

Toward a

# A science of science fiction

## Applying quantitative methods to genre individuation

Ryan Nichols<sup>a,b</sup>, Justin Lynn<sup>c</sup>, and Benjamin Grant Purzycki<sup>b</sup>

<sup>a</sup>Department of Philosophy, Cal State Fullerton / <sup>b</sup>Centre for Human  
Evolution, Cognition, and Culture, University of British Columbia /

<sup>c</sup>Department of Psychology, Cal State Fullerton

What is a genre? What distinguishes a genre like science fiction from other genres? We convert texts to data and answer these questions by demonstrating a new method of quantitative literary analysis. We state and test directional hypotheses about contents of texts across the science fiction, mystery, and fantasy genres using psychometrically validated word categories from the Linguistic Inquiry and Word Count. We also recruit the work of traditional genre theorists in order to test humanists' interpretations of genre. Since Darko Suvin's theory is among the few testable definitions of science fiction given by literary scholars, we operationalize and test it. Our project works toward developing a model of science fiction, and introduces a new method for the interdisciplinary study of literature in which interpretations of literary scholars can be put to the test.

**Keywords:** genre, science fiction, mystery, fantasy, corpus linguistics

### Literary theories of genre

"The term 'science fiction' resists easy definition", reads the first sentence of Adam Roberts' excellent *Science Fiction* (2006, p. 2), but repeating that this is a problem does not make it go away. Twenty years earlier Gary Wolfe recognized this problem and quoted thirty-three different definitions in his *Critical Terms for Science Fiction* (1986). Methods in traditional genre theory used by literary scholars proceed by drawing examples from texts, making a priori inferences, and generalizing. This method yields little knowledge of a general philosophical interest, and offers few benefits for the appreciation of science fiction literature in comparison with other methods of inquiry (Nichols et al., 2008). With a few key exceptions

from yesteryear, one of which is discussed below, writers of science fiction literary criticism resemble writers of fantasy literature as both groups are playing tennis without a net. By contrast, our project brings to science fiction literary criticism a scientific frame of mind.

Traditional literary criticism faces obstacles in defining genre. The two that impede the most progress are (i) that the ground rules used in offering definitions are unclear, and (ii) that the justification for methods of giving definitions are inadequate. Consider the adoption of *historiographical* criteria for genre differentiation. According to historiographical criteria, a literary work becomes a member of a set of works in a target genre only if the target story participates in the historical development of the genre. This technique risks omitting use of textual content for classificatory purposes. Use of historiographical criteria for genre differentiation leads to intractable disputes between literary scholars in which they trade anecdotal evidence, for example, ~~about whether a certain historical text~~ <sup>like</sup> *Frankenstein* <sup>ought to be considered the source text of the science fiction genre</sup> (cf. Aldiss, 1973; for criticism, see Kincaid, 2003).

According to *conceptual* criteria a literary work is sorted into one genre rather than another on the basis of the work's literary content. This method is also subject to debilitating problems. We risk omitting concerns with style and history (Suvin, 1978). If we use unfalsifiable a priori criteria, we risk begging questions against others who disagree (Chandler, 1997). For example, if we begin by stipulating that science fiction is enclosed within fantasy (Aldiss, 1973; Panshin, 1971), we beg the question against those who begin by stipulating that science fiction is "realistic" in opposition to fantasy (Heinlein, 1959).

Historiographical criteria and conceptual criteria are genera in the taxonomy of definitions, but there are many species in these and other genera. Some scholars advocate reading strategies as a means of providing a definition for a genre (Rawlins, 1982), a technique applied to science fiction (Delany, 1971). Ostensive definitions — <sup>like</sup> "science fiction is *that*" — meet with popularity amongst critics (Knight, 1967) but <sup>say</sup> only that science fiction is what people refer to when they point to it. They have met with harsh criticism (Fredericks, 1978), though few critics appear to have noticed. Closely related, a mutable definition is a definition according to which science fiction is a changing composite of people, practitioners, practices, and their shared history, in which the content of texts plays only one among many defeasible roles. The phylogeny of Rieder's (2010) award-winning but Lovecraftian offspring places it somewhere within this subspecies. We regard these definitions as flawed or seriously flawed, and of little general philosophical interest. Though we hope to situate our quantitative profile of science fiction in the context of these definitions and the history of the definitional project elsewhere,

we forego such a discussion in what follows in favor of presenting our methods, hypotheses, data, and results.

## Quantitatively profiling genres


We attempt to improve upon this state of affairs and respond to the obstacles described above by quantitatively profiling genres. To *quantitatively profile* a genre is to test predictions concerning the facts about the contents of written works in it in order to individuate a target genre from other genres. The ~~quantitative analysis~~ <sup>discussion</sup> of science fiction presented below states tractable, comparable, and testable hypotheses in terms of explicit parameters in several linguistic categories.

Our method compares and contrasts with compelling uses of quantitative analytics in literary inquiry from which we draw inspiration. Randall Stock in his “Rating the Canon” correlates Sherlock Holmes story lengths with readers’ ratings of quality (1999). Neil Goble used pioneering methods of word frequency counts and more in his study of the corpus of Isaac Asimov in *Asimov Analysed* (1972). In a methodologically pioneering work, Wu Yan collected data about people’s associations with the genre of science fiction and then ran a factor analysis on those word associations to identify six principle components in Chinese participants’ concept of science fiction: literature, exploration, science, cognition, atheism, and horror (2000).

Franco Moretti, Matthew Jockers, and The Literary Lab in which they participate produced fascinating quantitative explorations of genre concepts (Allison et al., 2011). Moretti has argued that for the big, interesting questions in the study of literature, answers “cannot be understood by stitching together separate bits of knowledge about individual cases” in the form of close readings of representative texts (Moretti, 2005, p.4). With its careful tone, Jockers’ book (2013) functions to gently ease the literature scholar into the digital humanities. With the help of Steven Ramsey, Jockers warns literary scholars using big data against presenting themselves as delivering the facts where other scholars only trade in opinion. Yet, following Moretti’s methods of “distant reading” (2013), Jockers motivates his use of quantitative methods on big data by saying that traditional “close reading” methods are “impractical as a means of evidence gathering in the digital library” (2013, p.7). The Literary Lab’s projects have ranged from examining play plots using mathematical models derived from network theory to creating a quantitative map of 19th Century British novels.

We depart from Goble, Stock, Moretti, Jockers, and others in our framing of the problem and in our end goals. Moretti discusses the study of genres and refers to them as “temporary structures within the historical flow” (Moretti, 2005).

