Mickiewicz entropy extravaganza

using entropic measures on Polish literature

Magdalena Rajtor

Project overview

1. Comparing entropy between literary gernes
2. Comparing entropy between two authors
3. Estimating the entropy rate for Polish

1) Entropy on different genres

	EPIC	DRAMA	POETRY	BALLADS
no. of works	7	12	83	14
tokens: letters	633k	251k	101k	56k
tokens: words	102k	42k	16k	9k

1) Entropy on different genres

	EPIC	DRAMA	POETRY	BALLADS
no. of works	7	12	83	14
tokens: letters	633k	251k	101k	56k
tokens: words	102k	42k	16k	9k
	6,2	5,9	6,3	6,2

average Polish word length: 6 (from NKJP samples)

Moździerz, T. (2020). Długość przeciętnego polskiego wyrazu w tekstach pisanych w świetle analizy korpusowej. Acta Universitatis Lodziensis. Kształcenie Polonistyczne Cudzoziemców, 27, 177–192

1) Entropy on different genres

	EPIC	DRAMA	POETRY	BALLADS
no. of works	7	12	83	14
tokens: letters	633k	251k	101k	56k
tokens: words	102k	42k	16k	9k

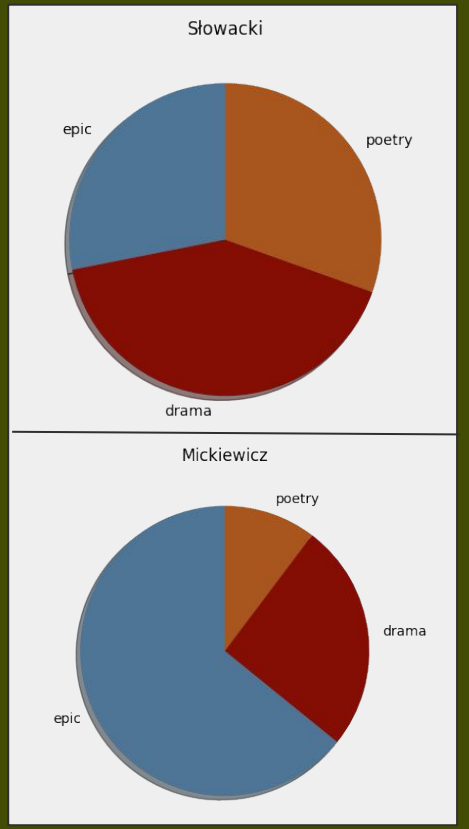
H (on letters)	4,48	4,48	4,48	4,46
H (on words)	11,91	11,39	11,16	10,57

2) Mickiewicz vs Słowacki

	EPIC	DRAMA	POETRY
no. of works	3	10	81
tokens: letters	657k	970k	712k
tokens: words	102k	162k	119k

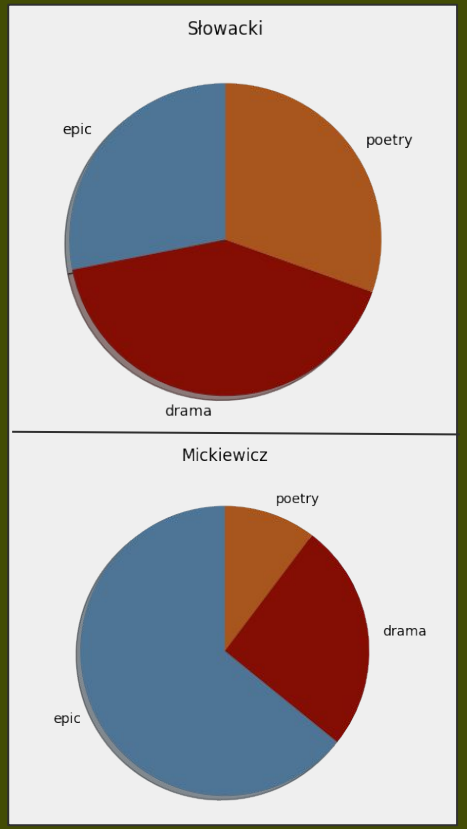
2) Mickiewicz vs Słowacki

	EPIC	DRAMA	POETRY
no. of works	3	10	81
tokens: letters	657k	970k	712k
tokens: words	102k	162k	119k



2) Mickiewicz vs Słowacki

	EPIC	DRAMA	POETRY
no. of works	3	10	81
tokens: letters	657k	970k	712k
tokens: words	102k	162k	119k



average word length: Słowacki - 6,12
Mickiewicz - 6,05

2) Mickiewicz vs Słowacki

	EPIC	DRAMA	POETRY
H (on letters)	4,50	4,47	4,47
H (on words)	11,71	11,66	**11,73***

