Fractality of informativity in 300 years of English scientific writing

Yuri Bizzoni
Center for Humanities Computing
Aarhus University
yuri.bizzoni@cc.au.dk

Stefania Degaetano-Ortlieb
Language Science and Technology
Saarland University
s.degaetano@mx.uni-saarland.de

Abstract

Scientific writing is assumed to have become more informationally dense over time (Halliday, 1988; Biber and Gray, 2016). Given that scientific writing is intended for communication between experts, we hypothesize a tendency towards optimizing language use by striving for a balance between highly informative content and a conventionalized style of writing. We study this by means of fractal analysis, asking whether the degree of informativity has become more persistent with predictable patterns of gradual changes between high vs. low informational content, indicating a trend towards an optimal code for scientific communication. Specifically, surprisal is used to measure informativity and the Hurst exponent is used as a long-term dependence measure for fractality analysis, quantifying the degree of persistence of informativity in scientific texts.

1 Introduction

Fractals are the product of dynamic systems and refer to structures that are self-similar, i.e. repeat themselves on every level of scale, indicating recurrent patterns. While fractality has its origin in mathematics by researcher Benoît Mandelbrot, it has been applied to a wide range of fields which consider complex dynamic systems, such as biology (Das et al., 2016) or music (Sanyal et al., 2016). Language is a complex dynamic system that shows inherent fractal patterns, especially in language evolution, language processing, change in language use, acquisition or development (e.g., Cordeiro et al. (2015); Gao et al. (2016); Mohseni et al. (2021)). This dynamic perspective on language allows us 'to tease out the processes through which a phenomenon unfolds' (Halliday, 2007, 362).

In this paper, we are interested in how English written scientific communication has evolved over a period of $\sim$ 330 years from its beginnings (1660s) up to modern science (1990s). One general hypothesis for the development of scientific writing is that it has become increasingly informationally dense over time (Halliday, 1988; Biber and Gray, 2016), moving from increased clausal to increased phrasal complexity (i.e. from a verbal towards a heavy nominal style). Given that scientific writing is intended for communication between experts, we hypothesize a tendency towards optimizing language use by striving for a balance between highly informative content and a conventionalized style of writing. We measure the informativity of scientific texts using the information-theoretic measure of surprisal, i.e. a word's predictability in context. The more predictable a word is in its context, the less its informativity (e.g. function words are low in informativity as they are quite predictable given particular contexts, consider e.g. the expression *on behalf of*, were *of* is quite predictable given *on behalf*), while lower predictability indicates high informativity (as e.g. for scientific terms). By means of fractal analysis, we compute the Hurst exponent (Riley et al., 2012) of informativity arcs, asking whether over time the degree of informativity has become more persistent with texts showing gradually increasing and decreasing patterns of informativity (higher Hurst exponents) which repeat themselves in a text (indicated as self-similarity) or whether informativity tends to fluctuate within a text (low Hurst exponents) without a recurrent pattern.

2 Data and Methods

2.1 The Royal Society Corpus

Our data set consists of the Proceedings and Transactions of the Royal Society of London from the RSC corpus (Kermes et al., 2016; Fischer et al., 2020; Menzel et al., 2021), which covers a vast amount of English scientific writing from its beginnings in 1665 up to 1996. Given the prominent role of the Royal Society of London in scientific publishing, its articles have not only been used for diachronic linguistic studies (Atkinson, 1999;

Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL2023), pages 38–44
May 5, 2023 ©2023 Association for Computational Linguistics

Moessner, 2009; Biber and Conrad, 2014; Feltgen et al., 2017; Degaetano-Ortlieb and Teich, 2019; Degaetano-Ortlieb, 2021), but also for historical and cultural analysis (Hunter et al., 1989; Purver, 2013; Moxham and Fyfe, 2018; Degaetano-Ortlieb and Piper, 2019). The corpus consists of 47,837 texts, Table 1 showing the no. of texts and tokens across 50-year time spans.[1]

