Science Fiction and Hyperchaos: Digital Humanities as Extro-Criticism

Graham Joncas[1] and Nora Li[2]

[1] Fudan University, School of Economics
[2] Shanghai International Studies University

Abstract. This paper compares Hugo and Nebula award-winning short stories using text mining and logistic regression. Science fiction is known for its radical singularity: each text is an 'event' in the philosophical sense, creating a universe unto itself. In this light, unlike traditional criticism, quantitative methods generalize in the absence of unifying conventions or topoi. Parallel to Meillassoux's concept of extro-science fiction, digital humanities acts as 'extro-criticism' within fields of radical contingency ('hyperchaos'). This asemic forensics not only traces 114 stories' lexical detritus to each award's institutional schemata, but presages xenographic re-mappings for conventional literary notions of 'code', 'genre', and 'text'.

Introduction

The two major awards for science fiction are the Hugo Award and Nebula Award. Their main difference is that the Nebula Award is selected by a committee of writers, while the Hugos are chosen by vote. While in some years, such as 2018, the same story wins both awards, an obvious question is whether the two awards systematically differ in any way, as well as how winning stories differ across time.

This paper uses digital humanities to compare winners of both awards for short stories. The Hugo award began in 1955, the corresponding Nebula award in 1966, giving a large corpus. The key advantage of using quantitative metrics is being able to analyze all winners for each award at once (cross-section), within the same award over time (time series), and across awards over time (panel).

Metadata such as word count are extracted from each winner, then used as inputs for logistic regression. If Hugo winners have value '1' and Nebula winners have '0', then a positive coefficient for a variable means stories with higher values of that variable will more likely win a Hugo award, and vice versa.

This paper analyzes various lexical parameters, how science fiction authors vary them for stylistic effect, and their interrelations. Regression analysis shows that three variables—average word length, average word frequency, and author gender—account for 20% of variation between Hugo winners and Nebula winners.

Such 'distant reading'—with a *quantitative* appeal of reading many books—is seen as the *raison d'être* for digital humanities. For science fiction, however, comparing stories is harrowing not only quantitatively, but *qualitatively*, as no common ground exists among disparate narratives. Thus the next section develops a qualitative motivation for digital humanities as 'extro-criticism' that, unlike conventional criticism, remains operative even in a state of absolute contingency.

1 Text Mining & Extro-Criticism

The philosophy of Quentin Meillassoux, who initiated the 'speculative realism' movement in avant-garde continental philosophy, lets us link science fiction to the structure of digital humanities as a discipline. Meillassoux (2015) isolates 'extro-science fiction' (XSF) as a parasitic anti-genre that violates the fundamental notion of 'science' on which the genre of science fiction (SF) is predicated.[1]

Whether at the borders of the galaxy, in hyper-advanced civilizations, or in extreme conditions such as black holes, SF's narratological force derives from the efficacy of science—embodied in constant laws and repeatable experiments.

Drawing from prior philosophical work, Meillassoux asks whether science fiction is possible under conditions of complete contingency—a state of 'hyperchaos' in which natural laws may change without warning, eliminating all grounds for science. This as-yet-unknown genre, Meillassoux names XSF.

1.1 Extro-Science Fiction

Asimov's story "The Billiard Ball" shows how SF's dramatic tension arises precisely from its internal notion of science, whose expectations are lost in XSF.

A theoretical physicist, Priss, always lived in the shadow of his rival Bloom, who became rich by applying Priss's theories. Priss develops a theory of anti-gravity, earning him a second Nobel, but claims it is impossible to realize in practice. Bloom swears he will find a way to apply it, and eventually invites the entire press to witness a public demonstration of his machine.

Bloom has not tested the machine, but is certain it will work. To humiliate Priss, he insists for his demonstration that Priss shoot a billiard ball into an antigravity ray atop a billiard table. Bloom predicts that on hitting the ray, the ball will rise, weightless. Priss aims, then strikes the ball. It bounces in a complex trajectory, then hits the ray. A thunderous noise is heard, and the onlookers see to their astonishment that the billiard ball has pierced through Bloom's heart.

Meillassoux (2015: 22) notes how the story only 'works' as SF, not XSF:

> If the story were Humean, i.e., extro-science fiction, there would be nothing more to say about this aberrant event, and the plot would leave us unsatisfied. But fortunately it is a story of science fiction, i.e., Popperian, and the plot finds a brilliant denouement.

Priss exposits at the story's end how this unforeseeable catastrophe arose: an object disconnected from gravity does not move with calm weightlessness, but "can only move at the speed of a massless object, i.e., the speed of a photon, the speed of light" (ibid.). The story ends as the protagonist speculates—what if Priss (otherwise famous for thinking slowly) had instantly understood what would happen, and Bloom's death had not been accident, but in fact, murder?

Meillassoux (2015: 49-56) later identifies a real XSF novel: René Barjavel's *Ravage*. In this story, electricity one day simply stops existing. A SF novel would,

in the course of its plot, attribute the disaster to some 'meta-law'; by contrast, Barjavel simply describes its consequences. No metaphysical 'closure' occurs.

