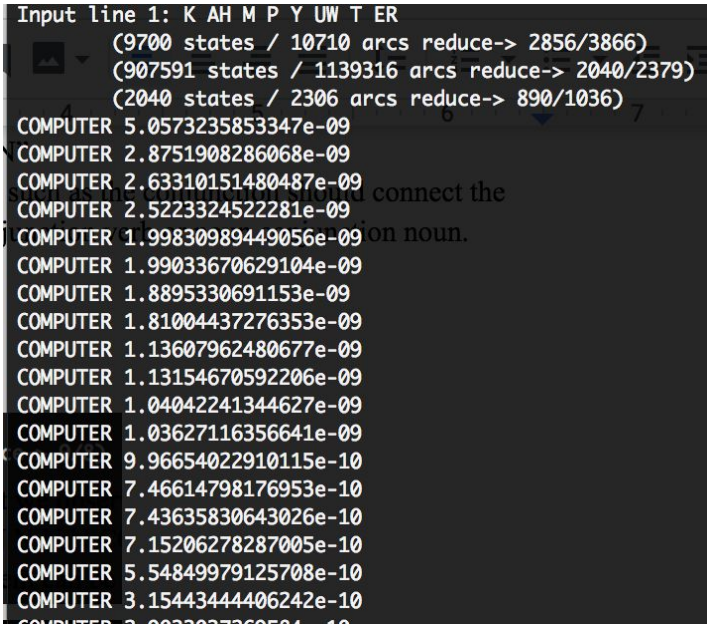
```
Input line 1: K AH M P Y UW T ER
        (9700 states / 10710 arcs reduce-> 2856/3866)
K U M P U T E R 1
C O M P U T E R 0.965622757841
C O M P U T E E R 0.7726948348
C O M P E W T E R 0.665114190009551
K U M P U T E R 0.568520242
C O M P U T E R 0.548976083968473
C U M P U T E R 0.52065148829
K U M P U T E R 0.520651184440792
C O M P U T E R 0.5027526325929
C O M P U T E R 0.481602882950816
C O M P E W T E R 0.477290848717625
K U M P U T E R 0.383843321976
C O M P U T E R 0.381548358292606
C O M P U T E R 0.38002599338
K U M P U T E R 0.373623130344
C O M P U T E R 0.360778997515961
C O M P U I T E R 0.346986342515288
```

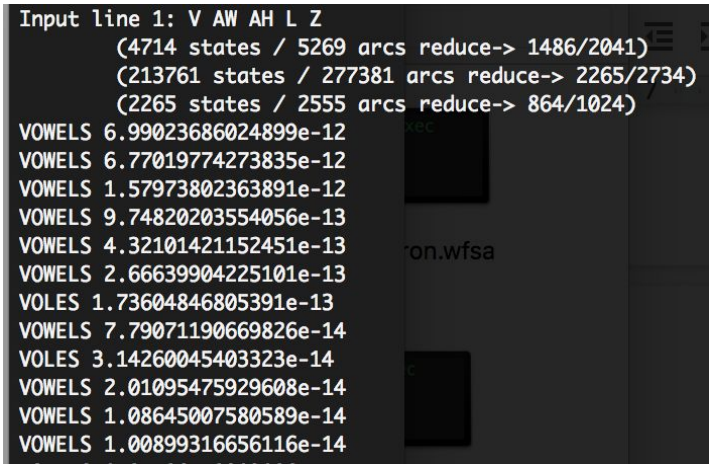
```
Input line 1: K AA M P Y UW T ER
        (9125 states / 9973 arcs reduce-> 2423/3271)
C O M P U T E R 1
C O M P U T E R 0.867683202474
C O M P U T E E R 0.6943232472
C O M P E W T E R 0.59765411048174
C O M P U T E R 0.568520242
C O M P U T E R 0.522215
C O M P U T E R 0.520651184440792
C O M P U T E R 0.493295464249854
C O M P U T E R 0.461343482355126
C O M P U T E R 0.459502737624
C O M P U T E R 0.451760287087467
C O M P U T E R 0.432755678556908
C O M P E W T E R 0.428880998054351
K O M P U T E R 0.384996
C O M P U T E R 0.373623130344
C O M P U T E E R 0.369168728667379
C O M P U T E R 0.324186514255243
C O M P E W T E R 0.317770158264373
C O M P U I T E R 0.311792797387632
C O M P U T E R 0.311480962696472
```

