Rupkatha Journal

On Interdisciplinary Studies in Humanities

An Online Open Access Journal
ISSN 0975-2935
www.rupkatha.com

Volume V, Number 3, 2013

Indexing and abstracting
Rupkatha Journal is an international journal recognized by a number of organizations and institutions. It is archived permanently by **www.archive-it.org** and indexed by **EBSCO**, **Elsevier**, **MLA International Directory**, **Ulrichs Web**, **DOAJ**, **Google Scholar** and other organisations and included in many university libraries

Canonical Values vs. the Law of Large Numbers: The Canadian Literary Canon in the Age of Big Data

Carolina Ferrer
Université du Québec à Montréal (UQAM), Canada

Abstract

In this article, I propose an alternative technique to the traditional method of constitution of the literary canon. Instead of basing the determination of the canon on different values, I scrutinize the *Modern Language Association International Bibliography* database in order to determine the most cited authors and literary works. Specifically, I study Canadian literature. Thus, through the process of data mining, I obtain a sample of over 25,000 references that allows us to observe the chronological evolution and the linguistic distribution of the critical bibliography about Canadian literature. This quantitative technique yields a corpus of 151 titles and 295 writers that are cited more than 10 times in the database. Consequently, this bibliography is not the result of subjective selection criteria, but is based on the law of large numbers. Furthermore, this study shows that the quantitative analysis of bibliographic databases is an effective way to bring new light to the field of literary studies.

[**Keywords**: literary canon, Canadian literature, bibliographic databases, bibliometrics, data mining, Big Data]

Canon formation

In spite of its apparent simplicity,the concept of canon is a difficult one to define. According to certain texts (Guillory 1995; Lentricchia and McLaughlin 1995), etymologically, the word finds its source in the Greek word kanon that signifies ruler or measuring stick. Initially, the term was used to identify those texts from the Old and New Testaments that are approved by the ecclesiastic authorities as the Word of God, thus, the texts that constitute the Sacred Scriptures. During the last decades, it has been considered that the canonization of literary texts operates in a way similar to the process of the constitution of the biblical canon.

In 1995, at the beginning of his book *The Western Canon. The Books and School of the Ages*, the American critique Harold Bloom determines that: "Originally the Canon meant the choice of books in our teaching institutions, and despite the recent politics of multiculturalism, the Canon's true question remains: What shall the individual who still desires to read attempt to read, this late in history?" (Bloom, 15).

As such, the fundamental question about the canon seems easy to deal with. However, it is a very complex concept that has caused considerable discussions. For instance, in 1983, *Critical Inquiry* published a volume completely dedicated to the concept of the canon. In the introduction, Robert von Halberg establishes that "Interest in canons is surely part of a larger inquiry into the institutions of literary studies and artistic production. 'Politics,' 'economics,' 'social,' 'authority,' 'power' –these are some of the terms that recur throughout these essays; we

Rupkatha Journal on Interdisciplinary Studies in Humanities (ISSN 0975-2935), Vol. V, No. 3, 2013.
Ed. Tirtha Prasad Mukhopadhyay.
URL of the Issue: http://rupkatha.com/v5n3.php
URL of the article: http://rupkatha.com/V5/n3/06_Canadian_literary_canon.pdf
Kolkata, India. © www.rupkatha.com

are most curious now about those points where art seems less private than social" (von Halberg, iii).

In 1999, Nel van Djik analyzes the canon formation from different viewpoints: nationalism, literature, and institutions. According to van Djik:

> The list of works that count as our western society's literary inheritance is no longer prescribed by the church and the state, but by authorized institutions such as literary criticism and literary education. In addition, scientific developments in the past decades have resulted in the widespread conviction that literary value is not an intrinsic but an attributed quality. This quality results from the consensus that exists between the members of a literary institution at a certain moment (121).

For him, the researches working on this subject are divided into two groups: those that consider it from an ideological-hermeneutic approach and those who prefer a sociological perspective.

A year later, Anderson and Zanetti (2000) determine that the discussion about the canon has been conducted according to two opposite poles. On the one hand, those that belong to the right wing of the political spectrum defend a traditional canon. On the other hand, those that have leftist ideas declare that the canon is an obsolete artefact. Between these two poles, we find a series of perspectives that open up the notion of the canon in order to include minorities or that consider the existence of multiple canons.

