P-LEADER MULTIFRACTAL ANALYSIS FOR TEXT TYPE IDENTIFICATION

R. Leonarduzzi[1]*, P. Abry*[1]*, S. Jaffard*[2]*, H. Wendt*[3]*,*
L. Gournay[4]*, T. Kyriacopoulou*[5]*, C. Martineau*[5]*, C. Martinez*[5]

[1] Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Lab de Physique, F-69342 Lyon, France
{roberto.leonarduzzi,patrice.abry}@ens-lyon.fr
[2] Univ Paris Est, LAMA, UPEC, CNRS, F-94010, Créteil, France jaffard@u-pec.fr
[3] Univ de Toulouse, IRIT-ENSEEIHT, CNRS, Toulouse, France herwig.wendt@irit.fr
[4] Univ Paris Est, IMAGER, UPEC, F-94010, Créteil, France lucie.gournay@u-pec.fr
[5] Univ Paris Est, LIGM, UPEM, F-77420, France tita.kyriacopoulou@univ-mlv.fr

ABSTRACT

Among many research efforts devoted to automated art investigations, the problem of quantification of literary style remains current. Meanwhile, linguists and computer scientists have tried to sort out texts according to their types or authors. We use the recently-introduced p-leader multifractal formalism to analyze a corpus of novels written for adults and young adults, with the goal of assessing if a difference in style can be found. Our results agree with the interpretation that novels written for young adults largely follow conventions of the genre, whereas novels written for adults are less homogeneous.

***Index Terms*—** Multifractal analysis, p-leaders, text analysis, stylometry

1. INTRODUCTION

Automated and computerized analysis of texts. The interest in automated and computerized art investigations increased recently [1–3]. This has notably been the case for the analysis of literary texts [4–8], with diverse goals, such as authorship detection [5], quantification of translation effects [4] or style [6, 7, 9].

Features used in text-resemblance measurement can be strictly quantitative (sentence length, word length or frequency, n-grams…) or they can be qualitatively motivated (lexical or syntactic cues…). Impressive progress has been made in genre classification, but time-saving methods capable of dealing with fuzzy genre-types or multi-genre texts remain to be developed [10]. In order to use mathematical and statistical tools introduced in signal processing, a first and key step consists in deriving from the text a digital signal that usually consists of quantified counts on text attributes. Such data have been analyzed using a wide range of tools. For instance, descriptive statistics [5] or heavy-tailed distributions [11], provide overall static information. Additionally, data may be analyzed as time series, where temporal dynamics are assumed to be irregular and complex and to convey crucial information for analysis and interpretation. In this case, suitable tools range from entropies [12] to multifractal analysis [4, 7].

Multifractal analysis. This stands out as a convenient method to quantify the irregularity of a time series; it is based on the study of the scale-invariance properties of the signal, i.e. assumes that no time-scale plays a predominant role in the description of the time-series' dynamics, but rather that all time-scales should be considered jointly. Though there exist several implementations of multifractal analysis (e.g. the *wavelet modulus maxima* [13], *multifractal detrended fluctuation analysis* [14] and *wavelet leaders* [15]) a new variant has recently been proposed: the p-leader multifractal formalism [16, 17]. This technique provides potentially richer information [16, 18] with improved and robust estimation performances [17].

Identification of writing styles. An important problem in the field of natural language processing is the identification of relevant, easily-selectable cues to determine similarities between text types. N-grams methods have shown that counter-intuitive digital signals were quite reliable to identify author or genre [9]. Yet, it is another challenge to automatically estimate resemblance between literary texts, i.e. rich, creative writing productions that most readers can intuitively identify according to audience target. One hypothesis is that young adult (YA) novels are not written in the same way than adult (AD) novels. Independently of topic content, the YA / AD classification, provided by publishers, seems to be grounded on narrative and linguistic features, linked to a scale of text difficulty (see [19] for automated evaluation of text difficulty). Yet, it is not clear what the actual reality of this alleged classification is.

Goals, contributions and outline. In this context, we pro-

This work was supported by ANR grant AMATIS #112432.

Table 1. **Corpus.** List of novels in the analyzed corpus.