Meillassoux ends by speculating upon even more radical XSF—whether a narrative in a state of hyperchaos, as the laws of nature mutate to the point of eliminating all extant life, could satisfy the barest conditions of a coherent plot.

1.2 The Arché-Text

Following Barthes, a text is seen as a tapestry of intertwined extra-literary codes, and literary criticism as the unweaving thereof. This is somewhat problematic in the case of SF, whose internal codes are judged the more exciting the less they correspond to extant ones. From a critic's view, then, the 'codes' of SF occur along a spectrum, from banal allegories (straightforward adoption, with some slight twist or other) to creating a whole new 'world'.

No doubt, much of SF is boilerplate. Yet, award-winning stories rule this out. Instead, 'great' SF qualitatively differs both from SF potboilers and great works of other genres, in that it relies far less on conventions, inhibiting criticism qua decoding. A great SF story constructs an autonomous system of codes, strictly incommensurate with any other such system (extant or fictional). The creation of such a system is thus an 'event' in the philosophical sense—radical contingency.

The challenge of SF criticism is mapping a genre (universe) whose conventions (laws) may change at any time. In hyperchaos, conventional criticism fails.

Meillassoux's philosophy arose as a way to answer how science can still be meaningful in a hyperchaotic world; his thoughts on SF are largely an allegory of this larger project. Thus, to construct an 'extro-criticism' for contingent literary matter, this section will sketch out a solution parallel to Meillassoux's own.

* * *

If there is a text, but nobody is around to read it, does it—so to speak—make a sound? This literary formulation of Berkeley's classic query is instructive in that a strict postmodernist ("It means whatever you want it to!") must answer no.

For Berkeley's idealist, a tree falling in the woods may well generate physical sound waves, but with no witnesses, a 'sound' as signifier cannot strictly exist. In the same way, a 'text' as semiotic entity only exists as a *correlation* between a reader and a written material. Such philosophical rhetoric may seem silly. Yet, suppose one day a book appears from nowhere. No-one has ever read it. Nor can one appeal to authorial intent. One need not be a hardline postmodernist to say it makes no sense to talk about the 'meaning' of this unwitnessed text.[2]

Pierre Bayard's *How to Talk About Books You Haven't Read* takes this Berkelian view even further, arguing for the radical non-equivalence of work and text. In one's internal impression of a work, it is simply impossible to remember *every* word, or even every feeling or event that the words evoke. Thus, given that any 'reading' of a text must be partial and incomplete, we cannot rigorously demarcate those who have 'read' a text from those who have not. As Bayard wryly notes, many French intellectuals have strong opinions about literary works they have not read—often stronger than the opinions of those who have read them!

Extending Bayard's view, a 'text' can exist even in the absence of a work—as with Abdul Alhazred's *Necronomicon*, or Ts'ui Pên's garden of forking paths. In many ways, these non-existent texts have more meaning than most real texts.

Our Berkelian query concerns the opposite claim: whether a text (work) can be said to have meaning even in the complete absence of a 'text' (social idea). Let us refer to this 'text'-less text as an **arché-text**.

A similar idea, called the *arché-fossil*, is what inspired Meillassoux. Scientific methods such as carbon dating make claims about matter that existed prior to any experiencing subject. Schools of thought such as phenomenology are irreconcilable with such claims, which imply "that manifestation itself emerged in time and space, and that consequently manifestation is not the *givenness of a world*, but rather an intra-worldly occurrence," or that the arché-fossil is "the givenness *in the present* of a being that is *anterior to givenness*" (2008: 15).

Much like carbon dating, methods from digital humanities make claims about texts, irrespective of readers' internal 'texts'. Recalling our unwitnessed Berkelian book, such claims would apply *even to a text that no human has ever read*. That is, digital humanities occupies itself exclusively with arché-texts.

1.3 Extro-Criticism

Meillassoux's project is not merely an exploration of odd premises, but is meant as a radically new answer to the question of how mathematics can provide an absolute description of the real (2011: 18). Unsurprisingly, this still a work-in-progress, yet his strategy for tackling this problem maps onto digital humanities in a striking way, yielding an insightful new angle on the discipline.

Crucial to Meillassoux's approach is the concept of the *kenotype*, or sign-devoid-of-meaning (2016: 166). In his view, to view mathematics as signifiers entirely misses the point. After all, mathematical models from game theory, for instance, can be applied to humans, computers and bacteria, without any change to the model's form. Thus, instead of saying that the model 'signifies' any or all of these objects, we can go a step further.

The kenotype "refers to neither meaning nor reference, but only to itself as a sign" (2011: 22). By so doing, it unifies the sign within its contingency (ibid.). For Meillassoux, the efficacy of mathematics arises due to its nature as kenotype.

In a similar way, digital humanities foregrounds that which is asemic within the literary sign. Within the signifier–signified relation of words evoking images, qualities such as letter count enter in only as asignifying detritus. This is shown by the applicability of such methods to texts that are strictly unreadable, such as the Voynich manuscript or *Codex Seraphinianus*.