In 2003, Jeffrey Insko picks up the debate about this concept, bringing back the tension between the imaginary canon (Guillory) and the pedagogical canon (Gallagher). The following year, Frank Kermode, in his book *Pleasure and Change. The Aesthetics of Canon*, pushes aside the ideological aspects of the canon constitution in order to introduce three new characteristics into the process: pleasure, change, and chance.

My purpose in mentioning these studies about the concept of the canon is not to carry on this debate, but to show that the constitution of a canonical corpus is a highly complex process. My aim is to introduce an alternative method that makes it possible to identify those literary works most frequently analyzed within a national literature[i]. In other words, instead of presenting a literary canon, I will trace a literary cartography. Specifically, I will focus on Canadian literature.

Theoretic and methodological approaches

With the purpose of developing this experimental method, I base my research in the articulation of several essential notions. From the theoretical viewpoint, I build my work, on one hand, upon the concept of literary field established by Pierre Bourdieu (1992), and, on the other hand, on scientometrics (De Solla Price, 1963; Garfield, 1980; Leydesdorff, 1998).

Methodologically, there are two essentials aspects: firstly, the analysis of field of knowledge (Albrechtsen, 1997; Hjorland, 2001; Hjorland and Albrechtsen, 1995), and, secondly, theories of citations (Kaplan, 1965; Moravcsik and Murugesan, 1975; Gilbert, 1977; Small, 1978, 1998; Leydesdorff, 1998; Enger, 2009).

Although several scientometrists consider that quantitative methods cannot be used in the humanities due to the differences in terms of citation across the disciplines (Archambault et al, 2006, Cole, 1983, Cozzens, 1985, Larivière et al, 2006, Nederhof et al, 1989), recently, we have witnessed several bibliometric analysis in the humanities in general (Linmans, 2010; Moed et al, 2002; Osca-Lluch and Haba, 2005), and in literary studies in particular (Ardanuy et al, 2009; Hammarfelt, 2011; Herubel and Goedeken, 2000). The latter indicate that, despite the abovementioned citing differences, scientometrics is a relevant approach to increase our understanding of the behaviour of the literary field.

In this research, I use the Modern Language International Bibliography database; from now on I will refer to it as MLAIB. This electronic bibliography, renown as the most important one in literary studies, contains over 2,107,000 references and includes approximately 4,400 journals. Besides the articles, the MLAIB database includes references to books, book chapters and thesis[ii]. In terms of chronology, it covers the literary critique from 1886 to the present.

Through the techniques of data mining (Han *et al*, 2012; Witten*et al*, 2011), and keywords (Callon *et al*, 1993), I initially obtain a sample of the critical bibliography about Canadian literature. Then, I extract the Canadian literary corpus as well as a list of the main Canadian writers. The data is directly transferred in parcels of 500 references each time to a RefWorks account. The latter data management tool allows me to automatically convert the series into a spreadsheet format. Thus, from that point on, I can efficiently work with the series in order to obtain indicators, graphics, and tables and to qualitatively query their contents. The cutting date for the analysis is 2010.

Literary cartography of Canadian literature

Through the application of the data mining technique, I obtained a sample of 25,102 references covering a period of 124 years, from 1886 until 2010. Figure 1 represents the results of this data mining process. The series begin to grow by the end of the 1950s. Since 2001, the average number of references has reached a volume of 792 publications per year. Journal articles are the most important type of documents, 67% of the sample, followed by book chapters, 26%. In terms of the language of publication, Figure 2, the sample is divided into 70% in English and 27% in French. Individually, other languages correspond to less than 1%, and altogether they only reach 3% of the references.

