CSC_51063 Final : Prediction and Entropy of Printed English (German, French, etc)

Yanting Liu, 2025-12-15

Prediction and Entropy of Printed English

By C. E. SHANNON

(*Manuscript Received Sept. 15, 1950*)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

1. Introduction

IN A previous paper[1] the entropy and redundancy of a language have been defined. The entropy is a statistical parameter which measures, in a certain sense, how much information is produced on the average for each letter of a text in the language. If the language is translated into binary digits (0 or 1) in the most efficient way, the entropy H is the average number of binary digits required per letter of the original language. The redundancy, on the other hand, measures the amount of constraint imposed on a text in the language due to its statistical structure, e.g., in English the high frequency of the letter E, the strong tendency of H to follow T or of U to follow Q. It was estimated that when statistical effects extending over not more than eight letters are considered the entropy is roughly 2.3 bits per letter, the redundancy about 50 per cent.

Since then a new method has been found for estimating these quantities, which is more sensitive and takes account of long range statistics, influences extending over phrases, sentences, etc. This method is based on a study of the predictability of English; how well can the next letter of a text be predicted when the preceding N letters are known. The results of some experiments in prediction will be given, and a theoretical analysis of some of the properties of ideal prediction. By combining the experimental and theoretical results it is possible to estimate upper and lower bounds for the entropy and redundancy. From this analysis it appears that, in ordinary literary English, the long range statistical effects (up to 100 letters) reduce the entropy to something of the order of one bit per letter, with a corresponding redundancy of roughly 75%. The redundancy may be still higher when structure extending over paragraphs, chapters, etc. is included. However, as the lengths involved are increased, the parameters in question become more

[1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, v. 27, pp. 379–423, 623–656, July, October, 1948.

Table of Contents

Problem Framing

Lecture 3 Example

THE ENGLISH LANGUAGE(I)

- Approximation of order 0 $\left(\text{all symbols, don't forget the "space", are i.i.d.}\right)$:

  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

  $H_0 = \log_2 27 \approx 4.76$

- Approximation of order 1 $\left(\text{the letters are chosen according to their frequency in English}\right)$

  OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

  $H_1 \approx 4.03$

- Approximation of order 2 : same distribution of the pairs as in English

  ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

- Approximation of order 3 : same frequency of the triplets as in English

  IN NO IST LAT WHEY CRATICT FROURE BERS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

Any idea to generate correctly some English test?

2) Entropy, Entropy Rate and Redundancy

N-gram entropies from text

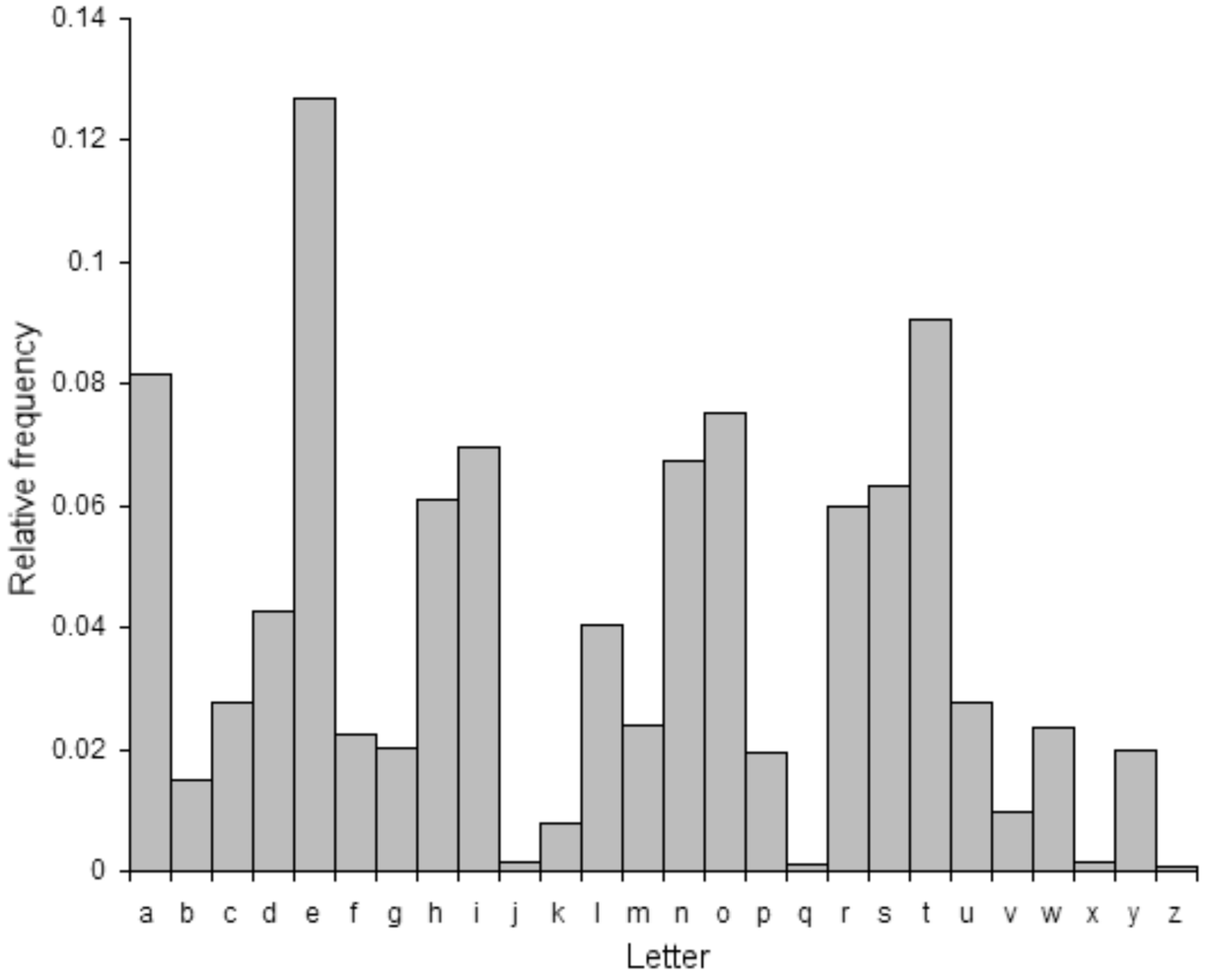
Introduction and Definitions

- Zero Order Model
- Maximum entropy (uniform letters) = 4.75 bits per letter
- given 26 letters + whitespace and assuming that all characters are equally likely