Years	Texts	Tokens
1665–1699	1.325	2.582.856
1700–1749	1.686	3.414.795
1750–1799	1.819	6.342.489
1800–1849	2.774	9.112.274
1850–1899	6.754	36.993.412
1900–1949	10.011	65.431.384
1950–1996	23.468	172.018.539

Table 1: Corpus size by texts and tokens according to approx. 50 years periods for the RSC corpus

In terms or processing, the RSC has been built in accordance with the FAIR principles (Wilkinson et al., 2016). While there is an extensive description of the corpus building procedure in Kermes et al. (2016) and Fischer et al. (2020) as well as a description of meta-data annotation in Menzel et al. (2021), we describe here some processing steps and information on meta-data relevant to this paper. Inspired by the principles of Agile Software Development (Cockburn, 2001; Voormann and Gut, 2008), corpus pre-processing, corpus annotation and linguistic analysis are intertwined and repeated cyclically. Thus, whenever problems with the corpus quality are detected by way of analysis, procedures are established to overcome these as good as possible. While we strive to maintain high quality data for the RSC, we recognize that there is always room for improvement, and we continually work towards enhancing the corpus dataset to the best of our ability. To tackle the problem of OCR-based text material, especially in the earlier periods, in version 6.0 of the RSC, we integrate the Noisy-Channel Spell Checker by Klaus et al. (2019) for a better recall and F-score at the cost of some loss in precision compared to a previously adopted method of pattern-based OCR post-correction. Considering meta-data, besides author and year of publication, which we use in this paper, there are various attributes on publication related information (e.g., journal, issn, text type), time slices (e.g., decades, 50-year periods) and textual information (e.g., pages, sentences) (cf. Fischer et al. (2020); Menzel et al. (2021)).

2.2 Informativity

Informativity is an information-theoretic notion measurable by surprisal (Shannon, 1948), which provides a useful tool for quantifying and analyzing informativity across linguistic contexts. Surprisal is defined as the negative log probability of an event measuring the amount of information conveyed by an event in bits:

$$S(w_t) = -\log P(w_t|w_{t-1}, w_{t-2}, w_{t-3}) \tag{1}$$

where $S(w_t)$ represents the surprisal of the current word w_t and $P(w_t|w_{t-1}, w_{t-2}, w_{t-3})$ represents the probability of the current word w_t given the previous three words $w_{t-1}, w_{t-2}, w_{t-3}$. The logarithm is typically taken in base 2, so that the unit of measure is bits of information. The intuition is that words with low probability convey more information than words with high probability. In the context of linguistic communication, an utterance with low surprisal conveys relatively little information, while an utterance with high surprisal conveys more information. We use surprisal to measure the degree of informativity of tokens in the RSC corpus. As the corpus presents noticeable variation in terms of corpus size and vocabulary size over time, we calculate the average surprisal value of each word in a given time period (here: decade), normalized by the vocabulary size:

$$\frac{\sum_{i=1}^{N} S(w_i)}{N} \tag{2}$$

where $S(w_i)$ is the surprisal value of word w_i, and N is the number of types in the corpus for the given time period. This controls for the effects of vocabulary size and corpus size on the average surprisal values, allowing us to make a fair comparison between time periods.

The probabilities needed for surprisal calculation are obtained by considering slices of decades, i.e. given a text, surprisal of each word in the text is calculated based on the probabilities of the words in context in the decade.[2]

2.3 Fractality

We measure the fractality of sequences with the Hurst exponent, computed on series of surprisal

[1] We excluded texts shorter than 200 sentences given that the Hurst exponent might not work well on very short sequences.

