CSE 510: Assignment 3 Analysis Report

Approximating Natural Language: From Shannon to Modern AI

---

1. Quality Comparison: The Progression from Chaos to Coherence

Claude Shannon discovered that language actually follows predictable patterns, and these patterns show up at different levels. To test this idea, we can watch how text gets better and better as we use smarter models, starting from simplest version upto the most advanced ones. For eg. when we tried it with Jane Austen's *Pride and Prejudice*, we see that the text transforms slowly from gibberish to something that sounds more and more realistic writing.

Character-Level Analysis

Level 0 (Char-0: Random Characters)

```
fzqxhsajij                                    sxgsoqyxgibzzuiynfmyjbtroqxlfuooe
tcyqyahsljfipppczvditzrubipuzafwzelniznkeftp zogzejsbrl
```

At the zero-order level, every character have equal probability from the alphabet. The output is 100% random, ie. no sensible English appears by anymeans. So we set it as our baseline: language is not random.

Level 1 (Char-1: English Character Frequencies)

```
etewr.eri ?tnfeoole letoroo tehc eoeiet h vaia  eetonaowlptn tfteunrdi.eqrne
r truwnowasrhau inp tmr
```

In this we use real frequency distributions between characters from English text and we can observe improvement in output. We can say this as it can be seen that spaces are appearing more frequently and letters grouped together are more naturally. Common letters like ‘e’ and ‘t’ appear more frequently. However, the text is still not readable, this is because individual letter frequencies alone cannot determine the output sufficiently.

Level 2 (Char-2: Character Bigrams)

```
a nshi hug witizawar tut ingirthan waco igides. ispis s canst m youre
bllitwad h auppeho winered he
```

In the second order, ie. By using two-character sequences (bigrams), we can now see word-like patterns generating. Recognizable letter combinations appears, for eg. , “nshi”, “witizawar” and “youre.” The condition that some characters are followed by others are more frequentl than random chance begins to impose structure. We can also see that spaces and punctuations looks more natural.

Level 3 (Char-3: Character Trigrams)

```
ns was an the hime reethersted thappyrint ingre towis am easeld spal a se of
sheyearcy's wer wo_ are
```

With trigrams, we can see actual words of English starts appearing, for eg. “was”, “an”, “the” and “am”. The model then learned enough context to produce fragments that resembles English. We can point out single words, although we were not able to figureout overall meaning of the sentence. This means that the model “knows” how English characters are typically clusters.

This progression shows us that Character-level generation follows Shannon’s core principle, ie. languages have a statistical structure. We can say that character level conditions (which letters follow which other letters) is sufficient to convert from absolute nonsense to something that looks like English.

Word-Level Analysis

The progression becomes even better when we move from character-level to word-level models, where the units are analyse are complete words rather than individual characters.

Level 1 (Word-1: Unigram Words)

```
Words. Those on robinson which her had i her cases as an what had and she
distress by kitty and it choice service his decorum safe much elizabeth been
of.
```

Word unigrams select words independently based on their frequency in the source text. The result is a “word salad”, every word is valid English, pulled directly from Austen, but the sequence makes no sense. Word order is completely random. This level reveals a critical constraint: knowing which words are common is not enough to generate coherent text.

Level 2 (Word-2: Word Bigrams)

```
That their inclinations of _that_. Enough and her own thoughts were well as
absolutely.
```

With bigrams, word pairs impose constraints. We see improvements: “That their,” “her own,” “were well” are all plausible phrases that could appear in Austen. The model has learned which words tend to follow which other words. However, the output is still fragmented and lacks narrative coherence. Sentences end abruptly or with grammatical awkwardness.

Level 3 (Word-3: Word Trigrams)

```
For been so well that was impossible not to be dancing while she remains with
us the house. Waste eating anything and persevered therefore in requiring an
explanation of what had been very miserable b...
```

At the third order with words, we suddenly have nearly coherent text. The model produces complete grammatical sentences: “it was impossible not to be dancing,” “she remains with us,” “what had been very miserable.” Reading this output, one might briefly forget it was machine-generated. The model has internalized enough of Austen’s word sequences to produce extended passages that follow English grammar and maintain themes across multiple words.