- First Order Model takes character frequencies into account

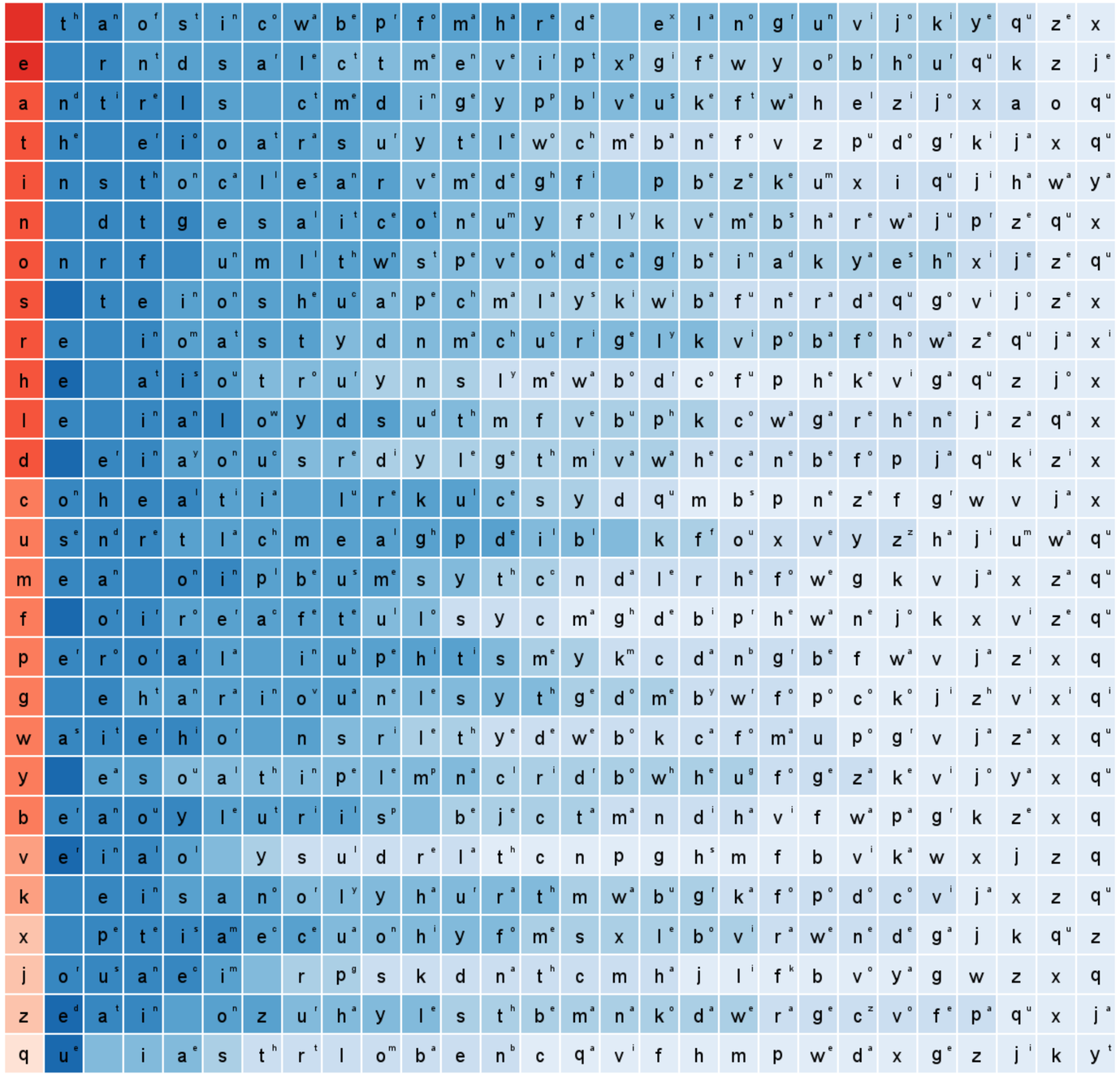
$$H_1 = H(X_t) = -\sum_x p(x) \log p(x)$$

- However this is assuming zero memory, ie characters appear independently from what was written before


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Relative frequency
a b c d e f g h i j k l m n o p q r s t u v w x y z
Letter

Letter and next-letter frequencies in English

measured across 1 million articles from Wikipedia



Colour key

a — Letters and spaces sorted by overall frequency. Ignores case and accents.

n^d — Next letter sorted by frequency. Small letter is the most common third letter following the pair.

Blank letters indicate spaces

Letter frequencies

- 31.6% to 100.0%
- 10.0% to 31.6%
- 3.2% to 10.0%
- 1.0% to 3.2%
- 0.3% to 1.0%
- 0.1% to 0.3%
- 0.0% to 0.1%

Markov generators

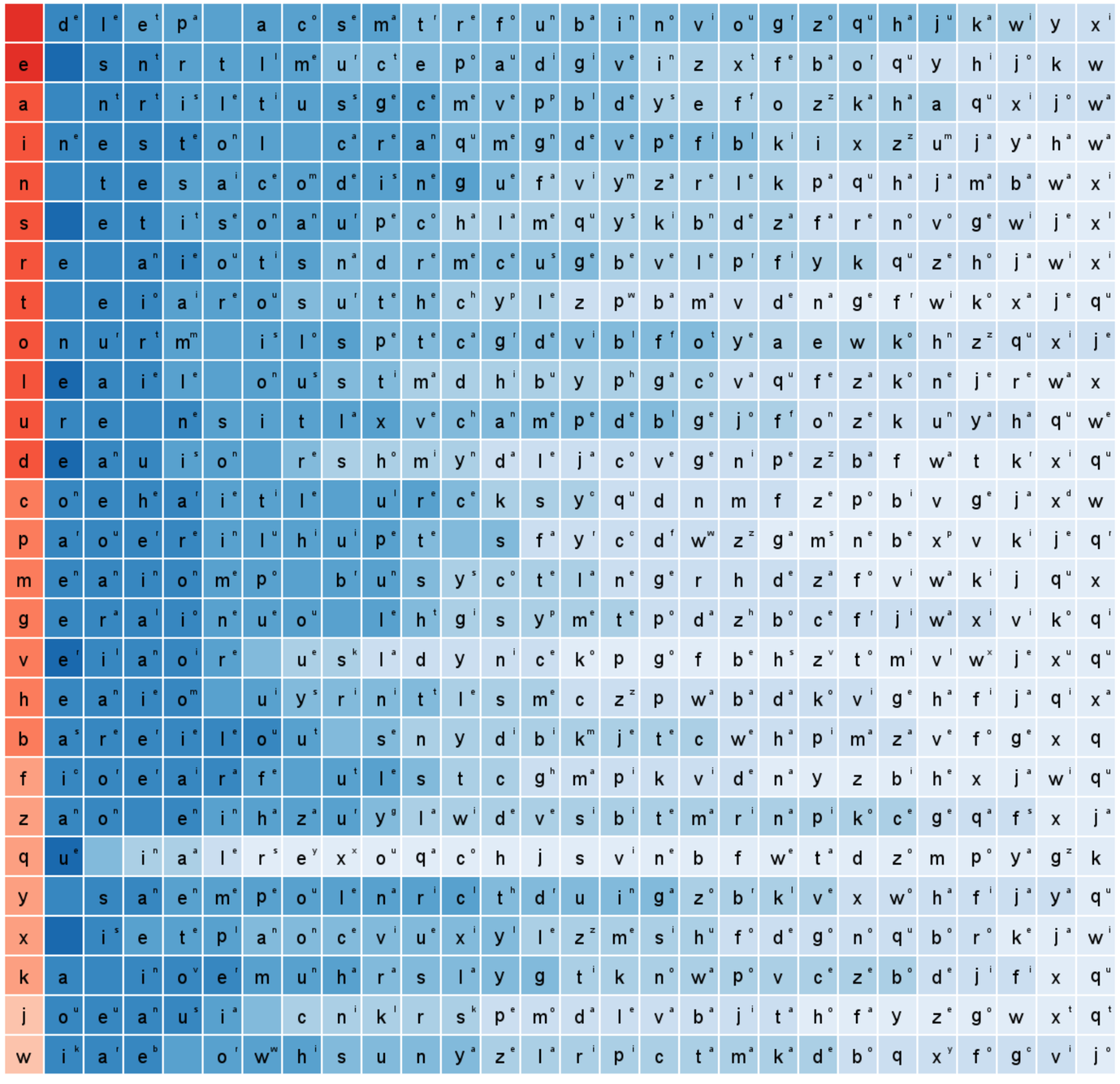
The letters distributions in the chart can be used to generate psuedowords such as the ones below. A similar approach, at the word level, is used for online parody generators.