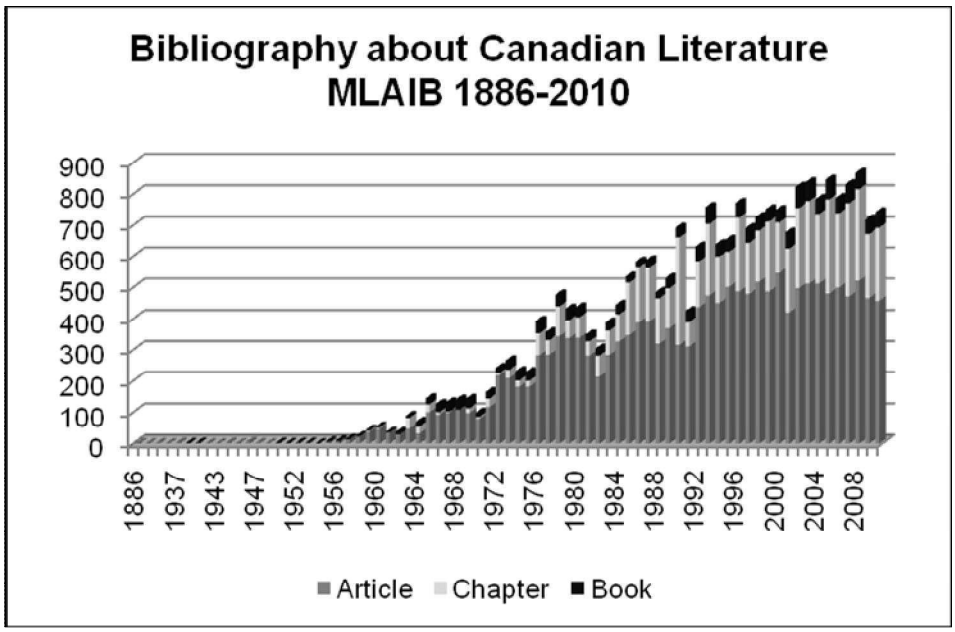

Fig. 1. Chronological evolution of publications about Canadian literature, MLAIB 1886-2010.

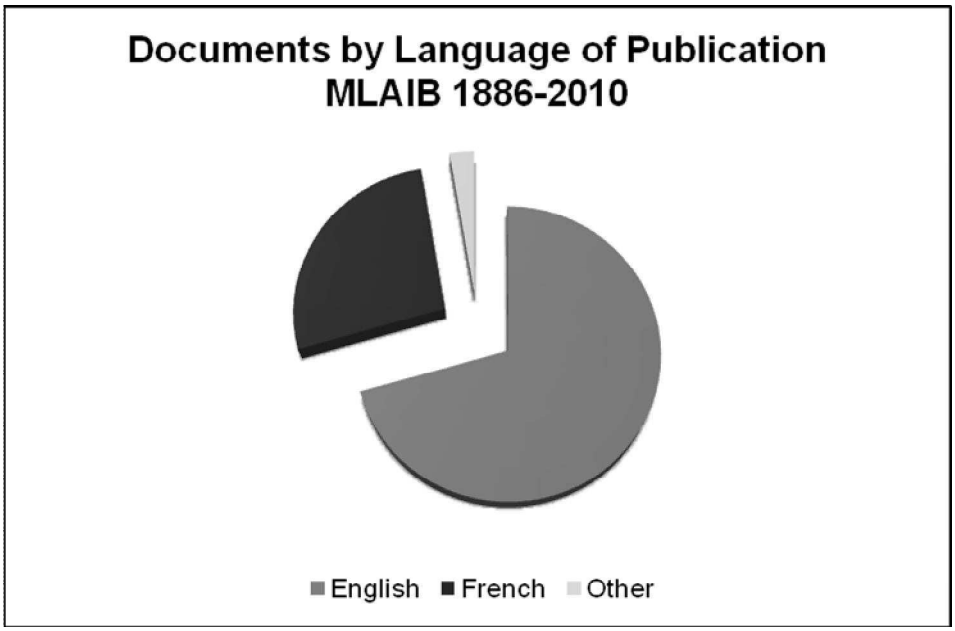

Fig. 2. Documents about Canadian literature by language of publication, MLAIB 1886-2010.

Writers and Texts

In order to identify the most studied Canadian literary works and writers, I interrogated again the MLAIB database. I selected those writers and works that have been the object of at least10 publications. The results of these interrogationswerea corpus of 151 titles and a list of 295 authors. Table 1 corresponds to the top 20 Canadian titles according to MLAIB.

Table 1. Corpus of the top 20 Canadian literary works, MLAIB 1886-2010.