While our study does not set out to prove that the historical sweep of science fiction is more than temporary, we appear to differ with previous researchers in quantitative literary studies in regard to the nature of genre. To continue our biological metaphor, we presume for the sake of argument that genres probably contain within them identifying linguistic markers that, with the right data analysis, can be identified, coded, and differentiated from the identifying linguistic markers of other genres. The majority of analyses conducted up until now have been exploratory in nature and lack hypothesis testing.



Researchers with a foot in statistical analysis, psychology, or corpus linguistics might wonder why we discuss literary theorists' previous definitions of science fiction, when we could have generated quantitative profiles of the three genres without the fuss. Likewise, literary scholars and philosophers of literature might wonder what we think we are doing by converting words into numbers, then examining variation between genres. In response, our enjoyment of singular features of science fiction literature such as its high new-ideas-per-page ratio, its sense of disciplined wonder, and its embrace of the future and technology inspire our focus on this genre as well as the thoroughgoing interdisciplinarity of our project. We explore differences that make science fiction special by pioneering a new, scientifically respectable method of study that tests theories about literature. But we also take seriously a subset of theories of literary scholars about genre. This is why we set out to empirically test them.

## Methods and hypotheses

We employ the Linguistic Inquiry and Word Count. The Linguistic Inquiry and Word Count (LIWC, pronounced 'luke') is a textual processing engine used by psychologists, mental health professionals, sociologists, academics and marketers. The LIWC2007 dictionary contains 4500 words and word stems. Each is filed into one or more subdictionaries. Individual subdictionaries represent one of the 55 variables or 'word categories' through which LIWC compiles the words of a target text. As explained by its developers, the word "cried" is part of "five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. Hence, if it is found in the target text, each of these five subdictionary scale scores will be incremented" (Pennebaker et al., 2007, p. 4; for examples of LIWC at work, see Tausczik & Pennebaker, 2010, p. 27).

LIWC's intended function of making inferences from texts to writers' psychological states originally found use in mental health diagnostics. But since its development LIWC has been used by a number of researchers to test a variety of hypotheses. Researchers have found that depressed participants use more first-person

pronouns than others (Rude, Gortner, & Pennebaker, 2004); that positive political ads use present and future tense verbs at higher frequencies than negative ads, which used more past tense verbs (Gunsch et al., 2000); and that positive emotion words are used more frequently when individuals write about positive events, and negative words more when writing about negative events (Kahn, Tobin, Massey & Anderson, 2007). Mean values across LIWC categories have been shown to correlate with big-five personality traits (Pennebaker & King, 1999; Mehl, Gosling, & Pennebaker, 2006). In further extending the application of LIWC, we apply the software to corpora of science fiction, fantasy, and mystery in order to test hypotheses about the content and style across these genres. Our hypotheses about contrasts between genres appear in Table 1.

Table 1. LIWC variables and hypotheses

LIWC variable	Definition	Dictionary examples	Omnibus hypotheses
1 COGMECH	cognitive processes	certainty, insight, causation	SF > M > F
2 SOCIAL	social words	talk, friends, home	M > F > SF
3 PERCEPT	perception words	see, hear, feel	M > F > SF
4 BIO	biological processes	body, health, ingestion	M > F > SF
5 RELIG	religious terms	God, altar, church	F > M > SF
6 PRONOUN	pronouns	I, them, it	M > F > SF
7 AUX	auxiliary verbs	am, will, be	M > SF > F

M = Mystery; F = Fantasy; SF = Science fiction

We explain below what leads us to these hypotheses, but before that we note that some of our hypotheses were inspired by one of the most fecund and philosophically interesting accounts of the science fiction genre. According to Darko Suvin, appropriate literary representations of *cognition* and *estrangement* are individually necessary and jointly sufficient for membership in the genre of science fiction. He writes, “Science fiction is a literary genre whose necessary and sufficient conditions are the presence and interaction of estrangement and cognition, and whose main formal device is an imaginative alternative to the author’s empirical experience”. More specifically, science fiction “is distinguished by the narrative *dominance* of a fictional novelty” or novum. It is differentiated from the fantastic genres by “the *presence* of scientific cognition as the sign or correlative of a method ... identical to that of a modern philosophy of science” (1978, pp. 46–47). A work of science fiction “should be defined as a fictional tale determined by the hegemonic literary device of a locus and/or dramatis personae that... are radically or at least significantly different from the empirical times, places, and characters of ‘mimetic’ or ‘naturalist’ fiction” (Suvin, 1979, p. viii). Suvin’s definition of “science fiction”

is put in terms of the uniquely individuating subject matter of literary works of science fiction. It calls for considerable reflection.

Subsequent scholars emphasize features of Suvin's formal analysis. Edward James highlights the importance of cognition for science fiction: "Estrangement is offered by the fairy tale and other literary genres as well, but sf is distinguished also by cognition, the process of acquiring knowledge and of reason. ... A cognitive — in most cases strictly scientific — element becomes a measure of aesthetic quality, of the specific pleasure to be sought in sf" (James, 1994, p. 108). This may be taken as a counterbalance for characterizations of science fiction as having great affinities with religious literature or fantasy (see Hartwell, 1996, p. 42). Though "cognition" is a term with even more pre-theoretic obscurity than "science fiction", here the term refers to cognition in the context of a "scientific ethos" (Rose, 1981, p. 20).

The dual necessary conditions Suvin uses to individuate science fiction from other genres can be operationalized in terms of the LIWC categories presented in Table 1, which leads us to our hypotheses. If Suvin is correct, then science fiction should contain higher frequencies of cognition words than fantasy (his principal point of contrast) and perhaps mystery. Since reasoned decision-making is constitutive of the resolution of typical forms of conflict in science fiction, Suvin's analysis suggests we will find significant variance in use of descriptors for mental action. LIWC's cognitive mechanisms category (LIWC2007 category name "COGMECH") includes subscales such as *tentative* (which in turn includes words such as "maybe", "perhaps", "guess"), *certainty* ("always", "never"), and *insight* ("think", "know"). This category and its subscales become a primary variable with which to test the hypothesis that science fiction has more cognition terms than fantasy or mystery. *Omnibus hypothesis 1* states that ~~mean values for~~ cognition terms will yield the following directional relationship: SF > M > F.