bastrabot	*dithely*
loctrion	*raliket*
calpereek	*amorth*
forliatitive	*asocult*
wasions	*quarm*
felogy	*winferlifterand*
sonsih	*uniso*
fourn	*hise*
meembege	*nuouish*
prouning	*guncelawits*
nown	*rectere*
abrip	*doesium*

/u/Udzu

Letter and next-letter frequencies in French

measured across a selection of articles from Wikipedia



Colour key

a — Letters and spaces sorted by overall frequency. Ignores case and accents.

n^d — Next letter sorted by frequency. Small letter is the most common third letter following the pair.

Blank letters indicate spaces

Letter frequencies

- 31.6% to 100.0%
- 10.0% to 31.6%
- 3.2% to 10.0%
- 1.0% to 3.2%
- 0.3% to 1.0%
- 0.1% to 0.3%
- 0.0% to 0.1%

Markov generators

The letters distributions in the chart can be used to generate psuedowords such as the ones below. A similar approach, at the word level, is used for online parody generators.

cillesil	*sulskini*
lidhemin	*plumeme*
bachogine	*crout*
taphie	*provicas*
copit	*odzzaccet*
extreiles	*pipiphien*
chetratagne	*outif*
suro	*extellages*
nans	*nutopune*
entote	*sporese*
zhiquis	*edes*
aliet	*randamelle*

/u/Udzu

N-gram entropies from text

Introduction and Definitions

- we now take ***context*** into account and list the most common digrams in english
- these are in order : th, he, in, en, nt, re, er, an, ti, es, on, at, se, nd …

- this is where conditional entropy comes into play !
- from the digrams we obtain $P(X_t \mid X_{t-1})$
- we can define $H_2 = H(X_t \mid X_{t-1}) = \Sigma_{x,y} p(x,y)\,[-\log p(y \mid x)]$
- Shannon calls this 2nd order entropy (1st order Markov chain)

N-gram entropies from text

Introduction and Definitions

- but this is still not enough, English is not 1st order Markov chain
- ‘tio’ almost always forces ’n’ where the context required is length 3
- also word-level and semantic constraints extend far beyond characters
- Shannon’s idea = if we don’t know the true dependency length -> approximate it with longer and longer finite contexts
- and the “true” entropy rate will be

$$H = \lim_{N \to \infty} H_N$$

- where H_N = conditional entropy of the next symbol given the previous N-1

$$H_N = H\left(X_t \mid X_{t-1}, \ldots, X_{t-(N-1)}\right) = \sum_{x_1, \ldots, x_N} p(x_1, \ldots, x_N)\left[-\log p(x_N \mid x_1, \ldots, x_{N-1})\right]$$

N-gram entropies from text

Introduction and Definitions

- Redundancy is the gap between Entropy and Encoding Efficiency
- Redundancy $R = 1 - \frac{H}{H_{\max}}$ where $H_{\max} = \log_2 |\mathcal{A}|$
- Redundancy means you can delete or corrupt part of the message and still recover it.
- eg : ths sntnc s stll rdbl -> this sentence is still readable

3) Character-level Entropy

Experiment Setting

- 30 books from project Gutenberg downloaded as .txt files
- Clean and tokenized either as char or words
- Markov chain of orders k=1,2,3 … fitted on the corpora
- Calculate entropy and generate text samples

English	Carroll - Alice's Adventures in Wonderland Fitzgerald - The Great Gatsby Hobbes - Leviathan Austen - Pride and Prejudice Shakespeare - Romeo and Juliet Machiavelli - The Prince Dickens - A Tale of Two Cities Tolstoy - War and Peace Dostoevsky - The Brothers Karamazov Dostoevsky - Crime and Punishment
French	Voltaire - Candide, ou l'Optimisme Leroux - Le Fantôme de l'Opéra Flaubert - Madame Bovary Dumas - Le Comte de Monte-Cristo Hugo - Notre-Dame de Paris Stendhal - Le Rouge et le Noir Dumas - Les Trois Mousquetaires Verne - Voyage au centre de la Terre Verne - L'Île mystérieuse Verne - De la Terre à la Lune Verne - Le Tour du monde en quatre-vingts jours Verne - Vingt mille lieues sous les mers
German	Kant - Der Wille zur Macht Kant - Kritik Hermann Hesse - Steppenwolf Struwwelpeter Kafka - Die Verwandlung Kafka - Das Urteil Mann - Zauberberg 1&2

N-gram entropies from text

Comparison with Result from Shannon

- Shannon also calculated N-gram entropies for small values of N
- Based on the book “Secret and Urgent” by Fletcher Pratt
- Unfortunately I couldn’t find a txt version online so i opted for another book by the same author
- The Blue Star by Fletcher Pratt written in 1950

The N-gram entropies F_N for small values of N can be calculated from standard tables of letter, digram and trigram frequencies.[2] If spaces and punctuation are ignored we have a twenty-six letter alphabet and F_0 may be taken (by definition) to be $\log_2 26$, or 4.7 bits per letter. F_1 involves letter frequencies and is given by

$$F_1 = -\sum_{i=1}^{26} p(i) \log_2 p(i) = \text{4.14 bits per letter.} \quad (3)$$

The digram approximation F_2 gives the result

$$\begin{aligned} F_2 &= -\sum_{i,j} p(i,j) \log_2 p_i(j) \\ &= -\sum_{i,j} p(i,j) \log_2 p(i,j) + \sum_i p(i) \log_2 p(i) \quad (4) \\ &= 7.70 - 4.14 = \text{3.56 bits per letter.} \end{aligned}$$

[2] Fletcher Pratt, “Secret and Urgent,” Blue Ribbon Books, 1942.