Critical Observation: The jump from word-2 to word-3 represents a threshold where quantity becomes quality. With enough context (three-word sequences), the model captures enough of language’s statistical structure to produce convincing output.

---

2. Author Distinctiveness: Recognizing Writing Styles

A secondary question emerges from examining multiple authors: Does an n-gram model capture author-specific style? And at what approximation level does author distinctiveness become apparent?

Comparing All Three Authors at Word-3 Level

Jane Austen (Word-3):

```
A made him a most delightful. Backwards her mother's rapacity for whist players and in the guardianship of miss.
```

Austen’s output, even generated, carries her stylistic fingerprints. We see focus on social hierarchies (“whist players,” “guardianship”), domestic concerns (“her mother”), and formal vocabulary (“rapacity,” “delightful”). The broken grammar (“A made him”) aside, the vocabulary is distinctly Austen’s preoccupation with romance and social propriety.

Mark Twain (Word-3):

```
Tom's when this thought broke her down the street and. Won't and he see she.
```

Twain’s output is noticeably shorter and more colloquial. The fragmented nature (“Tom’s when,” “Won’t and he see she”) suggests Twain’s conversational, direct style. Though the output is grammatically broken, the vocabulary and brevity are recognizable.

Arthur Conan Doyle (Word-3):

```
Pathway place whence it had dropped in my life has been for better men before you come with. Long down completely he has not been wasted since it was not sure that you will find it rather disconnected i fear that i had gained the cover was a chestnut tired-looking.
```

Doyle’s output is longer, denser, and more convoluted. Phrases like “Pathway place whence it had dropped,” “I fear that I had gained the cover,” and “chestnut tired-looking” reflect Doyle’s Victorian mystery narrative style. The model captures Doyle’s propensity for elaborate description and deduction-focused language.

When Does Distinctiveness Emerge?

Character-Level (Char-0 through Char-3): No author distinctiveness whatsoever. All three authors produce equally nonsensical output at character levels. The character-level constraints are too weak to preserve author identity.

Word-1 (Unigram): Minimal distinctiveness. All three authors' unigrams produce word salads, though a careful reader might notice that Austen's pool includes more names like "Elizabeth" and "Bennet," while Twain's includes "Tom," and Doyle's includes "Watson" and "Holmes."

Word-2 (Bigram): Distinctiveness begins to emerge. Austen's bigrams contain more formal speech patterns ("their inclinations," "own thoughts"), Twain's contains more action-oriented phrases, and Doyle's contains more analytical constructions.

Word-3 (Trigram): Full distinctiveness. At the trigram level, the model captures enough linguistic context to produce author-specific outputs that are readily distinguishable.

Why This Matters

This demonstrates that **author style is not purely in vocabulary but in sequential patterns**. The trigram-based Markov chain learns not just which words each author uses, but the *order and combination* in which words appear. Austen's preference for certain phrase patterns, Twain's narrative rhythm, and Doyle's descriptive elaboration are all encoded in the statistical structure of word sequences.

---

3. Anchor Word Integration: Maintaining Naturalness While Constraining Generation

The challenge of integrating required anchor words while maintaining natural flow reveals important limitations of n-gram models.

Austen with Anchors (Word-2, Requested: "elizabeth," "bennet," "pride")

```
She. Room pride sir.. To bennet long.
```

Doyle with Anchors (Word-2, Requested: "elementary," "watson," "deduce")

```
Out deduce encompass.. Young lady on the unconscious man dark room he 'here's
another. That elementary there.
```

Analysis: The anchor word integration reveals a fundamental tension. The requested anchor words do appear in the output ("pride," "bennet," "elementary," "deduce," "watson"), but their inclusion seriously degrades text quality. Sentences become fragmented and ungrammatical ("She. Room pride sir..").

The technical challenge is that anchor words may not form natural word-pair or word-triplet sequences in the source text. When forced to include "pride" at word-2 level, the model must deviate from its learned patterns, resulting in awkward constructions. At

word-3, the model likely cannot find valid trigrams that naturally integrate these words, so it breaks the sentence structure to accommodate them.