Author	Title	References
Atwood, Margaret (1939-)	*The Handmaid's Tale* (1985)	164
Ondaatje, Michael (1943-)	*The English Patient* (1992)	126
Atwood, Margaret (1939-)	*Surfacing* (1972)	94
Kogawa, Joy Nozomi (1935-)	*Obasan* (1981)	93
Montgomery, L. M. (1874-1942)	*Anne of Green Gables* (1908)	84
Atwood, Margaret (1939-)	*Cat's Eye* (1989)	70
Atwood, Margaret (1939-)	*Oryx and Crake* (2003)	65
Laurence, Margaret (1926-1987)	*The Diviners* (1974)	61
Hébert, Anne (1916-2000)	*Les Fous de Bassan* (1982)	60
Hébert, Anne (1916-2000)	*Kamouraska* (1970)	57
Ondaatje, Michael (1943-)	*Anil's Ghost* (2000)	55
Roy, Gabrielle (1909-1983)	*Bonheur d'occasion* (1945)	55
King, Thomas (1943-)	*Green Grass, Running Water* (1993)	54
Ondaatje, Michael (1943-)	*Running in the Family* (1982)	54
Atwood, Margaret (1939-)	*Alias Grace* (1996)	52
Frye, Northrop (1912-1991)	*Anatomy of Criticism* (1957)	50
Ondaatje, Michael (1943-)	*In the Skin of a Lion* (1987)	50
Hémon, Louis (1880-1913)	*Maria Chapdelaine* (1916)	45
Atwood, Margaret (1939-)	*Lady Oracle* (1976)	44
Atwood, Margaret (1939-)	*The Robber Bride* (1993)	42

As we can observe, Margaret Atwood is the most outstanding author with 7 titles on the corpus. Next on the list are Michael Ondaatje, with 4 titles, and Anne Hébert, with 2 titles. Among this list, there are only 4 titles in French: *Les fous de Bassan* and *Kamouraska* by Anne Hébert, *Bonheur d'occasion* by Gabrielle Roy, and *Maria Chapdelaine* by Louis Hémon. This short list includes works from 1916 until 2000. However, the complete corpus spans from 1832 to 2008.

In Table 2, I present the top 20 Canadian authors. Again, Margaret Atwood is by far the writer with the highest number of references. In this case, we observe a perfect equilibrium between English and French Canadian literature, since there are 10 writers that represent each language.

Table 2. List of the top 20 Canadian writers, MLAIB 1886-2010.

Author	References
Atwood, Margaret (1939-)	1096
Frye, Northrop (1912-1991)	398
Ondaatje, Michael (1943-)	395
Hébert, Anne (1916-2000)	376
Laurence, Margaret (1926-1987)	366

Munro, Alice (1931-)	360
Roy, Gabrielle (1909-1983)	340
Montgomery, L. M. (1874-1942)	223
Tremblay, Michel (1942-)	198
Aquin, Hubert (1929-1977)	192
Blais, Marie-Claire (1939-)	176
Brossard, Nicole (1943-)	161
Kroetsch, Robert (1927-2011)	161
Davies, Robertson (1913-1995)	140
Findley, Timothy (1930-2002)	139
Ferron, Jacques (1921-1985)	137
Grove, Frederick Philip (1879-1948)	136
Ducharme, Réjean (1942-)	135
Maillet, Antonine (1929-)	135
Saint-Denys Garneau, Hector de (1912-1943)	127

Concluding remarks

At the beginning of this research, I presented several aspects that, according to various scholars, make the process of identifying canonical literary works very difficult. Then, I introduced an alternative method based on data mining. Thus, I was able to obtain a corpus of literary works that reflect the interest that the academic critique has expressed for Canadian literature through their publications for over 120 years. It seems to me that this method overcomes the different tensions involved in the process of canon constitution and, at the same time, it responds to the issues signified by Nel van Djik as well as to Harold Bloom's essential question about the canon.

According to van Djik, the canon is prescribed "by authorized institutions such as literary criticism and literary education" (van Djik, 121). In this sense, the method I propose here is based on a very significant part of the critical activity of scholars. When using and exploiting the MLAIB database with scientometric methods, the bibliography obtained is not the result of subjective selection criteria, but is based on the law of large numbers. Actually, in the case of Canada, the results compiled in this research are the reflection of more than 25,000 references obtained with quantitative techniques that can be reproduced and verified.

At the same time, the results presented here respond to Harold Bloom's question. Adapted to the literature analyzed in this case: "What shall the individual who still desires to read [Canadian literature] attempt to read, this late in history?" (Bloom,15). It seems to me that the corpus obtained provides an appropriate answer to what Bloom considers to be the interrogation that lays at the center of canon constitution.

Moreover, I consider that the corpus here obtained is clearly superior to any canonical list. Firstly, these results may be classified according to different parameters: geographic, chronologic, and linguistic. Secondly, these lists are dynamic as they can be actualized by periodically interrogating the MLAIB database. Finally, the corpus presented here, that includes 151 titles, is much larger than Bloom's list for Canada, which contains only 9 titles (Bloom, 530).