[2] Note that the RSC provides surprisal annotation at the token level based on decades, 50-year periods, and the whole corpus and given the pre-processed material.

values averaged on sentences for each RSC text. Recently, fractality has been applied to analyze dynamics in language use such as sentiment arcs in stories (Gao et al., 2016), on stylometric and sentiment features finding correlations to the perceived 'beauty' of a text (Cordeiro et al., 2015; Bizzoni et al., 2022), and for determining differences between fiction and non-fiction texts (Mohseni et al., 2021). A Hurst exponent >0.5 indicates relatively smooth transitions between highly informative and less informative sentences, i.e. rather gradual changes in informativity pointing to a relatively uniform information distribution (cf. uniform information density hypothesis (UID) (Jaeger and Levy, 2007; Jaeger, 2010)). These transitions will form patterns which are repeatedly encountered in a text (i.e. self-similarity of persistent trends). In our specific case of studying scientific writing, we would expect a rather high Hurst exponent, which would confirm our hypothesis towards striving for a balance between highly informative content and a conventionalized style of writing for expert-to-expert communication. On the other hand, a Hurst exponent <0.5 would suggest rather abrupt changes between more vs. less informative sentences (i.e. anti-persistent trends), which we would not assume to be the case.

We estimate the Hurst exponent by Adaptive Fractal Analysis (AFA) (Gao et al., 2011), by which time series (here: sentences in a text) are partitioned into overlapping segments of length $w = 2n + 1$, where neighboring segments overlap by $n + 1$ points. In each segment, the time series is fitted with the best polynomial of order M using standard least-squares regression. The fitted polynomials in the overlapping regions are then combined to yield a single global smooth trend (cf. Riley et al. (2012) for an introduction). The Hurst exponent is estimated based on the fluctuations around this trend and the scale at which these fluctuations occur. The original time series is denoted by $x_1, x_2, ..., x_T$ and the fitted polynomials for the i^{th} and $(i + 1)^{th}$ segments are denoted by $y_i(l_1)$ and $y_{i+1}(l_2)$, respectively, where $l_1, l_2 = 1, 2, ..., 2n + 1$. The fluctuations around the smooth trend's mean m can be measured by the residuals: $residual_i = x_i - m$. The scale-dependent fluctuations can be measured by the fluctuation function $F(n)$, which is given by:

$$F(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} residual_i^2} \qquad (3)$$

The Hurst exponent H is estimated as the slope of the regression line of $\log F(n)$ against $\log n$.

3 Analysis

Using the Hurst exponent, we ask whether the distribution of informativity within the RSC texts becomes more persistent over time, which would indicate a more uniform and smooth distribution of information within articles, i.e. a more persistent *informativity profile*.

Informativity is measured by surprisal (cf. Section 2.2). Example (1) shows a sentence with low informativity on average (4.6 of surprisal), where words are relatively predictable given their previous context (such as *Newton* given *Sir Isaac*). Example (2) shows a sentence which is higher in informativity on average (surprisal of 8.1), where *illustrated* is relatively unpredictable given its previous context (compare to the more explicit version *which occurred and which illustrated* which would lower the average informativity). Example (3) presents a highly informative sentence (surprisal of 10.9) due to the terms used in it.

