Experiment 4

Analysing the frequency of POS tags in Formal and Informal text corpora

The formal.txt and informal.txt files are taken from the Formal sentences from the EPIE dataset.

Text Corpus Link here

The following points explain our methodology and results:

1. Introduction - In this analysis, we embarked on a comprehensive examination of the frequency distribution of Part-of-Speech (POS) trigrams within formal and informal sentences. Our primary objective was to gain insights into the linguistic patterns that distinguish these two styles of communication. To achieve this, we harnessed the power of the spaCy library for natural language processing and utilized matplotlib for data visualization. .

2. Data Collection - We initiated the analysis by collecting two distinct sets of sentences: one comprising formal sentences retrieved from "formal.txt" and the other consisting of informal sentences sourced from "informal.txt." These sentences were processed to eliminate any extraneous whitespace characters, ensuring clean and consistent data.

3. POS Tagging and Trigram Extraction - Next, we engaged in POS tagging, a crucial step that involves assigning a grammatical category (POS tag) to each word in the sentences. For our analysis, we chose to focus on five specific POS tags: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), and PRON (pronouns). These tags were chosen as they play pivotal roles in sentence structure and meaning. To unveil patterns within the text, we extracted trigrams – sequences of three consecutive POS tags – from the tagged sentences. These trigrams served as our basic units of analysis.

4. Frequency Distribution and Filtering - Having successfully extracted the trigrams, we proceeded to calculate their frequency of occurrence within both the formal and informal sentence datasets. However, to pinpoint the most significant linguistic distinctions between the two styles, we applied a filter to the trigrams. Specifically, we retained trigrams with a frequency count greater than 200, considering lower-frequency trigrams as less indicative of notable linguistic distinctions.

5. Comparative Analysis - For the sake of visual analysis, we leveraged matplotlib to create bar charts illustrating the frequency distribution of the filtered trigrams in both formal and informal sentences. These charts enabled us to perform a side-by-side comparison, making it easier to identify variations in trigram usage patterns.

6. Cumulative Frequency and Scaling - Going beyond individual trigram frequencies, we delved into the cumulative frequencies of trigrams. Cumulative frequencies provide a broader perspective on the distribution of linguistic patterns. To facilitate

comparison, we normalized and scaled these cumulative frequencies. Normalization involved dividing the cumulative frequencies by their respective maximum values, while scaling returned these normalized values to their original scale. This allowed us to observe how linguistic patterns evolved throughout the datasets..

7. Notable Differences - While examining the comparative analysis and cumulative frequency plots, one salient observation was the divergence in the frequencies of certain trigrams, such as "VERB - PRONOUN - NOUN." These differences merit further investigation and potentially reveal key distinctions in linguistic expression between formal and informal communication styles.

8. Conclusion - By visualizing trigram frequencies, exploring cumulative frequencies, and identifying disparities we gained insights into differences between text corpora featuring informal and formal sentences.

   Additionally, we created a sorted distribution of trigrams based on frequency differences. We stored this data in "diff.txt". This file helps us understand which combinations of POS tags occur in formal text and how this differs from the informal text.

Results from experiment showcasing "diff.txt"