Trade-off: Perfect anchor word inclusion comes at the cost of coherence. A more sophisticated approach would weight anchor word integration probabilistically rather than requiring hard constraints, but this is beyond what n-gram models can express naturally.

---

4. Modern Connections: From n-Grams to Transformers

Shannon's foundational insight that language can be modeled as a stochastic process using n-gram frequencies directly led to modern language models. However, contemporary systems like GPT, BERT, and Claude have fundamentally improved upon n-gram models by addressing their core limitations.

The Relationship Between n-Grams and Modern LLMs

Both n-gram models and modern neural language models operate on the same principle: **predicting the next token given previous context**. An n-gram model computes P(word_n | word_{n-1}, word_{n-2}, ...). A transformer-based language model does the same thing, but using learned attention mechanisms rather than fixed-window statistics.

Key Limitations of n-Gram Models Our Experiments Reveal

1. Fixed Context Window: Our word-3 model has access only to the previous two words when predicting the next word. As we see in the Austen output, long-range dependencies are completely missed: "For been so well that was impossible not to be dancing while she remains with us the house." The model cannot maintain narrative coherence across 15+ words because it has no memory of what was said 10 words ago.

Modern solution: Transformer models with attention mechanisms can attend to relevant tokens anywhere in the input, not just the immediate previous context. GPT-3 can attend to tokens up to 2,048 positions ago; GPT-4 extends this to 8,000-32,000 tokens.

2. Data Sparsity and Exponential Growth: To generate word-4 models, we would need to store every four-word sequence and its frequency. The number of possible four-word sequences in English is approximately 14 billion (if using a 100,000-word vocabulary). Most of these sequences never appear in even massive training corpora, making higher-order n-grams impractical.

Modern solution: Neural networks learn dense, continuous representations that generalize across similar contexts without requiring every sequence to be explicitly observed in training data. A transformer can effectively generalize to novel combinations of words because it learns semantic relationships, not just surface statistics.

3. No Semantic Understanding: The word-3 model treats "king" and "queen" as completely unrelated tokens. It has no knowledge that queens are female and kings are

male, or that they occupy similar positions in society. It cannot reason about relationships between words, it only knows which sequences of words typically follow each other.

Modern solution: Neural embeddings (like those in BERT or GPT) capture semantic similarities. "King" and "queen" are represented as nearby points in a high-dimensional vector space, encoding their conceptual similarity.

4. No World Knowledge or Reasoning: Our model cannot understand whether a statement is true or plausible. It could generate "The sun rose in the west" if such a sequence appeared in Austen (it didn't, but if it had, the model would happily reproduce it).

Modern solution: Large language models, trained on billions of tokens from diverse sources, develop some form of implicit world knowledge through pattern recognition at scale. While not perfect, they have absorbed enough information about the world that they are less likely to generate obvious contradictions.

5. Repetition and Loop Entrapment: In our generated samples, we see incomplete or trailing text ("what had been very miserable b…"). With n-gram models, once a word sequence exists in the training data, that sequence becomes perpetually likely. The model can get "stuck" generating the same phrase repeatedly.

6. Incoherence Over Long Passages: Even our best model (word-3 Austen: "For been so well that was impossible not to be dancing while she remains with us the house. Waste eating anything…") loses coherence mid-way through. The second sentence ("Waste eating anything") has nothing to do with the first. The model has no global "planning" mechanism.

Modern solution: Transformer models, through mechanisms like multi-head attention and careful architectural design, can maintain higher-level coherence across entire documents or conversations.

Why Neural Networks Win

The gap between our n-gram model and GPT-3 fundamentally comes down to **learning capacity and representational power**. An n-gram model is essentially a lookup table: given context X, output the most frequent completion. A neural network learns a complex, multi-layered function that captures intricate patterns of language, including long-range dependencies, semantic relationships, and even reasoning patterns.

Put another way: our trigram model memorizes; modern LLMs understand patterns in a way that transfers to novel situations.