H (on letters)	4,48	4,48	4,48
H (on words)	11,91	11,39	**11,16**

*11 “poematy” = narrative poetry (written in metered verse)

2) Mickiewicz vs Słowacki

	EPIC	DRAMA	POETRY	TOTAL
H (on letters)	4,50	4,47	4,47	4,48
H (on words)	11,71	11,66	**11,73***	12,09

H (on letters)	4,48	4,48	4,48	4,48
H (on words)	11,91	11,39	**11,16**	12,10

2) Mickiewicz vs Słowacki

	EPIC	DRAMA	POETRY	TOTAL
H (on letters)	4,50	4,47	4,47	4,48
H (on words)	11,71	11,66	**11,73***	12,09

H (on letters)	4,48	4,48	4,48	4,48
H (on words)	11,91	11,39	**11,16**	12,10

\+ add Jensen-Shannon divergence?

3) Estimating Polish information rate

information rate (source entropy) = the limit of the conditional entropy of each consecutive letter given the previous ones

$$h(X) = \lim_{n \to \infty} H(X_{n+1} | X_1, \ldots, X_n)$$

3) Estimating Polish information rate

information rate (source entropy) = the limit of the conditional entropy of each consecutive letter given the previous ones

$$h(X) = \lim_{n\to\infty} H(X_{n+1}|X_1, \ldots, X_n)$$

```
model = {
  ('w', 'h'): {'y':25, 'o':12, 'a':16, ...},
  ('t', 'h'): {'i':15, 'a':18, 'e':34, ...},
  ...
}
```

https://pit-claudel.fr/clement/blog/an-experimental-estimation-of-the-entropy-of-english-in-50-lines-of-python-code/

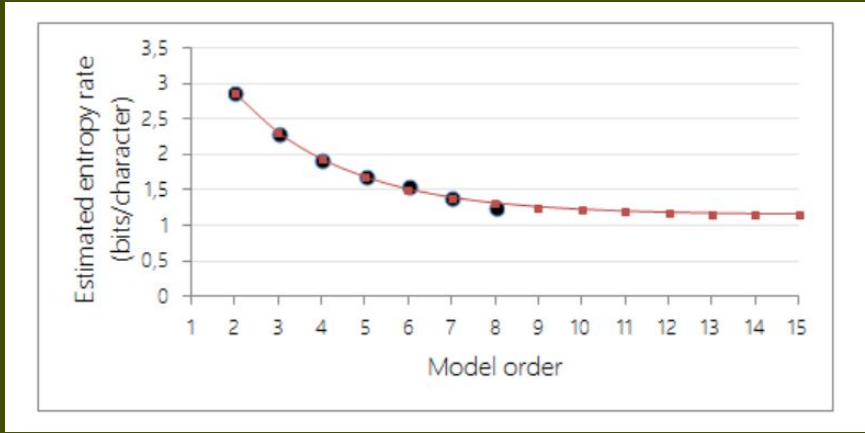
3) Estimating Polish information rate

information rate (source entropy) = the limit of the conditional entropy of each consecutive letter given the previous ones

$$h(X) = \lim_{n\to\infty} H(X_{n+1}|X_1, \ldots, X_n)$$

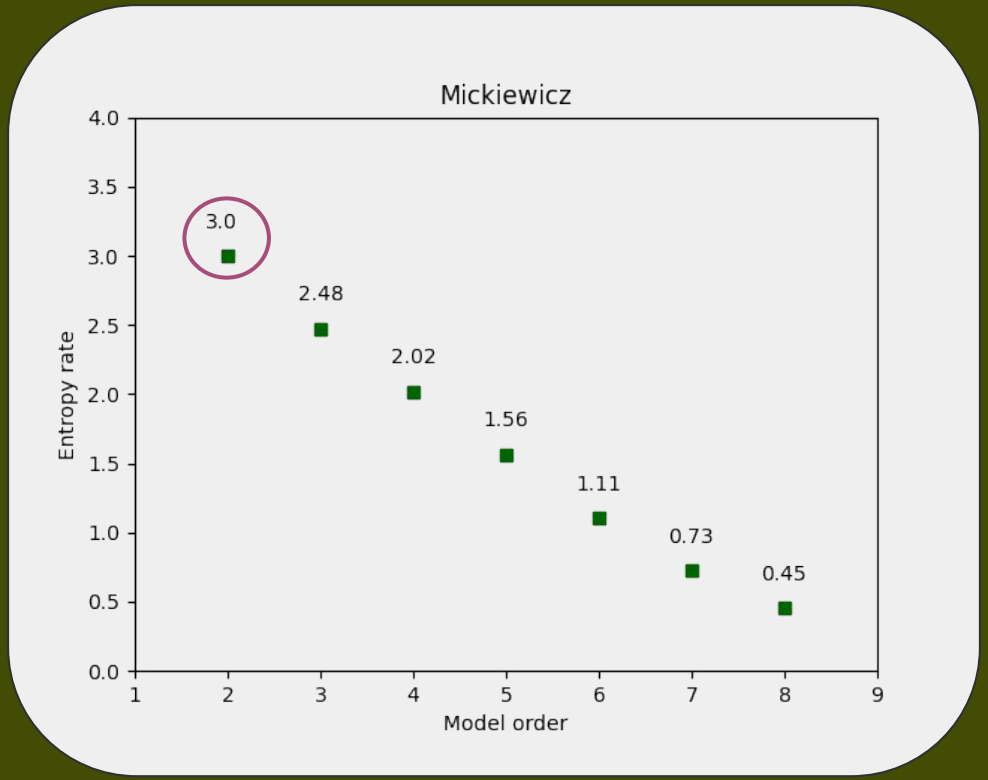
```
model = {
   ('w', 'h'): {'y':25, 'o':12, 'a':16, ...},
   ('t', 'h'): {'i':15, 'a':18, 'e':34, ...},
   ...
}
```