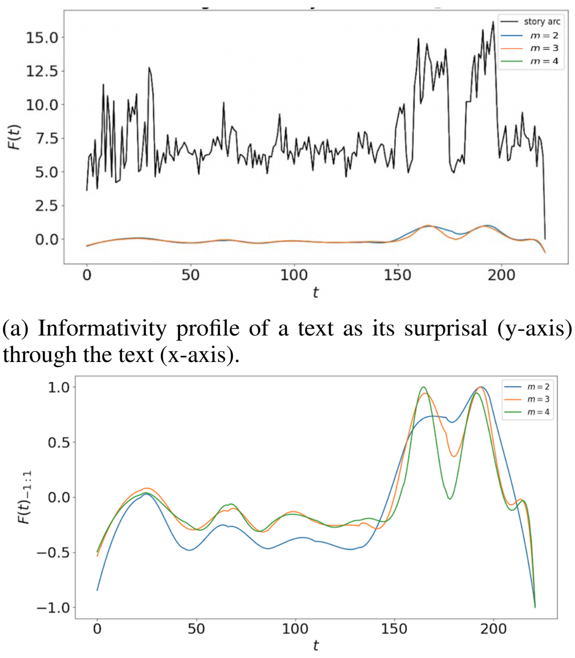
(1) And$_{5.29}$ that$_{4.97}$ something$_{8.78}$ like$_{4.40}$ this$_{7.12}$ must$_{8.52}$ be$_{0.94}$ the$_{4.49}$ Case$_{6.92}$,$_{2.84}$ appears$_{12.91}$ from$_{2.61}$ what$_{0.92}$ Sir$_{5.25}$ Isaac$_{0.11}$ Newton$_{0.48}$ has$_{3.86}$ said$_{8.84}$ upon$_{1.98}$ this$_{2.42}$ Subject$_{3.67}$.$_{3.14}$ (J.T. Desaguliers 1724, average sentence surprisal 4.6)

(2) Another$_{10.09}$ effect$_{9.16}$ which$_{6.91}$ occurred$_{10.39}$ illustrated$_{15.13}$ the$_{4.10}$ same$_{5.28}$ point$_{7.44}$.$_{4.61}$ (M. Faraday 1846, average sentence surprisal 8.1)

(3) He$_{7.49}$ also$_{4.45}$ accepted$_{9.85}$ Van$_{12.92}$ Slyke$_{9.12}$ amino-N$_{23.48}$ determinations$_{16.54}$.$_{3.54}$ (R.L.M. Synge et al. 1990, average sentence surprisal 10.9)

We compute the Hurst exponent on the articles' sentence-based informativity arcs. Figure 1 and 2 show the lines of two arcs with extremely high ($H = 0.92$) vs. low ($H = 0.27$) exponents. Informativity is on the y-axis of Figure 1a and 2a with the time series of the texts (i.e. sentences) on the x-axis. Figure 2a shows high fluctuations around the mean with rather abrupt changes in surprisal values, while Figure 1a shows a much smoother trend, with smaller and gradual changes in informativity from low to high and vice versa. Figure 1b and 2b present a globally smooth trend signal, represented as a polynomial fit on the detrended informativity profile[3]. A detrended profile is a profile where the datapoints' values are subtracted to the mean;

[3] With linear trend (m1), quadratic trend (m2), and cubic trend (m3).



(a) Informativity profile of a text as its surprisal (y-axis) through the text (x-axis).

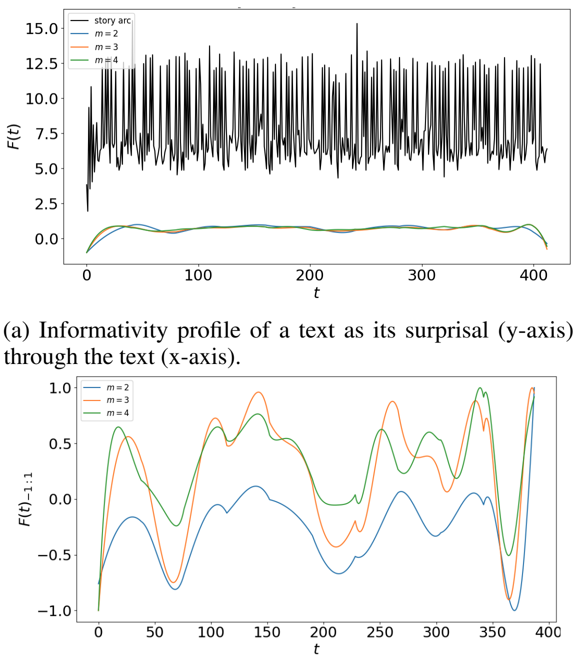
(b) Detrended profile: three alternative polynomial fits.