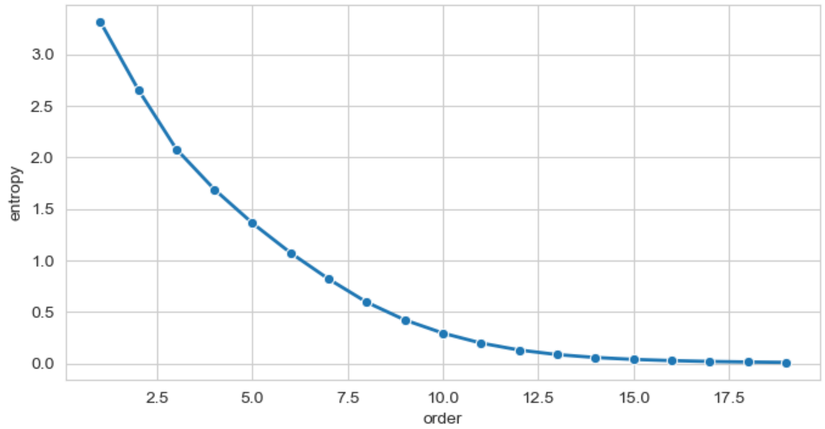
	F_0	F_1	F_2	F_3	F_{word}
26 letter	4.70	4.14	3.56	3.3	2.62
27 letter	4.76	4.03	3.32	3.1	2.14

Results from Shannon Paper

N-gram entropies from text

Empirical Estimates from Text

- Import and tokenize the text
- For all N-grams count distinct next step tokens and normalize to probability
- Calculate entropy

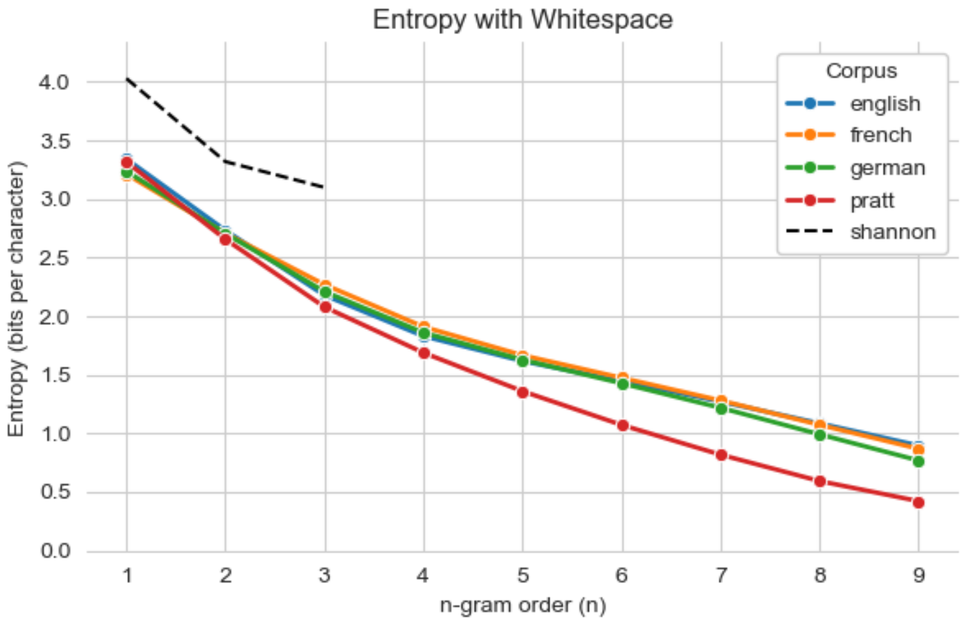


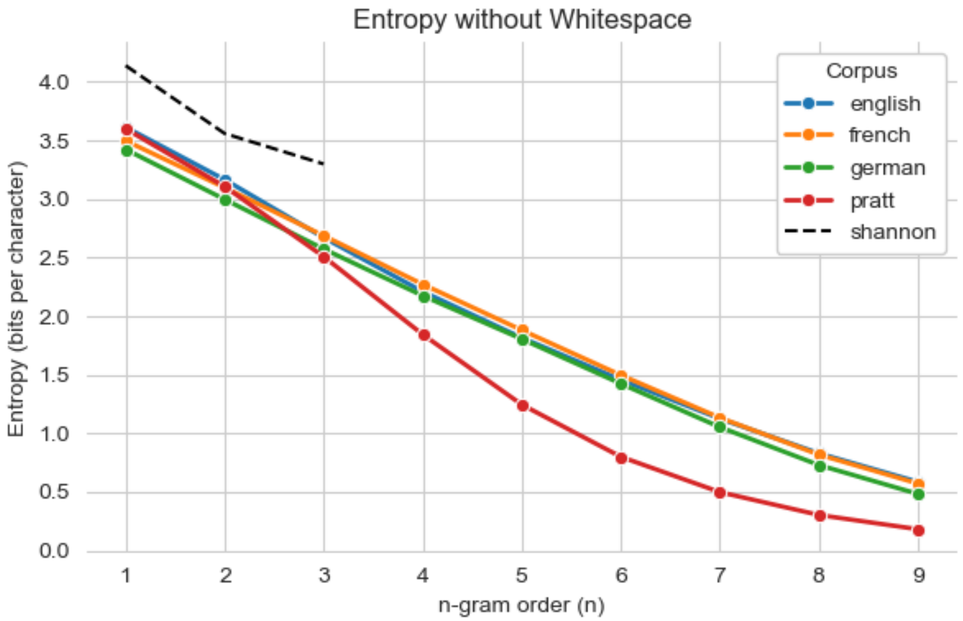

```
Alice's Adventures in Wonderland

                              CHAPTER I

                        Down the Rabbit-Hole

  Alice was beginning to get very tired of sitting by her sister
on the bank, and of having nothing to do:  once or twice she had
pee

Total tokens: 143407
Unique tokens: 31
['\n' ' ' ',' '.' '?' 'a' 'b' 'c' 'd' 'e' 'f' 'g' 'h' 'i' 'j' 'k' 'l' 'm'
 'n' 'o' 'p' 'q' 'r' 's' 't' 'u' 'v' 'w' 'x' 'y' 'z']
```

```
----- Model Statistics -----
Order of N-gram Model (N)               : 2
Computation Time (s)                    : 0.0241
Number of Unique Contexts:              : 544
Total Observed N-grams (Transitions)    : 143,405
Unique (Context, Token) N-grams         : 4,211
Conditional Vocab Size (Next Tokens)    : 31

Most likely token after ('s', ' ') :
   token     proba
5      t  0.138748
0      a  0.131980
8      s  0.097011
16        0.077270
1      i  0.068246

 Model Entropy: 2.5795 bits
```