ID	Title	Author	Class
1	The Amber Spyglass	P. Pulman	YA
2	Bloody Red Baron	K. Newman	AD
3	The Bad Beginning	L. Snicket	YA
4	A Bend on the Road	N. Sparks	AD
5	The Blade Itself	J. Abercrombie	AD
6	The Dante Club	M. Pearl	AD
7	Divergent	V. Roth	YA
8	Hunger Games	S. Collins	YA
9	Legend	M. Lu	YA
10	Little Friend	D. Tartt	AD
11	The Luminaries	E. Catton	AD
12	Missing You	H. Coben	AD
13	Never Let Me Go	K. Ishiguro	AD
14	Oracle Night	P. Auster	AD
15	The Order of the Phoenix	J.K. Rowling	YA
16	The Lightning Thief	R. Riordan	YA
17	Seconds Away	H. Coben	YA
18	Selection	K. Cass	YA
19	The Silent Boy	L. Lowry	YA
20	Struck by Lightning	C. Colfer	YA
21	Trading in Danger	E. Moon	AD
22	Twilight	S. Meyer	YA
23	The Casual Vacancy	J.K. Rowling	YA
24	Whale Talk	C. Crutcher	YA
25	Youth: Scenes from Provincial Life II	J.M. Coetzee	AD

pose to use the p-leader multifractal formalism [16, 17] to provide quantitative information that enables to distinguish between novels written for Adults (AD) and Young Adults (YA). We focus on time series consisting of the word count of sentences. In that regard, we use a corpus of 25 contemporary novels in English written for AD and YA (described in Section 2). From these novels, we extracted time series consisting of sentence lengths (in words), as described in Sec. 2. Then, we applied the p-leader multifractal formalism, described in Sec. 3, to obtain the results detailed and discussed in Sec. 4.

2. MATERIALS

Book collection. We analyzed a corpus of 25 best-sellers, written in English, and published between 1989 and 2014; 13 are targeted at YA, while the remaining 12 are targeted at AD. Table 1 shows information on these novels. The left-column tags are used as identifiers for the plots in Sec. 4.

Preprocessing. We identified the sentences in all novels automatically by means of the *Unitex* corpus processing suite[1]. This software provided landmarked sentences for each novel. Each sentence was further split at whitespace characters into a list of words, the length of which was measured to form the sentence-length time-series that we analyze. Table 2 shows information on the length of the time series for each class. Further, Fig. 1 shows examples of such time series for both YA (top) and AD (bottom) novels, exhibiting highly irregular dynamics in both cases. In order to use the tools of multifractal analysis, we now interpret these discrete sequences as functions of a continuous real parameter t which have been discretized.

[1]Originally developed by Sébastien Paumier, and now a community project. See `http://www-igm.univ-mlv.fr/~unitex/`.

Table 2. **Novel lengths.** Statistics on the number of sentences for the novels in the corpus.

	mean	std	min	max
YA	10603.2	5225.54	3271	20298
AD	9713.85	7222.78	1666	29247

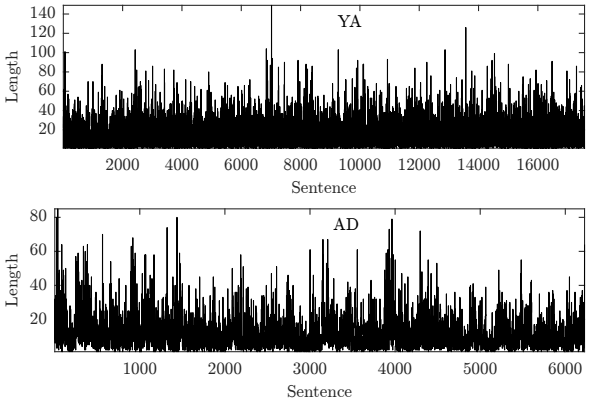

Fig. 1. **Example data.** Sentence length time series for YA (top) and AD (bottom).

3. P-LEADER MULTIFRACTAL ANALYSIS

Local regularity. Let $X \in L^p_{loc}(\mathbb{R})$ for $p \geq 1$. X is said to belong to $T^p_\alpha(t)$, with $\alpha > -1/p$, if there exist $C, R > 0$ and a polynomial P_t (with $deg(P_t) < \alpha$) such that $\forall \alpha < R$, $\left(\frac{1}{a}\int_{t-a/2}^{t+a/2} |X(u) - P_t(u-t)|^p \, du\right)^{1/p} \leq Ca^\alpha$. The *p-exponent* of X at t is defined as $h_p(t) = \sup\{\alpha : X \in T^p_\alpha(t)\}$, and provides a measure of the regularity of X at time t. The particular case where $p = \infty$ corresponds to the well-known *Hölder exponent* [20]. The p-exponent can be considered as a natural extension of the Hölder exponent that allows to deal with functions that are not bounded (but belong to L^p), and for which h_p may admit negative values (larger than $-1/p$).