Digital humanities, seen as extro-criticism, operates at the barest degree of code *as* code. In this sense, by treating literary matter solely at the level of arché-text, reducing stories to numbers allows comparison of disparate code-systems while refusing any commensurability among them (e.g. common themes).

The remainder of this paper aims to illustrate this approach, deciphering a series of literary parameters within an institutionally-imposed state of hyperchaos. The next section outlines each variable, followed by a quantitative analysis.

2 Hugos vs. Nebulas – Variables

The Nebula Award was founded in 1966 by Science Fiction and Fantasy Writers of America, founded by Damon Knight in 1965. This organization has over 2000 members, where membership requires having published science fiction, being professionally involved in science fiction or fantasy, representing a group based on science fiction or fantasy (e.g. a university library), or being a legal representative for a deceased science fiction author's estate. These members decide the Nebula Award by vote. Unlike the Hugo Award, it is more welcome to avant-garde material, so that less popular works ('boring' or 'abstruse') can win the prize.

The Hugo Award was set up by the World Science Fiction Society in 1953, and is named after the author Hugo Gernsback, considered the father of science fiction. Its award for short stories began in 1955. In contrast to the Nebulas, the Hugo Award is decided by public vote at the World Science Fiction Convention. In other words, anyone with a ticket to the convention has the right to vote. Accordingly, Hugos tend to be awarded to more popular works, such as those with more exciting scenes and plots.

Measuring differences between awards requires extracting textual variables from each story. Regression analysis produces numbers called *coefficients* that identify each variable's sign (positive or negative), magnitude (large or small), and significance (whether the coefficient is noticeably different from zero). In general, variables with large positive coefficients favor the Hugos, and variables with negative or small positive coefficients favor the Nebulas.[3]

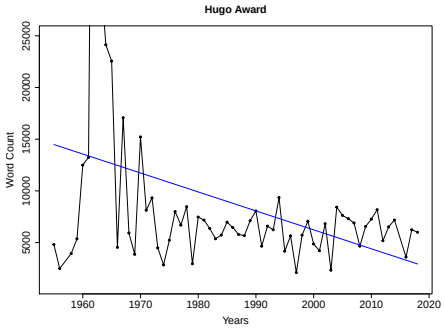
The present section describes each variable used in our regression, giving summary statistics, illustrating stories where a given variable plays a large role in its style, and predicting the coefficient's sign and magnitude.

2.1 Word Count

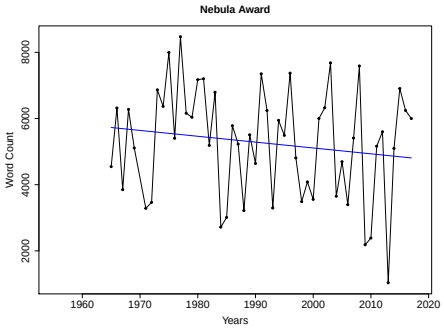
For both awards, a short story is defined as having less than 7,500 words, with no lower bound. Naturally, stories with fewer words cannot fit as much information, so all else being equal, require more technique to produce a 'complete' story. Thus we expect winners of the Nebula award, which favors technique, to have fewer words, giving the coefficient a negative value.

On average, Nebula winners have $5,250$ words, with a standard deviation of $1,700$.[4] Early in its history, the Hugo awards had several winners with very high word counts, the most obvious being the 1962 winner *Hothouse* by Brian Aldiss, with 82,600 words. Even after dropping the latter from the sample, several other early winners were more novellas than short stories, raising the mean to $7,500$ and standard deviation to $5,200$. Beyond these early stories, word count has been largely stable over time, with a slight downward trend.

Also noteworthy is the exceptionally short 2013 Nebula winner, "If You Were a Dinosaur, My Love" by Rachel Swirsky, at a mere 1,035 words. The story begins in the style of a whimsical nursery rhyme, growing sillier and sillier until—not to ruin the ending—it becomes much more than a nursery rhyme. Here, shortness is crucial to the work, which would otherwise feel belaboured.



(a) Hugos – Word Count

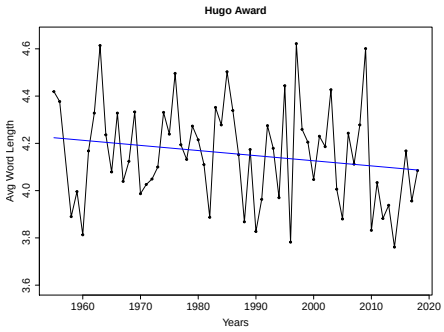


(b) Nebulas – Word Count

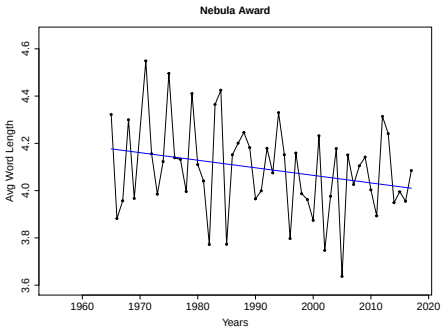
Fig. 1: Word Count (note: different y-axes)

2.2 Average Word Length

Mean word length helps to measure a text's general verbosity, e.g. using 'obtain' or 'acquire' instead of 'get', or using complicated scientific terminology. We expect this variable to decline over time for both awards. Likewise, we expect the committee-selected Nebula awards to have higher mean word length, and the popular-vote Hugos to have less, giving a negative regression coefficient.