Figure 1: Informativity profiles (raw and detrended) of a text with high Hurst exponent ($H = 0.92$).

polynomial fits on such a signal allow us to estimate the series' underlying systematic components and to forecast its long-term behaviour (see also Riley et al. (2012) on obtaining global lines, i.e. detrended lines).

In general, texts with a low Hurst exponent contain sentences with strongly varying averages in surprisal (low-high-low etc), while texts with a relatively high Hurst exponent are built up by sequences of sentences presenting a gradual increase or decrease in surprisal (low-lower-high-higher-highest-higher-high-lower-low etc). Thus, a Hurst exponent of >0.5 indicates long-term trends in the informativity profile of a text and more persistent patterns (e.g. a gradual increase in informativity followed by a gradual decrease or vice versa). A Hurst exponent of 0.5 indicates abrupt changes with unpredictable peaks and troughs in informativity, while a value of <0.5 suggests an anti-persistent trend, i.e. a trend reverting constantly to the mean through a "zig-zag"-like behaviour.

Figure 3 shows an overall trend of the Hurst exponent for the RSC texts averaged over each year, with the averaged exponent value on the y-axis and years on the x-axis. For almost all years the Hurst exponent is >0.5. From the 1650s to 1800 there



(a) Informativity profile of a text as its surprisal (y-axis) through the text (x-axis).

(b) Detrended profile: three alternative polynomial fits.

Figure 2: Informativity profiles (raw and detrended) of a text with low Hurst exponent ($H = 0.27$).[4]

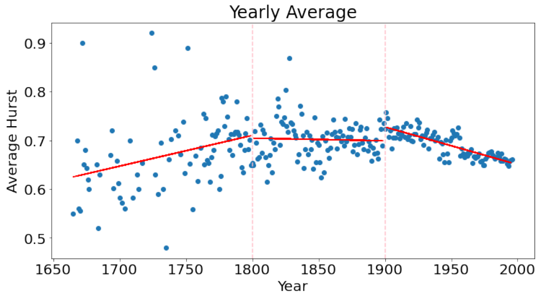


Figure 3: Hurst exponent of RSC texts over time averaged on each year.

is an increase, followed by a plateau until 1900 and a slight decrease until the 1990s. Spearman's correlation between average Hurst and the articles' publication date is positive and significant until roughly the late 18th century (0.43), negative and significant from the beginning of the 20th century (-0.79), and non-significant in-between. This indicates that the RSC texts show rather persistent informativity sequences, a tendency that becomes stronger through the 18th century, i.e. changes in informativity across the texts become more coherent showing recurrent patterns of change in a text. This is in line with psycholinguistic accounts on language processing. Given that smoother signals of informativity have shown to be related to less processing effort (Jaeger and Levy, 2007; Jaeger, 2010), the observed change might indeed indicate

[4] To best show mean-reverting patterns, (b) excludes the first and last ten sentences of the article.

change towards more coherent texts in English scientific writing to serve expert-to-expert communication, where a smoother signal indicates a more uniform distribution of information that goes beyond the sentence unit.

However, after a relatively stable period, informativity profiles become slightly less persistent in the 20th century. This could indicate a development of a scientific code at pressure given the highly demanding process of increased specialization, i.e. growing specialized domains, in which formulaic language and grammatical consolidation are combined with an increasingly diverse, domain-specific use of terminology. These high demands might introduce sharper changes in the informativity profiles of more contemporary articles, leading to the slight disrupt of the "smoothness" of the informativity trends, slightly lowering the overall Hurst score.

Finally, it is worth noting that while the texts' average Hurst exponent oscillates through macro-periods (here centuries), its variance and standard deviation decline steadily (Table 2 and Figure 4), i.e. the differences between the articles' informativity profiles becomes smaller – scientific authors converging on a particular range of informativity profiles.