Multifractal spectrum. For multifractal models, estimation of the function $h_p(t)$ itself (which is highly erratic) is not useful, and the distribution of its values is more pertinent. One thus considers the *multifractal spectrum* $D_p(h)$, defined as the Hausdorff dimension of the set of points t where $h_p(t)$ takes a given value: $D_p(h) = \dim_{\mathrm{H}}(\{t : h_p(t) = h\})$

[16]. Since neither Hausdorff dimensions nor pointwise p-exponents can be estimated directly from their definitions, a procedure termed *multifractal formalism* is used for the estimation of the multifractal spectrum [21].

Wavelet coefficients. Let ψ be a *mother wavelet*, i.e. a zero-average compactly-supported function with N_ψ vanishing moments (i.e. orthogonal to polynomials of order less than N_ψ), and such that the $\{\psi_{j,k}(t) = 2^{-j}\psi(2^{-j}t - k)\}_{(j,k)\in\mathbb{N}^2}$ (obtained by dilations and translations of ψ) form an orthonormal basis of $L^2(\mathbb{R})$. The (L^1-normalized) discrete wavelet transform coefficients are defined as $c_{j,k} = 2^{-j}\langle\psi_{j,k}|X\rangle$ (cf. e.g. [22], for details on wavelet transforms).

Wavelet *p*-leaders. *p-leaders* are defined as local L^p-norms of scaled wavelet coefficients [16, 17]

$$\ell^{(p)}_{j,k} := \Big(\sum_{\lambda'\subset 3\lambda} |c_{\lambda'}|^p \; 2^{j'-j}\Big)^{\frac{1}{p}}, \qquad (1)$$

with $\lambda = \lambda_{j,k} = [k2^j, (k+1)2^j)$, $c_\lambda = c_{j,k}$ and $3\lambda = \bigcup_{m\in\{-1,0,1\}} \lambda_{j,k+m}$. That is, the local norm considers all wavelet coefficients in a narrow time neighbourhood of $t = 2^j k$, and for all finer scales $j' \le j$. p-leaders allow to measure the p-exponent since $\ell^{(p)}_{j,k} \sim 2^{h_p(t)}$ for $j \to \infty$ [16, 18].

Multifractal formalism. The p-leader multifractal formalism allows to estimate $D_p(h)$ from the scale-invariance properties of the $\ell^{(p)}$. It is defined as follows [17]: First, structure functions are defined as sample moments of p-leaders:

$$S_p(j,q) = 2^j \sum_{k=1}^{2^{-j}} \Big(\ell^{(p)}_{j,k}\Big)^q \sim 2^{j\zeta_p(q)}, \quad j \to -\infty. \qquad (2)$$

The *scaling function* $\zeta_p(q)$ provides information on the scale-invarance properties of X, and can be estimated by linear regressions of $\log_2 S_p(j,q)$ versus j. Then, the concave hull of $D_p(h)$ is obtained as the Legendre transform of $\zeta_p(q)$ [20]:

$$D_p(h) \le \mathcal{L}_p(h) := \min_q(1 + qh - \zeta^{(p)}(q)). \qquad (3)$$

In practice, the function $\mathcal{L}_p(h)$ is the only accessible quantity and is used as the estimate of $D_p(h)$.

Log-cumulants. Instead of computing the function $\zeta(q)$ directly from the p-leaders, we will follow [23] and compute a polynomial approximation instead: $\zeta(q) = \sum_{m=1}^{\infty} c_m q^m/m!$. Interestingly, the coefficients c_m, which are called *log-cumulants*, are related to the decay across the scales of the cumulants $C^{(p)}_m(j) = \mathrm{Cum}_m \ln \ell^{(p)}_{j,\cdot}$

$$C^{(p)}_m(j) = C^{(p)}_0 + c^{(p)}_m \ln 2^j, \quad j \to -\infty. \qquad (4)$$

Therefore, $c^{(p)}_m$ can be estimated by linear regressions of $C^{(p)}_m(j)$ against $\ln 2^j$, for all scales $j \in [j_1, j_2]$. The coefficient c_1 measures the second order (or correlation) properties of X, whereas $c^{(p)}_m$, $m \ge 2$, characterize the higher-order statistical behavior of X. In practice, the energy of the cumulants decreases rapidly with m, so that most of the multifractal information is contained in the first two *log-cumulants*: $c^{(p)}_1$ measures the dominant regularity, while $c^{(p)}_2$ estimates the range of p-exponents present in the data.