Char-level Entropy Across Languages

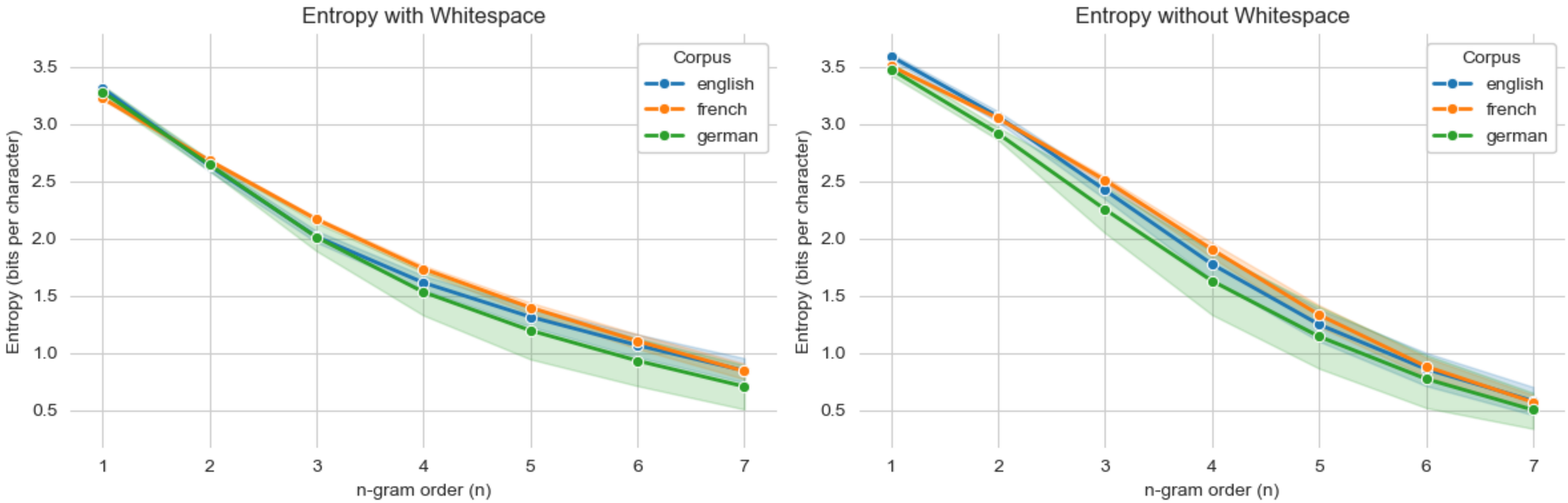
Comparison with tabulated values from Shannon