3.5. F

```
Input line 1: W ER L D
        (3925 states / 4275 arcs reduce-> 992/1342)
        (158901 states / 217056 arcs reduce-> 1852/2205)
        (1852 states / 2052 arcs reduce-> 846/993)
WORLD 4.59941105854103e-09
WORLD 1.9549317084350 2e-09
WORLD 3.88355190221048e-10
WORLD 1.65066323891318e-10
WORLD 4.7880250707014e-12
WORLD 6.7424185216011e-14
WORLD 2.86579468363471e-14
WALD 2.46370947344287e-14
WELD 2.22750297605393e-14
WALD 1.0471740204784e-14
WELD 9.46776911890658e-15
```


Input line 1: M IH L K
(5432 states / 5875 arcs reduce-> 1149/1592)
(365131 states / 490797 arcs reduce-> 4029/4933)
(4029 states / 4515 arcs reduce-> 1372/1656)
MILK 2.81081968444271e-10
MILK 1.04182438482352e-10
MILK 7.1046463303416e-11
MILK 7.86509770846403e-12
MILK 2.80298985825729e-13
MILK 1.91147873869604e-13
MILE 1.80593595928241e-13
MILLS 1.66775113077043e-13
MILLS 1.43986443806815e-13
MILE 1.23154501592139e-13
MILE 1.36336440017105e-14
MILES 3.92154705136433e-15
MILO 3.7773494447953e-15


Input line 1: K AH M P Y UW T ER
(9700 states / 10710 arcs reduce-> 2856/3866)
(907591 states / 1139316 arcs reduce-> 2040/2379)
(2040 states / 2306 arcs reduce-> 890/1036)
COMPUTER 5.0573235853347e-09
COMPUTER 2.8751908286068e-09
COMPUTER 2.63310151480487e-09
COMPUTER 2.5223324522281e-09
COMPUTER 1.99830989449056e-09
COMPUTER 1.99033670629104e-09
COMPUTER 1.8895330691153e-09
COMPUTER 1.81004437276353e-09
COMPUTER 1.13607962480677e-09
COMPUTER 1.13154670592206e-09
COMPUTER 1.04042241344627e-09
COMPUTER 1.03627116356641e-09
COMPUTER 9.96654022910115e-10
COMPUTER 7.46614798176953e-10
COMPUTER 7.43635830643026e-10
COMPUTER 7.15206278287005e-10
COMPUTER 5.54849979125708e-10
COMPUTER 3.15443444406242e-10


Input line 1: V AW AH L Z
(4714 states / 5269 arcs reduce-> 1486/2041)
(213761 states / 277381 arcs reduce-> 2265/2734)
(2265 states / 2555 arcs reduce-> 864/1024)
VOWELS 6.99023686024899e-12
VOWELS 6.77019774273835e-12
VOWELS 1.57973802363891e-12
VOWELS 9.74820203554056e-13
VOWELS 4.32101421152451e-13
VOWELS 2.66639904225101e-13
VOLES 1.73604846805391e-13
VOWELS 7.79071190669826e-14
VOLES 3.14260045403323e-14
VOWELS 2.01095475929608e-14
VOWELS 1.08645007580589e-14
VOWELS 1.00899316656116e-14

```
Input line 1: S AY AH N S
        (6188 states / 6940 arcs reduce-> 2027/2779)
S I A H N C E 0.924076
S I A H N S 0.835345501725
S I O N S 0.690172373427237
C Y N C E 0.450408654683
S I N C E 0.439732385790475
C Y A N C E 0.297782033266414
S I A N C E 0.290723551984004
S A Y A N C E 0.276828113256158
C Y A N S 0.269188770170011
S I A N S 0.262808049765768
S A Y A N S 0.250246861902647
P S Y N C E 0.2487505351
C Y O N S 0.192400872027514
C Y E N S 0.190106796012128
S I O N S 0.187840294819317
```

3.6.

```
Input line 1: M IH L K
        (12 states / 12 arcs reduce-> 8/7)
        (8 states / 7 arcs)
MILK 3.1434858e-05
```

```
Input line 1: K AH M P Y UW T ER
        (33 states / 40 arcs reduce-> 32/39)
        (29 states / 33 arcs)
COMPUTER 0.00057585406
COMPUTE ER 6.609210176108 39e-12
COMPUTE ERR 5.14049662297977e-12
COME PUGH TER 1.2338211973203e-16
COME PEW TER 1.2338211973203e-16
CUM PUGH TER 5.5720949812637e-19
CUM PEW TER 5.5720949812637e-19
```

```
Input line 1: V AW AH L Z
        (20 states / 20 arcs reduce-> 9/8)
        (9 states / 8 arcs)
VOWELS 5.4198027e-07
```

```
Input line 1: S AY AH N S
        (25 states / 29 arcs reduce-> 9/8)
        (9 states / 8 arcs)
SCIENCE 0.00010758309
```