https://pit-claudel.fr/clement/blog/an-experimental-estimation-of-the-entropy-of-english-in-50-lines-of-python-code/

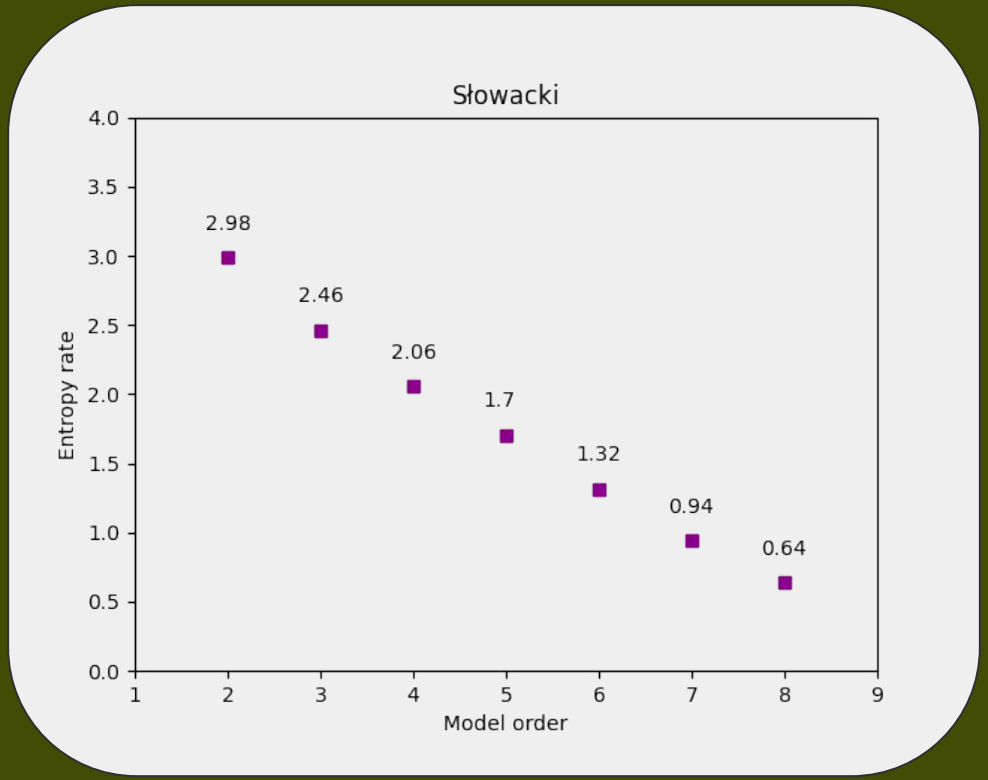


Shannon: 0.6 and 1.3 bits per letter

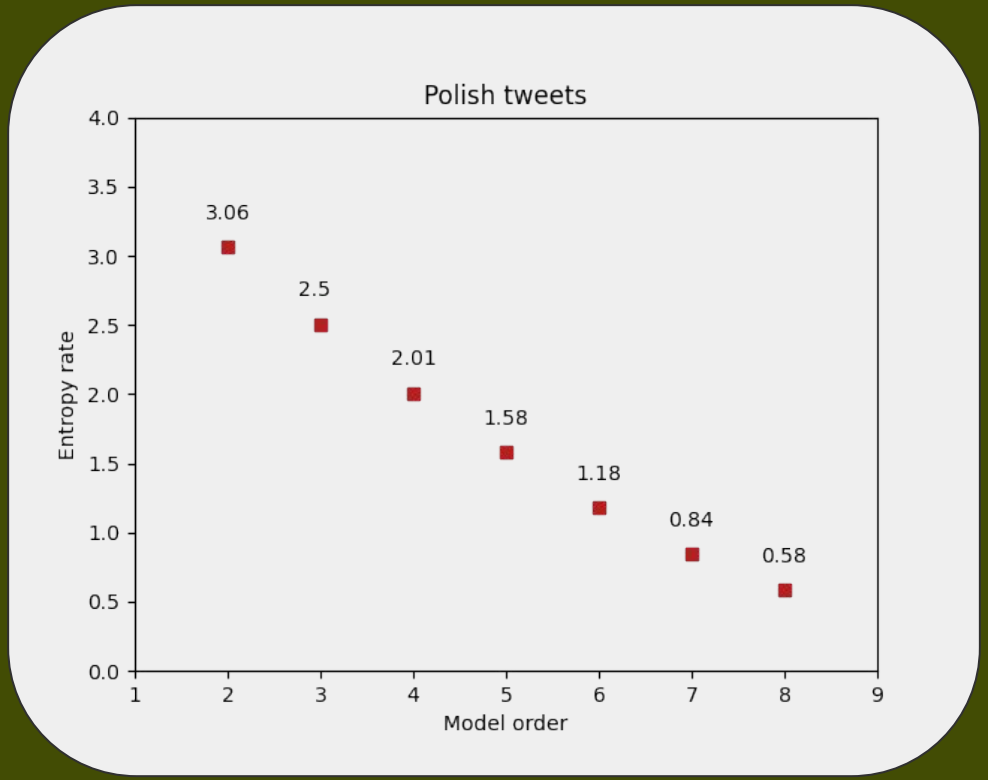
3) Estimating Polish entropy rate



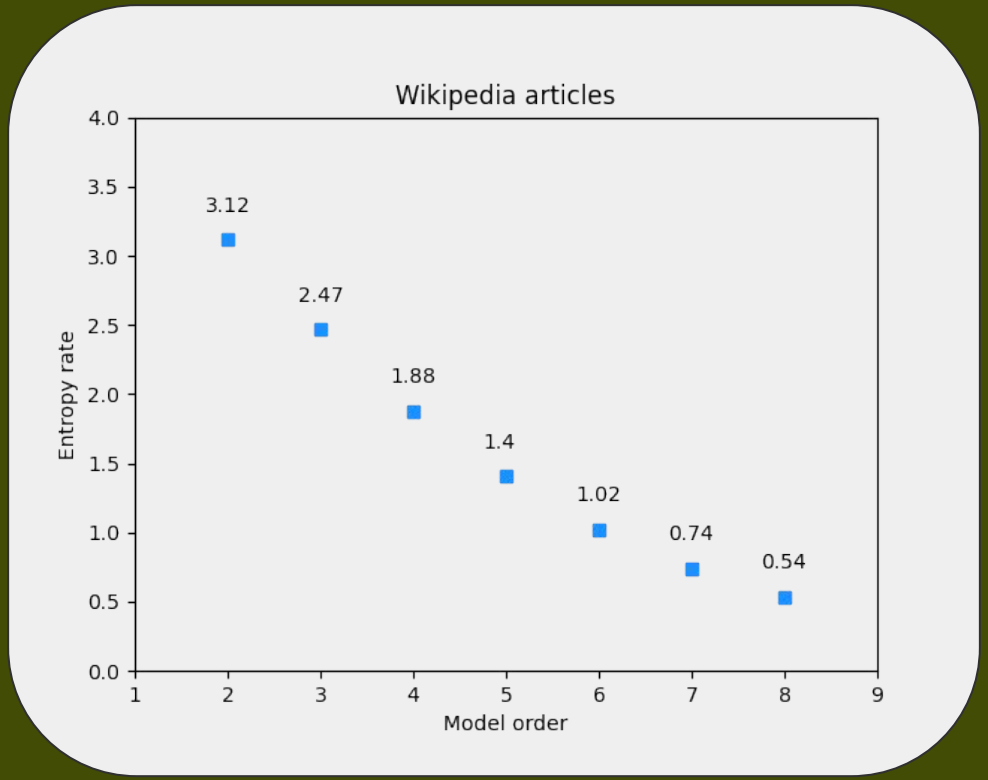
3) Estimating Polish entropy rate



3) Estimating Polish entropy rate



3) Estimating Polish entropy rate



Is entropy a good measure for poetry?

Love shines, warm and light,
*Brightening both **day and night**.*
*Roses are **red, violets are blue**,*
*Sugar is sweet, and so **are you**.*

HIGH entropy
easy to predict

Blue skies, blue seas, blue dreams,
Whispers, whispers, lost in streams.
Time ticks, ticks time, in reverse,
Echoes, echoes, a cryptic verse.

LOW entropy
hard to predict

THANK YOU!