Multilingual Noun Gender Analysis

CSCE 685: Directed Studies – Prof James Caverlee

Student: Sunayna Ray

OUTLINE

1. Introduction
2. Problem Statement
3. Approach
4. Dataset
5. Data Preprocessing
6. Analysis and Results – Hindi
7. Analysis and Results - French
8. Challenges
9. References

INTRODUCTION

- Human beings have made great strides in many fields of arts and sciences. But one could argue that the basis of all these advancements is language.
- If hunting was the first most essential skill a man learnt, then a common communication method is the same for any society or civilization.
- Naturally, we have uncountable languages today in the world.
- It is easy to spot the differences among them – as they were developed in geographically far location.
- But are there any innate similarities or common patterns in them? If so, then why and how?

Presentation Outline

PROBLEM STATEMENT

- This opens doors to much deeper analysis of how humans perceive concepts and sounds.
- For this study, we focus on how the languages classify nous as masculine or feminine in case of non-gendered languages.
- Do we have a natural tendency to classify some words as male or female based on how they sound or what they mean?

Presentation Outline

APPROACH

- Consider 2 gendered languages – Hindi, French
- Study their gender classification for a set of nouns and the phonetics for the same.
- Which nouns to consider?
- Top 1500 most commonly used nouns
- Source: https://www.talkenglish.com/vocabulary/top-1500-nouns.aspx

Presentation Outline

DATASET

- The first step was to translate each word to Hindi and French languages and note down the gender they were classified to.

- For this I have used to python to make web requests to 2 language translation websites:
  1. BoltiDictionary (Hindi)
  2. Linguee (French)

word	translation	gender
history	इतिहास	masculine
way	ढंग	masculine
art	कला	feminine
world	जगत	masculine
information	जानकारी	feminine
map	नक़्शा	masculine
government	सरकार	feminine
health	स्वास्थ्य	masculine

word	translation	gender
people	personnes	masculine
history	histoire	feminine
way	façon	feminine
art	art	masculine
world	monde	masculine
information	information	feminine
map	carte	feminine
family	famille	feminine

Presentation Outline

DATASET PREPROCESSING

- Since data is in different languages and even different scripts as well!
- Hence a valid way to compare it is using their phonetic translations (IPA)
- For this I have used 2 translation translation websites:
  1. https://www.fontconverter.in/index.php?q=Devanagari-to-IPA (Hindi)
  2. https://unalengua.com/ipa-translate?hl=en&ttsLocale=fr-CA&voiceId=Chantal&sl=fr&text= (French)

word	translation	gender	ipa
history	इतिहास	masculine	ɪt̪ihɑ:sə
way	ढंग	masculine	ɖʱⁿgə
art	कला	feminine	kəlɑ:
world	जगत	masculine	ɟəgət̪ə
information	जानकारी	feminine	ɟɑ:nəkɑ:ri:
map	नक़्शा	masculine	nəkə̤̤ɕɑ:
government	सरकार	feminine	sərəkɑ:rə
health	स्वास्थ्य	masculine	svɑ:st̪ʰjə

word	translation	gender	ipa
people	personnes	masculine	personnes
history	histoire	feminine	histoire
way	façon	feminine	façon
art	art	masculine	art
world	monde	masculine	monde
information	information	feminine	information
map	carte	feminine	carte
family	famille	feminine	famille

Presentation Outline

DATASET PREPROCESSING

- For each of the languages, I calculated the following for each of the entries using Spark and Scala:
- First Phonetic symbol, First 2 Phonetic Symbols
- Last Phonetic symbol, Last 2 Phonetic Symbols
- First Translated symbol, First 2 Translated Symbols
- Last Translated symbol, Last 2 Translated Symbols
- Translation length, IPA length