It seems to me that this analysis is a clear demonstration of the relevance of using scientometric methods in the study of literature as they allow us to increase and deepen our knowledge about the literary field.

Notes

[i] In his book *Dialogues with/and Great Books* (2012), David Fishelov uses bibliometric and Webometric indicators, not to propose an alternative method to canon formation, but to confirm the importance of canonical lists that already exist.

[ii] In this study, I omit the thesis, since the MLAIB database essentially includes thesis published in the USA. See Fitz-Enz, 2008.

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada.

References

Albrechtsen, H. (1997) "Knowledge organization in the humanities", *Knowledge Organization*, 24(2): 61-63.

Anderson, E. R. and Zanetti, G. (2000) "Comparative Semantic Approaches to the Idea of a Literary Canon", *The Journal of Aesthetics and Art Criticism*, 58(4): 341-360.

Archambault, E., Vignola-Gagné, E., Côté, G., Larivière, V., and Gingras, Y. (2006) "Benchmarking scientific output in the social sciences and humanities: The limits of existing databases", *Scientometrics* 68(3): 329-342.

Ardanuy, J., Urbano, C., and Quintana, L. (2009) "A Citation Analysis of Catalan Literary Studies (1974-2003). Towards a Bibliometrics of Humanities Studies in Minority Languages", *Scientometrics* 81(2): 347-366.

Bloom, H. (1995) *The Western Canon: the Books and School of the Ages*, New York: Riverhead.

Bourdieu, P. (1992) *Les règles de l'art.Genèseet structure du champ littéraire*, Paris: Seuil.

Callon, M., Courtial, J-P., and Penan.H. (1993) *La Scientométrie*. Que Sais-Je? 2727, Paris: Presses Universitaires de France.

Childers, J., and Hentzi, G. (Eds.).(1995) *Columbia Dictionary of Modern Literary and Cultural Criticism*.New York: Columbia University Press.

Cole, S. (1983) "The Hierarchy of the Sciences", *American Journal of Sociology* 89(1): 111-139.

Cozzens, S. E. (1985) "Using the Archive - Price, Derek Theory of Differences among the Sciences", *Scientometrics* 7(3-6): 431-41.

Doan, A., Halevy, A., and Ives, Z. (2012) *Principles of Data integration*, Waltham: Morgan Kaufmann.

Enger, K. B. (2009) "Using citation analysis to develop core book collections in academic libraries", *Library & Information Science Research*, 31(2): 107-112.

Fishelov, D. (2010) *Dialogue with/and Great Books. The Dynamics of Canon Formation*, Brighton: Sussex Academic Press.

Fitz-Enz, D. (2008) *MLA International Bibliography Database Guide*, CSA Illumina, www.csa.com.

Garfield, E. (1980) "Is Information-Retrieval in the Arts and Humanities Inherently Different from that in Science – Effect that ISIS-Citation-Index-for-the-Arts-and-Humanities is Expected to Have on Future Scholarship", *Library Quarterly*, 50(1): 40-57.

Garfield, E. (2005) "A Prospective View of Citation Indexing Retrieval in the 21st. Century". *On the occasion of being presented the ASIS&T Los Angeles Chapter's Contributions to Information Science & Technology Award*, 2004. Los Angeles, California.

Gilbert, G. N. (1977) "Referencing as Persuasion", *Social Studies of Science* 7(1): 113-22.

Guillory, J. (1995) "Canon", In: Childers J. and Hentzi G. (Eds.) *Columbia Dictionary of Modern Literary and Cultural Criticism*: 233-249, New York: Columbia University Press.

Hammarfelt, B. (2011) "Interdisciplinarity and the Intellectual Base of Literature Studies: Citation Analysis of Highly Cited Monographs", *Scientometrics* 86(3): 705-725.

Han, J., M. Kamber and J. Pei. (2012) *Data Mining.Concepts and Techniques*. Waltham: Morgan Kaufmann.

Herubel, J., and Goedeken, E. A. (2000) "Metadisciplinarity, Belles Lettres, and André Malraux: A bibliometric exploration of knowledge formation", *Serials Librarian*, 37(4): 51-68.

Hjorland, B. (2001) "Towards a theory of aboutness, subject, topicality, theme, domain, field, content ... and relevance", *Journal of the American Society for Information Science and Technology*, 52(9): 774-778.

