

Observing Surprisal through the blurry lense of Language Models

Peter Fankhauser

Goals of this Talk

- Gain a better understanding of „Entropy Rate“
- Influence of the domain of Language Models
- Interaction of Entropy Rate with other variables
 - Register
 - Field
 - Texttype
 - Sentence Length
 - Sentence Position

Entropy Rate: Genzel/Charniak 2002

- Probability of a sentence S (2 Gram context):

$$p(S) = p(w_1)p(w_2|w_1) \times \prod_{i=3}^n p(w_i|w_{i-1}w_{i-2})$$

- Entropy: $H(S) = -\log(p(S))$
- Average Entropy per word in S : $H(S)/|S|$
- Entropy Rate: $H(Y_i)$

$$H(Y_i) = H(X_i|L_i) - I(X_i|C_i, L_i)$$

i ... position in text, j ... sentence number for X_i

C_i ... context up to sentence S_j , L_i ... local context in S_j

$$H(X_i|L_i) = H(S_j)/|S_j|$$

$I(X_i|C_i, L_i)$... mutual information of X_i and C_i, L_i

- Entropy Rate Constancy Principle:

$H(Y_i)$ is **constant** - for reasons of communicative efficiency

$I(X_i|C_i, L_i)$ cannot be estimated, but is assumed to increase with i – **thus**:

$H(X_i|L_i)$ also should increase with i .

Language Models

- Cache (unigram only)
 - LM up to current position in text
 - No smoothing; exponential decay (Clarkson/Robinson 1997, $e=0.005$)
 - „**Best**“ **bet at MI with Context.**
- Document
 - LM of the Document
 - (Interpolated) Witten-Bell Smoothing
 - Blurry lense at (lexical) **cohesion** (cf. Eisenstein/Barzilay 2008)
- **Corpus**
 - LM of Corpus **excluding** current Document
 - (Interpolated) Witten-Bell Smoothing
 - Domain LM (cf. Genzel/Charniak 2002,2003, Keller 2004, Qian/Jaeger 2010)
- Web
 - Microsoft N-Grams (2013/Body) (Huang et al. 2010)
 - Backoff Absolute Discount Smoothing, Frequency cutoff: 20
 - „General“ LM ... with all caveats applying.

Example: 3gram context, corpus, FLOB-J (learned)

More recent studies have examined effects of moderate intensity noise on cognitive vigilance tasks.

An example of this type of task is the Bakan task where single digits are presented on a screen and subjects have to detect a particular sequence such as three successive odd digits.

Poulton (1977) suggested that all detrimental effects of noise could be explained either by masking of acoustic cues or, in some cases, cognitive tasks, by masking of inner speech.

The first part of the view has been strongly opposed (see Broadbent, 1978). The second part led to an interest in effects of noise on cognitive vigilance, for post-experimental reports from subjects having completed a Bakan task clearly suggest that they say the digits to themselves.

If noise masks internal speech such tasks should be impaired and this issue was examined by Jones, Smith & Broadbent (1979). They carried out four experiments using slightly different versions of the Bakan task.

In each case there was evidence of a detrimental effect of noise, although the exact nature of the noise effect depended on specific features of the task.

This, and the finding that noise had no differential effect on digits which were rehearsed together and those which were rehearsed separately, argues against noise producing effects solely by interfering with subvocal articulation.

These results do, however, show that vigilance tasks involving memory are susceptible to noise below 95 dB, which is consistent with results obtained by Benignus, Otto & Knelson (1975). This finding has also been confirmed by subsequent studies by Miles, Auburn & Jones (1984) and Smith (1988a & b).

The importance of the nature of tasks will be discussed again in a later section, after results from other attention paradigms have been summarized.

Noise in the five-choice serial reaction time task.

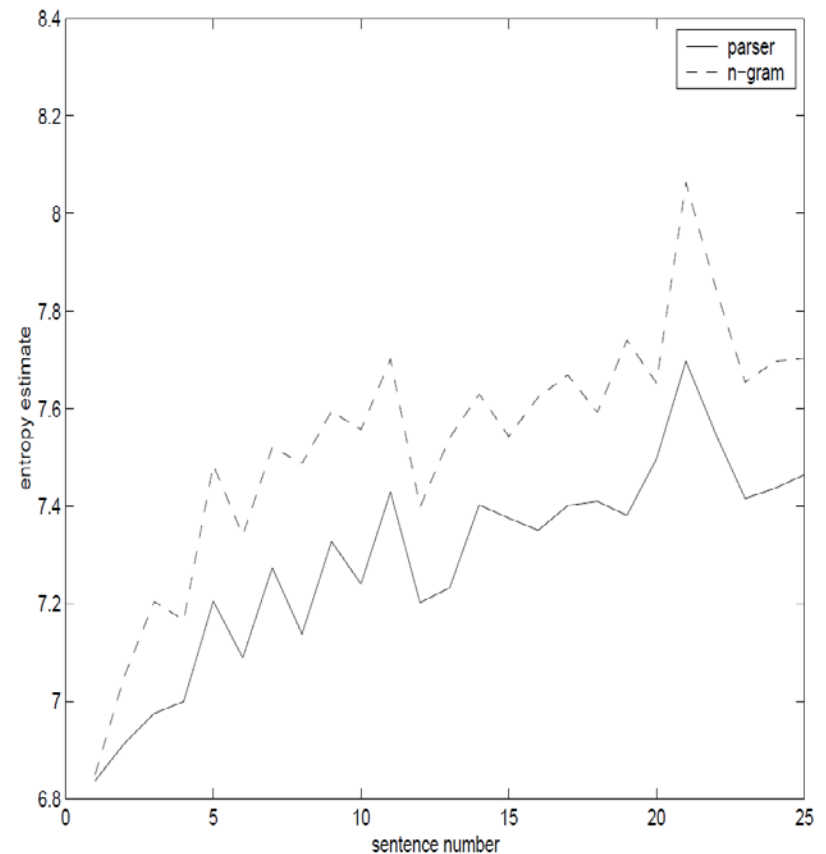
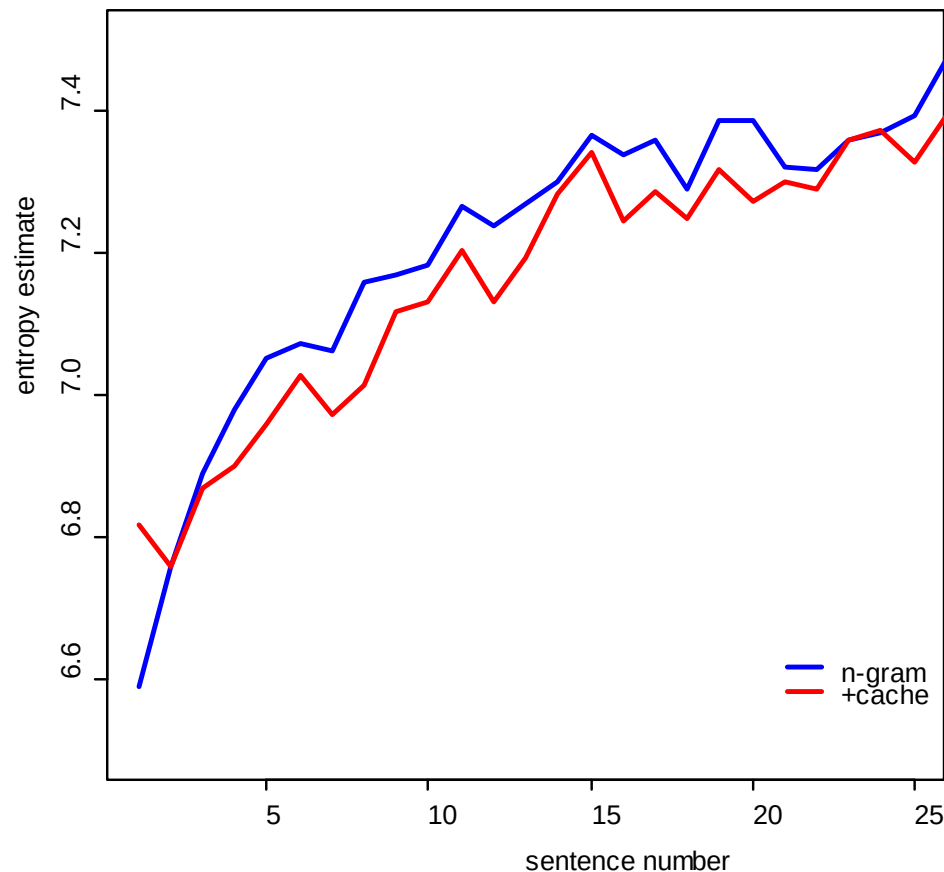
Another task which has been widely used to study the effects of noise is the five-choice serial reaction time task.

Broadbent (1979) has reviewed effects of very loud noise on this task and concluded that noise increases momentary inefficiency, so that there are more errors and gaps (occasional very long reaction times) than in quiet.