(a) Hugos – Word Length



(b) Nebulas – Word Length

Fig. 2: Average Word Length

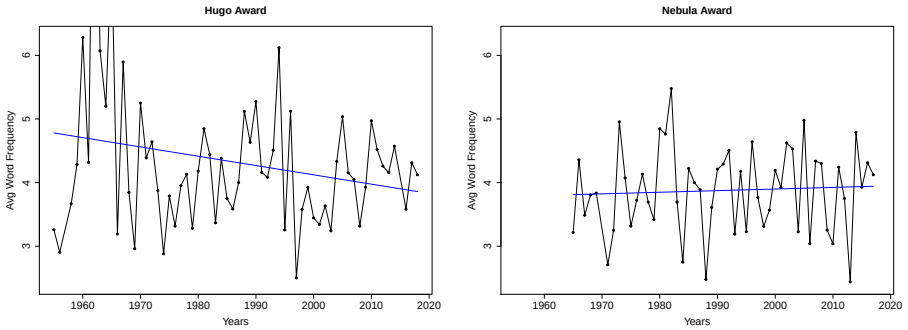
Hugo winners tend to have slightly longer words (4.15 letters, $\sigma = 0.21$) than Nebula winners (4.09 letters, $\sigma = 0.19$), though the difference is not significant ($p = 0.16$). For both awards, word length is approximately normally distributed.

The story with the smallest mean word length (3.64) is Carol Emshwiller's "I Live with You", the 2005 Nebula winner. It's hard to imagine how any story can have such small words, but this story makes frequent use of "I" and "you", as well as contractions, which our program splits into two words (e.g. "don"+"t").

The Hugo winners for 1997, 1963, and 2007 have the highest word length. The 1997 winner is a faux literary analysis of Emily Dickinson in light of H.G. Wells' *War of the Worlds*, using verbosity to establish the "author's" personality. Likewise the 1963 story takes place in a medieval-esque setting, using long words to establish an 'archaic' ambiance. Yet, the 2007 story has nothing odd about it—the author simply likes long words, though the story has no sense of verbosity.

2.3 Average Word Frequency

Mean word usage measures how often words are re-used. Since we already control for story length, this variable has the straightforward interpretation of word repetition, and thus of (lack of) lexical variety. We expect the popular Hugos to have higher mean word usage, giving a positive coefficient, though this variable's evolution over time is unclear.



(a) Hugos – Word Frequency **(b)** Nebulas – Word Frequency

Fig. 3: Average Word Frequency

Hugo winners re-use words slightly more frequently, with a mean of 4.2 and standard deviation of 0.94, compared to Nebula winners with mean 3.9 and standard deviation 0.67. Again, the difference fails to be significant ($p = 0.12$).

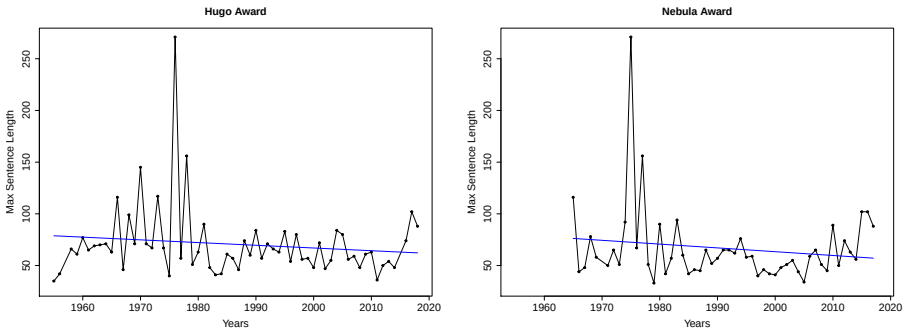
A slight problem is that this index (as shown in fig. 3) shows the average of *absolute* word frequency—that is, not proportional to text length, so that longer stories tend to have higher values, hence the very large values for early Hugos. It would be easy enough to divide each value by the word count in that story, but this new index gives poor values in our regression later on, so we omit it.

Lowest word frequency for Hugos is found in the aforementioned review of Dickinson, doubtless as a conscious literary device. For the Nebulas, this occurs in 1988: James Morrow's "Bible Stories For Adults, No. 17: The Deluge". The story takes place in a Noah's ark-like setting, and is written in King James-style biblical English, which is crucial for setting up its tone and twist ending.

2.4 Sentence Length

Sentence length is a well-known parameter for setting the tone of a piece of writing, with short, clipped sentences giving a far different atmosphere than long, flowing ones. Further, literary works abound with stylistic abuses of sentence length, such as Joyce's 4,391-word sentence in *Ulysses*, and Faulkner's 1,288-word sentence in *Absalom, Absalom!*

Thus, one variable worth experimenting with is maximum sentence length, to encode how many authors indulge in this literary flourish, and to what extent. The largest sentences in the Hugos have on average 70.5 words ($\sigma = 35$), while the Nebulas' largest sentences have on average 66.5 words ($\sigma = 36.8$).