Given science fiction's use of estranging devices to introduce unfamiliar settings, characters, species, and actions, Suvin's analysis leads us to expect fantasy and mystery to contain significantly more social and family terms than science fiction. This hypothesis has *prima facie* justification irrespective of whether Suvin is correct that science fiction predominantly estranges the familiar, or Clute that science fiction makes familiar the strange. LIWC contains a social category (SOCIAL) that contains subscales such as *family* and *home*. The contents of the *family* subscale are obvious; *home* includes words for concepts of home ("house", "apartment") and words for rooms ("kitchen", "bathroom"). *Omnibus hypothesis 2* states that mean values for social terms will yield the following directional relationship: M > F > SF.

Actions in science fiction frequently consist of knowledge-based resolution to conflicts and scientific knowledge is gathered at some stage via perceptual processes. It would be natural to predict that science fiction contains higher rates of

comparing the  
frequencies of  
^

perceptual terms (PERCEPT) than fantasy for this reason. However, science, and so science fiction, often contain forms of knowledge acquisition lacking embodied perceptual components. When a ship's sensor detects and analyzes subatomic particles, no perceptual faculties embodied in a human being are likely to be described. Acquisition of information in science fiction narratives is less likely to be portrayed as embodied perceptual processes than it is in fantasy or mystery due to science fiction's settings, characters and technologies. Characters in fantasy, even non-human characters, typically possess traditional perceptual faculties; in mystery they certainly do. Given this, more characters in fantasy and mystery are likely to see, hear, and touch than are characters in science fiction. *Omnibus hypothesis 3* states that mean ~~values for~~ perceptual terms will yield the following directional relationship:  $M > F > SF$ .

Related, a significantly higher number of characters in mystery and fantasy than characters in science fiction have humanoid bodies to begin with, for which reason we expect mystery and fantasy to contain higher rates of terms describing bodies and their biological actions (BIO) than science fiction. This is so despite the fact that fantasy, much more than science fiction, is likely to contain supernatural beings. We regard the population of fantasy with spirits, ghosts, etc., as a red herring for the purposes of a quantitative analysis of the texts because to interact with physical, embodied characters these supernatural beings must take on apparent physical manifestations and be described in physical ways allowing relationships with humans. Wraiths appear to wear black and have deep voices, etc. *Omnibus hypothesis 4* states that mean values for biological terms will yield the following directional relationship:  $M > F > SF$ .

For very similar reasons we predict fantasy and mystery will contain more religion (RELIG) words. Fantasy ought to have many more religion terms than science fiction. This is because fantasy is likely to be more socially oriented than science fiction, and because traditional fantasy settings, whether in the Earth's or another planet's past, likely appeal to the religious worldviews and practices of that world more often than secular notions contained within science fiction. Mystery likely has more religious terms than science fiction, ~~for similar reasons.~~ We posit these hypotheses despite the fact that science fiction appears more often concerned with the transcendent than mystery or fantasy. *Omnibus hypothesis 5* states that mean values for religion terms will yield the following directional relationship:  $F > M > SF$ .

It is unclear whether use of word frequencies across LIWC word categories is a means of assessing what most literary critics would consider stylistic features of texts. As a result, we regard our hypotheses about broadly stylistic features of texts as more speculative than our hypotheses about content-based features of texts stated above. By 'style' here we only refer, crudely, to properties of parts of speech

by virtue of  
science fiction's  
lack of them.

frequencies of  
^



and verb tenses. With that qualification on the table, we offer tentative predictions about pronoun use and about auxiliary verb use between genres.

We are neutral about the relative rates at which stories across science fiction, mystery, and fantasy are written from the first-person or third-person perspectives. This increases uncertainty about the relative rates of personal pronouns between the genres. However, plot, action, and characterization in mystery stories focus on relationships between human agents and their alleged actions. More gendered human agents populate mystery than fantasy, and fantasy than science fiction. Plot, action, and characterization in fantasy more closely resembles mystery than science fiction: there are perhaps more characters in fantasy than science fiction and so more need for use of pronouns. As a result, we expect greater rates of pronouns (PRONOUN) in mystery due to its greater rates of personal pronouns than the rates found in fantasy, and greater rates in fantasy than science fiction. *Omnibus hypothesis 6* states that mean values for pronouns will yield the following directional relationship:  $M > F > SF$ .

predict  
^

Auxiliary or helping verbs (AUX) typically express mood, aspect, tense or voice. With auxiliary verbs writers convey nuanced facts about action. Reports of LIWC2007's output on genre-based test corpora indicate emotional writing has the highest rate of auxiliary verb usage (Pennebaker, Chung, et al., 2007, p. 11). Since mystery is concerned with nuance, mood, and aspect more than fantasy and science fiction, and since it represents fear and other emotions, especially relating to mortality salience, more than fantasy or science fiction, we predict it will yield the highest rates auxiliary verbs. A second supporting reason for this prediction has to do with the nature of action in stories. Fantasy appears more driven by setting than by action, and as a result we tentatively infer that it will have the lowest rate of helping verbs, with science fiction in between fantasy and mystery. We suspect that fantasy will yield more auxiliary verbs than science fiction for similar reasons. *Omnibus hypothesis 7* states that mean values for religion terms will yield the following directional relationship:  $M > SF > F$ .

predict  
^

### Dataset

In these analyses we deferred to the community of experts by compiling volumes from preeminent anthologies of science fiction, fantasy, and mystery edited by industry-leading editors including Gardner Dozois, David Hartwell, Kathryn Cramer and Ed Gorman. The purpose of populating our dataset with stories selected in this way is that it allows us a response to skeptics who would accuse us of begging the question. We have in mind a question such as: <sup>Q1</sup>Why do you think that the literature you take to be representative of fantasy is *in fact* representative of fantasy? You've pre-judged what qualifies as fantasy. This is why you get the







results you predict. Our reply is that in fact we are not presupposing what is and is not fantasy; rather, we are letting the experts make that determination for us. As a result of this approach, a brief statement of the credentials of these editors is in order to show that it is likely that they know what is and is not science fiction, fantasy, and mystery.

Dozois has won 15 Hugo Awards for Best Professional Editor, and a book-length interview with him (Swanwick, 2002) won the Locus Award, not to mention his Nebula and Sidewise awards as an author. Hartwell has won the World Fantasy Award for Best Anthology, the Hugo Award for Best Professional Editor and Best Editor Long Form multiple times, has edited several Hugo and Nebula award-winning best novels, and edits *The New York Review of Science Fiction*. Cramer has won a World Fantasy Award for an anthology, was nominated for other World Fantasy Awards in her capacity as editor, and was runner-up for a Pioneer Award. Gorman's editing work has won him multiple nominations for the Bram Stoker award for Best Fiction Collection and as a writer he has won a Spur Award.