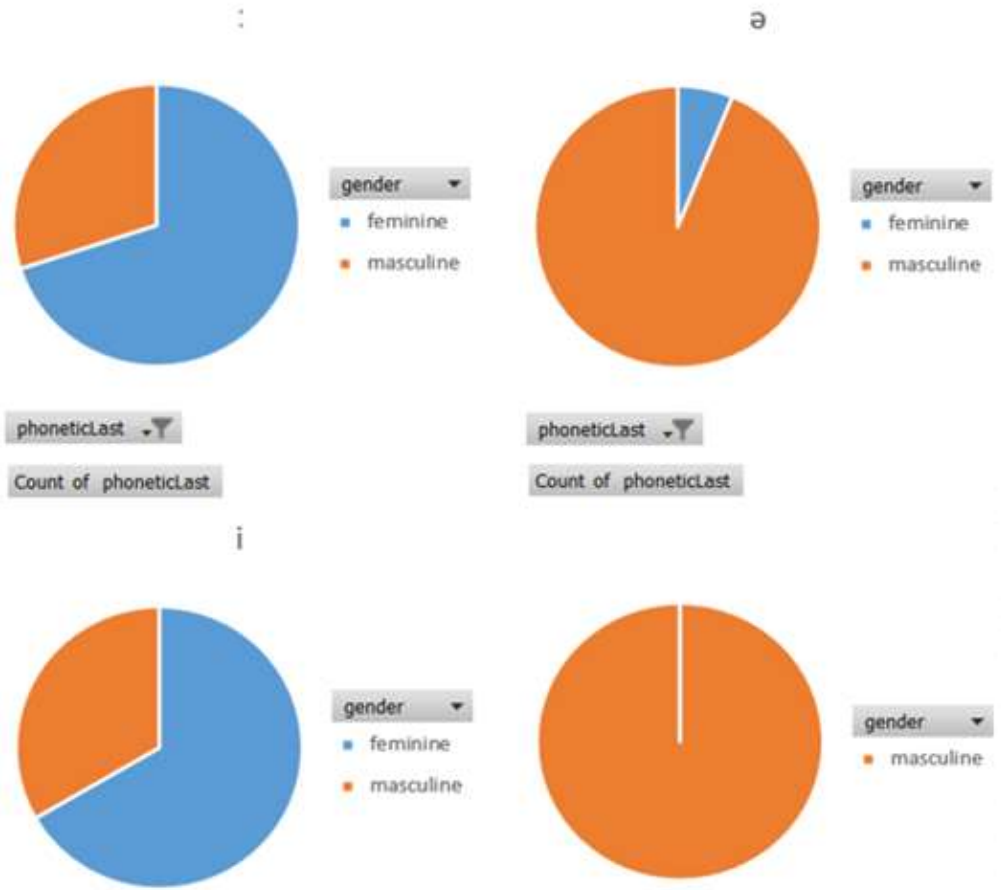
word	translation	gender	phonetic Last	phonetic Last2	phonetic First	phonetic First2	translation Last	translation Last2	translation First	translation First2	phonetic Length	translation Length
history	इतिहास		ə	sə	i	it	स	ास	इ	इत	9	6
way	ढंग		ə	gə	ɖ	ɖʰ	ग	ंग	ढ	ढं	5	3
art	कला	feminine	:	ɑ:	k	kə	ा	ला	क	कल	5	3
world	जगत		ə	ˌə	ʄ	ʄə	त	गत	ज	जग	7	3
information	जानकारी	feminine	:	i:	ʄ	ʄɑ	ी	री	ज	जा	11	7
map	नक़्शा		:	ɑ:	n	nə	ा	शा	न	नक	9	6
government	सरकार	feminine	ə	rə	s	sə	र	ार	स	सर	9	5
health	स्वास्थ्य		ə	jə	s	sv	य	्य	स	स्	10	9
system	व्यवस्था	feminine	:	ɑ:	v	vj	ा	था	व	व्	11	8
computer	कंप्यूटर		ə	rə	k	kⁿ	र	टर	क	कं	10	8
meat	मांस		ə	sə	m	mɑ	स	ंस	म	मा	6	4