3.7.

```
flip1 ~/CS539/hw2-data 1040$ echo APPLECIDER | carmel -sliOEQk 5 eword-epron.wfst
Input line 1: APPLECIDER
       (0 states / 0 arcs)
Empty or invalid result of composition with transducer "eword-epron.wfst".
0
0
0
0
0
No derivations found for 1 of 1 inputs
```

```
Input line 1: TIMBERHAWK
       (0 states / 0 arcs)
Empty or invalid result of composition with transducer "eword-epron.wfst".
0
0
0
0
0
No derivations found for 1 of 1 inputs
```

```
Input line 1: DURIAN
       (0 states / 0 arcs)
Empty or invalid result of composition with transducer "eword-epron.wfst".
0
0
0
0
0
No derivations found for 1 of 1 inputs
```

3.8.

```
Input line 1: A P P L E C I D E R
        (766505 states / 767135 arcs reduce-> 1420/2050)
        (15997 states / 165664 arcs reduce-> 13892/150980)
AH P L EH S AH D ER 8.34598476329027e-12
AH P L EH S AH D ER 7.63561977759942e-12
AH P L EH CH IY D ER 3.83054824167259e-12
AH P L IH S AH D ER 3.38470671705732e-12
AH P L IH S AH D ER 3.09661882727279e-12
Derivations found for all 1 inputs
```

```
Input line 1: T I M B E R H A W K
        (765128 states / 765675 arcs reduce-> 1229/1776)
        (14049 states / 135723 arcs reduce-> 12614/122881)
T IH M B ER HH AO K 2.55196134950528e-11
T IH M B ER HH AO K 2.23206966910135e-11
T IH M B ER HH AO K 1.68888429445642e-11
T IH M B ER HH AO K 1.65215356094057e-11
T IH M B ER HH AO K 1.17968666095071e-11
Derivations found for all 1 inputs
```

```
Input line 1: D U R I A N
        (487436 states / 487818 arcs reduce-> 929/1311)
        (10195 states / 103898 arcs reduce-> 8610/94407)
D UH R IY AH N 1.32761626770593e-07
D UH R IY AH N 1.31967818941064e-07
D UH R IY AH N 5.72241512492152e-08
D UH R IY AH N 4.81856712761284e-08
JH UH R IY AH N 4.70105023119895e-08
Derivations found for all 1 inputs
```

3.9.

```
Input line 1: AH P L EH S AH D ER
        (39 states / 46 arcs reduce-> 36/43)
UP LES UH DURR 1
UP LES UH DER 1
UP LES UHDE UR 1
UP LESS UH DURR 1
UP LES UHDE OR 1
Derivations found for all 1 inputs
```

```
Input line 1: AH P L EH S AH D ER
        (9814 states / 10936 arcs reduce-> 2950/4072)
A P P L E S U D D E R 0.7619571672
A P P L E S S D E R 0.746877677344
A P P L E S S D E R 0.683307497715475
A P P L E C I D E R 0.471599848725768
A P P L E S S U D D E R 0.440612045061162
Derivations found for all 1 inputs
```

```
Input line 1: T IH M B ER HH AO K
        (23 states / 27 arcs)
TIMBER HAWKE 1
TIMBER HAWK 1
TIMM BURR HAWKE 1
TIMBER HAUCK 1
TIMM BURR HAWK 1
Derivations found for all 1 inputs
```

```
Input line 1: T IH M B ER HH AO K
        (8933 states / 9700 arcs reduce-> 2077/2844)
T E M B E R H A W K 0.801004
T I M M B E R H A W K 0.7696113856
T I M B E R H A W K 0.7387103556
T I M B E R H A W K 0.646112049975
T I M M B E R H A L K 0.64353437025792
Derivations found for all 1 inputs
```

```
Input line 1: D UH R IY AH N
       (0 states / 0 arcs)
Empty or invalid result of composition with transducer "eword-epron.wfst".
0
0
0
0
0
No derivations found for 1 of 1 inputs
```

```
Input line 1: D UH R IY AH N
        (7393 states / 8246 arcs reduce-> 2432/3285)
D U R I O N 0.999001
D U E R R I O N 0.965622378588
D U E R R E A N 0.954109828145918
D U R I A N 0.861633156168
D U R I A N 0.856481282375962
Derivations found for all 1 inputs
```

3.10. It changes a word to it's pronunciation, and then find all of it's possible homophones. At last, it filters the result and retrieve just the word that exists in the word data.
Eg. 'BEAR' => 'B EH R' => all possible homophones => 'BEAR'

BAER
BAHR
BARE
BEAR
BEHR

3.11. We have observed from 3.1 that it will have just one output which is the correct answer if we use the whole word to find its pronunciation. Because there are epron-eword.data.