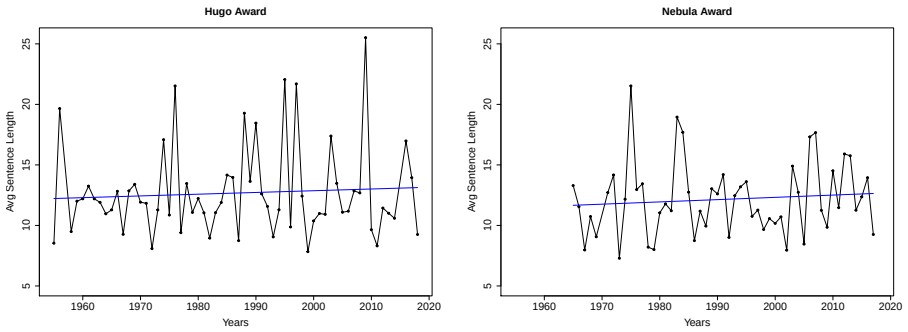
(a) Hugos – Sentence Length **(b)** Nebulas – Sentence Length

Fig. 4: Maximum Sentence Length

The longest sentence out of all the stories is 271 words, from "Catch That Zeppelin!" by Fritz Leiber, which won both the 1975 Nebula award and 1976 Hugo award.[5] Here the excessively long sentence is meant to give a panoramic view of the culture on a German zeppelin, in an alternate timeline (circa 1937) where this is a primary means of travel. One can imagine how breaking this into smaller sentences would give the dull impression of an itemized list.

The runner-up for longest sentence (156) is also a dual-winner, of the 1978 Hugo and 1977 Nebula award: "Jeffty is Five" by Harlan Ellison. Here, the long sentence occurs when the main character has entered the magical 'world' of the titular character (an ageless five-year-old), which involves watching TV and radio shows cancelled long ago (or involving actors who died long ago) as if they have run up to the present day. Its length in part mimics a childish garrulousness, mixed with the main character's sense of nostalgia for a time that never was.

Perhaps a more general picture, however, is given by average sentence length, which encodes the general atmosphere given by either terseness or loquacity. In the Hugos, the average sentence has 12.7 words, with a standard deviation of 3.7, while for the Nebulas the average sentence has 12.16 words ($\sigma = 3$).



(a) Hugos – Sentence Length **(b)** Nebulas – Sentence Length

Fig. 5: Average Sentence Length

The clearest outlier is Ted Chiang's "Exhalation", the 2009 Hugo winner. Given its magnitude, Chiang's use of long sentences is clearly a deliberate literary device—the main character is part of a race of sentient robots, so this loquaciousness forms an implicit contrast with robots' typical terseness in other science fiction stories. Interestingly, there are no outliers for short sentences.

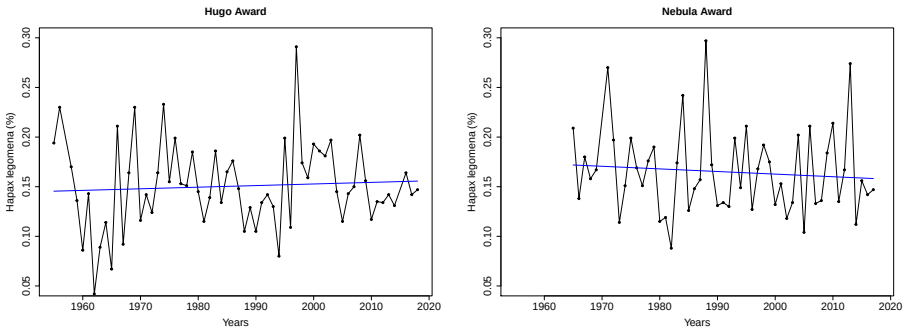
2.5 Hapax Legomena

Hapax legomena are words that appear in a text only once. This serves as a broad index of a text's lexical variety, such as diction or non-standard spelling. Longer texts tend have more unique words, so this index can be normalized by dividing by the text's length (Jockers, 2014: 69), yielding the percentage of hapax legomena as a fraction of the text. We expect the less 'popular' Nebula awards to have higher lexical variety, giving a negative coefficient. Yet, this variable's evolution over time is unclear, as the relative verbosity of older writing styles may be counterbalanced by stylistic irregularities in spelling (e.g. for accents).

Hapax legomena make up a larger part of Nebulas (16.5%, $\sigma = 0.044$) than Hugos (15%, $\sigma = 0.042$). While this difference is visible via the regression line of figure 6, it is not strong enough to be statistically significant ($p = 0.2$).

The Hugo with the smallest hapax count is the extremely long 1962 winner mentioned above, suggesting that words in longer stories are more likely to be reused. This is confirmed by the runner-up: the 1965 winner "Soldier, Ask Not" by Gordon R. Dickson has 22,551 words, and further stars a military-like journalist whose speech and internal monologues stand out as markedly cold and laconic.