Presentation Outline

ANALYSIS AND INSIGHTS - HINDI

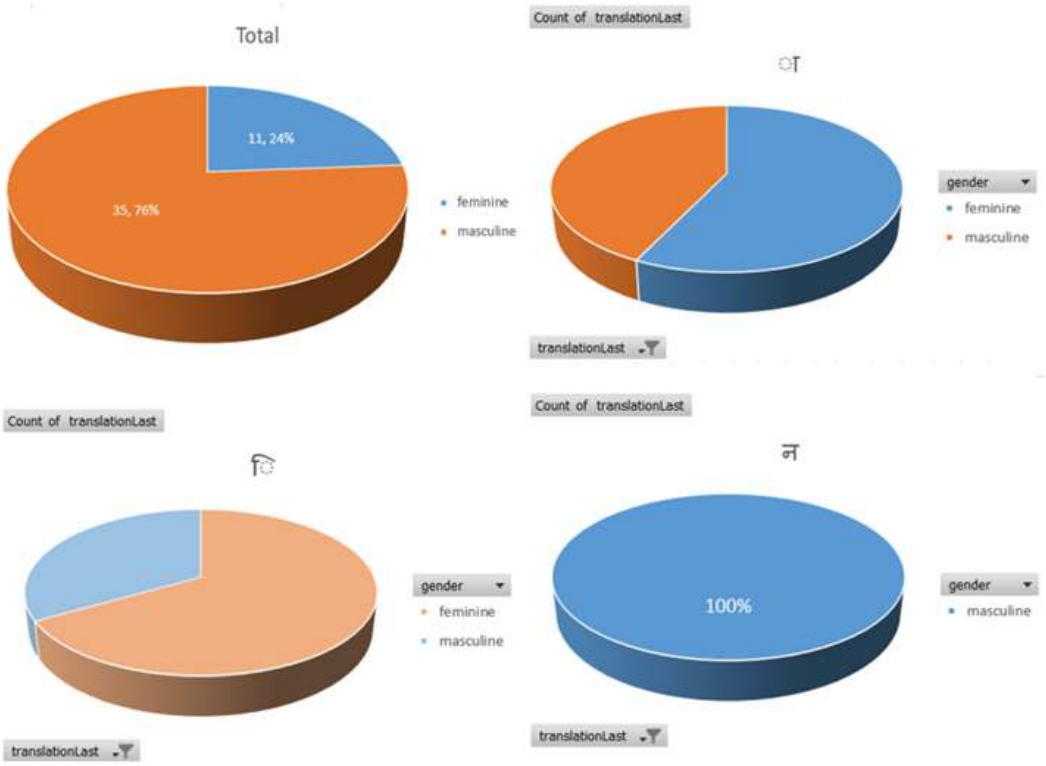
- We can observe that there is a clear model on how some ending phonetics tend to be more feminine inclined than others

- This is even more interesting since the proportion of female words is much lower compared to male words



ANALYSIS AND INSIGHTS - HINDI

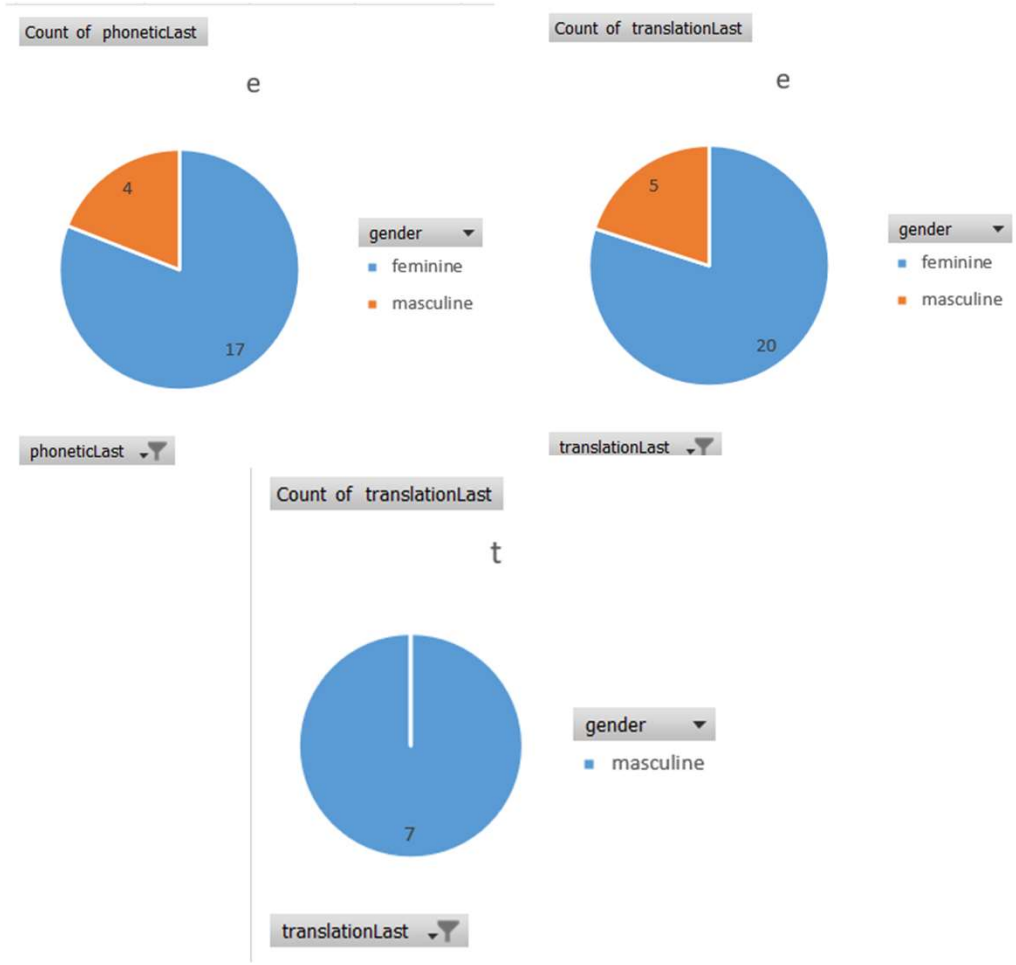
- We can see the original split in words as well. The words seem to be mostly masculine rather than feminine.
- Similar split can be observed on the actual translation as well. The words ending with “n” sounding symbol in lower right quadrant are all masculine.