And for 3.2, we found the pronunciation of the word by checking each character's pronunciation, since each character can be derived from different pronunciations, we got many different results.

In 3.3, after using epron.wfsa, the pronunciation has been revised by lookup the whole word.

In 3.4, the input is a pronunciation of a word, by using epron-espell.wfst in the forward direction, it would have many different restored spellings of the word.

One way to improve the result in 3.4, 3.5 we transferred the result to possible words by using espell-eword.wfst, and used word.wfsa to find words which are existed.

Another way to improve the result in 3.4, 3.6 we directly used epron-eword.wfst to find out all possible words, and got the filtered result by using word.wfsa.

It is not a surprise that the output of 3.6 is less than 3.5, because in 3.6 we produced words from just one pronunciation while we produced words from many spellings in 3.5. For example, when we input 'S AY AH N S', in 3.5 we can have lots of results just like in 3.2, and in 3.6 we only get 'SCIENCE'.

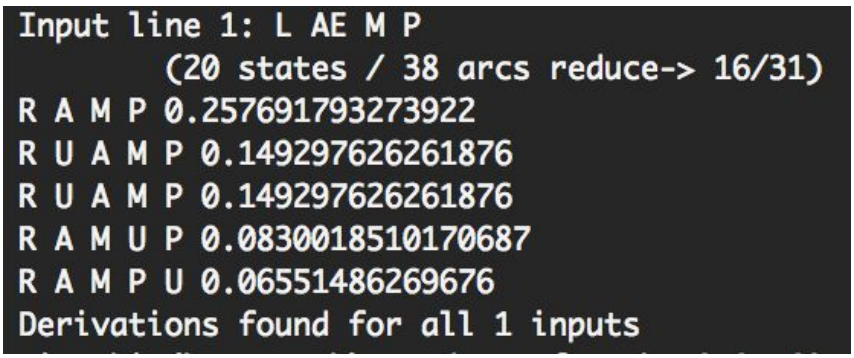
In 3.7, we tried to input the words are not in epron-eword.data to get their pronunciations. Since the words are not included in the eword, we got nothing. On the other hand, in 3.8 we used the epron.wfsa and epron-espell.wfst in reverse direction to get possible pronunciations from the letter sequences of those same words, and we did get some results because they were produced based on characters but not words.

After we got the pronunciation of the words in 3.8, we used the pronunciation which has the highest probability as the input for running eword-epron.wfst and epron-espell.wfst in 3.9. The result of epron-espell.wfst is better when compared with the result of eword-epron.wfst. In conclusion, eword-epron.wfst looks up for the whole world, and espell-epron.wfst considers the characters. So, eword-epron.wfst always has better results no matter the input is spelling or pronunciation. Because it's just like a dictionary records correct pronunciation of words. On the contrary, if the input word doesn't exist in the dataset, eword-epron.wfsa will become useless, it might find nothing or turn out ridiculous results.

4. Decoding English Words from Japanese Katakana (80 pts)

4.1. estimate.py and epron-jpron.probs.

4.2.



4.3. Some make sense, but others do not! The answer should be ‘RANPU’. Usually ‘L’ will become ‘RA’, since Japanese doesn’t have sound ’L’. The last character ‘P’, in Japenese have to combine with vowels.

4.4. By using bigram and trigram will improve it to become more Japanese like. For example, if we look ‘MIL’ for ‘MILK’ at the same time. We can output ‘MIRU’, and if we know ‘K’ is the last character, the output will be ‘KU’ or ‘KO’ something like this which ‘KU’ is the answer.