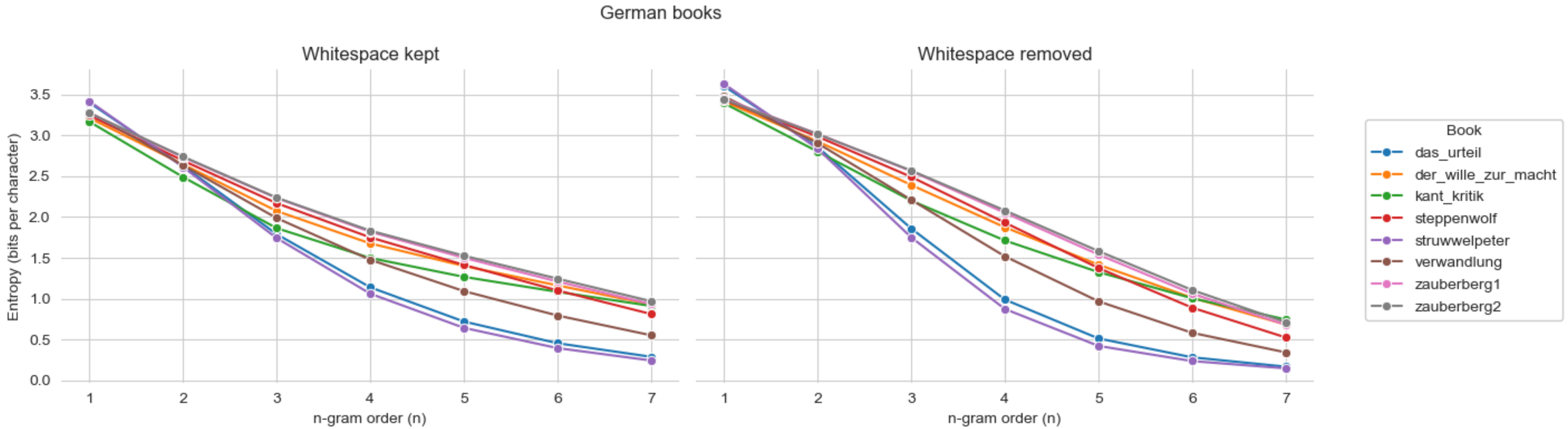
Char-level Entropy Across Languages

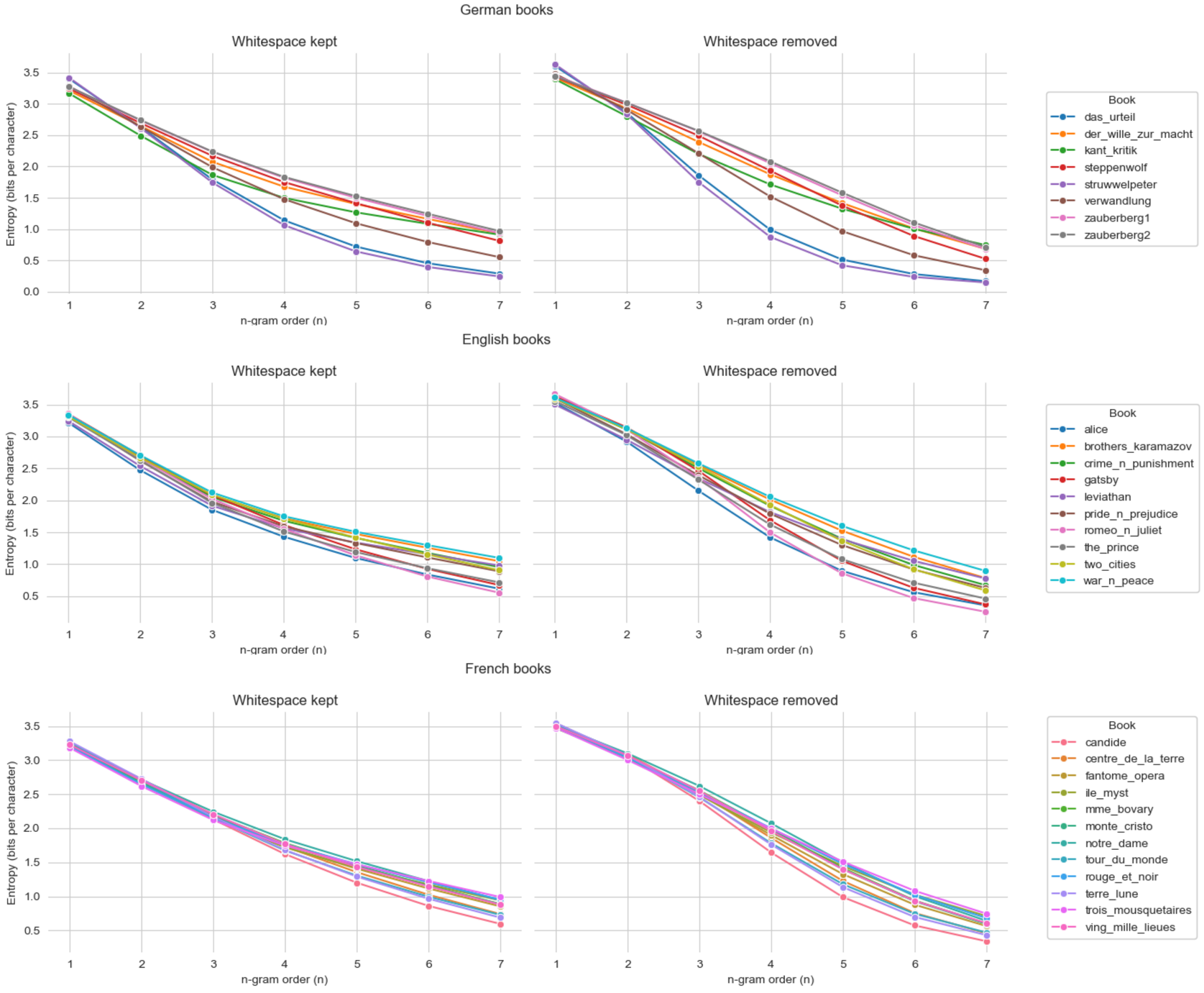
Book level analysis for each language



Char-level Entropy Across Languages

Book level analysis for each language




German books
Whitespace kept
Whitespace removed
Entropy (bits per character)
n-gram order (n)
Book
das_urteil
der_wille_zur_macht
kant_kritik
steppenwolf
struwwelpeter
verwandlung
zauberberg1
zauberberg2
English books
Whitespace kept
Whitespace removed
Entropy (bits per character)
n-gram order (n)
Book
alice
brothers_karamazov
crime_n_punishment
gatsby
leviathan
pride_n_prejudice
romeo_n_juliet
the_prince
two_cities
war_n_peace
French books
Whitespace kept
Whitespace removed
Entropy (bits per character)
n-gram order (n)
Book
candide
centre_de_la_terre
fantome_opera
ile_myst
mme_bovary
monte_cristo
notre_dame
tour_du_monde
rouge_et_noir
terre_lune
trois_mousquetaires
ving_mille_lieues

4) Word-level Entropy and Zipf’s Law

N-gram entropies from text

Comparison with Result from Shannon

- Shannon also calculated word-level N-gram entropies for small values of N
- But only based on estimation from tabulated word frequencies and obtaining 11.82 bits per word
- and then by taking the average word length to be 4.5 letters
- 11.82 / 4.5 = 2.62

The N-gram entropies F_N for small values of N can be calculated from standard tables of letter, digram and trigram frequencies.[2] If spaces and punctuation are ignored we have a twenty-six letter alphabet and F_0 may be taken (by definition) to be $\log_2 26$, or 4.7 bits per letter. F_1 involves letter frequencies and is given by

$$F_1 = -\sum_{i=1}^{26} p(i) \log_2 p(i) = 4.14 \text{ bits per letter.} \quad (3)$$

The digram approximation F_2 gives the result

$$\begin{aligned} F_2 &= -\sum_{i,j} p(i,j) \log_2 p_i(j) \\ &= -\sum_{i,j} p(i,j) \log_2 p(i,j) + \sum_i p(i) \log_2 p(i) \quad (4) \\ &= 7.70 - 4.14 = 3.56 \text{ bits per letter.} \end{aligned}$$

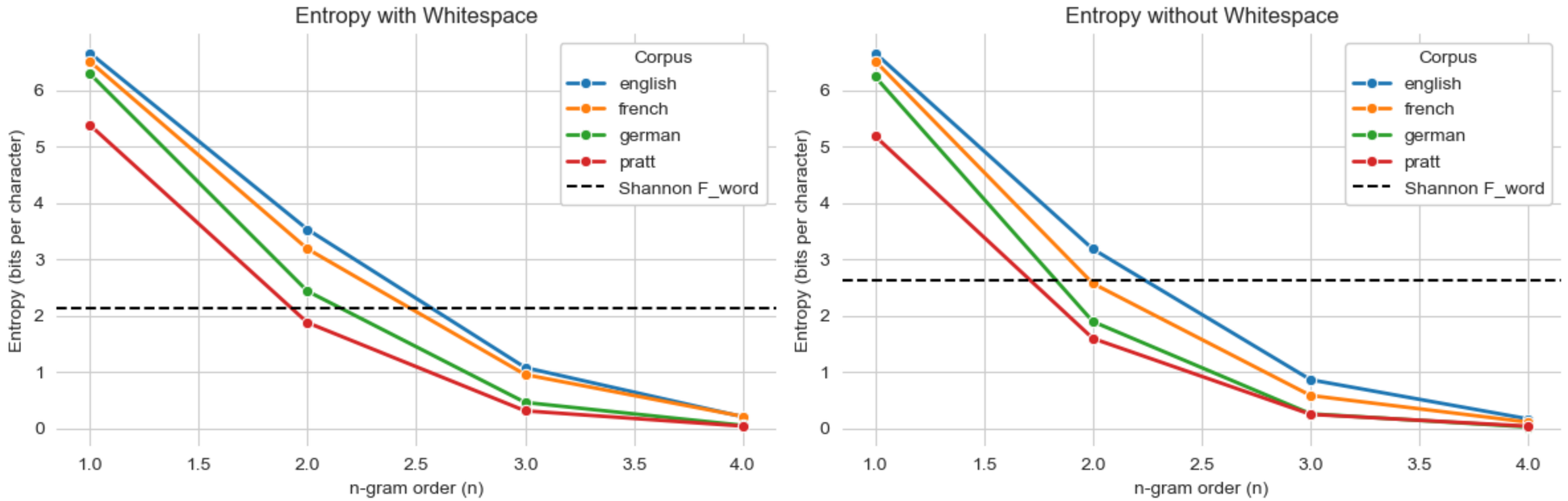
[2] Fletcher Pratt, "Secret and Urgent," Blue Ribbon Books, 1942.