The science fiction ( $n=157$ ) data were drawn from *The Year's Best Science Fiction*, fantasy data ( $n=147$ ) were drawn from *The Year's Best Fantasy*, and mystery data ( $n=161$ ) were drawn from *The World's Finest Mystery and Crime Stories*. Paper copies of these books were purchased, disassembled, and scanned with a Xerox WorkCentre 5150 with maximum resolution. The resulting high-dpi scans supported optical character recognition by Adobe Acrobat Pro 10. Books were separated into individual files representing one story each. Resulting files were manually inspected for spelling errors caused by the optical character recognition process with Microsoft Word 14. Files were then processed with LIWC2007. Gender information for authors whose stories appear in the dataset was researched online. Coding for gender was manually entered into our dataset, as was year of publication (Table 2).

All stories fall within a shared narrow range of years, which functions in our analyses as a control of effects of temporal change on writing. As is obvious, we do not propose to create a pan-historical quantitative profile of science fiction. That would require compiling thousands of novels and short fiction written across many centuries. All stories are written in the English language, which functions as a control on the effects of different languages on writing and genre.

Table 2. Database characteristics

Source	Publication	Stories	Male authors	Female authors^	Total authors*
Science Fiction (Dozois YBSF)					
YBSF 15	1999	28	26	4	30
YBSF 16	2000	24	19	5	24
YBSF 17	2001	27	23	4	27
YBSF 18	2002	23	19	5	24
YBSF 20	2004	25	20	5	25
YBSF 23	2007	30	26	6	32
Subtotal		157	133	29	162
Fantasy (Hartwell YBF)					
YBF 1	2001	23	16	7	23
YBF 2	2002	22	14	9	23
YBF 3	2003	29	19	10	29
YBF 4	2004	21	12	9	21
YBF 5	2005	24	15	9	24
YBF 9	2009	28	19	10	29
Subtotal		147	95	54	149
Mystery (Gorman WFMCS)					
WFMCS 1	2000	38	25	14	39
WFMCS 2	2001	42	30	12	42
WFMCS 3	2002	39	21	18	39
WFMCS 4	2003	42	27	15	42
Subtotal		161	103	59	162
Grand Total		465	330	142	472

^ Selections co-authored by one male and one female author were coded as female due to technical limitations. \* Total = total number of authors rather than number of unique authors.

Results

Assumption tests

We then compiled all texts in our three corpora with LIWC2007 and gathered data about mean values of word use across target word categories. Table 3 reports the descriptive statistics for each variable by genre including Kolmogorov-Smirnov tests for normality. The mean scores (*M*) represent word frequency data for the category whose name appears in the leftmost column of Table 3.

Variances were equal across groups for COGMECH  $F(2, 462) = 2.34, p = 0.10$ , SOCIAL,  $F(2, 462) = 2.43, p = 0.09$ , and PRONOUN,  $F(2, 462) = 0.50, p = 0.61$ .

However, PERCEPT,  $F(2, 462) = 15.57, p < 0.001$ , BIO,  $F(2, 462) = 3.57, p = 0.03$ , RELIG,  $F(2, 462) = 16.30, p < 0.001$ , and AUXVERB,  $F(2, 462) = 3.98, p = 0.02$  all violated the assumption of homogenous variance. As such, Welch's  $F$ -ratios are reported for all analyses in these cases.

Table 3. Word frequencies for LIWC variables

	<i>Science fiction (n = 157)</i>				<i>Fantasy (n = 147)</i>				<i>Mystery (n = 161)</i>			
	M (SD)	Mdn	D		M (SD)	Mdn	D		M (SD)	Mdn	D	
1 COGMECH	14.81 (1.53)	14.75	0.40		14.30 (1.77)	14.17	0.06		14.57 (1.56)	14.39	0.06	
2 SOCIAL	11.05 (2.30)	11.09	0.06		11.87 (2.77)	11.85	0.05		12.96 (2.23)	12.60	0.08**	
3 PERCEPT	3.68 (0.75)	3.65	0.09**		4.20 (1.18)	4.23	0.03		3.71 (0.86)	3.63	0.08*	
4 BIO	2.43 (0.76)	2.37	0.09**		2.64 (0.89)	2.63	0.06		2.43 (0.73)	2.35	0.07*	
5 RELIG	0.31 (0.37)	0.22	0.23***		0.42 (0.48)	0.24	0.19***		0.25 (0.21)	0.19	0.17***	
6 PRONOUN	14.57 (2.43)	14.52	0.06		14.77 (2.65)	14.55	0.06*		15.99 (2.54)	15.71	0.08	
7 AUXVERB	7.66 (1.38)	7.58	0.05		7.01 (1.63)	7.01	0.04		8.19 (1.29)	8.08	0.05	

\*\*\* $p \leq 0.001$ , \*\* $p \leq 0.01$ , \* $p \leq 0.05$

### Results from content categories

In order to test for any potential effects that the gender of author may have on mean LIWC values, we conducted initial ANOVAs controlling for gender. Gender only showed to have significant effects for the perception category ( $F(1, 460) = 12.97, p < 0.001$ ) and the biology category ( $F(1, 460) = 5.30, p = 0.02$ ). However, they were not driving the results since overall effects of genre were still significant. Gender showed no significant effects for any other variables ( $p > 0.05$ ). As such, the following analyses were conducted without such controls.

All results are reported in Table 4 (see Figures 1–5 for visual comparison). For the cognitive mechanism category (COGMECH), there were significant differences across genres,  $F(2, 462) = 3.75, p = 0.02, \omega = 0.11$ . However, there were no significant linear trends,  $F(1, 462) = 1.74, p = 0.19, \omega = 0.04$ . Planned contrasts 1 (SF > M > F) and 3 (SF > F) were supported. However, contrasts 2 (SF > M) and 4 (M > F) showed no significant differences between genre types. Of special note, *with respect to Suvin's analysis and its explicit contrast between science fiction and fantasy in regards to the representation of cognition, contrast 3 (SF > F) was supported at a high level of significance.*

Overall, there were significant differences across genre for frequencies in the social category (SOCIAL),  $F(2, 462) = 24.60, p < 0.001, \omega = 0.09$  with significant linear trends,  $F(1, 462) = 48.86, p < 0.001, \omega = 0.09$ . All planned contrasts were ~~con-~~ ~~firmed~~. In other words, mystery has significantly higher frequency of social terms than fantasy, and fantasy in turn has higher a higher frequency than science fiction.

supported  
^

For words falling in LIWC's perception category (PERCEPT) there were significant differences overall,  $F(2, 294.25) = 11.05, p < 0.001, \text{adj. } \omega = 0.20$ . For planned contrast 3 (M > SF), there were no significant differences. However, contrasts 1 (M > F > SF), 2 (M > F) and 4 (F > SF) showed significant effects in the *opposite* pattern predicted. In other words, while mystery and science fiction had no differences, fantasy was shown to have significantly more perception words than the other two genres.