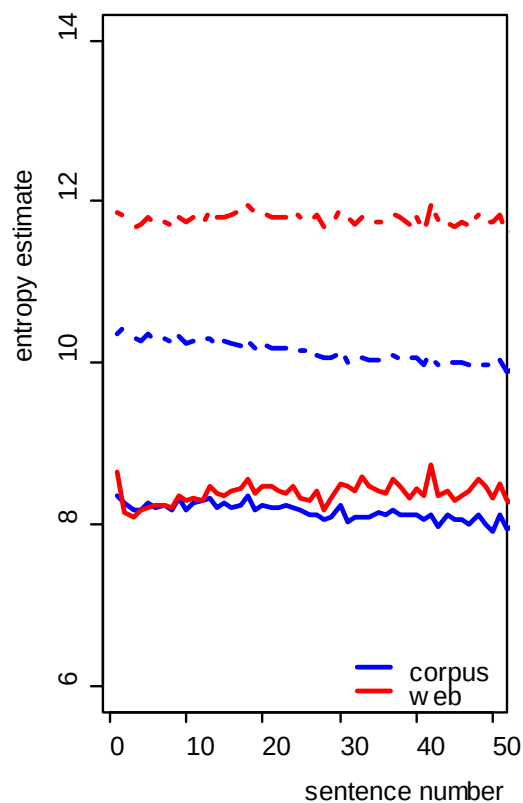
Corpora

- FLOB: British English, 1991
 - 15 Registers (3 Press, 6 NonFiction, 6 Fiction)
 - 1587 Documents, 56 K Sentences, 1.16 Mio Tokens
- Wall Street Journal Corpus (WSJ)
American English, 1987 – 1989
 - 2499 Documents, 53 K Sentences, 1.29 Mio Tokens
- Darmstadt Scientific Text (DaSciTex), 1998 - 2007
 - 9 Disciplines
 - Annotated for Head, Abstract, Intro, Main, Conclusion
 - 2111 Documents, 920 K Sentences, 18.8 Mio Tokens

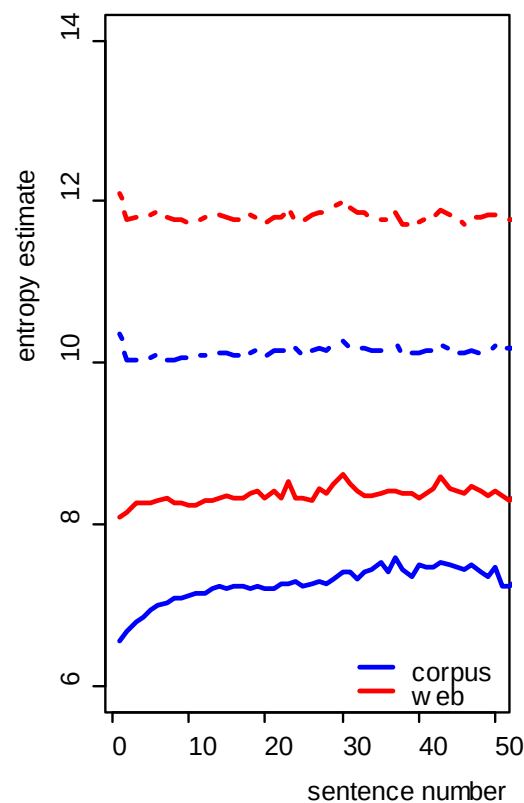
WSJ: Comparison with Genzel/Charniak 2002



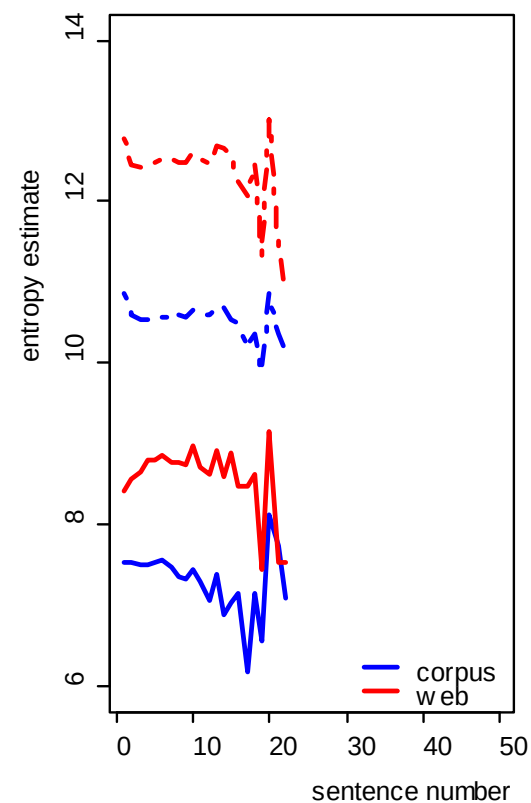
Entropy Rate Comparison



FLOB



WSJ



DaSciTex Abstracts

— 3 Context
- - - - - No Context

WSJ: (Near) Duplicates

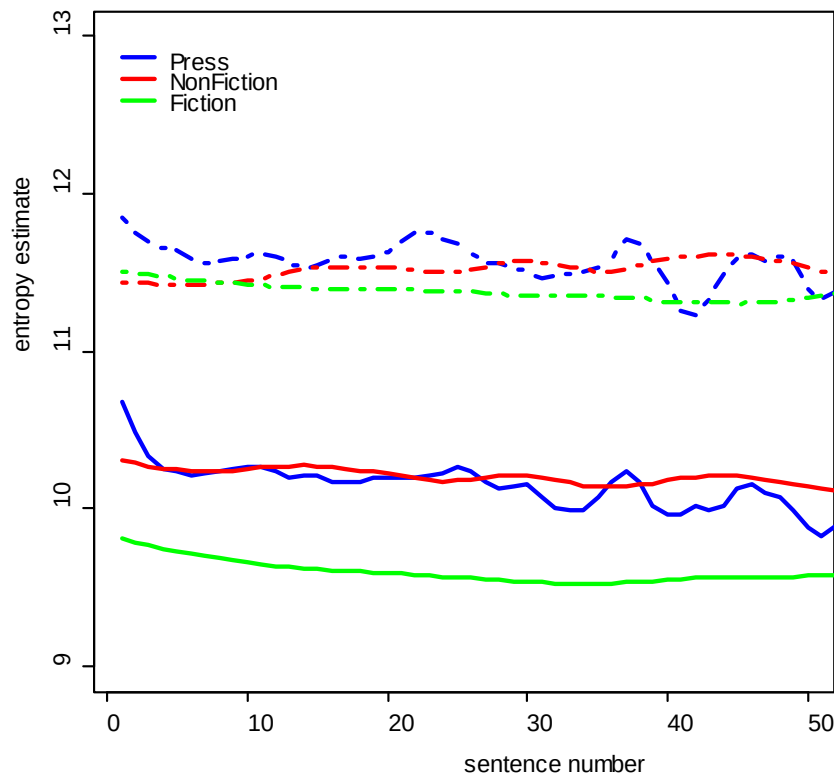
1-5: 217 Occurrences of 44 Duplicates in 10777 Sentences

n	corpus	web	sentnum	std	sentence
27	2.22	11.87	1 8/27	0.82	Terms were n't disclosed .
14	0.96	6.28	2	0.00	The base rate on corporate loans at large U.S. money center commercial banks .
14	0.98	4.82	4	0.00	Reserves traded among commercial banks for overnight use in amounts of \$ 1 million or more .
14	1.74	13.54	5	0.00	Source : Fulton Prebon (U.S.A .
14	1.86	9.61	1	0.00	PRIME RATE : 10 1\ % .
13	0.84	8.40	3	0.00	The percent difference compares actual profit with the 30-day estimate where at least three
13	0.85	7.37	1	0.00	The companies are followed by at least three analysts , and had a minimum five-cent change in ...
13	1.38	10.06	2	0.00	Estimated and actual results involving losses are omitted .
13	1.40	9.42	4	0.00	Otherwise , actual profit is compared with the 300-day estimate .
4	2.53	9.27	1 1/2	0.58	The price was n't disclosed .