	F_0	F_1	F_2	F_3	F_{word}
26 letter	4.70	4.14	3.56	3.3	2.62
27 letter	4.76	4.03	3.32	3.1	2.14

Results from Shannon Paper

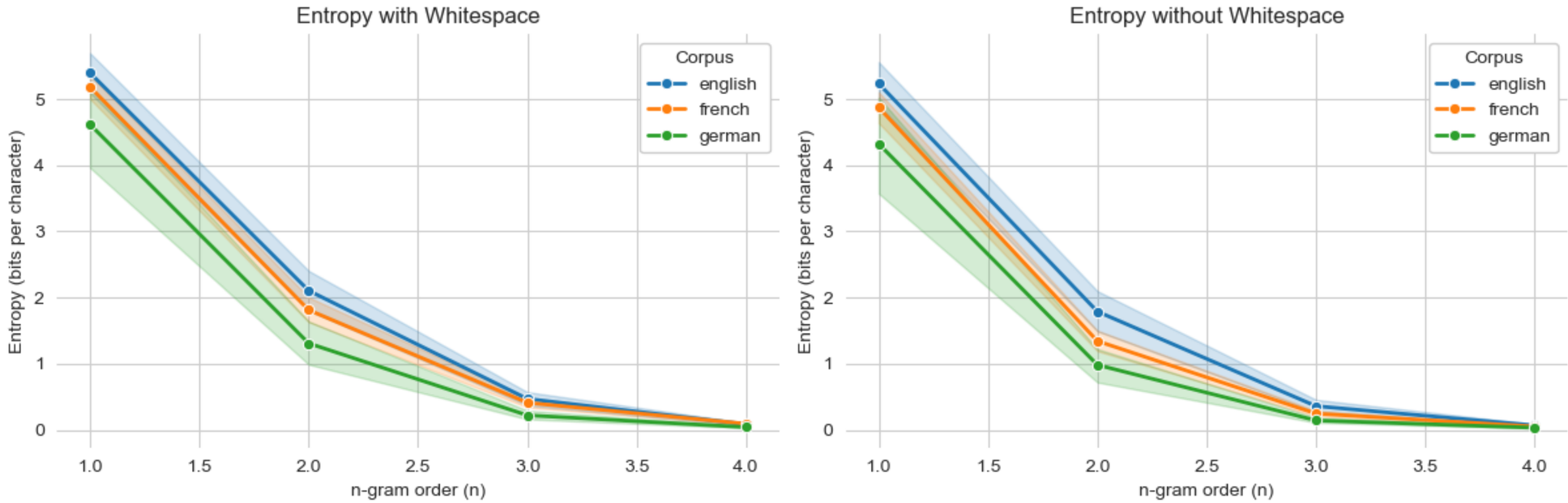
Char-level Entropy Across Languages

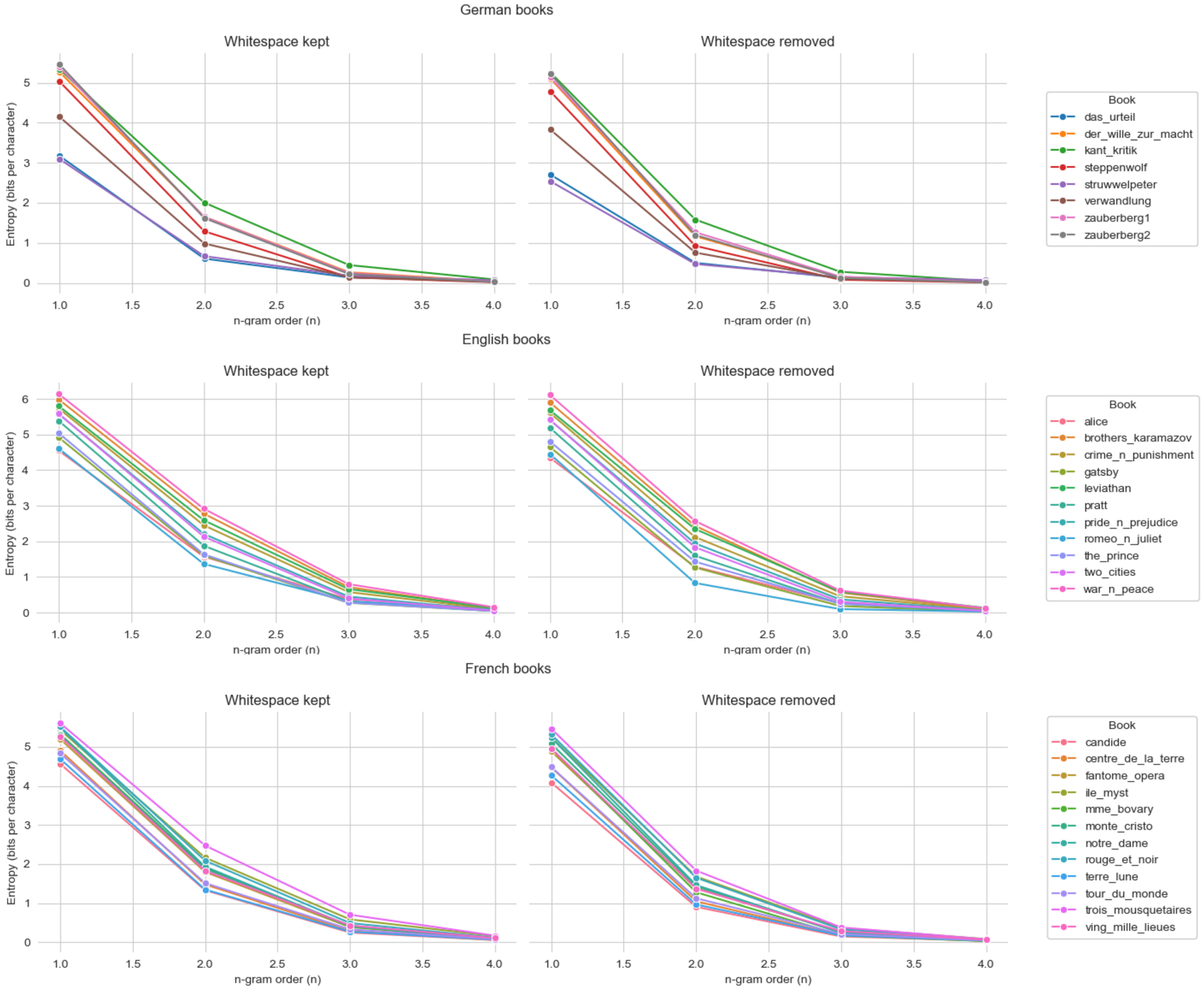
Comparison with tabulated values from Shannon



Char-level Entropy Across Languages

Book level analysis for each language




German books
Whitespace kept
Whitespace removed
Entropy (bits per character)
n-gram order (n)
Book
das_urteil
der_wille_zur_macht
kant_kritik
steppenwolf
struwwelpeter
verwandlung
zauberberg1
zauberberg2
English books
Whitespace kept
Whitespace removed
Entropy (bits per character)
n-gram order (n)
Book
alice
brothers_karamazov
crime_n_punishment
gatsby
leviathan
pratt
pride_n_prejudice
romeo_n_juliet
the_prince
two_cities
war_n_peace
French books
Whitespace kept
Whitespace removed
Entropy (bits per character)
n-gram order (n)
Book
candide
centre_de_la_terre
fantome_opera
ile_myst
mme_bovary
monte_cristo
notre_dame
rouge_et_noir
terre_lune
tour_du_monde
trois_mousquetaires
ving_mille_lieues

Excursion to Chinese

- Problem = Chinese does not have whitespaces between words
- therefore character-level n-grams are possible, but word-level n-grams are slightly more involved
- luckily there are libraries that do word segmentation by most frequently occurring words -> eg jieba
- 我来到巴黎综合理工学院 = I have come to École Polytechnique in Paris
  - 我 + 来到 + 巴黎 + 综合 + 理工 + 学院
  - 我 + 来到 + 巴黎 + 综合 + 理工学院