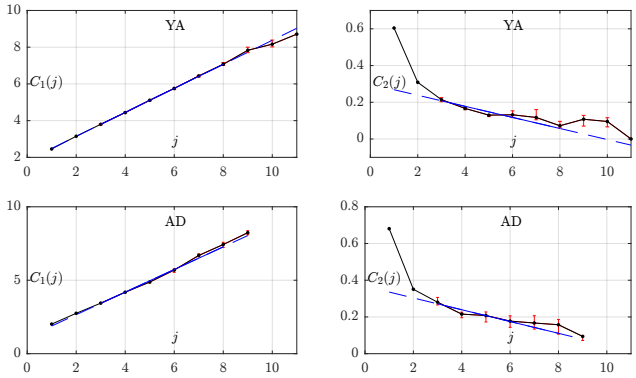

Fig. 2. **Sample logscale diagrams,** for $C^{(p)}_1(j)$ (left column) and $C^{(p)}_2(j)$ (right column), and for both classes (top and bottom rows). The regression line is shown in blue.

4. RESULTS AND DISCUSSION

Analysis parameters. The analysis was performed using a Daubechies wavelet with $N_\psi = 3$ vanishing moments. The value $p = 2$ was used for p-leaders, since it offers improved estimation performance as shown in [17]; we checked that this choice satisfies the minimum regularity conditions detailed in [16,17]. We verified that results do not change substantially for different choices of p. All regressions were performed in the range of scales $j \in [3, 8]$, which corresponds to $2^3 = 8$ to $2^8 = 256$ sentences.

Scale invariance. Fig. 2 shows examples of logscale diagrams (cf. (4)) for $C_1(j)$ (left column) and $C_2(j)$ (right column), corresponding to the two time series in Fig. 1. All plots display a good linear behavior in the range of scales $j \in [3, 8]$, evidencing the scale-invariant nature of the data (in agreement with results in [7] for a different corpus). This indicates that the analysis of a single sentence length is insufficient to fully characterize the dynamics; rather, the relationship between all sentence lengths needs to be considered. Thus, analysis of logscale diagrams in these two representative cases highlights the relevance of the multifractal paradigm for this kind of data. A similar behavior was seen in all time series under analysis.

Multifractal spectra. Fig. 3 shows the multifractal spectra corresponding to the two time series in Fig. 1. Both spectra have a wide support, showing clearly the multifractal nature of the data. Moreover, they suggest that the data only have singularities of exponent $h > 0.5$, indicating the presence of

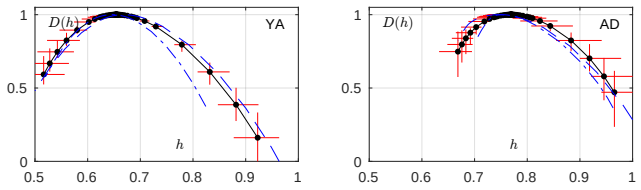

Fig. 3. Sample multifractal spectra (black solid line with points), for YA (right) and AD (left). The blue dashed and dashed-dotted lines correspond to the spectra computed from the first and second halves of the time series, respectively.

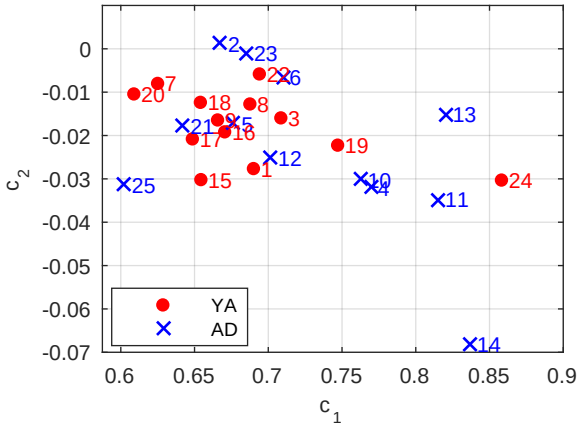

Fig. 4. **Multifractal features.** The number tags correspond to those listed in Table 1.

positive correlations in the lengths of sentences. This characteristic is shared by all analyzed texts.