We encountered the highest-hapax Hugo before—the excessively verbose analysis of Emily Dickinson, where the large variety of words again helps enforce the story's naïvely pedantic tone. Likewise for the highest-hapax Nebula, which is the Noah's ark story written in grandiloquent biblical prose.



(a) Hugos – Hapax Legomena **(b)** Nebulas – Hapax Legomena

Fig. 6: Percentage of Hapax Legomena

2.6 Dummy Variables

Dummy variables encode binary (either 'this' or 'that') properties, such as labeling stories as Hugo winners (1) or Nebula winners (0). It is likewise possible to include dummies as explanatory variables, though since our dependent variable is itself a dummy, the sign is easy to predict just by cursory inspection.

For example, only 13 Hugo winners (21%) have been women, whereas 29 Nebula winners (56%) are women. Such a large difference implies a strong and significant negative coefficient for gender. The first female author nominated for a Hugo was Pauline Ashwell in 1961, but it was not until 1974 that Ursula Le Guin became the first female Hugo-winning author.

Conversely, the author Jane Beauclerk was nominated during the Nebulas' opening year of 1965 (albeit among 31 stories authored by men), and in 1968 the Nebula award was won by Kate Wilhelm. As a further curious fact, since 2001 only two Nebula winners have been male. It would be interesting to see if this reflects changes in the Nebula committee's gender composition over time, but such data are currently unavailable.

Another dummy variable is tense, with 1 for first-person and 0 for third-person. One way to automate this dummy would be to check whether "I" is in the top 20-or-so most frequent words. However, this is liable to error in stories such as Terry Bisson's "macs" (2000 Nebula winner), which consists of monologue-type responses to the main character, with little use of the personal pronoun.

Curiously, the 2018 winner (Rebecca Roanhorse) is the only author to use the second-person tense, though for reasons of practicality this is coded as 0. Unlike for gender, choice of tense is approximately even for both awards: 33 Hugos (53%) are first-person, while 31 Nebulas (59%) are first-person. Thus we should expect a negative but small and insignificant coefficient for tense.

Another dummy variable is whether the author has won an award multiple times. This is encoded using two lists of authors for each story (one for Hugos, one for Nebulas), and setting '1' if the author appears more than once on the

list, '0' otherwise. For the Hugos, 26 stories (42%) are by repeated winners, while for the Nebulas this occurs for 21 stories (40%). The most frequent winner of both awards is Harlan Ellison, who has won 4 Hugos and 3 Nebulas. Since the difference between awards is minuscule, we expect an insignificant coefficient.

Other possible dummy variables might include the main character's gender, the venue in which the story was first published, data about competing stories, or data about nominations. However, the large corpus size limits dummies related to content, such as whether the story takes place on Earth.

3 Results

Our corpus spans 62 Hugo-winners and 52 Nebula winners. Our dependent variable is the award won by each story, and our regression has 9 regressors, 3 being dummy variables. Figures 1–6 give time series diagrams for word count, average word length, average word frequency, average sentence length, maximum sentence length, and hapax legomena percentage, showing the awards side-by-side.

The advantage of regression analysis is to show how much variation in award can be explained by a given variable, holding the other variables constant. For example, all else being equal, longer stories can be expected to have more hapax legomena; thus, to use it as an index for lexical variety, we should control for word count. Note that after controlling for word count, both hapax legomena and average word frequency act as indices for lexical variety, hence including both dulls their respective effects, so it is preferable to only include one.

3.1 Logistic Regression

Here, we use logistic regression, designed for binary dependent variables such as award.[6] Once again, award is coded as 1 for Hugo winners and 0 for Nebula winners, which implies positive coefficients if a quality is associated with Hugos, negative coefficients if linked to Nebulas. Thus our initial regression is as follows:

$$\begin{aligned}\text{award} = \beta_0 &+ \beta_1 \cdot \text{word\_count} + \beta_2 \cdot \text{hapax\_legomena} + \beta_3 \cdot \text{word\_length} \\ &+ \beta_4 \cdot \text{word\_frequency} + \beta_5 \cdot \text{avg\_sentence} + \beta_6 \cdot \text{max\_sentence} \\ &+ \beta_7 \cdot \text{author\_gender} + \beta_8 \cdot \text{first\_person} + \beta_9 \cdot \text{multiple\_wins} + \varepsilon\end{aligned}$$

Coefficients are shown in table 1. Regression (1) includes all variables at once, and the only significant coefficient is for author_gender at the 1% level, though word_length ($p = 0.07$) and word_frequency ($p = 0.095$) are significant at the 10% level.[7] Using the McFadden pseudo-R^2, we find that these regressors explain 22% of variation between Hugo and Nebula winners.

Regression (2) drops word_count, hapax_legomena, and multiple_wins due to their lack of significance, as well as max_sentence because its coefficient is so small. It is somewhat surprising that word_count has little value as a control variable, but this is likely due to its high variance.