21-25: 73 Occurrences of 12 Duplicates in 4200 Sentences

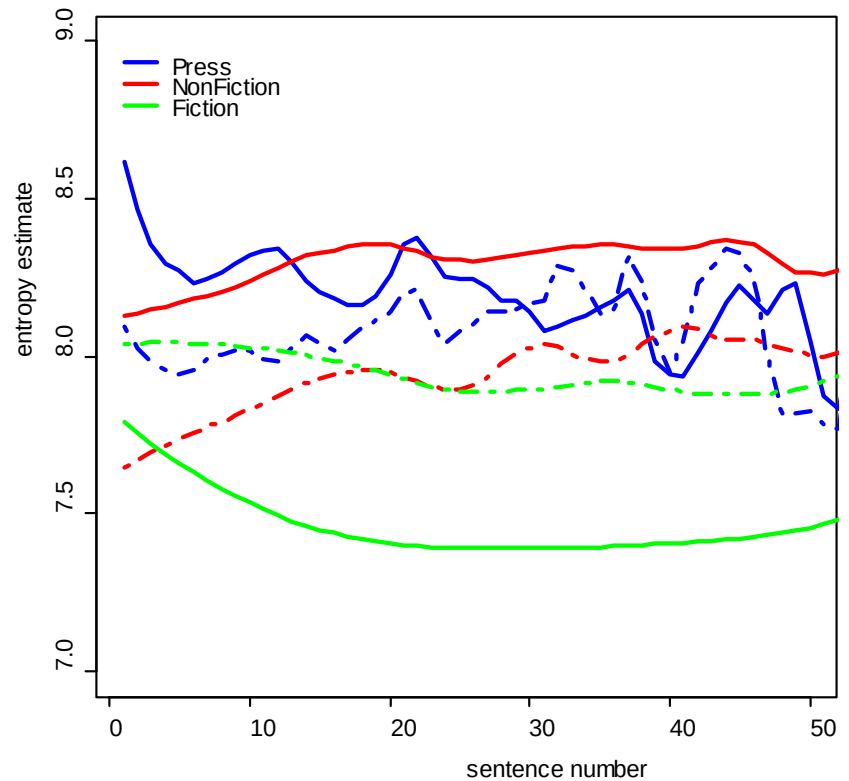
n	corpus	web	sentnum	std	sentence
14	0.83	4.65	21	0.00	The average of interbank offered rates for dollar deposits in the London market based on ...
14	0.87	7.15	23	0.00	These rate indications are n't directly comparable ; lending practices vary widely by location .
13	2.17	5.32	25	0.00	FEDERAL HOME LOAN MORTGAGE CORP .
8	0.64	12.56	22	0.00	FOREIGN PRIME RATES : Canada 13.50 % ; Germany 9 % ; Japan 4.875 % ; Switzerland 8.50 % ; ...
6	0.71	13.26	22	0.00	FOREIGN PRIME RATES : Canada 13.50 % ; Germany 8.50 % ; Japan 4.875 % ; Switzerland 8.50 % ; ...
4	0.66	7.96	24	0.00	TREASURY BILLS : Results of the Monday , October 16 , 1989 , auction of short-term U.S. ...
4	0.69	7.89	24	0.00	TREASURY BILLS : Results of the Monday , October 23 , 1989 , auction of short-term U.S. ...
2	0.78	7.94	24	0.00	TREASURY BILLS : Results of the Monday , October 30 , 1989 , auction of short-term U.S. ...
2	1.85	8.11	24 1/2	0.71	San Antonio , Texas -- \$ 640 million of gas and electric revenue refunding bonds , ...
2	2.58	10.76	24	0.00	Estimated volume was three million ounces .

FLOB: Entropy by Sentence Number/Register



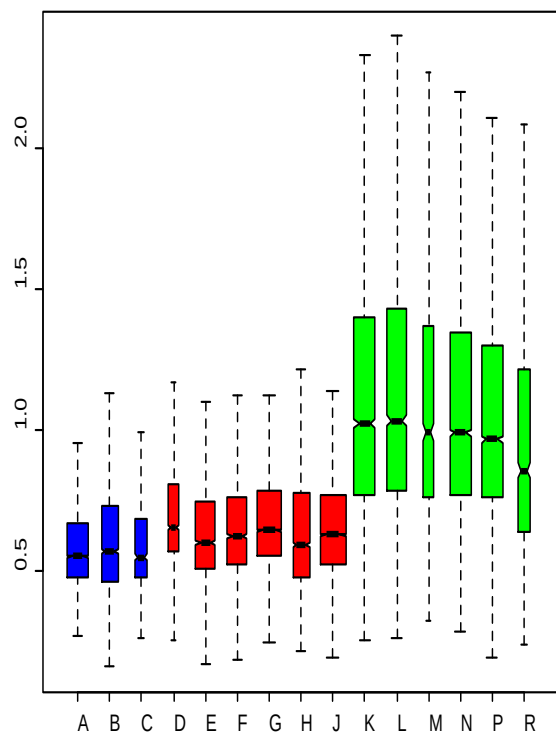
No Context

— Corpus
- - - Web

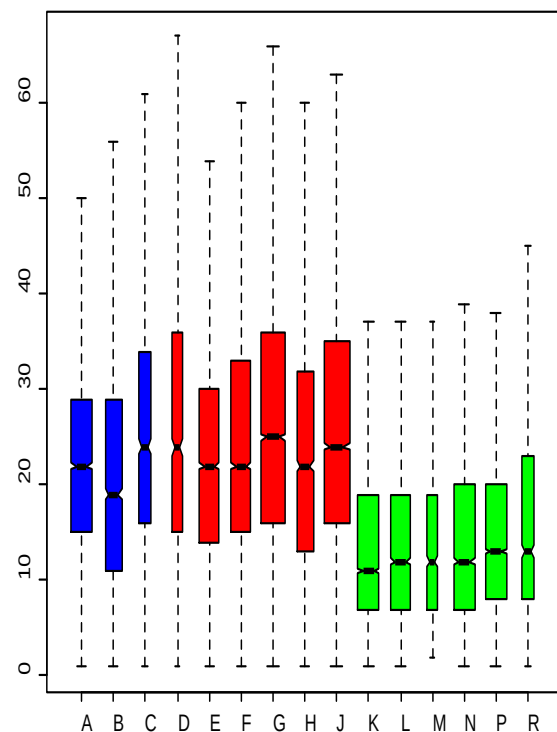


3 Context

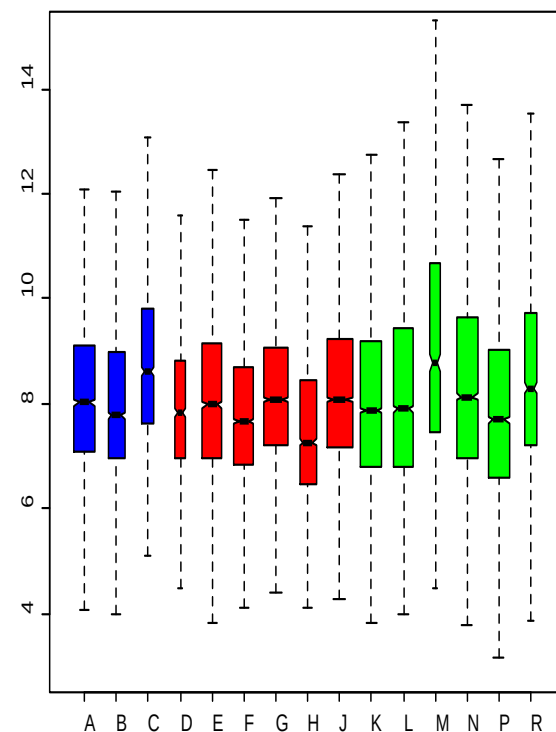
FLOB: Entropy by Register (3 Ctx)



Document

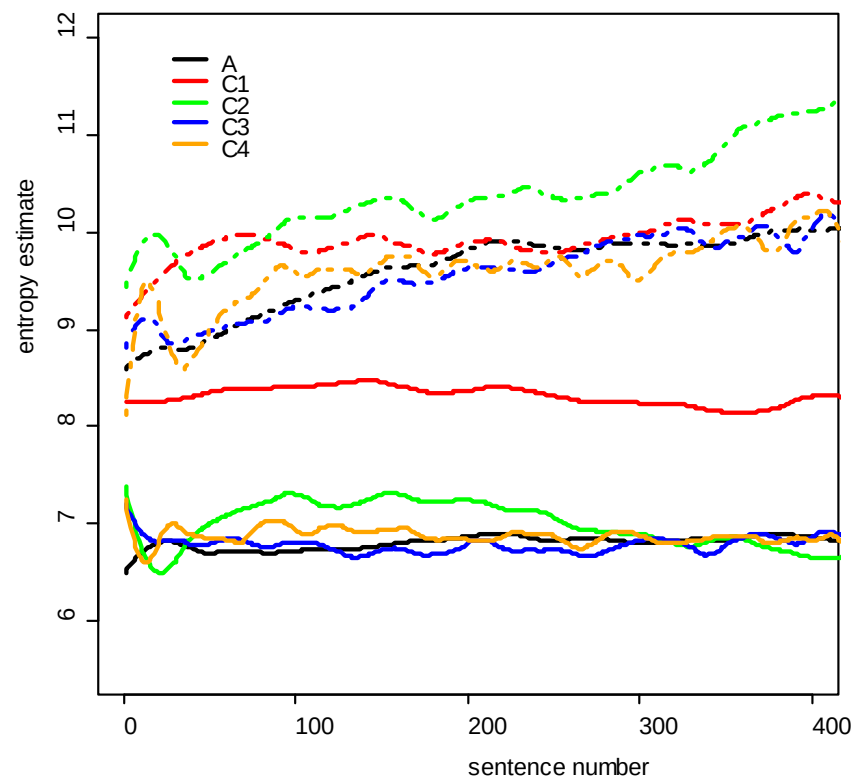
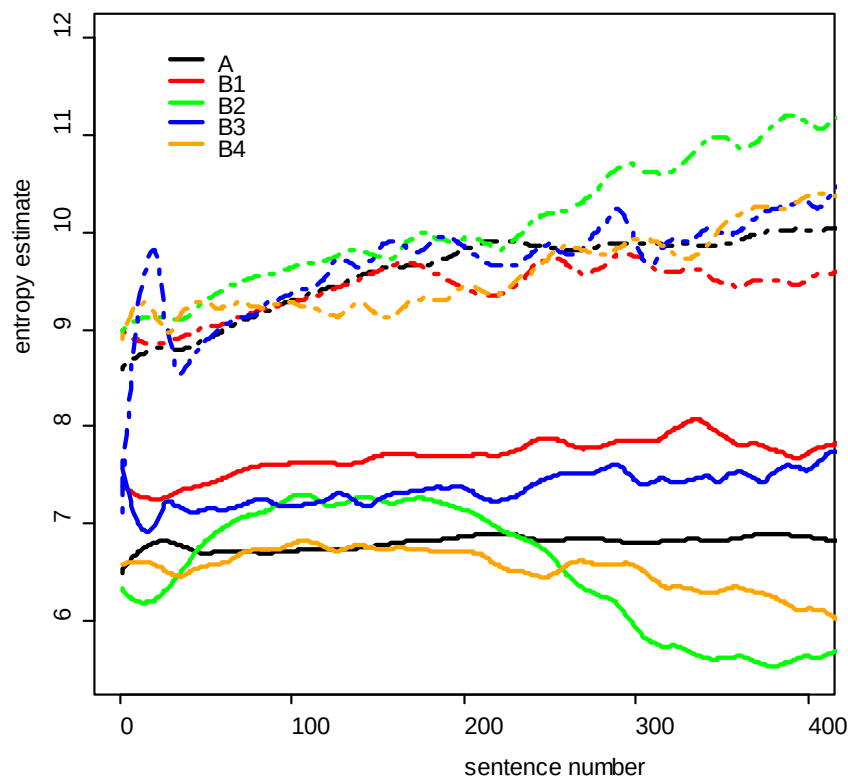