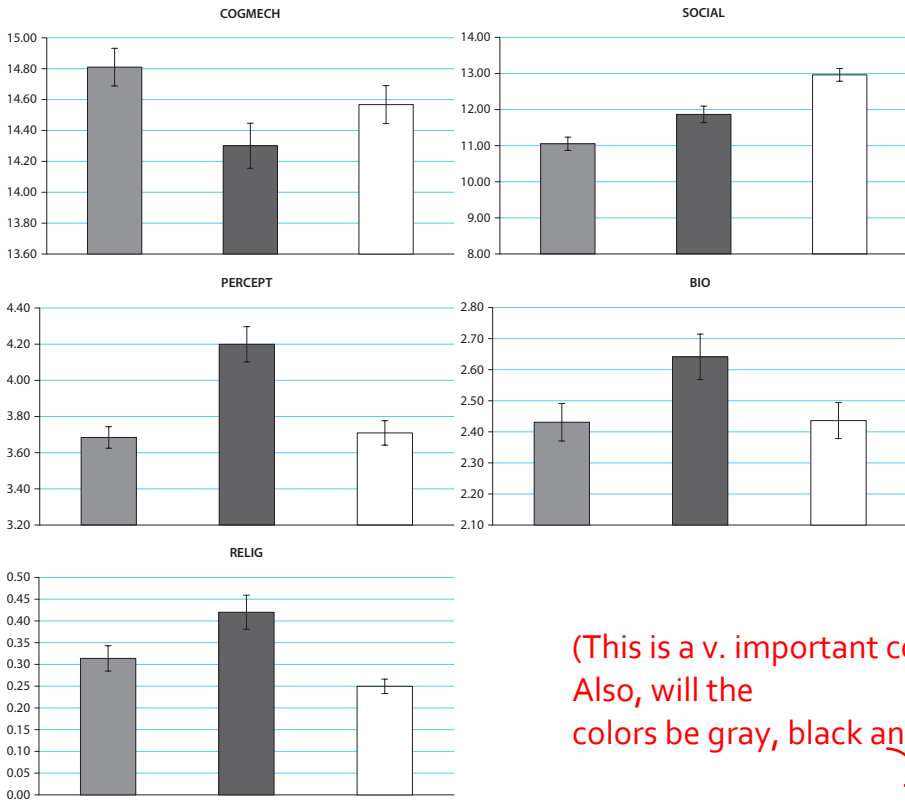
Genre also showed significant overall effects on the biological category (BIO),  $F(2, 302.44) = 3.01, p = 0.05, \text{adj. } \omega = 0.09$ . While contrasts 1 (M > F > SF) and 3 (F > SF) showed no significant differences between genres, fantasy had more biological terms than mystery (contrast 2) and science fiction (contrast 4). In other words, while science fiction does not distinguish itself from mystery in relation to rates of biological terms, fantasy stood out as having more than the other genres.

There were also significant effects across genre for frequencies of words in LIWC's religion category (RELIG),  $F(2, 268.12) = 8.63, p < 0.001, \text{adj. } \omega = 0.18$ . All

Table 4. ANOVA summary for LIWC variables 1

Variable	Contrast	Prediction	<i>t</i>	<i>df</i>	<i>p</i>	<i>r</i>
COGMECH	1	SF > M > F	2.59**	462	0.010	0.12
	2	SF > M	1.34	462	0.180	0.06
	3	SF > F	2.74***	462	0.003	0.13
	4	M > F	1.44	462	0.150	0.07
SOCIAL	1	M > F > SF	6.79***	462	0.000	0.30
	2	M > F	3.94***	462	0.000	0.18
	3	M > SF	6.99***	462	0.000	0.31
	4	F > SF	2.91***	462	0.002	0.13
PERCEPT	1	M > F > SF	-1.73*	336.75	0.040	0.09
	2	M > F	-4.13***	264.22	0.000	0.25
	3	M > SF	0.28	311.98	0.390	0.02
	4	F > SF	4.51***	243.39	0.000	0.28
BIO	1	M > F > SF	-0.87	347.27	0.190	0.05
	2	M > F	-2.20*	283.78	0.015	0.13
	3	M > SF	0.06	314.95	0.476	0.00
	4	F > SF	2.21**	287.57	0.014	0.13
RELIG	1	F > M > SF	2.89***	221.16	0.002	0.19
	2	F > M	3.99***	197.58	0.000	0.27
	3	F > SF	2.17*	273.55	0.020	0.13
	4	M > SF	-1.91*	248.84	0.030	0.12

\* $p \leq 0.05$ , \*\* $p \leq 0.01$ , \*\*\* $p \leq 0.001$



Figures 1–5. Frequencies of LIWC content categories across genre, science fiction (gray), fantasy (striped), and mystery (white)

(This is a v. important correction. Also, will the colors be gray, black and white?)

black

planned contrasts showed significance. However, planned contrast 4 ( $M > SF$ ) was significant in the *opposite* predicted direction,  $t(248.84) = -1.91$ ,  $p = 0.03$ ,  $r = 0.12$ .