Table 1: Logit Regression – Hugos (1) vs. Nebulas (0)

	(1)	(2)	(3)	(4)
word_count	0.1			
	(0.2)			
hapax_legomena	1.4			
	(1.8)			
word_length	3.5*	3.7**	4.5***	4.5***
	(1.9)	(1.6)	(1.5)	(1.4)
word_frequency	2.2*	1.5***	1.5***	1.3***
	(1.3)	(0.5)	(0.5)	(0.4)
avg_sentence	0.12	0.07		
	(0.1)	(0.08)		
max_sentence	-0.01			
	(0.01)			
author_gender	-1.8***	-1.7***	-1.7***	
	(0.5)	(0.5)	(0.5)	
first_person	-0.55	-0.5		
	(0.5)	(0.5)		
multiple_wins	-0.25			
	(0.5)			
pseudo-R^2	0.22	0.21	0.20	0.11

Note: *** – 1%; ** – 5%; * – 10%

As noted above, hapax_legomena and word_frequency both act as indices of lexical variety, which explains the strong effect on word_frequency of dropping hapax_legomena. Auxiliary regressions (not shown) make clear that word_length's sharp increase in significance is due to dropping hapax_legomena.

Regression (3) further drops avg_sentence and first_person, both of which were insignificant in regression (2). The only difference from (2) is in word_length; only dropping max_sentence from (2) does not change increases its coefficient, but dropping only first_person raises it up to 4.2 ($\sigma = 1.6$). This may reflect how first-person pronouns such as "I" and "my" have fewer letters than third person pronouns ("he/she", "his/her", etc.), which thereby reduces the average.

We are left with three highly significant variables—average word length, average word frequency, and author's gender—which together explain a fifth of the variation between Hugo and Nebula winners. As we expected simply from observing the summary statistics, author_gender is strong and negative. However, the coefficient for word_length is quite high, contradicting our earlier hypothesis that popular-vote Hugos would tend to have smaller words.

Recall that for Hugos the mean word length is 4.15 ($\sigma = 0.21$), and for Nebulas the mean is 4.09 ($\sigma = 0.19$). Likewise, the Hugos' mean word frequency is 4.2 ($\sigma = 0.94$) and the Nebulas' mean is 3.9 ($\sigma = 0.67$). Given these small differences, it's surprising that these variables turn out to be the most important.

To isolate the interrelations among each variable, we run another series of regressions (table 2). One striking observation is the change when word_length and word_frequency are put together, rather than separately. This likely reflects how short words such as "the" or "a" tend to be most frequent. Thus the two variables control for one another in a complementary way—word_length becomes an index for verbosity, while word_frequency becomes an index for lexical variety.

Table 2: Logit Regression – Effects of Main Variables

	(1)	(2)	(3)	(4)	(5)	(6)
word_length	1.54			4.5***	1.11	
	(0.95)			(1.4)	(1.02)	
word_frequency		0.53**		1.3***		0.68**
		(0.25)		(0.4)		(0.29)
author_gender			-1.56***		-1.5***	-1.7***
			(0.42)		(0.4)	(0.4)
pseudo-R^2	0.017	0.037	0.095	0.115	0.103	0.140

Last, although author_gender is strong and highly significant in regression 1.3, regression (4) drops it to see how this affects the other two variables. Curiously, word_frequency decreases, suggesting that female authors reuse words at a higher rate than men. A simple regression of word_frequency on author_gender yields insignificant results (not shown), but adding word_count as a control variable indeed gives a positive coefficient (0.31, $\sigma = 0.12$) significant at the 1% level.

Our hypotheses about the signs of these coefficients relied on the Hugos being chosen by popular vote, versus the Nebulas being chosen by a team of experts. We initially predicted that more 'popular' qualities would tend to be favored in the Hugos, i.e. have positive coefficients. The variable for word frequency, taken as indicating lack of lexical variety, appears in line with this hypothesis. However, the positive coefficient for word length (interpreted earlier as an index of verbosity) contradicts this hypothesis. Last, with the exception of author gender, the other variables seem not to be strongly favored by either award.

3.2 Next Steps

Several methodological issues arise that we hope to fix as the project progresses.

First, using a binary dependent variable means that qualities favoring Hugos (1) will have positive coefficients, those favoring Nebulas (0) will have negative coefficients. In this framework, however, a coefficient near zero can mean that a quality is either helpful for both awards, or for neither. It would be preferable to disaggregate these two effects, perhaps by incorporating data for nominees.

Second, six stories have won both awards, and our solution (the easiest) is to re-use the text twice, as if they were separate texts. A common view is that dual winners are (or will be) classics, so this may allow an auxiliary analysis comparing dual winners to single winners. It's not clear, however, whether the sample size of dual winners is large enough to yield meaningful results.

Last, our analysis has the advantage of covering the entire population of winners. Nevertheless, our scanned versions are not perfect, and so our analysis faces standard issues involving transcription errors. Still, we care more about the relative magnitudes of our regression coefficients rather than literal magnitudes, and since typos are random, they will simply be accounted for by the error term.

* * *

The goal of our analysis is to find systematic differences between winners of the two awards, using regression analysis to make these differences explicit.