Sentence Length



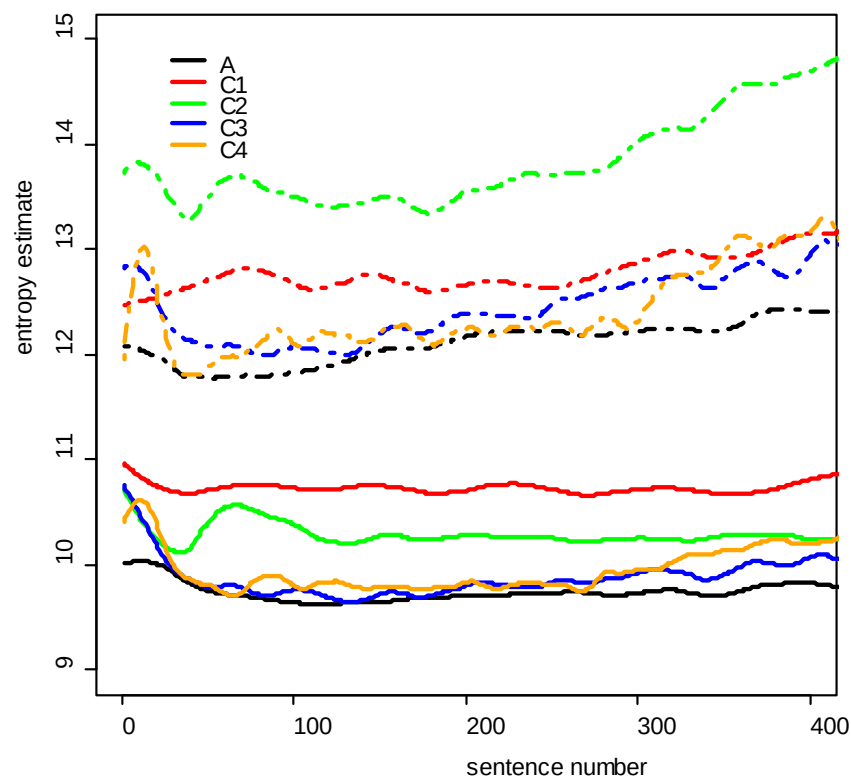
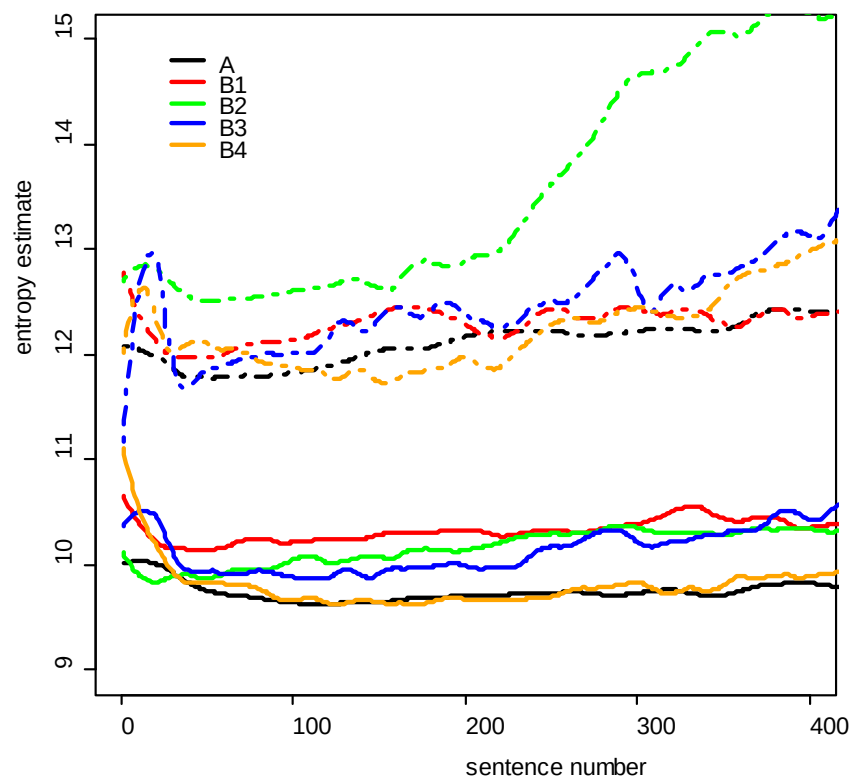
Web

DaSciTex: Entropy by Sentence Number/Discipline (3 Ctx)



— Corpus
- - - Web

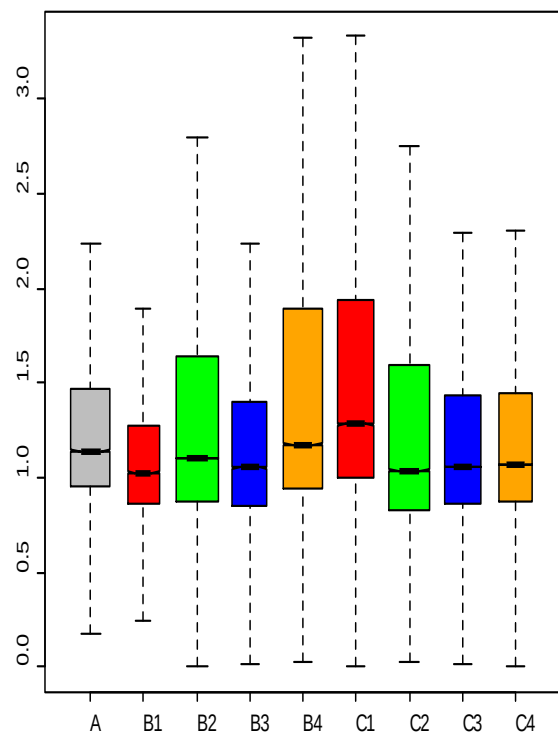
DaSciTex: Entropy by Sentence Number/Discipline (0 Ctx)