4.5.

```
HH IH R AH L IH K L IH NG T AH N 2.44385112533544e-17
HH IH R AH L IH K L IH N T AH N 1.94656644327155e-17
HH IH L AE R IH K L IH NG T AH N 1.69807768826539e-17
F IH L AE R IH K L IH NG T AH N 1.40761058970867e-17
HH IH L AE R IH K L IH N T AH N 1.35254599258448e-17
D N AH L D T R AE M P 2.17714466571863e-19
D N AH L D T R AE M P 2.17714466571863e-19
D N AH L D T R AH M P 9.41766040556432e-20
D N AH L D T R AH M P 9.41766040556432e-20
D N AH L D AH T R AE M P 7.48392925360527e-20
V IH D IY OW T EY P 3.53468209749603e-16
V IH D IY OW T EY P 3.53468209749603e-16
B IH D IY OW T EY P 3.38999748233113e-16
B IH D IY OW T EY P 3.38999748233113e-16
V IH D IY AH T EY P 2.67047710882556e-16
HH OW M ER SH IH M P S AH N 2.12790515284886e-17
HH OW M ER SH IH M P S AH N 2.12790515284886e-17
HH AA M ER SH IH M P S AH N 5.07610674047621e-18
HH AA M ER SH IH M P S AH N 5.07610674047621e-18
HH AH M ER SH IH M P S AH N 3.59180492559891e-18
R AE P T AA P 4.49855484430511e-12
L AE P T AA P 3.56363661316989e-12
R AH P T AA P 2.7428789262193e-12
R AE P T OW P 2.43626323733618e-12
L AE P T OW P 1.92994354239825e-12
SH EY V IH NG G AH K R IY M 1.24451082007779e-20
SH EY V IH NG G AH K R IY M 1.24451082007779e-20
S EY V IH NG G AH K R IY M 8.98142558168043e-21
S EY V IH NG G AH K R IY M 8.98142558168043e-21
SH EY V IH NG G AH K L IY M 6.94116712973799e-21
```

```
CH AY L D SH IY T 1.31622475060612e-14
CH AY L D SH IY T 1.31622475060612e-14
CH AY L D SH IY T 1.31622475060612e-14
CH AY L D SH IY T 1.31622475060612e-14
CH AY L D SH IY T OW 3.61295007002713e-15
SH IY T B EH L T 5.76272005465039e-14
SH IY T B EH L T 5.76272005465039e-14
S IY T B EH L T 3.41523011281742e-14
S IY T B EH L T 3.41523011281742e-14
SH IY T AH B AH L T 2.27799377237603e-14
SH IH NG G AH L R UW M 3.45185273822278e-16
SH IH NG G AH L R UW M 3.45185273822278e-16
SH IH NG G AH L R UW M 3.45185273822278e-16
SH IH NG G AH L R UW M 3.45185273822278e-16
SH IH NG G AH L R UW M 3.45185273822278e-16
G ER L F R EH N D 5.14601213769284e-14
G ER L F R EH N D 5.14601213769284e-14
G ER L F R EH N D 5.14601213769284e-14
G ER L F R EH N D 5.14601213769284e-14
G AE L F R EH N D 4.36335725415427e-14
T R AH B AH L ER D UW CH EH K 4.17021668800174e-22
T R AH B AH L ER D UW CH EH K 4.17021668800174e-22
T R AH B AH L ER Z AH CH EH K 2.85660752096888e-22
T R AH B AH L ER Z AH CH EH K 2.85660752096888e-22
T R AH B AH L ER D AH CH EH K 2.67464325118583e-22
B EH V IY SH IY T ER 8.01804238381053e-15
B EH V IY SH IY T ER 8.01804238381053e-15
B EH V IY S IH T ER 2.88123618540634e-15
B EH V IY S IH T ER 2.88123618540634e-15
B EH V IY S IY T ER 9.43738914604613e-16
```

```
S K OW T R AE N D 2.7019380933071e-12
S K OW T R AE N D 2.7019380933071e-12
S K OW T R AE N D 2.7019380933071e-12
S K AA T R AE N D 2.43919284434248e-12
S K AA T R AE N D 2.43919284434248e-12
V AY AH L IH N K AA N CH EH L T 8.52306350664637e-21
V AY AH L IH N K AA N CH EH L T 8.52306350664637e-21
V AY AH L IH N K AA N CH EH L T 8.52306350664637e-21
V AY AH L IH N K AA N CH EH L T 8.52306350664637e-21
B AY AA L IH N K AA N CH EH L T 6.83737076873024e-21
AH P R UW M AH K B UH K S P R OW 7.70285497101109e-24
AH P R UW M AH K B UH K S P L OW 2.64264608264009e-24
AH P R AH M AH K B UH K S P R OW 2.46639912620405e-24
AE P AH L M AH K B UH K S P R OW 2.3844190575644e-24
AE P AH L M AH K B UH K S P R OW 2.3844190575644e-24
K AA M P Y UW T AH S AY AH N Z 2.83756790517874e-20
K AA M P Y UW T AH S AY AH N Z 2.83756790517874e-20
K AA M P Y UW T AH S AY AH N Z 2.83756790517874e-20
K AA M P Y UW T AH S AY AH N Z 2.83756790517874e-20
K AA M P Y UW T AH S AY AH N Z 2.83756790517874e-20
HH IH JH IH K AH L T R EY N IH NG Z 6.9266563898196e-22
HH IH JH IH K AH L T R EY N IH NG Z 6.9266563898196e-22
HH IH JH IH K AH L T R EY N IH NG Z 6.9266563898196e-22
HH IH JH IH K AH L T R EY N IH NG Z 6.9266563898196e-22
F IH JH IH K AH L T R EY N IH NG Z 5.61843905575872e-22
HH IH JH IH K AH L EH K S AH S AY Z 3.23145482021864e-22
HH IH JH IH K AH L EH K S AH S AY Z 3.23145482021864e-22
HH IH JH IH K AH L EH K S AH S AY Z 3.23145482021864e-22
HH IH JH IH K AH L EH K S AH S AY Z 3.23145482021864e-22
HH IH JH IH K AH L EH K S AH S AY Z 3.23145482021864e-22
```

```
R IH S K R IY M 4.28335608522476e-12
R IH S K R IY M 4.28335608522476e-12
R IH S K R IY M 4.28335608522476e-12
AY S K R IY M 2.86665046705486e-12
AY S K R IY M 2.86665046705486e-12
HH AA T M IH L K 4.32926120179773e-14
HH AA T M IH L K 4.32926120179773e-14
F OW T M IH L K 8.1682396383895e-15
F OW T M IH L K 8.1682396383895e-15
HH AA T M AH L K 7.88406657499306e-15
T R IH P L UH R UW M 4.69315113358454e-16
T R IH P L UH R UW M 4.69315113358454e-16
T R IH P L UH R UW M 4.69315113358454e-16
T R IH P AH L R UW M 3.94546051167219e-16
T R IH P AH L R UW M 3.94546051167219e-16
K R AW N P R AH DH ER HH OW T EH L 9.30214573933018e-23
K R AW N P R AH DH ER HH OW T EH L 9.30214573933018e-23
K R AW N P R AH DH ER HH OW T EH L 9.30214573933018e-23
K R AW N P R AH DH ER HH OW T EH L 9.30214573933018e-23
K R AW N P R AA Z AH F OW T EH L 6.57908881066391e-23
HH EY S B UH K L IH S ER CH IH Z AY AH N T IH S T 2.21639216251738e-34
HH EY S B UH K L IH S ER CH IH Z AY AH N T IH S T 2.21639216251738e-34
HH EY S B UH K L IH S ER CH IH Z AY AH N T IH S T 2.21639216251738e-34
HH EY S B UH K L IH S ER CH IH Z AY AH N T IH S T 2.21639216251738e-34
HH EY S B UH K L IH S ER CH IH Z AY AH N T IH S T 2.21639216251738e-34
W AO L F G AE NG G AH M OW T AH L T 1.9418716987656e-25
W AO L F G AE NG G AH M OW T AH L T 1.9418716987656e-25
W AO L F G AE NG G AH M OW T AH L T 1.9418716987656e-25
W AO L F G AE NG G AH M OW T AH L T 1.9418716987656e-25
W OW L F G AE NG G AH M OW T AH L T 1.20900204705081e-25
```