Each of these stories is a universe unto itself. Receiving a Hugo or Nebula award means that a story is like nothing else the world has ever seen. Almost by definition, there can be no common factors on the order of *content*. Nevertheless, despite the myriad reasons for selecting a given text as a winner, this study finds that a fifth of the variation between these two awards can be explained by three variables: average worth length, average word frequency, and the author's gender.

Further extensions to this project can take three directions. The first is to develop more detailed variables for each story, such as topics or sub-genres; the key here is figuring out how to automate these tasks, e.g. by topic modeling.

A second extension is to examine each award's evolution over time using time series and panel methods. The difficulty is that such methods do not work for binary dependent variables, which instead are used to group the data. Rather, we can only ask questions using our regressors (e.g. explaining word_count with word_length), which is far less intuitive than simply comparing the two awards.

A third extension is to see whether a support vector machine can successfully classify stories according to the award they won. If it cannot, this will be a major negative result; if it can, it will be worthwhile to inspect any cases that it misclassifies, as well as how it classifies stories that have won both awards.

Conclusion

The premise behind this research project is that science fiction's marginalization within literary theory is due not to any inherent quality as 'low-art', but because its structure as a genre resists the theme-based analysis of conventional literary criticism. Conversely, digital humanities lets us make general statements about a corpus of stories noted for their singularity. Thus, we hope that our project can give insight both into the nature and evolution of science fiction, and also into the structure of digital humanities as a methodology that can open up otherwise-neglected vistas of textual inquiry.

References

1. Bayard, P.; Mehlmann, J. (trans.). (2007). *How to Talk About Books You Haven't Read*. New York: Bloomsbury.
2. Jockers, M. (2014). *Text Analysis with R for Students of Literature*. Heidelberg: Springer.
3. Meillassoux, Q.; Brassier, R. (trans.). (2008). *After Finitude: An Essay on The Necessity of Contingency*. New York: Continuum.
4. Meillassoux, Q.; Lozano, B (trans.). (2011). "Contingency and The Absolutization of the One." Retrieved from `https://www.scribd.com/document/81307810/Contingency-and-Absolutization-of-the-One`
5. Meillassoux, Q.; Edlebi, A. (trans.). (2015). *Science Fiction and Extro-Science Fiction*. Minneapolis, MN: Univocal.
6. Meillassoux, Q.; Mackay, R. & Gansen, M. (trans.). (2016). "Iteration, Reiteration, Repetition: A Speculative Analysis of the Meaningless Sign," in Malik, S. & Avanessian, A. (Eds.) (2016). *Genealogies of Speculation*. New York: Bloomsbury.

Notes

[1]Meillassoux's French title is *Métaphysique et fiction des mondes hors-science*, which literally translates to: "Metaphysics and Fiction of Worlds Outside/Beyond Science." This *hors* is famously difficult to translate, as in Derrida's "Il n'y a pas de hors-texte" (There is nothing outside the text), whence the awkward neologism 'extro-science'.

[2]A less esoteric example might involve the reams of prose generated by spambots and AI, which will be less and less distinguishable from human texts as NLP improves.

[3]Taking word count as an example, a story obviously must have some words, but since Nebulas tend to have fewer words the coefficient will be positive, but small.

[4]These figures are rounded, to account for transcription errors in the text files.

[5]This 271-word sentence is part of an internal monologue early in the story:

> I stole another glance up at the Ostwald, which made me think of the matchless amenities of that wondrous deluxe airliner: the softly purring motors that powered its propellers—electric motors, naturally, energized by banks of lightweight TSE batteries and as safe as its helium; the Grand Corridor running the length of the passenger deck from the Bow Observatory to the stern's like-windowed Games Room, which becomes the Grand Ballroom at night; the other peerless rooms letting off that corridor—the Gesellschaftsraum der Kapitän (Captain's Lounge) with its dark woodwork, manly cigar smoke and Damentische (Tables for Ladies), the Premier Dining Room with its linen napery and silverplated aluminum dining service, the Ladies' Retiring Room always set out profusely with fresh flowers, the Schwartzwald bar, the gambling casino with its roulette, baccarat, chemmy, blackjack (vingt-et-un), its tables for skat and bridge and dominoes and sixty-six, its chess tables presided over by the delightfully eccentric world's champion Nimzowitch, who would defeat you blindfold, but always brilliantly, simultaneously or one at a time, in charmingly baroque brief games for only two gold pieces per person per game (one gold piece to nutsy Nimzy, one to the DLG), and the supremely luxurious staterooms with costly veneers of mahogany over balsa; the hosts of attentive stewards, either as short and skinny as jockeys or else actual dwarfs, both types chosen to save weight; and the titanium elevator rising through the countless bags of helium to the two-decked Zenith Observatory, the sun deck wind-screened but roofless to let in the ever-changing clouds, the mysterious fog, the rays of the stars and good old Sol, and all the heavens.

[6]We also repeated the regression using a probit model, which gives essentially the same results. Although the absolute magnitude of the probit coefficients differ, their sign, relative magnitudes, and significance are basically identical to logistic regression.

[7]Both robust standard errors and clustering by years make little difference, so for the present analysis we simply opt for homoskedastic errors.