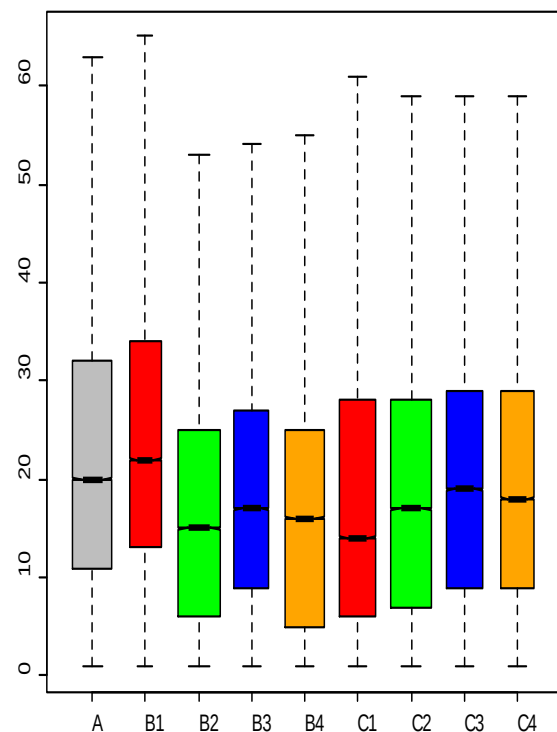
— Corpus
- - - Web

DaSciTex: Entropy by Discipline

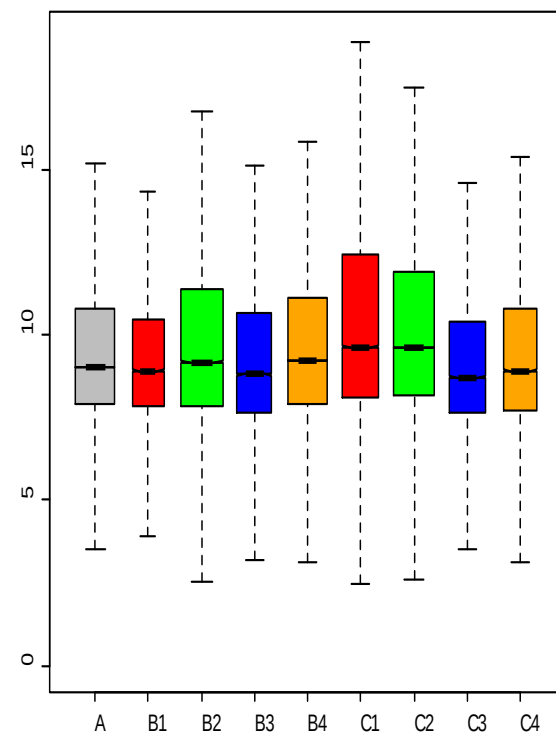
(3 Context)



Document

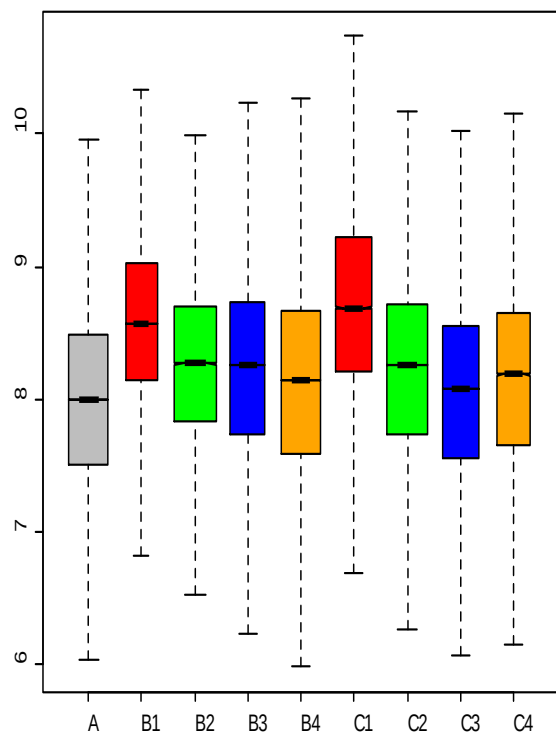


Sentence Length

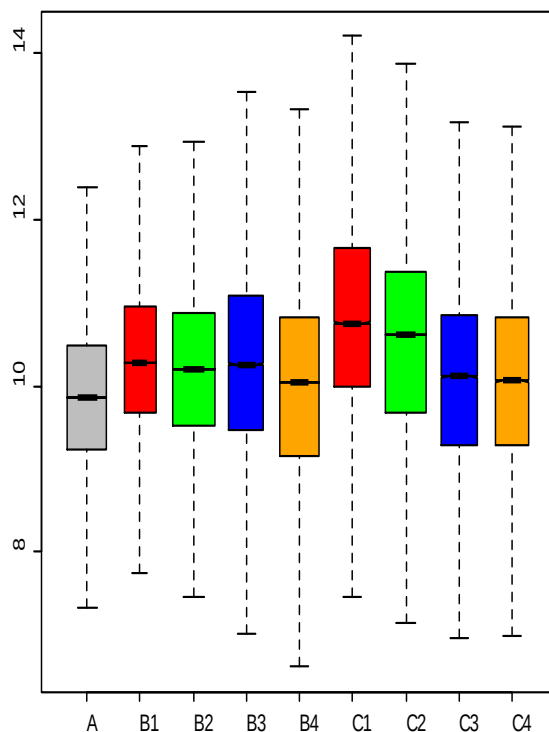


Web

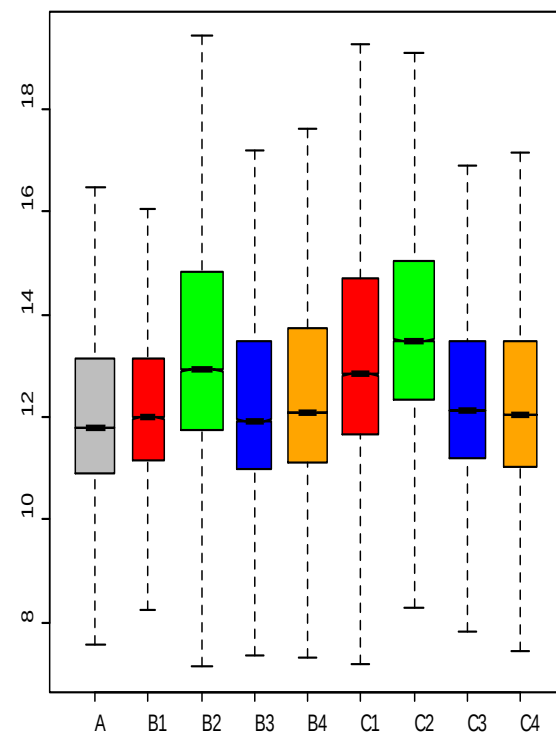
DaSciTex: Entropy by Discipline (No Context)



Document



Corpus



Web

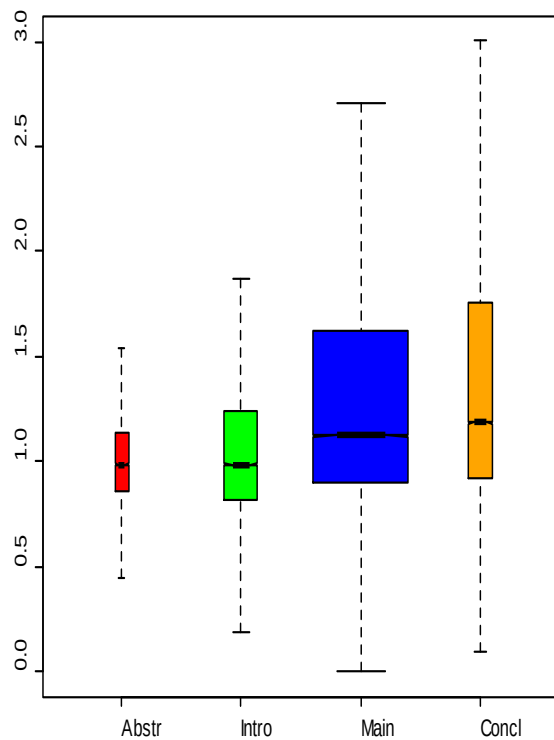
Reasons for low Corpus Entropy

- Abbreviations: „Fig .“ (12950), „2 .“ (4086)
- Boilerplate: Freq 63, Ent 0.364
„Users are entitled to use , reproduce , disseminate , or display“
- References: Freq 16, Ent 1.203
„(1999) Molecular classification of cancer : class discovery and class ...“
also prominent reason for **low** Web Entropy
- Duplicate(s): Holzer2006, Rocher 2006 (B4)
- Formulaic Expressions: Freq 46, Ent 2.222
„This paper is organized as follows .“
„The organization of the rest of this paper is as follows .“

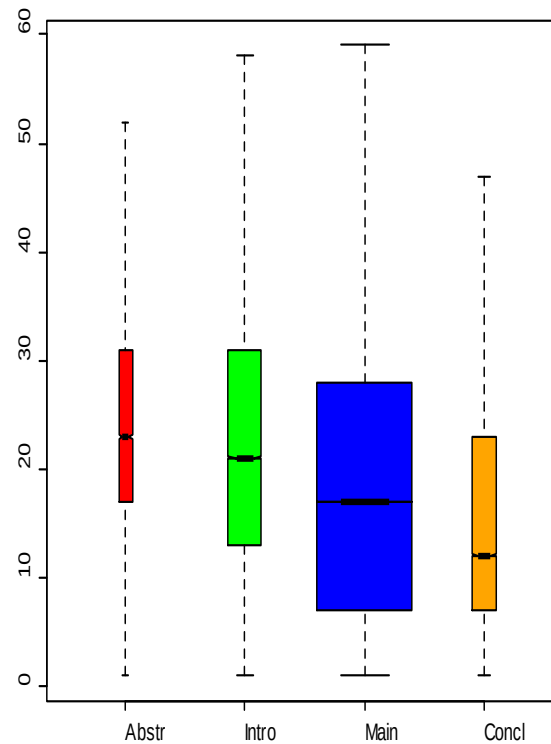
Reasons for high Corpus Entropy

- **Out of Vocabulary terms**
- Long lists of numbers
- Tokenization Issues
- Foreign Language (Ent 15.7)
 - „Je schneller Hans rennt , umso schneller wird er müde .“
 - “Pero es sabida la periodicidad de los terremotos de magnitud e intensidad superiores en esos territorios....”
- “Dense” Constructions (Ent 13.4)
 - “Leishmania donovani , possesses unusual heterodimeric type IB topoisomerase
.”
 - “Signatures spanning consecutive exons confirmed 142 predicted introns . “