Measure	Spearman corr.	Kendall corr.
Mean	0.160	0.076
Variance	-0.664* [-0.83, -0.4]	-0.517* [-0.74, -0.2]
Std	-0.619* [-0.8,-0.33]	-0.476* [-0.71,-0.13]

Table 2: Correlations between time and Hurst's decade-based average, variance, and standard deviation. All measures were taken starting from 1700, to avoid distortions induced by the data scarcity of the 17th century. *indicates p-values <0.05; confidence intervals for alpha=0.05 are in brakets.

4 Conclusion

We have modeled fractality for English scientific texts given their informativity over time on the RSC Corpus, showing a general trend towards the use of smoother informativity profiles. This is in line with previous accounts on information density, which hypothesize uniform distributions to ease processing cost (Jaeger and Levy, 2007; Jaeger, 2010). Here, we have shown that this also adheres at the textual level. Considering that scientific writing is meant for expert-to-expert communication, being subject to increased processes of specialization, more contemporary scientific writing shows trends towards slightly less smoothed signals, with stronger alternations between more conventionalized, formulaic vs. highly specialized informational content. At the same time, we observed a strong converging trend indicated by a significant reduction in variance and standard deviation of the Hurst exponent. In future, we would like to see whether this notion of an optimized informativity signal is observable in the development of other registers. Also, there might be discipline-specific trends which would shed light on processes of specialization and diversification among scientific disciplines. Moreover, as we continuously work on enhancing corpus quality, we would like to have a through analysis of possible confounds that might have an impact on surprisal as well as fractality calculation. This will lead us to uncover more comprehensively source for changes in fractality. Also, as we here have applied one of the most simple ways of calculating fractality, we want to experiment with other measures in order to evaluate their performance for this task.

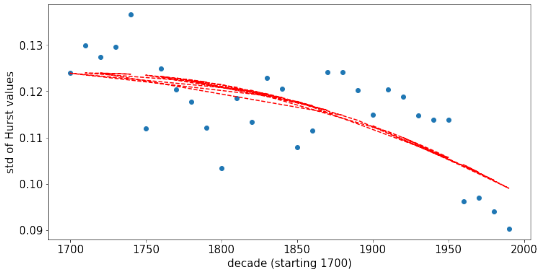


Figure 4: Standard deviation for each decade in our corpus starting 1700, with a polynomial fit. The standard deviation diminishes almost linearly, which seems to indicate a progressive stylistic convergence.

5 Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 and by *Velux Foundation* - Project *Fabula-Net* - Project-ID 35361. We would also like to express our sincere gratitude to the reviewers for their insightful comments and constructive feedback, which greatly contributed to the improvement of this paper.

References

Dwight Atkinson. 1999. *Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. Erlbaum, New York.

Douglas Biber and Susan Conrad. 2014. *Variation in English: Multi-dimensional studies*. Routledge.

Douglas Biber and Bethany Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Studies in English Language. Cambridge University Press, Cambridge, UK.

Yuri Bizzoni, Telma Peura, Mads Thomsen, and Kristoffer Nielbo. 2022. Fractal Sentiments and fairy tales - Fractal scaling of narrative arcs as predictor of the perceived quality of Andersen's fairy tales. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Alistair Cockburn, editor. 2001. *Agile Software Development*. Addison-Wesley Professional, Boston, USA.

Joao Cordeiro, Pedro RM Inácio, and Diogo AB Fernandes. 2015. Fractal beauty in text. In *Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings 17*, pages 796–802. Springer.

Nandan Kumar Das, Rajib Dey, Semanti Chakraborty, PK Panigrahi, and Nirmalya Ghosh. 2016. Probing multifractality in depth-resolved refractive index fluctuations in biological tissues using backscattering spectral interferometry. *Journal of Optics*, 18(12):125301.

Stefania Degaetano-Ortlieb. 2021. Measuring informativity: The rise of compounds as informationally dense structures in 20th century scientific english. In Elena Soave and Douglas Biber, editors, *Corpus Approaches to Register Variation*, chapter 11, pages 291–312. John Benjamins Publishing Company.