ANALYSIS AND INSIGHTS - FRENCH

- The preference for words ending with e to be female is a very prominent trend and seems almost like a rule – This was true for phonetic as well as semantic translations

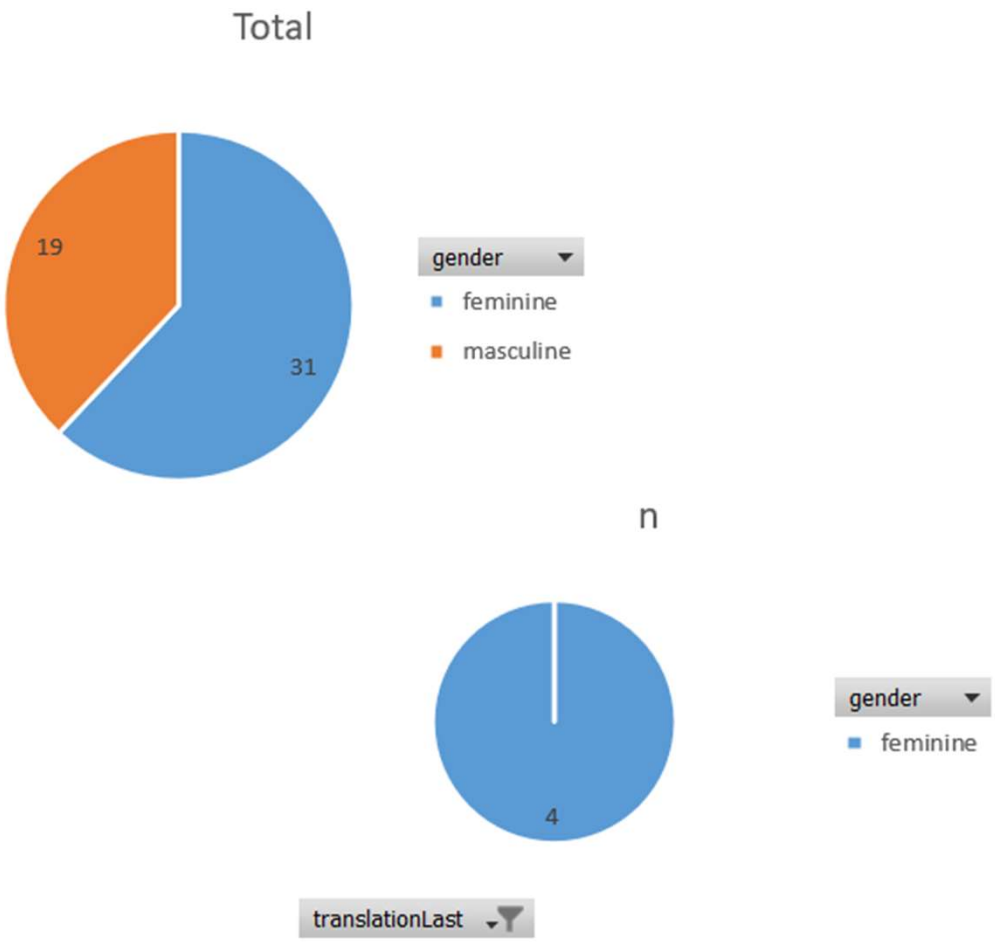
- All words ending with ‘t’ are masculine. This is even more interesting since the proportion of male words is much lower compared to female words



ANALYSIS AND INSIGHTS - FRENCH

- The words seem to be mostly feminine rather than masculine, which is in sharp contrast to Hindi. Again, we do not discount the sample bias for the same, which can be done in further research work

- The words ending with “n” sounding symbol are all feminine, which is again opposite to what we observed for Hindi.