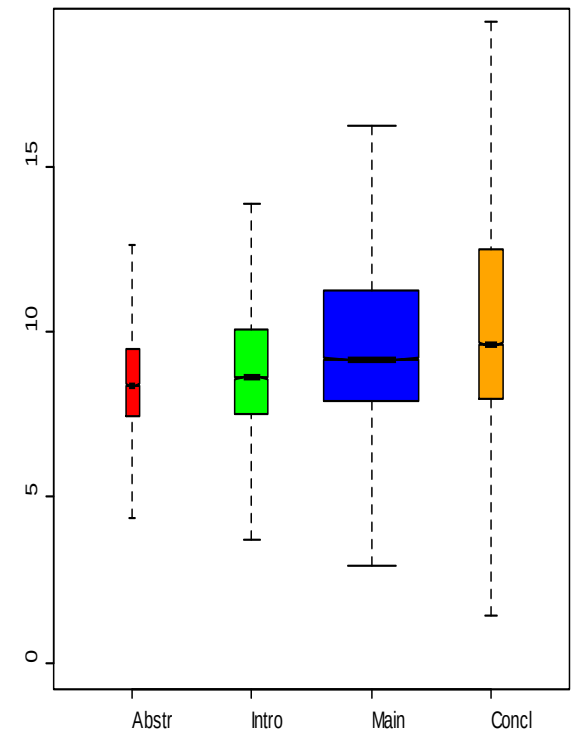
DaSciTex: Entropy by Section (3 Context)



Document

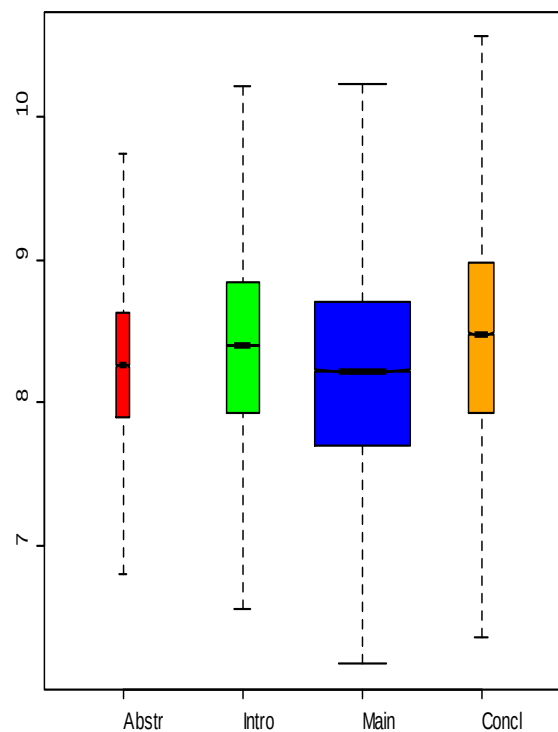


Sentence Length

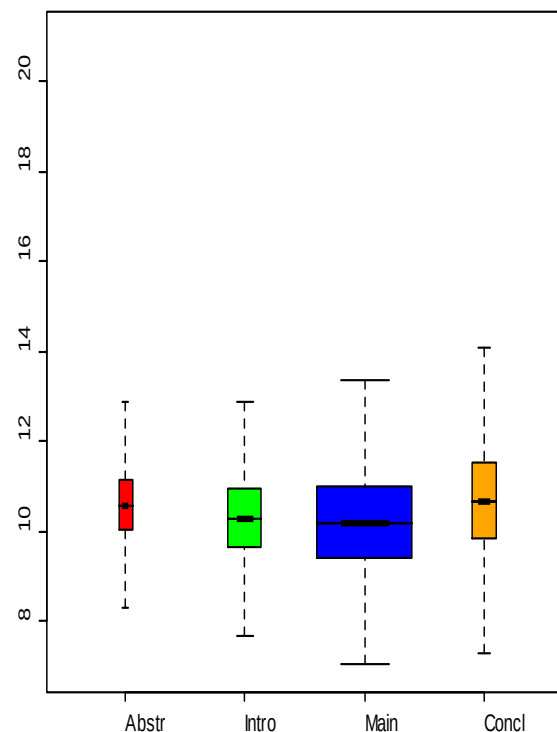


Web

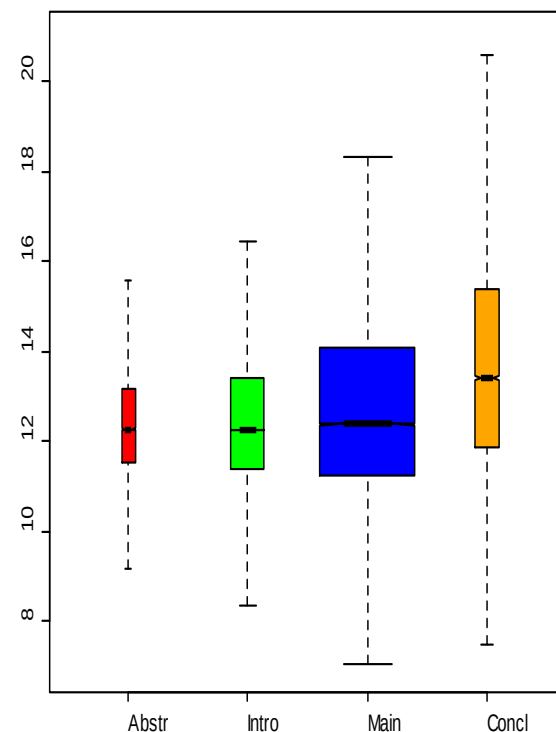
DaSciTex: Entropy by Section (No Context)