---

5. Shannon's Insight: Language as a Stochastic Process

The Core Insight

When Claude Shannon published "A Mathematical Theory of Communication" in 1948, he made a radical claim: **language is not fundamentally deterministic or rule-based, but probabilistic**. English speakers do not follow explicit grammar rules (most cannot

articulate them), yet they produce comprehensible sentences. Why? Because language has statistical structure ie. certain patterns are far more probable than others.

Shannon demonstrated this by showing that increasingly sophisticated n-gram models could approximate English text with startling fidelity. Zero-order approximation (char-0) produces gibberish. First-order (char-1) includes English letter frequencies. Higher orders produce recognizable words and then plausible sentences.

What Our Experiments Prove

Our implementation validates Shannon's insight at every level:

Char-0 proves: Pure randomness is unintelligible. Language is not random.

Char-1 proves: English letter frequencies matter. Knowing that 'e' appears ~13% of the time while 'z' appears ~0.07% of the time is already useful information.

Char-2/Char-3 prove: Context is crucial. Letters don't appear independently; they are constrained by their neighbors. The fact that char-3 produces recognizable English words shows that three-character context is sufficient to capture basic English phonotactics.

Word-1 proves: Word frequency follows a power law. A tiny number of words (the, of, and, to, a) dominate, while most words are rare. Yet simply knowing word frequencies produces gibberish order matters more than frequency.

Word-2/Word-3 prove: Language is deeply contextualized. Two-word and three-word sequences encode enough structure to approximate coherent English. The fact that word-3 generates grammatical sentences with plausible vocabulary reveals that Austen's writing, and by extension, English itself and is substantially predictable from local context.

Philosophical Implications

Shannon's insight has profound implications:

1. Understanding vs. Pattern Recognition: If we can approximate human language using only n-gram statistics—without any explicit understanding, no semantic knowledge, no reasoning—what does this tell us about language itself? It suggests that a significant portion of what we think of as "understanding language" might actually be sophisticated pattern recognition. We may understand language better than our model, but not in a fundamentally different way—just with more context and more patterns.

2. Language is Redundant: The fact that we can reproduce plausible English using only short-range statistics proves that English is highly redundant. A huge amount of information is encoded in sequential patterns. This redundancy is why we can understand a sentence even with typos or missing words—the surrounding context constrains what could come next.

3. The Boundary Between Structure and Randomness: Language sits on an interesting boundary. It is not purely deterministic (you cannot predict what a speaker will say next

with certainty), nor is it purely random (random text is gibberish). Instead, it is constrained chaos—probabilistic processes with strong patterns.

Connection to Modern AI

Shannon's work is the direct ancestor of modern deep learning approaches to NLP. Transformers, attention mechanisms, and large language models are sophisticated elaborations on Shannon's basic insight: language is predictable from context. Modern models simply have:

- Longer context windows (attended to relevant parts of much longer histories)
- Richer representations (learned embeddings that capture semantic similarity)
- Multiple layers of abstraction (allowing the capture of hierarchical patterns)
- Orders of magnitude more data (enabling learning of complex statistical relationships)

But the fundamental principle that language is a stochastic process whose patterns can be learned and exploited for generation is Shannon's.

Conclusion

This assignment demonstrates why Shannon's work remains foundational to artificial intelligence. By building n-gram models at increasing orders of sophistication, we have shown that:

1. Language exhibits hierarchical statistical structure, from individual character frequencies to multi-word patterns.
2. Higher-order models capture progressively more of language's structure, producing qualitatively better approximations.
3. Author-specific styles are encoded in sequential patterns and emerge at sufficient approximation levels.
4. Yet even our best models reveal the limitations of local, context-free approaches: they cannot maintain coherence over long passages, reason about meaning, or adapt to novel contexts.

These limitations motivate the development of neural approaches that learn richer representations and can attend to broader context. The path from Shannon's 1948 insight to modern large language models is a direct line: each generation of models addresses the limitations of the previous generation while preserving the core principle that language is a learnable statistical process.

Our implementation brings Shannon's beautiful ideas to life, demonstrating both their power and their limits. In doing so, it illustrates how foundational theory, implemented and tested empirically, drives progress in AI.