4.6.

```
HILARY CLINTON 7.58045209187269e-15
HILLARY CLINTON 2.24605981000383e-15
HERE EARLY CLINTON 5.69126827453591e-16
HILL EARLY CLINTON 3.15959232668614e-16
HIGH RALLY CLINTON 2.69071811872948e-16
DONALD TRUMP 3.39444604276982e-14
DONALD TRUMP 3.39444604276982e-14
DONALD TRAMP 1.87896078314198e-15
DONALD TRAMP 1.87896078314198e-15
DONALD TRUMPS 1.08598121448787e-16
VIDEOTAPE 5.64978139944713e-09
VIDEOTAPE 5.64978139944713e-09
VIDEOTAPES 6.80835604342525e-10
VIDEOTAPES 6.80835604342525e-10
VIDEOTAPED 7.45601459156095e-11
HOMER SIMPSON 4.19288889481375e-16
HOMER SIMPSON 4.19288889481375e-16
HOME RUSSIAN PUSAN 3.49329387799809e-18
HOME RE SIMPSON 2.48626063098372e-18
HOME RUSHING PUSAN 2.19258411422064e-18
LAPTOP 1.5880963588933e-08
LAPTOPS 1.30084330436547e-10
WRAPPED UP 1.31716686351624e-11
WRAP TOP 1.2782162636438e-11
RAP TOP 1.18691510195496e-11
SAVINGS CREAM 4.38370011676579e-14
SAVINGS CREAM 4.38370011676579e-14
SAVING SCREAM 3.21501993945868e-16
SAVING SCREAM 3.21501993945868e-16
SAVING SCREAMS 7.07836162746074e-17
```

```
CHILD SHEET 5.5314903292533e-12
CHILD SHEET 5.5314903292533e-12
CHILD SHEET 5.5314903292533e-12
CHILD SHEET 5.5314903292533e-12
CHILD SEAT 1.76534499643474e-12
SHEET BELT 3.13619480665801e-12
SHEET BELT 3.13619480665801e-12
SHEET BUILT 1.07865531917376e-12
SHEET BUILT 1.07865531917376e-12
SEAT BELT 1.00089948282088e-12
SINGLE ROOM 3.09576848470528e-12
SINGLE ROOM 3.09576848470528e-12
SINGLE ROOM 3.09576848470528e-12
SINGLE ROOM 3.09576848470528e-12
SINGLE ROOM 3.09576848470528e-12
GIRLFRIEND 4.5689621538819e-08
GIRLFRIEND 4.5689621538819e-08
GIRLFRIEND 4.5689621538819e-08
GIRLFRIEND 4.5689621538819e-08
GIRLFRIENDS 4.75569231123923e-10
TRAVELERS CHECK 2.01766655996513e-14
TRAVELERS CHECK 2.01766655996513e-14
TRAVELERS CHECK 2.01766655996513e-14
TRAVELERS CHECK 2.01766655996513e-14
TRAVELERS CZECH 1.82633615568587e-15
BABY SITTER 4.70382403217975e-15
BABY SITTER 4.70382403217975e-15
BABY SHUTTER 3.84207172909087e-15
BABY SHUTTER 3.84207172909087e-15
BABY SEATER 2.9964343531341e-15
```