Further, Fig. 3 shows the multifractal spectra computed from the first and second halves of each time series (blue dashed and dash-dotted lines, respectively). These spectra are remarkably similar to the one computed from the entire time series. This clearly shows that multifractality and scale invariance provide robust and consistent characterizations of temporal dynamics across the entire books, and that they are not an artifact caused by nonstationarity.

Multifractal features. Fig. 4 shows the scatter plots of coefficients c_1 and c_2 for all texts, with YA and AD shown in red dots and blue crosses, respectively. Interestingly, texts for YA mostly group in a well-formed cluster around point $(0.65, -0.02)$, while texts for AD appear to be less homogeneous and scattered throughout the plot. This result might be understood in the sense that novels for YA are probably less "creative" in terms of narrative style, meaning that the rules of the genre prevail over the author's style.

A detailed analysis of metadata and literary information concerning the novels is useful to better understand some points that appear *a priori* as outliers. Novel 19, *"The Silent Boy"*, is categorized by its publisher as both a young adult novel and a historical fiction. Novel 24, *"Whale talk"* uses coarse language and features "adult" content. Contrary to other YA novels, both 19 and 24 are realistic fictions studied in class for their literature quality. Thus, both books have been perceived by critics and readers to not fully belong to the YA class because of either their theme or style. Interestingly, these differences are picked up also by the multifractal features, which clearly makes such novels stand out.

Two AD novels, 5 and 21, fall clearly into the cluster of YA novels. Interestingly, novel 5, *"The Blade Itself"*, is a fantasy novel whose author recently published a young adult novel series. Novel 21, *"Trading in Danger"*, is a science fiction novel whose author is also prolific in fantasy novels.

Two authors, H. Coben and J.K. Rowling were represented in the corpus in both classes (HC: novel 12 (AD), 17 (YA); JKR: novel 23 (AD), 15 (YA)). It is interesting to notice that their novels appear in the same cluster, even though Rowling's AD novel is located at the margin. On the whole, the multifractal features allow to identify the most unconventional and literary works independently of the AD/YA classification. Thus, novels by E. Catton (11), K. Ishiguro (13), P. Auster (14), L. Lowry (19), C. Crutcher (24), JM Coetzee (25) are located outside the cluster. Yet, the vicinity of N. Sparks' and D. Tartt's novels is surprising, the first one being a typical romance and the second an original literary novel. As for the AD/YA classification – independently of topic content – it seems correlated to literary quality and convention only.

5. CONCLUSIONS

We have presented preliminary results on the use of multifractal features to quantify the differences in style of contemporary novels targeted at adults or young adults. We have used the state-of-the-art p-leader multifractal formalism to analyze a moderately-sized corpus of novels. Our results show that the style of novels targeted at young adults is more homogeneous and consistent, suggesting that in this case the rules of the genre predominate over the style of the author. These encouraging results will serve as a starting point for a more extensive analysis, where the corpus will be extended with other novels to test again the author parameter (eg. Cohen, Rowling) and the fiction genre (eg. romance, fantasy, realistic) parameter. Also, other digital signals than sentence length will be taken into account to improve resemblance measurement.

6. REFERENCES

[1] C. R. Johnson Jr., E. Hendriks, I. J. Berezhnoy, et al., "Processing for artist identification: Computerized analysis of Vincent van Gogh's painting brush- strokes," *IEEE Signal Process. Mag.*, vol. 25, pp. 37–48, 2008.

[2] P. Abry, H. Wendt, and S. Jaffard, "When Van Gogh

meets Mandelbrot: Multifractal classification of painting's texture," *Signal Proc.*, vol. 93, no. 3, SI, pp. 554–572, 2013.

[3] P. Abry, S. G. Roux, H. Wendt, et al., "Multiscale anisotropic texture analysis and classification of photographic prints: Art scholarship meets image processing algorithms," *IEEE Signal Process. Mag.*, vol. 32, no. 4, pp. 18–27, July 2015.

[4] M. Ausloos, "Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series," *Phys Rev E*, vol. 86, no. 3, pp. 031108, 2012.