Document

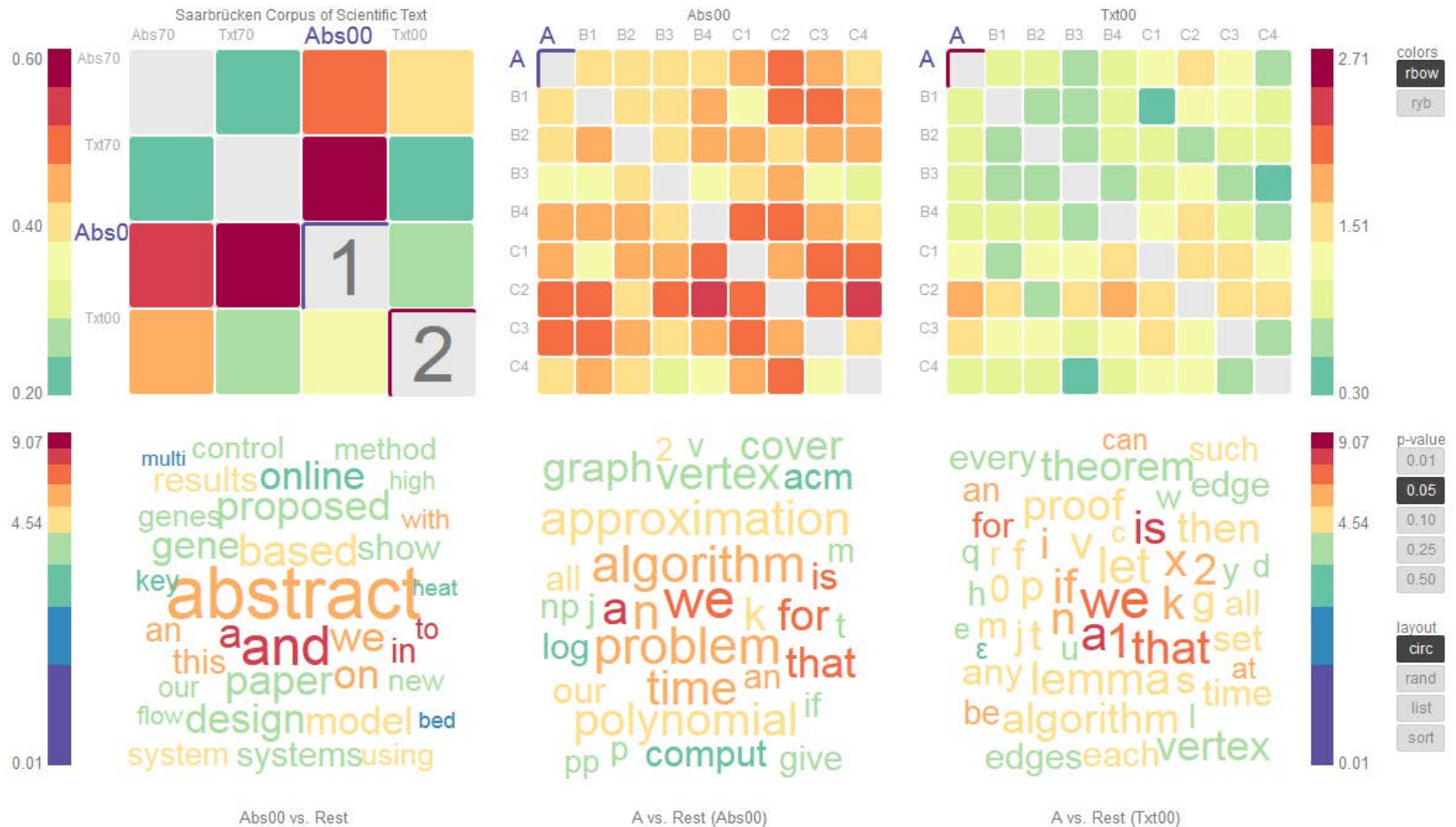


Corpus

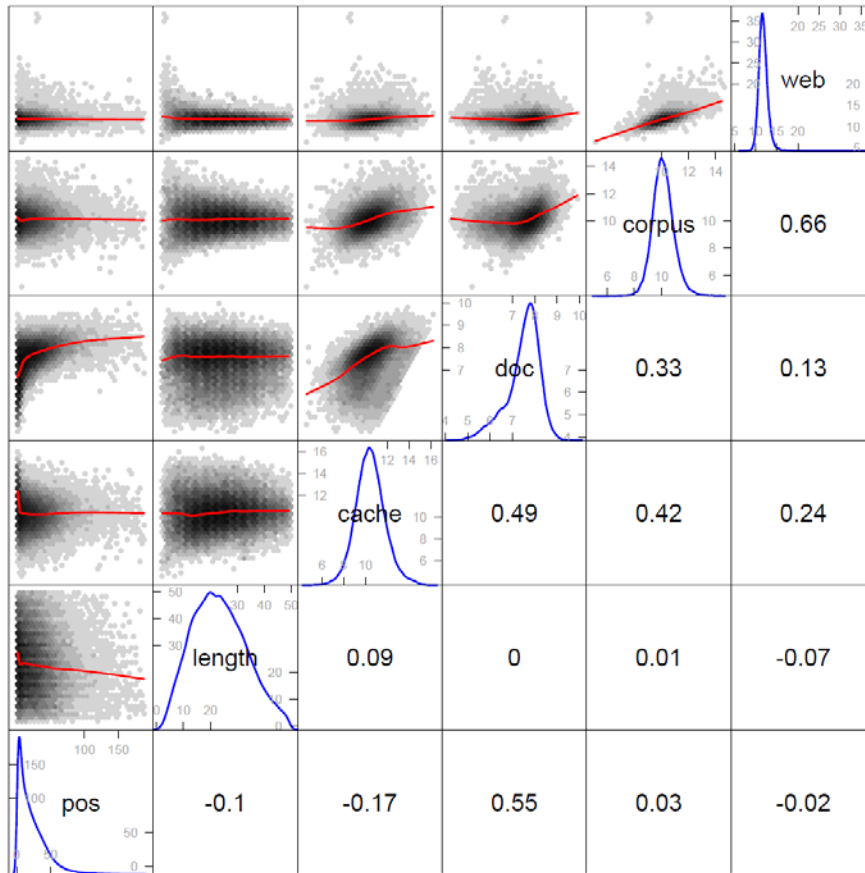


Web

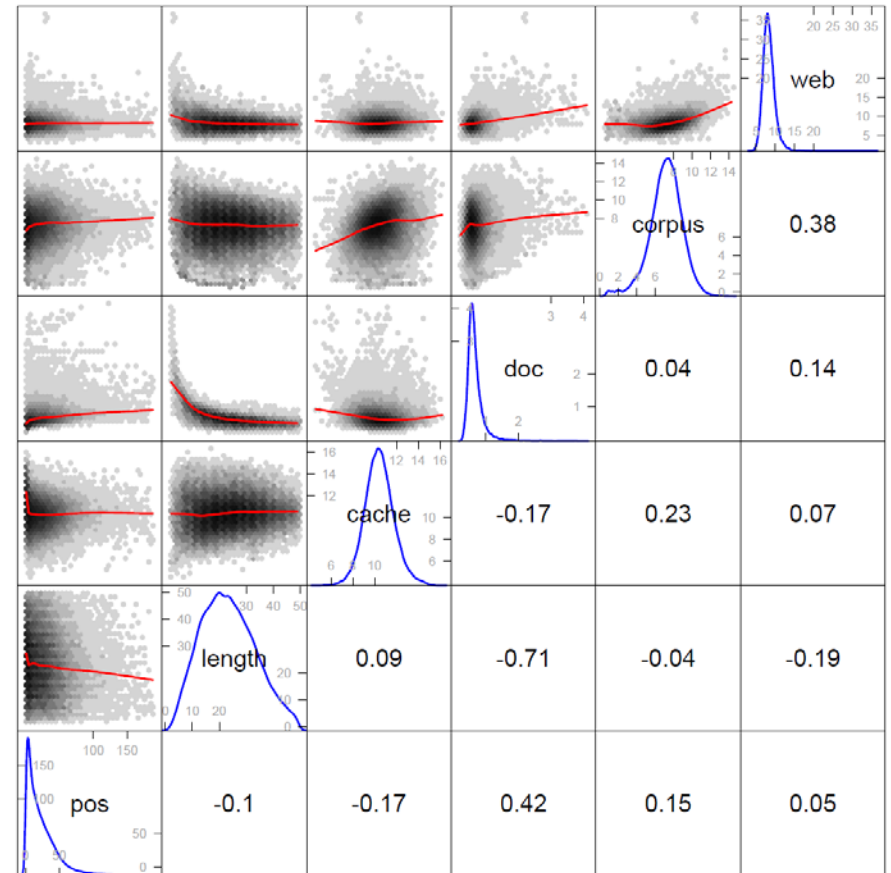
Relative Entropy: Abstracts vs. Text



WSJ: SPLOM

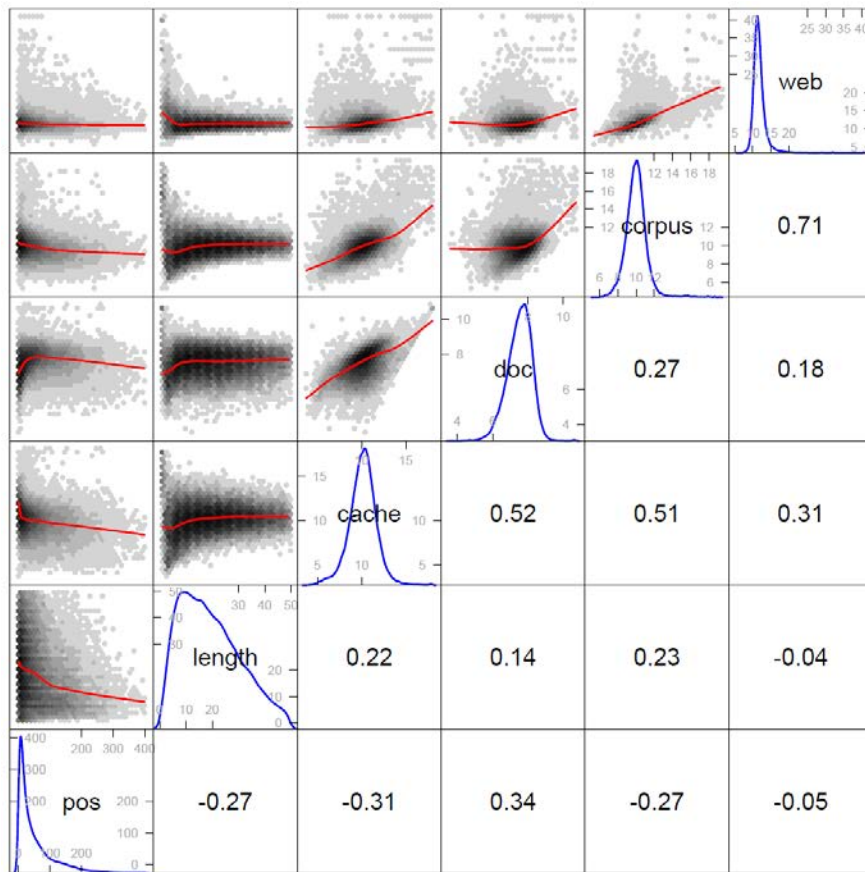


WSJ: no context

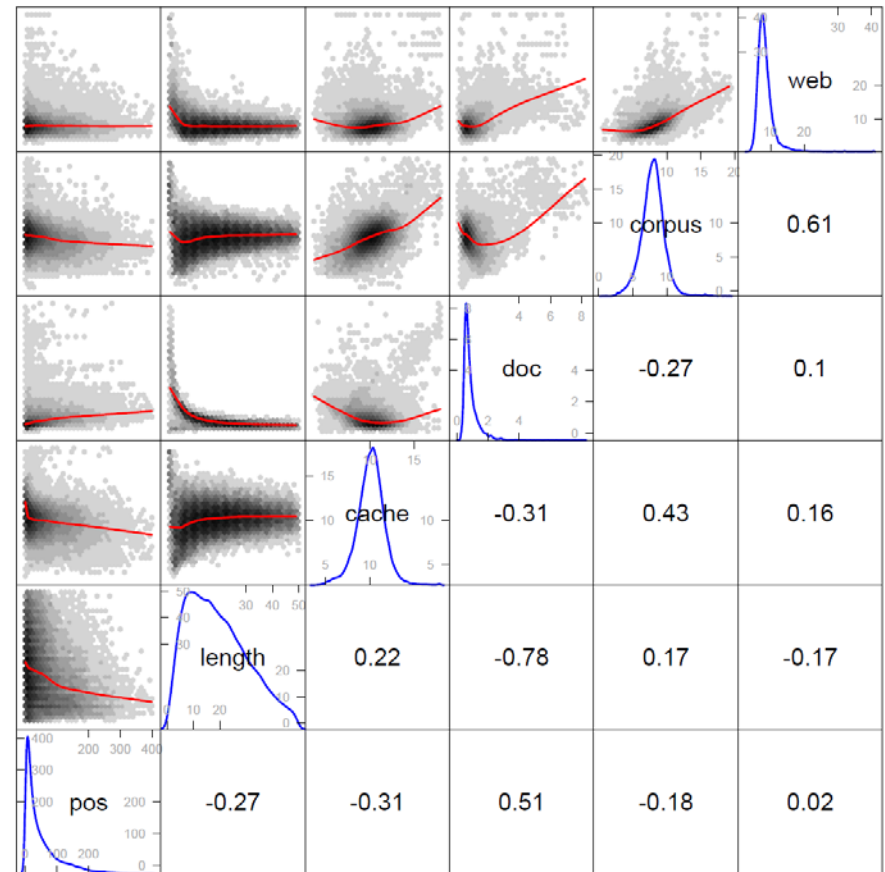


WSJ: 3gram context

FLOB: SPLOM

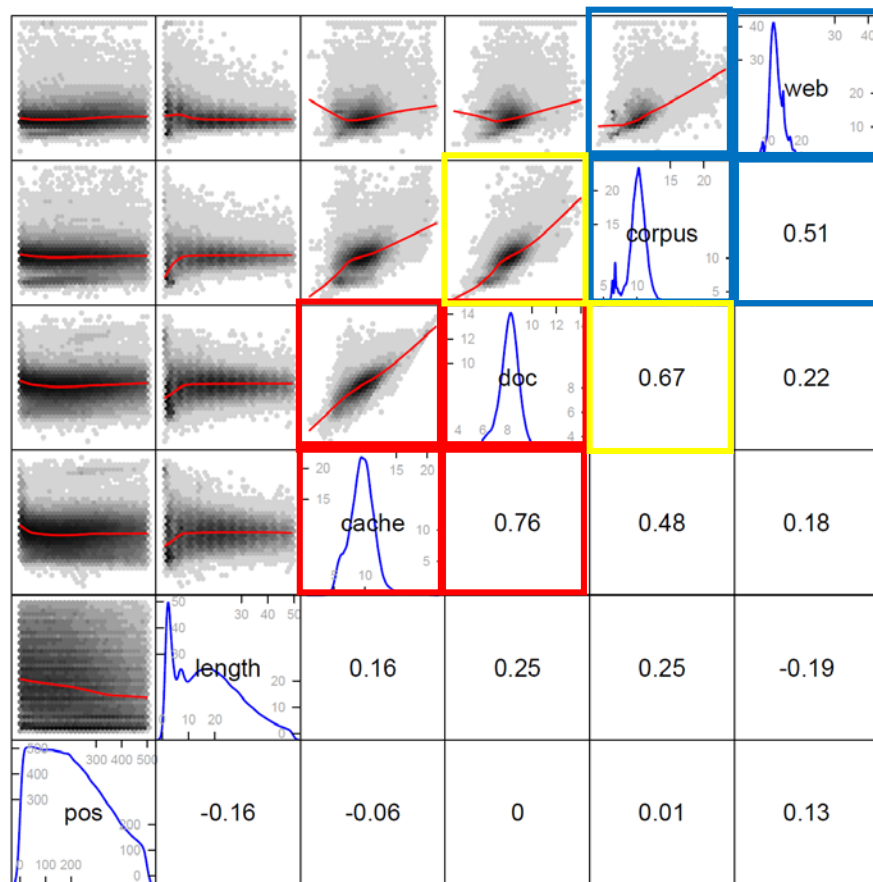


Flob: no context

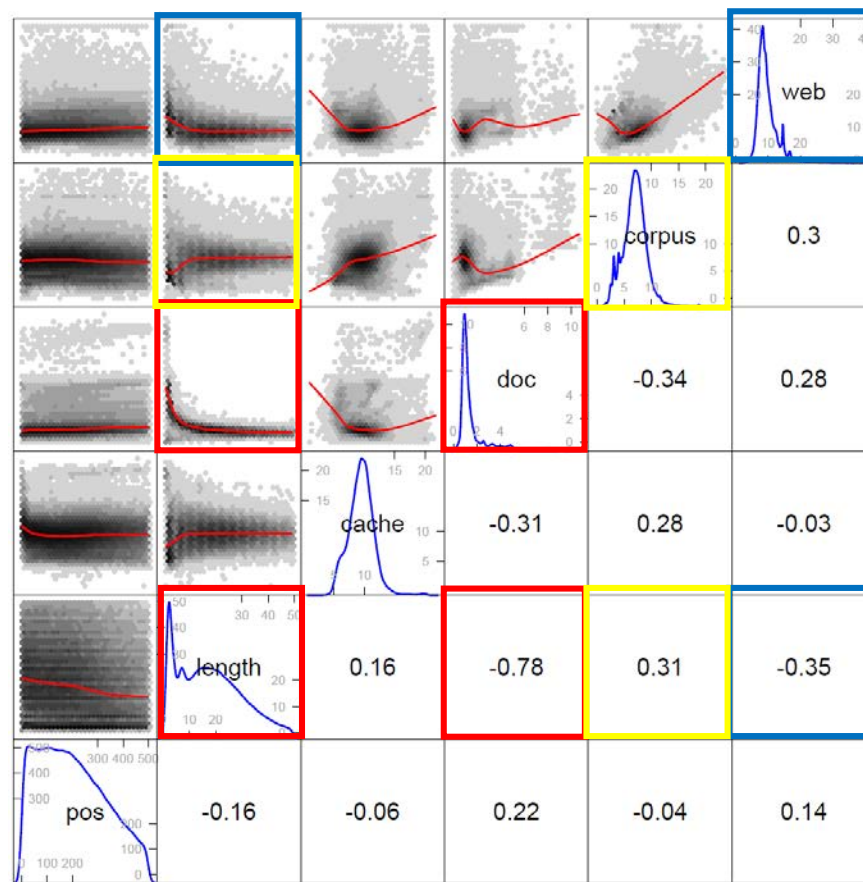


Flob: 3gram context

DaSciTex: SPLOM



DaSciTex: no context



DaSciTex: 3gram context

Summary

- Entropy & Frequency
 - Yet another good predictor for dirt.
- Entropy Rate Constancy Principle
 - Maybe not as universal as assumed
- Entropy Rate vs. Other Variables
 - Sentence position: inconclusive (only Cache LM)
 - Sentence length (Doc, 3Gram): strong correlation, by construction
- Correlation of LMs
 - Doc with Cache: by construction
 - Corpus with Web: not always
 - Doc with Corpus: in DaSciTex/Unigrams

Future Work

- Cleansing ☹️
- Better Language Models
 - Phrases: Segmenting n-grams to maximize sentence entropy
 - Longer Ranges: (Unsupervised) Dependency Parsers
 - ...
- Other contextual variables
 - POS (low hanging fruit)
 - Time
- Linguistically informed hypothesis testing

References

P.R. Clarkson and A.J. Robertson. 1997. Language Model Adaptation Using Mixtures And An Exponentially Decaying Cache. In Proceedings of ICASSP-97

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 199-206.

Dmitriy Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In Proceedings of the 2003 conference on Empirical methods in natural language processing (EMNLP '03). Association for Computational Linguistics, Stroudsburg, PA, USA, 65-72.

Frank Keller. 2004. The Entropy Rate Principle as a Predictor of Processing Effort: An Evaluation against Eye-tracking Data. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 317-324. Barcelona.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian Unsupervised Topic Segmentation. In Proceedings of EMNLP.

Ting Qian and T. Florian Jaeger. 2010. Close = Relevant? The Role of Context in Efficient Language Production. In Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics, ACL 2010, pages 45–53, Uppsala, Sweden, 15 July 2010.

Jian Huang, Jiangeng Gao, Jiangbo Miao, Xiaolong Li, Kuansan Wang, Fritz Behr: 2010. Exploring Web Scale Language Models for Search Query Processing. In Proceedings of WWW 2010.