SCOTLAND 2.66088826702266e-08
SCOTLAND 2.66088826702266e-08
SCOTLAND 2.66088826702266e-08
SCOTT LAND 3.59784772666574e-11
SCOTT LAND 3.59784772666574e-11
VIOLIN CONCERTO 8.11620583965411e-18
VIOLIN CONCERTO 8.11620583965411e-18
VIOLIN CONCERTO 8.11620583965411e-18
VIOLIN CONCERTO 8.11620583965411e-18
VIA RIM CONCERTO 3.70510014198154e-20
APPLE MAKE BOOK PRO 1.84707628146325e-22
APPLE MAKE BOOK PRO 1.84707628146325e-22
APPLE MAC BOOK PRO 1.79098814708455e-22
APPLE MAC BOOK PRO 1.79098814708455e-22
OPERA MAKE BOOK PRO 9.09876011245134e-23
COMPUTER SCIENCE 2.03767007278584e-14
COMPUTER SCIENCE 2.03767007278584e-14
COMPUTER SCIENCE 2.03767007278584e-14
COMPUTER SCIENCE 2.03767007278584e-14
COMPUTER SCIENCE 2.03767007278584e-14
PHYSICAL TRAINING GU 1.40923434711295e-19
PHYSICAL TRAINING GU 1.40923434711295e-19
PHYSICAL TRAINING GU 1.40923434711295e-19
PHYSICAL TRAINING GU 1.40923434711295e-19
PHYSICAL TRAINING GOO 6.03957568843812e-20
PHYSICAL EXERCISE 9.799927453285e-16
PHYSICAL EXERCISE 9.799927453285e-16
PHYSICAL EXERCISE 9.799927453285e-16
PHYSICAL EXERCISE 9.799927453285e-16
PHYSICAL EXERCISE 9.799927453285e-16

ICE CREAM 4.47342795008632e-12
ICE CREAM 4.47342795008632e-12
ICE CREAM 4.47342795008632e-12
ICE CREAM 4.47342795008632e-12
ICE CREAM 4.47342795008632e-12
HOT MILK 7.97792136735865e-12
HOT MILK 7.97792136735865e-12
FOUGHT MILK 2.39228403885395e-12
FOUGHT MILK 2.39228403885395e-12
HUT MILK 3.69101719938599e-13
TRIPLE ROOM 2.41054062685351e-12
TRIPLE ROOM 2.41054062685351e-12
TRIPLE ROOM 2.41054062685351e-12
TRIPLE ROOM 2.41054062685351e-12
TRIPLE ROOM 2.41054062685351e-12
CROWN PLAZA HOTEL 2.51647416168154e-18
CROWN PLAZA HOTEL 2.51647416168154e-18
CROWN PLAZA HOTEL 2.51647416168154e-18
CROWN PLAZA HOTEL 2.51647416168154e-18
CROWN PLAZA HOTEL 1.23832486014311e-18
FACE BOOK RESEARCH SCIENTIST 2.54238617326893e-22
FACE BOOK RESEARCH SCIENTIST 2.54238617326893e-22
FACE BOOK RESEARCH SCIENTIST 2.54238617326893e-22
FACE BOOK RESEARCH SCIENTIST 2.54238617326893e-22
FACE BOOK RESEARCH SCIENTIST 2.54238617326893e-22
WALL WHO GANGS MOZART 2.55632147266913e-25
WALL WHO GANGS MOZART 2.55632147266913e-25
WALL WHO GANGS MOZART 2.55632147266913e-25
WALL WHO GANGS MOZART 2.55632147266913e-25
WOLF GANGS MOZART 2.81051306164613e-26

Yes, almost correct. The reason is that eword.wfsa works as a filter to make it easier to build up the word.