Stefania Degaetano-Ortlieb and Andrew Piper. 2019. The scientization of literary study. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 18–28, Minneapolis, USA. Association for Computational Linguistics.

Stefania Degaetano-Ortlieb and Elke Teich. 2019. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*, 0(0):1–33. Online print.

Quentin Feltgen, Benjamin Fagard, and Jean-Pierre Nadal. 2017. Frequency patterns of semantic change: Corpus-based evidence of a near-critical dynamics in language change. *Royal Society Open Science*, 4(11):170830.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0. providing 300+ years of scientific writing for humanistic study. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC) 2020, Marseille, France, May 2020*, pages 794–802.

Jianbo Gao, Jing Hu, and Wen-wen Tung. 2011. Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering. *PLOS ONE*, 6(9):1–8.

Jianbo Gao, Matthew L Jockers, John Laudun, and Timothy Tangherlini. 2016. A multiscale theory for the dynamical evolution of sentiment in novels. In *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, pages 1–4. IEEE.

M.A.K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy, editor, *Registers of Written English: Situational Factors and Linguistic Features*, pages 162–177. Pinter, London.

M.A.K. Halliday. 2007. On the concept of 'Educational linguistics'. In Jonathan J. Webster, editor, *The collected works of M. A. K. Halliday, Volume 9: Language and education*, pages 354–367. Continuum, London. Originally published in: Giblett, R., & O'Carroll, J. (Eds.). (1990). Discipline, dialogue, difference: Proceedings of the Language in Education conference, Murdoch University, December 1989. Murdoch, Australia: Duration, pp. 23–42.

Michael Cyril William Hunter et al. 1989. *Establishing the new science: The experience of the early Royal Society*. Boydell & Brewer Ltd.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

T. Florian Jaeger and Roger P. Levy. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schölkopf, John C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 849–856. MIT Press.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The Royal Society Corpus: From uncharted data to corpus. In *Proceedings of the 10th LREC*, Portorož, Slovenia.

Carsten Klaus, Dietrich Klakow, and Peter Fankhauser. 2019. OCR post-correction of the Royal Society Corpus based on the noisy channel model. In *Proceedings of the 41. Jahrestagung der Deutschen Gesellschaft fuer Sprachwissenschaft (DGfS2019)*, University of Bremen, Germany.

Katrin Menzel, Jörg Knappen, and Elke Teich. 2021. Generating linguistically relevant metadata for the Royal Society Corpus. *Research in Corpus Linguistics*, 9(1):1–18.

Lilo Moessner. 2009. The influence of the Royal Society on 17th-century scientific writing. *ICAME journal*, 33:65–87.

Mahdi Mohseni, Volker Gast, and Christoph Redies. 2021. Fractality and variability in canonical and non-canonical English fiction and in non-fictional texts. *Frontiers in Psychology*, 12.

Noah Moxham and Aileen Fyfe. 2018. The Royal Society and the prehistory of peer review, 1665–1965. *The Historical Journal*, 61(4):863–889.

Margery Purver. 2013. *The Royal Society: Concept and Creation*. Routledge.

Michael A. Riley, Scott Bonnette, Nikita Kuznetsov, Sebastian Wallot, and Jianbo Gao. 2012. A tutorial introduction to adaptive fractal analysis. *Frontiers in Physiology*, 3:371.

Shankha Sanyal, Archi Banerjee, Anirban Patranabis, Kaushik Banerjee, Ranjan Sengupta, and Dipak Ghosh. 2016. A study on improvisation in a musical performance using multifractal detrended cross correlation analysis. *Physica A: Statistical Mechanics and its Applications*, 462:67–83.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Holger Voormann and Ulrike Gut. 2008. Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4:235–251.

Mark Wilkinson, Michel Dumontier, and IJsbrand Aalbersberg. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Nature – Scientific Data*, 3(160018).