Table 1: Trigram Frequency Comparison

Trigram	Formal Frequency	Informal Frequency	Difference
('VERB', 'PRON', 'NOUN')	472	863	391
('DET', 'NOUN', 'PUNCT')	781	1011	230
('ADP', 'DET', 'NOUN')	1682	1909	227
('DET', 'ADJ', 'NOUN')	1417	1196	221
('PRON', 'NOUN', 'PUNCT')	373	592	219
('ADP', 'PRON', 'NOUN')	390	589	199
('ADJ', 'NOUN', 'PUNCT')	1110	920	190
('ADP', 'DET', 'ADJ')	770	614	156
('VERB', 'DET', 'NOUN')	1048	1196	148
('NOUN', 'ADP', 'NOUN')	669	816	147
('ADP', 'NOUN', 'PUNCT')	480	620	140
('PRON', 'NOUN', 'ADP')	214	336	122
('DET', 'NOUN', 'ADP')	1262	1380	118
('VERB', 'DET', 'ADJ')	519	412	107
('VERB', 'PRON', 'PUNCT')	239	137	102
('VERB', 'ADP', 'PRON')	315	416	101
('NOUN', 'ADP', 'PRON')	349	425	76
('NOUN', 'PUNCT', 'PUNCT')	284	345	61
('NOUN', 'VERB', 'ADP')	237	292	55
('ADJ', 'NOUN', 'ADP')	745	691	54
('VERB', 'ADP', 'DET')	677	730	53
('ADJ', 'ADJ', 'NOUN')	269	217	52
('NOUN', 'CCONJ', 'NOUN')	288	236	52
('PART', 'VERB', 'PRON')	343	394	51
('VERB', 'NOUN', 'ADP')	287	336	49
('AUX', 'VERB', 'PRON')	314	363	49
('ADJ', 'NOUN', 'CCONJ')	203	156	47
('DET', 'NOUN', 'CCONJ')	217	263	46
('ADP', 'NOUN', 'ADP')	253	297	44
('NOUN', 'CCONJ', 'VERB')	234	274	40
('NOUN', 'NOUN', 'PUNCT')	363	403	40
('PRON', 'VERB', 'ADP')	276	316	40
('VERB', 'NOUN', 'PUNCT')	224	186	38
('DET', 'NOUN', 'VERB')	285	322	37
('PRON', 'VERB', 'PRON')	410	447	37
('NOUN', 'PRON', 'VERB')	277	241	36
('NOUN', 'PRON', 'AUX')	281	246	35
('DET', 'NOUN', 'PRON')	266	233	33
('PART', 'VERB', 'DET')	336	367	31
('DET', 'NOUN', 'NOUN')	294	325	31
('AUX', 'VERB', 'PRON')	314	363	49
('VERB', 'PART', 'VERB')	569	539	30
('NOUN', 'ADP', 'PROPN')	334	362	28

Table 1 – Continued from previous page

Trigram	**Formal Frequency**	**Informal Frequency**	**Difference**
('NOUN', 'PUNCT', 'VERB')	318	344	26
('NOUN', 'PUNCT', 'CCONJ')	363	388	25
('VERB', 'ADJ', 'NOUN')	206	181	25
('SCONJ', 'PRON', 'AUX')	459	436	23
('NOUN', 'PART', 'VERB')	261	281	20
('ADP', 'DET', 'PROPN')	303	322	19
('PRON', 'AUX', 'ADV')	299	281	18
('AUX', 'ADV', 'VERB')	272	254	18
('NOUN', 'SCONJ', 'PRON')	194	211	17
('AUX', 'AUX', 'VERB')	268	252	16
('PRON', 'AUX', 'PART')	286	270	16
('PRON', 'AUX', 'AUX')	255	240	15
('AUX', 'PART', 'VERB')	309	295	14
('NOUN', 'PUNCT', 'PRON')	334	348	14
('NOUN', 'ADP', 'DET')	1251	1264	13
('PROPN', 'PROPN', 'PUNCT')	371	383	12
('ADJ', 'NOUN', 'NOUN')	202	212	10
('PRON', 'ADJ', 'NOUN')	203	194	9
('ADP', 'ADJ', 'NOUN')	379	371	8
('NOUN', 'AUX', 'VERB')	343	351	8
('VERB', 'ADP', 'NOUN')	275	283	8
('DET', 'NOUN', 'AUX')	292	284	8
('PART', 'VERB', 'ADP')	266	259	7
('PUNCT', 'PRON', 'VERB')	409	415	6
('NOUN', 'ADP', 'VERB')	201	206	5
('NOUN', 'PUNCT', 'NOUN')	225	220	5
('SCONJ', 'PRON', 'VERB')	325	330	5
('ADP', 'PRON', 'PUNCT')	210	205	5
('PUNCT', 'DET', 'NOUN')	234	238	4
('PUNCT', 'PRON', 'AUX')	431	427	4
('PRON', 'AUX', 'VERB')	863	866	3
('CCONJ', 'PRON', 'AUX')	222	220	2
('ADP', 'PROPN', 'PUNCT')	215	217	2
('PUNCT', 'CCONJ', 'PRON')	212	211	1
('NOUN', 'ADP', 'ADJ')	239	238	1
('AUX', 'VERB', 'DET')	261	260	1
('VERB', 'PRON', 'ADP')	274	275	1
('PRON', 'VERB', 'DET')	348	349	1

Conlusion

From the differences obtained in the diff.txt file, we observe noticeable differences in a few combinations of POS tagging when we sort in descending order of difference. Specifically in 'VERB', 'PRON', 'NOUN' combinations.

However, this is very data specific and by plotting the cumulative frequency graphs, we can conclude that there does not seem to be a very stark difference between the tags of informal and formal text.