When the input is “H O TT O M I R U K U”, it is transferred to all possible English pronunciations. Then, we generated all possible English words based on . Finally it filt out to the word in data.

4.7.

If there are no specific words in the data, it can not find it. For example, the last one should be “Wolfgang Mozart”, but there are no Wolfgang in the data, so it only find the closest one like “WALL WHO GANGS”.

4.8.

```
TOMATO SOUTH 3.84577098725048e-14
TOMATO SOUTH 3.84577098725048e-14
TOMATO SOUTH 3.84577098725048e-14
TOMATO SAUCE 1.10604120065469e-14
TOMATO SAUCE 1.10604120065469e-14
AIR MAIL 3.99249724635438e-10
AIR MAIL 3.99249724635438e-10
AIR MALE 2.27989795996946e-10
AIR MALE 2.27989795996946e-10
AIR MEL 1.79611669483982e-10
MICROPHONE 6.13880946748028e-09
MICROPHONE 6.13880946748028e-09
MICRO PHONE 1.04992395225827e-10
MICRO PHONE 1.04992395225827e-10
MICRO HONG 1.01766206267942e-10
TOM FIDDLE STONE 4.98641119426556e-19
TOM FIDDLE STONE 4.98641119426556e-19
TOM FIDDLE STONE 4.98641119426556e-19
TOM FIDDLE STONE 4.98641119426556e-19
TOM FIDDLE STONE 4.98641119426556e-19
HOT DOG 5.41624022558522e-12
FOUGHT DOG 1.62413045273672e-12
HOT DOUG 4.06720749972171e-13
HUT DOG 2.50584518298657e-13
HOT DUG 1.58169174985959e-13
```

4.9.

1. *carmel -sriIEQk 5 eword.wfsa eword-epron.wfst epron.wfsa epron-jpron.wfst*

```
 I K U R I N T O N' | carmel -sriIEQk 5 eword.wfsa eword-epron.wfst epron.wfsa e
pron-jpron.wfst
Input line 1: H I R A R I K U R I N T O N
        (59 states / 142 arcs reduce-> 19/91)
        (477 states / 2570 arcs reduce-> 409/2115)
        (2172 states / 3862 arcs reduce-> 1391/2606)
        (1218 states / 2236 arcs reduce-> 1156/2140)
HILARY CLINTON 9.92572207124635e-29
HILLARY CLINTON 2.94095459733815e-29
HERE ALLY CLINTON 3.29122620094901e-31
HILL ARA CLINTON 2.45224663616878e-31
HILL EARLY CLINTON 2.04450124942026e-31
Derivations found for all 1 inputs
```

This way creates more paths to filter the English pronunciations which translated from Japanese pronunciations by using epron.wfsa. The speed is slower than running without epron.wfsa but the results are smoother and more accurate.

2. *carmel -sriIEQk 5 eword.wfsa eword-epron.wfst epron-eword.wfst eword-epron.wfst epron-jpron.wfst*

```
 I K U R I N T O N' | carmel -sriIEQk 5 eword.wfsa eword-epron.wfst epron-eword.
wfst eword-epron.wfst epron-jpron.wfst
Input line 1: H I R A R I K U R I N T O N
        (59 states / 142 arcs reduce-> 19/91)
        (1749 states / 2448 arcs reduce-> 1094/1649)
        (121341247 states / 128842152 arcs reduce-> 10237/13083)
        (59047 states / 89210 arcs reduce-> 33203/48472)
        (31244 states / 44242 arcs reduce-> 29341/41734)
0
0
0
0
0
No derivations found for 1 of 1 inputs
```

We have tried to add eword-epron.wfst and epron-eword.wfst between eword.wfsa and eword-epron.wfst, but it didn't find a result.

4.10.

There are lots of long sounds in katakana, and also voiced sound. So, it is more complicated in Japanese. We would like to add long sounds and voiced sound wfst. Moreover, bigram and trigram might be useful, too. It can decide which katakana is most likely beside the others. It would help jpron more specific.