[5] X. Hu, Y. Wang, and Q. Wu, "Multiple authors detection: a quantitative analysis of dream of the red chamber," *Advances in Adaptive Data Analysis*, vol. 06, no. 04, pp. 1450012, Oct. 2014.

[6] J. M. Hughes, N. J. Foti, D. C. Krakauer, and D. N. Rockmore, "Quantitative patterns of stylistic influence in the evolution of literature," *PNAS*, vol. 109, no. 20, pp. 7682–7686, 2012.

[7] S. Drożdż, P. Oświęcimka, A. Kulig, et al., "Quantifying origin and character of long-range correlations in narrative texts," *Information Sciences*, vol. 331, pp. 32–44, Feb. 2016.

[8] S. Régis, R. Meynet, R. Calif, A. Doncescu, and R. Emilion, "Proprietés fractales des structures issues de la rhétorique biblique et sémitique: premiers exemples," in *12 Journées Internationales d'Analyse Statistique des Données Textuelles*, 2014, pp. 567–579.

[9] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, Mar. 2009.

[10] M. Santini, "Characterizing Genres of Web Pages: Genre Hybridism and Individualization," in *40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007*, Jan. 2007, pp. 71–71.

[11] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, "Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words," *PLOS ONE*, vol. 4, no. 11, pp. e7678, Nov. 2009.

[12] A. N. Pavlov, W. Ebeling, L. Molgedey, A. R. Ziganshin, and V. S. Anishchenko, "Scaling features of texts, images and time series," *Physica A*, vol. 300, no. 1–2, pp. 310–324, Nov. 2001.

[13] J. F. Muzy, E. Bacry, and A. Arneodo, "Multifractal formalism for fractal signals: The structure-function approach versus the wavelet-transform modulus-maxima method," *Phys Rev E*, vol. 47, no. 2, pp. 875–884, Feb. 1993.

[14] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. E. Stanley, "Multifractal detrended fluctuation analysis of nonstationary time series," *Physica A*, vol. 316, no. 1–4, pp. 87–114, Dec. 2002.

[15] H. Wendt, P. Abry, and S. Jaffard, "Bootstrap for Empirical Multifractal Analysis," *IEEE Signal Proc Mag*, vol. 24, no. 4, pp. 38–48, July 2007.

[16] S. Jaffard, C. Melot, R. Leonarduzzi, H. Wendt, P. Abry, S. G. Roux, and M. Torres, "p-exponent and p-leaders, Part I: Negative pointwise regularity.," *Physica A*, vol. 448, pp. 300–318, 2016.

[17] R. Leonarduzzi, H. Wendt, P. Abry, S. Jaffard, C. Melot, S. G. Roux, and M. Torres, "p-exponent and p-leaders, Part II: Multifractal Analysis. Relations to Detrended Fluctuation Analysis.," *Physica A*, vol. 448, pp. 319–339, 2016.

[18] S. Jaffard, P. Abry, C. Melot, R. Leonarduzzi, and H. Wendt, "Multifractal analysis based on p-exponents and lacunarity exponents," in *Fractal Geometry and Stochastics V (Progress in probability)*, M. Z. C. Bandt, K. Falconer, Ed., vol. 70, pp. 279–313. Birkhäuser, 2015.

[19] A. C. Graesser, D. S. McNamara, Z. Cai, M. Conley, H. Li, and J. Pennebaker, "Coh-metrix measures text characteristics at multiple levels of language and discourse," *The Elementary School Journal*, vol. 115, no. 2, pp. 210–229, 2014.

[20] S. Jaffard, "Wavelet techniques in multifractal analysis," in *Fractal Geometry and Applications: A Jubilee of Benoit Mandelbrot, M. Lapidus and M. van Frankenhuijsen, Eds., Proc. Symposia in Pure Mathematics*. 2004, vol. 72(2), pp. 91–152, AMS.

[21] G. Parisi and U. Frisch, "Fully developed turbulence and intermittency," in *Turbulence and Predictability in geophysical Fluid Dynamics and Climate Dynamics*, M. Ghil, R. Benzi, and G. Parisi, Eds., Amsterdam, 1985, Proc. of Int. School, p. 84, North-Holland.

[22] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA, 1998.

[23] B. Castaing, Y. Gagne, and M. Marchand, "Log-similarity for Turbulent Flows?," *Physica D*, vol. 68, no. 3-4, pp. 387–400, Oct. 1993.