### Results from stylistic categories

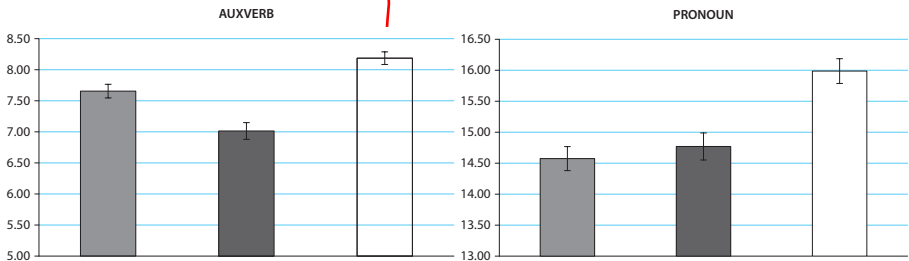
Table 5 reports analyses for variables relevant for an assessment of stylistic differences across genres (see Figures 6–7). Frequencies in LIWC's pronoun category mostly confirm our predictions by showing both general effects,  $F(2, 462) = 14.43$ ,  $p < 0.001$ ,  $\omega = 0.23$ , as well as linear trends,  $F(1, 462) = 24.79$ ,  $p < 0.001$ ,  $\omega = 0.22$ . Planned contrast 4 ( $F > SF$ ) shows no statistically significant differences. However, planned contrasts 1 ( $M > F > SF$ ), 2 ( $M > F$ ) and 3 ( $M > SF$ ) are all supported. See Table 6.

For auxiliary verb frequencies (AUXVERB), genre shows significant effects,  $F(2, 301.18) = 24.44$ ,  $p < 0.001$ , adj.  $\omega = 0.30$ . Planned contrasts are consistent with support

Table 5. ANOVA Summary for LIWC Variables 2

Variable	Contrast	Prediction	<i>t</i>	<i>df</i>	<i>p</i>	<i>r</i>
PRONOUN	1	M > F > SF	5.37***	462	0.000	0.24
	2	M > F	4.21***	462	0.000	0.19
	3	M > SF	4.97***	462	0.000	0.23
	4	F > SF	0.67.	462	0.251	0.03
AUX	1	M > SF > F	6.83***	346.89	0.000	0.34
	2	M > SF	3.54***	313.21	0.000	0.20
	3	M > F	6.96***	277.36	0.000	0.39
	4	SF > F	3.69***	286.67	0.000	0.21

\*\*\**p* ≤ 0.001, \*\**p* ≤ 0.01, \**p* ≤ 0.05



Figures 6–7. Frequencies of LIWC style categories across genre, science fiction (gray), fantasy (striped), and mystery (white)

our prediction that mystery contains higher rates of auxiliary verbs than science fiction and fantasy.

Table 6. [INSERT TABLE 6 HERE]

Typesetter's note: No Table 6 was found in the MS

Thank you for your attention to this detail. We removed all references to Table 6, so this table area too should be removed. Apologies.

placeholder for Table 6

## Summary

While science fiction writers use cognition words more frequently than do fantasy writers, mystery writers use cognition words only slightly less frequently than science fiction writers, as hypothesized. Notably, science fiction has the lowest frequency of social terms amongst our three genres. Science fiction ~~has the highest~~ mean number of perception words, which was inconsistent with our hypothesis. Our hypothesis that the prevalence of non-embodied forms of perception in science fiction — perception with a device's sensors for example — would significantly reduce the rate of perception terms relative to the other genres was falsified. Fantasy writers use significantly more biological terms in their writing than do writers of mystery and science fiction, which is consistent with our predictions, but they also use more words in the biological category than mystery writers, which is inconsistent with our predictions. In this case, science fiction is no different from mystery.

Though fantasy has significantly more religious terms than any other genre, science fiction uses more religious terms than mystery. Science fiction authors might have a higher rate of religion terms in their work than mystery authors because they are more concerned to discuss religious subject matter than are mystery authors. This is contrary to our hypothesis, which conflated the *religiosity* of the genres with the genre's concern to *discuss* issues of religion. For example, Greg Egan's "Oceanic" is clearly concerned with religion, albeit from skeptical point of view. This story alone contains 45 uses of "God", almost 1/20th of the total uses of that token found in our science fiction dataset. Alternatively, we might have obtained this result due to the unusual breadth of LIWC's RELIG category. RELIG includes a number of word stems that might be better denominated as members of an as yet nonexistent MORALITY subcategory (immoral\*, moral, morality, morals). Consistent with its philosophical elements, our science fiction dataset contains more tokens of these four terms (81) than mystery (44) and fantasy (36) put together.

In terms of literary style, science fiction and fantasy both use lower rates of pronouns than does mystery, as hypothesized. Science fiction writers use a lower rate of auxiliary verbs than mystery writers, but a significantly higher rate than fantasy. Our omnibus hypotheses specifying directionality across all three genres were confirmed for categories of cognition, social, religion, as well as for categories of pronoun and auxiliary verb, but not for categories of perception and biology. Of the 18 ~~bi~~directional contrasts tested, our hypotheses were disconfirmed in three cases in which the directionality of the data was the reverse of our prediction (PERCEPT M > F, BIO M > F, and RELIG M > SF). In an additional five ~~bi~~directional contrasts, though our prediction about directionality was correct, the results were not statistically significant. ~~This suggests a need for further testing.~~

did not have  
the lowest  
^

(please delete  
'bi' and leave  
'directional'  
here)



## General discussion

Considerable skepticism about what our method adds to the value of the study of genre is natural. After all, Jockers anticipates a bias from literary studies against “the usefulness of quantification” (2013, p. 30). In response to this skepticism, first note that traditional genre theory is not immune to this criticism. As we briefly argued above, traditional genre theory has made little tractable progress answering the questions “What is genre?” and “How is one genre distinguished from others?” in the previous decades. Though the questions still receive answers from literature scholars, over time those answers have a decreasing relationship to testability.

To say only this in response to skepticism about our methods is to invite the accusation that we commit a *tu quoque* fallacy, so we add the following points. At least two features set our project apart both from traditional genre theory and from quantitative studies of genre. First, we use *quantitative, statistical techniques to study literature*, in contrast to traditional literary theory. Rather than *tell* readers about a new a priori criterion purporting to individuate the science fiction genre from others and support it with examples, cherry-picked or not, we articulate planned contrasts in terms of mean variance across word categories by genre and then *show* readers how we come to our quantitative profiles of three genres.

Second, our method aims to *test hypotheses of literary scholars*. Neither traditional genre theory nor, that we know of, recent quantitative literary research tests hypotheses. In our case, we deployed an ANOVA for this purpose as opposed to the more exploratory forms of data mining found in previous quantitative textual analysis such as factor analysis. This allowed us to put Suvin’s formal analysis *to the test* by operationalizing it in terms of cognition words and social, family and home words (which represent a lack of estrangement). Put in other terms, our quantitative methods have vindicated the most important traditional literary theory about the science fiction genre. We have shown not only *that* literary scholars’ ideas about genre can be held to account with data, but, by demonstrating a new method, also *how* they can be.