CHALLENGES

- Collecting the data was extremely challenging because the web pages are dynamic and different for many words. Hence the Regex patterns had to be robust.
- Further for the IPA translations, there were no open endpoints for either of the languages. Hence the entire work had to be done manually
- All of these tasks along with the analysis over MS Excel were to be done with Unicode embedding. This led to lot of errors and workarounds.
- Finally for the analysis itself, I could not find many decision tree models for string data. Hence, I had to come up with my own ideas for features and 1-hot encode them over excel (no packaged tool was present on Spark-Scala).
- It was an interesting challenge to figure out the best ways to summarize and visualize results from this data.

Presentation Outline

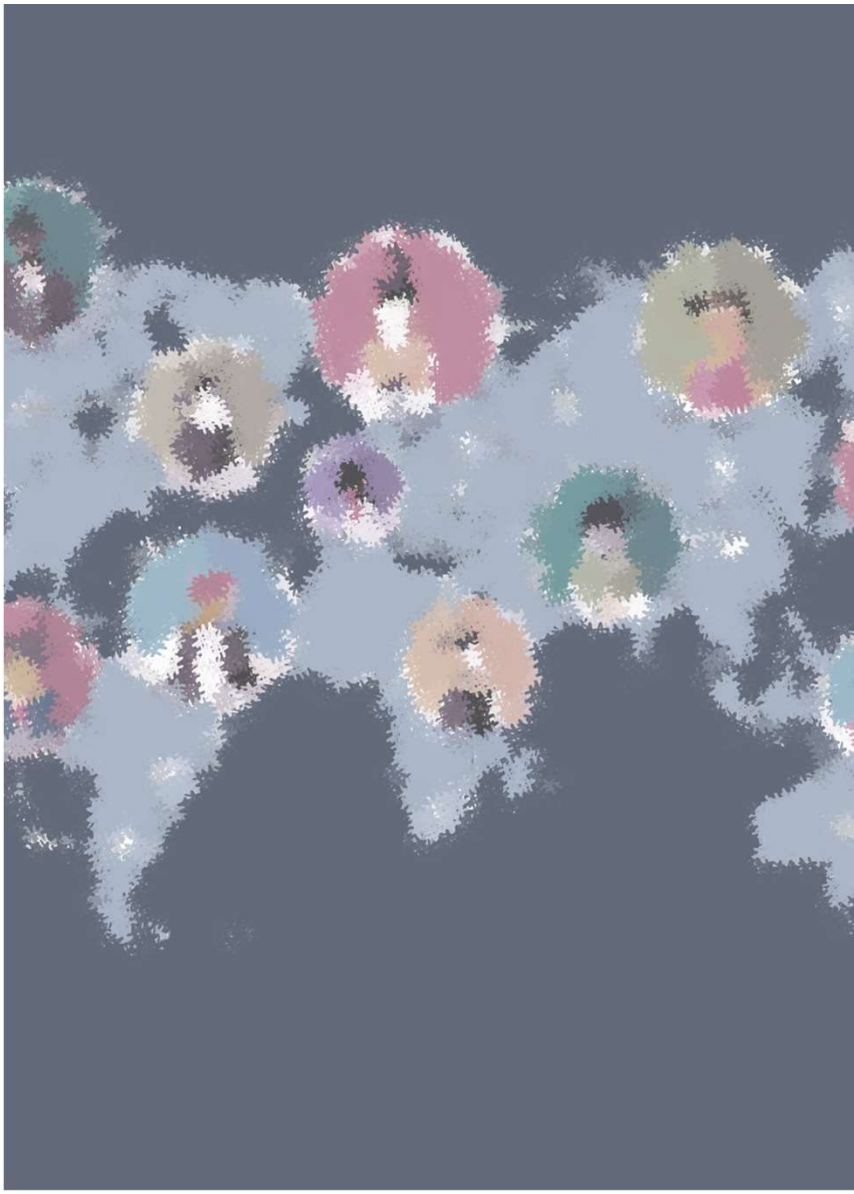
REFERENCES

A. Following papers shared by the professor were very helpful in guiding my work:

1. https://aclanthology.org/W18-5118/
2. https://aclanthology.org/C16-1234/

B. Further I leaned on various web sources from linguistics discussion forums etc.

C. I have shared most of the links in the presentation slides, but one more important site was: http://www.elinguistics.net/




END