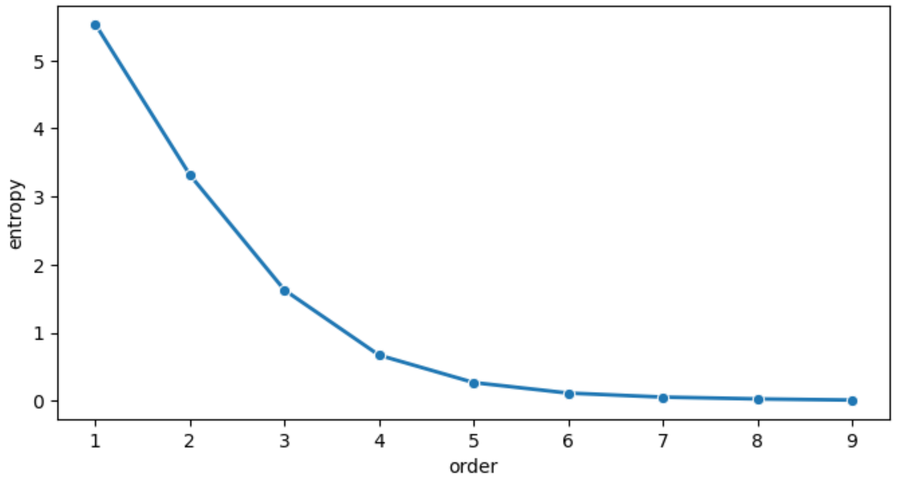
Example Model Output

《红楼梦》Dream of the Red Chamber (18th-century Chinese novel)

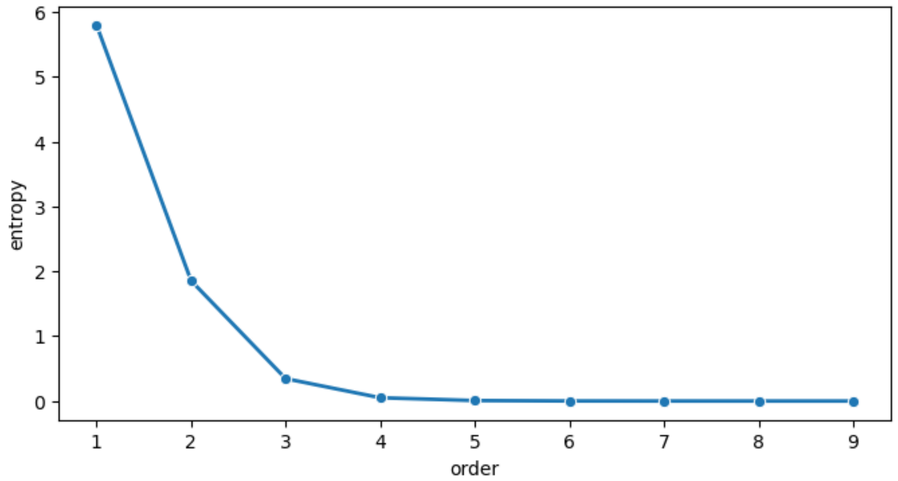
```
----- Model Statistics -----
Order of N-gram Model (N)               : 2
Computation Time (s)                    : 0.3320
Number of Unique Contexts:              : 144,941
Total Observed N-grams (Transitions)    : 812,964
Unique (Context, Token) N-grams         : 426,849
Conditional Vocab Size (Next Tokens)    : 4,270

Sample counts:
[(('紅', '樓'), Counter({'夢': 9})),
 (('樓', '夢'),
  Counter({'第': 2, '。': 2, '仙': 1, '稿': 1, '引': 1, '通': 1, '，': 1, '話': 1}))]

Most likely token after ('來', '，') :
    token     proba
64      我  0.035526
21      說  0.032566
11      只  0.029934
39      便  0.028618
26      一  0.027632
..    ...       ...
331     階  0.000329
332     謝  0.000329
335     淘  0.000329
338     賴  0.000329
539     執  0.000329
```

Chinese Char- and Word-level Entropy

《红楼梦》Dream of the Red Chamber (18th-century Chinese novel)



Char-level Entropy in Chinese



Word-level Entropy in Chinese

5) Prediction, Illustration, and Limits

Generated Text

Char-level, Order: 0

nncs oafa s hiw dhwdnaeahtuf haideflfs t v ieiasi hd ali sbnrtslssr kwfhdiia dtyo nosoe h an iieare h ilfly l oo nd o twxai leeip r rtns astase nneeeiuhniifq tsyharghtltfymnkuet tr ttb t yr a

Char-level, Order: 1

ate ysne as nd chiclangod mo rrer shinido n ar t len s anowoupyot r nd wen tenere ins terayerureacomokixpond rsestofi santhe ju gherouthe inghad tupo s ghi wh the t tr s herlincaise ieg s rdid tshensof

Char-level, Order: 2

gforyphould lit isher for to the se of threepeal he cousenchou ke mark there it alselself wed the the sh sheriethey sing oreme begaid if a cre why quithe crome the ort of ful off to up a lown what wed v

Char-level, Order: 3

o me more queen head if you not at as she word his it nothe said the the can the madedly you cook ture here thatter dormoutoff anxiouse doubtful zigzag willar and one enought to get minings despeaking ab

Char-level, Order: 4

oung close the play cross tallest explain sigh her own only because of it is was out chin his so herself to pleasant alices at made and the tarts the said the duchess she queen want to caller and of its t

Generated Text

Word-level, Order: 0

if matters skurried must may a i perfectly with long or shouting said the here loud it but patiently alice the always youve in answer baby of what to till came duchess was dry its above into tone not that howling and it come said me out quite there a

Word-level, Order: 1

pigeon i did with pink eyes very curious as it had never said the cat remarked the dodo in a red rosetree she sat silent for i wonder she ran round the twentieth time of very deep and the while the duck its a thing said the blame on with the

Word-level, Order: 2

out you know and he called the queen left off sneezing by this time youre nothing but outoftheway things had happened lately that alice said nothing perhaps it was said the caterpillar well perhaps your feelings may be different said alice very humbly you had been broken to pieces please then said

Word-level, Order: 3

just at first the two creatures got so close to her ear youre thinking about something my dear and that makes you forget to talk i cant tell you just now what the moral of that is but i shall remember it in a few minutes she heard a little animal she couldnt

Word-level, Order: 4

she was surprised to find quite a large crowd collected round it there was a dispute going on between the executioner the king and the executioner ran wildly up and down looking for it while the rest of the party went back to the table half hoping she might find another key on it

Limitations

N-gram Markov Model

- computational cost grows exponentially
- empirical transition probabilities no smoothing
- entropy estimator $\hat{H}_N \approx \Sigma_c P(c)\, H(X \mid c)$
- finite sample bias due to data sparsity at large N