Establishing a hypothesis-testing method with humanities sources becomes important in reference to wider methodological challenges voiced by scientists in a variety of popular and semi-academic venues in terms of the failure of the humanities. Since hypothesis testing is an essential feature of the scientific method, we infer that a study like this takes a step toward a science of science fiction and a science of literature more generally. This is why our primary contribution to traditional debate about genre is foremost a *methodological* contribution. Since our results are falsifiable, they promote further testing and extrapolation. Given the serious limitations of our study, further research and testing is needed. Perhaps in long-form writing one would find more evidence of omniscient narration in

science fiction, in which case our model's commitment to variance in rates of pronoun use across the genres might not hold. Perhaps by using texts from early twentieth-century science fiction, with its prevalence of "psi-powers", one would find evidence to vindicate our hypotheses about BIO and PERCEPT<sup>3</sup> hypotheses that the tests on the current dataset falsified. Perhaps by examining a parallel set of Chinese science fiction, mystery, and fantasy texts we could uncover key cross-cultural differences in the nature and meaning of the science fiction genre.

—  
(an em-dash)

In addition to these three methodological outcomes of our project, we have illuminated genuine differences between science fiction, mystery, and fantasy by confirming the majority of our directional hypotheses. Our results provide answers to recurrent questions posed by not only genre scholars but fans the world over.

Our method has limitations. First, our data processing procedures resulted in errors in the encoding of some of the words in the texts. Though we attempted to catch these errors, no doubt some slipped through our quality control efforts. The use of a large dataset is intended to compensate for this problem, but it remains a problem. Second, our sampling of three genres was constructed carefully so as to control for confounds such as date of composition. Were dates of publication to vary widely — from 1890 to 2010, for example — we would expect changes in literary expression over time to swamp the unique literary profile of each genre. However, controlling for date of composition implies that we quantitatively profile our genres as represented in a narrow set of years at the dawn of the millennium. Related, literature sampled for our studies was drawn exclusively from the English language. Ongoing collaboration with Wu Yan, the dean of Chinese science fiction scholarship, might offer a unique opportunity to conduct further research addressing the language limitation we note.

A less obvious ~~third~~ limitation arises through the use of LIWC2007. LIWC presents users with a fixed array of categories and subcategories. Thus our hypothesis testing activity was limited to LIWC's native categories and subcategories. Though a limitation, we selected hypotheses that ported well into the LIWC environment.

Where does the science of genre go from quantitative profiling? For our research group, this paper was conceived as a proof-of-concept project in an effort to establish the reliability of methods applied to texts in order to test other hypotheses drawn from cognitive science and psychology. But the method we have introduced can be used to test additional hypotheses, and answer additional research questions. Given LIWC's original use in mental health diagnostics, and Pennebaker's (2011) discussion of correlations between LIWC data and the mental health of writers, this method might be repurposed to make guarded inferences about the psychological and emotional profiles of writers and readers. In light of the recent push toward understanding how evolved psychology interacts with and expresses itself in literature (e.g., Boyd, 2009; Gottschall & Wilson, 2005), and the

use of texts by cognitive scientists to test predictions about salience and transmission of textual content (e.g., Barrett, Burdett, & Porter, 2009; Norenzayan, Atran, & Schaller, 2006), new horizons await a scientific study of genre. Methods in this paper, supplemented with customized word subscales, can generate an evidence-based characterization of the differences between other genres and subgenres; allow us to understand what draws individuals to particular genres; help understand why exemplar works have achieved their status (see Jockers 2013, pp. 154–170); and identify diachronic psychological and linguistic changes within and between genres. This paper is a small step toward a science of science fiction.

## Acknowledgments

This paper was presented to audiences at the World Science Fiction Convention in Reno, August 2011, at the Eaton Conference in Riverside, April 2013, and at Beijing Normal University, December 2012. We thank participating audience members for feedback, questions or comments, and we give thanks to Wu Yan, Nolan Belk, Cory Doctorow, Mark Silcox, Amy Coplan, two anonymous referees, Anthony Davis Jr., and Fred Lerner. Special thanks go to David Hanauer for his careful reading and constructive criticism of previous versions of this paper. Research for this project was supported by California State University, Fullerton, by the Social Sciences and Humanities Research Council (SSHRC) of Canada, and the John Templeton Foundation through the Cultural Evolution of Religion Research Consortium (CERC) at University of British Columbia.

## References



Amison, S., Heuser, R., Jockers, M., Moretti, F., & Witmore, M. (2011, January 15). Stanford literary lab- quantitative formalism: An experiment. Retrieved from <http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf>. Accessed 7/31/11.



Belamy, S. R. (1971). About five thousand one hundred and seventy five words. In T. D. Claerson (Ed.), *SF: The other side of Realism* (pp. 130–146). Bowling Green: Bowling Green State University Press.

Fredericks, S. C. (1978). Problems of fantasy. *Science Fiction Studies*, 14, 33–34.

Goble, N. (1972). *Asimov analysed*. Baltimore: Mirage Press.



Gunsch, M. A., Brownlow, S., Haynes, S. E., & Mabe, Z. (2000). Differential linguistic content of various forms of political advertising. *Journal of Broadcasting & Electronic Media*, 44, 27–42. Doi: 10.1207/s15506878jobem4401\_3

James, E. (1994). *Science fiction in the 20th century*. Oxford: Opus.

Hartwell, D. (Ed.). (1996). *The year's best science fiction: Thirteenth annual collection*. New York: St. Martin's Griffin.



Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

- Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *American Journal of Psychology*, 120, 263–286.
- Kincaid, P. (2003). On the origins of genre. *Extrapolation*, 44, 409–419.
- Knight, D. (1967). *In search of wonder: Essays on modern science fiction*. Chicago: Advent Publishing.
- Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90, 862–877. Doi: 10.1037/0022-3514.90.5.862
- Moretti, F. (2013). *Distant reading*. New York, NY: Verso.
- Moretti, F. (2005). *Graphs, maps, trees*. New York, NY: Verso.
- Nichols, R., Smith, N., & Miller, F. (2008). *Philosophy through science fiction*. NY: Routledge.
- Pennebaker, J. W. (2011). *The secret life of Pronouns: What our words say about us*. NY: Bloomsbury Press. Doi: 10.1093/llc/fqt006
- Pennebaker, J. W., Chung, C., Ireland, M., Gonzales, A., & Booth, R.J. (2007). The development and psychometric properties of LIWC2007.” Retrieved from [www.liwc.net/LIWC2007LanguageManual.pdf](http://www.liwc.net/LIWC2007LanguageManual.pdf). Accessed 7/31/11.
- Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77, 1296–1312. Doi: 10.1037/0022-3514.77.6.1296
- Rawlins, J. P. (1982). Confronting the alien: Fantasy and anti-fantasy in science fiction film and literature. In G. E. Slusser, E. S. Rabkin, & E. R. Scholes (Eds.), *Bridges to fantasy* (Vol. 2). Carbondale, Illinois: Southern Illinois University Press.
- Rieder, J. (2010). On defining SF, or not: Genre theory, SF, and history. *Science Fiction Studies*, 37(2), 191–209.
- Roberts, A. C. (2006). *The history of science fiction*. New York: Palgrave Macmillan.
- Rose, M. (1981). *Alien encounters: Anatomy of science fiction*. Cambridge MA: Harvard.
- Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression vulnerable college students. *Cognition & Emotion*, 18, 1121–1133. Doi: 10.1080/02699930441000030
- ~~Ryman, G. (2006). The mundane fantastic: Interview excerpts. Retrieved from <http://www.10-cusmag.com/2006/Issues/01Ryman.html>. Accessed 9/23/2007.~~
- ~~Ryman, G. (2007). Take the third star on the left and on till morning. *New York Review of Science Fiction*, 19, 4–7.~~
- Stock, R. (1999). Rating the canon. *The Baker Street Journal*, 49, 5–11.
- Suvín, D. (1978). On what is and is not an SF narration: With a list of 101 Victorian books that should be excluded from SF bibliographies. *Science Fiction Studies*, 5, 45–57.
- Suvín, D. (1979). *Metamorphoses of science fiction*. New Haven: Yale UP.
- Swanwick, M. (2002). *Being Gardner Dozois*. Baltimore: Old Earth Books.
- Tausczik, Y., & Pennebaker, J. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 24–54. Doi: 10.1177/0261927X09351676
- Wolfe, G. K. (1986). *Critical terms for science fiction: A glossary and guide to scholarship*. Westport CT: Greenwood.
- Yan, Wu. (2000). Reader’s expectation model about science fiction. *Public understanding of science: International conference of science communication*. Heifei: University of Science and Technology of China Press: 63–70.



### Appendix 1.

The full dataset contains multiple stories by single authors within the same genre. In order to address concerns regarding results using the full dataset, we reduced each corpus so as to allow only one story for each author in each genre. We did this by retaining only the same-author story with the highest word count. This significantly reduced the number of works for each genre (SF = 96, F = 108, M = 105). We retested our hypotheses using the reduced sample. Results appear in the Table A1 below.

Table A1. Results with reduced dataset (duplicate authorship removed)

Variable	Contrast	Prediction	<i>t</i>	<i>df</i>	<i>p</i>	<i>r</i>
COGMECH	1	SF > M > F	2.04*	306	0.042	0.59
	2	SF > M	0.52	306	0.606	0.22
	3	SF > F	2.47*	306	0.014	0.22
	4	M > F	2.00*	306	0.047	0.22
SOCIAL	1	M > F > SF	6.19***	306	0.000	0.30
	2	M > F	4.72***	306	0.000	0.18
	3	M > SF	5.81***	306	0.000	0.31
	4	F > SF	1.24	306	0.215	0.13
PERCEPT	1	M > F > SF	-2.59**	213.18	0.010	0.32
	2	M > F	-3.74***	200.74	0.000	0.14
	3	M > SF	-1.33	198.80	0.185	0.11
	4	F > SF	2.74**	189.90	0.008	0.13
BIO	1	M > F > SF	-0.79	209.30	0.433	0.28
	2	M > F	-1.54	210.86	0.126	0.11
	3	M > SF	-0.26	197.94	0.792	0.11
	4	F > SF	1.26	201.52	0.211	0.11
RELIG	1	F > M > SF	2.30	185.95	0.047	0.16
	2	F > M	3.24***	146.72	0.001	0.05
	3	F > SF	1.42	201.72	0.224	0.06
	4	M > SF	-1.73*	132.36	0.054	0.05

\*\*\**p* ≤ 0.001, \*\**p* ≤ 0.01, \**p* ≤ 0.05

Results using the reduced dataset are largely consistent with the full dataset, though several differences arise (see Table A2). First, no analyses produced results in any category that were significant in the opposite direction hypothesized. Second and more important, the study of genre with the reduced set does not differentially weigh highly successful, genre-defining authors more than it does stories of single contributions. Consider that the full dataset contains six science fiction stories by Greg Egan. Egan has received the John W. Campbell Memorial Award, the Hugo Award, the Locus Award, the Asimov's Readers Award, the Kurd-Laßwitz-Preis, the Seiun Award, and the Ditmar Award. Whereas the full dataset contains all six stories, the reduced dataset contains one. Parallel arguments can be made that the reduced datasets for the mystery

and fantasy genres are just as unrepresentative of the genre as is the reduced dataset for science fiction. Clark Howard and Gene Wolfe, multiple-award winning writers in mystery and fantasy, each have five stories in their respective genres in the full dataset but only one in the reduced dataset. Thus the reduced dataset gives as much weight to Egan (and Howard and Wolfe) as it does to an author who has a single story represented in only one of the six volumes of *The Year's Best Science Fiction* tested. Our quantitative profiles aim at representations (of a time-slice) of a genre. The best authors sell the most books and short stories, win the most awards, have their stories optioned to television and film, reach the largest numbers of readers, are promoted most at conventions, are discussed most frequently by reviewers, are most influential on making a genre what it is, and are most often anthologized in "best of" books.

**Table A2.** Comparison between full and reduced datasets

	Full	Reduced
COG 4	M > F (1.44)	M > F (2.00)*
SOC 4	F > SF (2.91)***	F > SF (1.24)
BIO 2	M > F (-2.20)*	M > F (-1.54)
BIO 4	F > SF (2.21)**	F > SF (1.26)
RELIG 1	F > M > SF (2.89)***	F > M > SF (2.30)
RELIG 2	F > SF (2.17)*	F > SF (1.42)

\*\*\* $p \leq 0.001$ , \*\* $p \leq 0.01$ , \* $p \leq 0.05$ ; F-ratios in ellipses

### *Corresponding author's address*

Ryan Nichols  
 Department of Philosophy  
 Cal State Fullerton  
 Fullerton CA 92834-6868.  
 rnichols@fullerton.edu.