Hjorland, B. and Albrechtsen, H. (1995) "Toward a New Horizon in Information-Science - Domain-Analysis". *Journal of the American Society for Information Science* 46(6): 400-25.

Insko, J. (2003) "Generational Canons", *Pedagogy*, 3(3): 341-358.

Kaplan, N.(1965) "The norms of citation behavior - Prolegomena to the footnote", *American Documentation* 16(3): 179-84.

Kermode, F. (2004) *Pleasure and Change. The Aesthetics of Canon*, Oxford: Oxford University Press.

Larivière, V., Archambault, E., Gingras, Y. and Vignola-Gagné, E. (2006) "The Place of Serials in Referencing Practices: Comparing Natural Sciences and Engineering with social Sciences

and Humanities", *Journal of the American Society for Information Science and Technology* 57(8): 997-1004.

Lentricchia, F., and McLaughlin, T. (Eds.) (1995) *Critical Terms for Literary Studies*, Chicago: The University of Chicago Press.

Leydesdorff, L. (1998) "Theories of citation?" *Scientometrics* 43(1): 5-25.

Linmans, A. J. M. (2010) "Why with bibliometrics the Humanities does not need to be the weakest link. Indicators for research evaluation based on citations, library holdings, and productivity measures", *Scientometrics*, 83(2): 337-354.

Modern Language Association International Bibliography, www.mla.org.

Moed, H. F., Luwei, M., and Nederhof, A. J. (2002) "Towards research performance in the humanities", *Library Trends*, 50(3): 498-520.

Moravcsik, M. J., and Murugesan, P. (1975) "Some Results on Function and Quality of Citations", *Social Studies of Science* 5(1): 86-92.

Nederhof, A. J., Zwaan, R. A., Debruin, R. E. and Dekker P. J. (1989) "Assessing the Usefulness of Bibliometric Indicators for the Humanities and the Social and Behavioral-Sciences – A Comparative Study", *Scientometrics* 15(5-6): 423-435.

Osca-Lluch, J., and Haba, J. (2005) "Dissemination of Spanish social sciences and humanities journals", *Journal of Information Science* 31(3): 230-237.

Price, D. D. S. (1963) *Little Science, Big Science*, New York: Columbia University Press.

RefWorks. http://www.refworks.com

Small, H. (1998) "Citations and consilience in science - Comments on Theories of citation?" *Scientometrics* 43(1): 143-48.

Small, H. (1978) "Cited documents as concept symbols", *Social Studies of Science* 8: 327-40.

Thompson, J. W. (2002) "The death of the scholarly monograph in the Humanities? Citation patterns in literary scholarship", *Libri* 52(3): 121-136.

Van Dijk, N. (1999) "Research into canon formation: nationalism, literature, and an institutional point of view", *Poetics Today*, 20(1): 121-132.

Von Halberg, R. (1983) "Editor's introduction", *Critical Inquiry*, 10(1): iii-vi.

Witten, I.H., Frank, E., and Hall, M.A. (2011) *Data Mining. Practical machine Learning Tools and Techniques*, Waltham: Morgan Kaufmann.

Carolina FERRER is Associate Professor at the Department of Literary Studies of the University of Quebec at Montreal (UQAM), Canada. Her research covers Spanish American literature and culture, cultural dynamics, semiotic approaches to database systems, literature and electronic archives, film studies, epistemocriticism. Currently, she works on the propagation processes of ideas across disciplinary fields as well as on the interdiscoursive relations between literature, cinema, and socio-political context. In 2008, she inaugurated Babel Borges <www.babelborges.org>, a research group dedicated to the study of the diffusion of Jorge Luis Borges's work through culture. Among her main publications, we should mention the coedited work with Lucille Beaudry and Jean-Christian Pleau: *Art et politique. La représentation en jeu* (Québec: Presses de l'Université du Québec, 2011). Since 2012, she is the Director of the PhD program in semiotics (UQAM) <http://www.doctorat-semiologie.uqam.ca>.

Literature in the literary canon must be able to be passed down by generations; they must possess excellent literary techniques, appeal to a variety of people, and provoke thought. What is a canon as it relates to literature? In literary terms, a canon is a highly respected body of work.Â A literary canon reflects the values of those who compiled it. There may have been other books contemporary with the canonical ones which were influential but which the compilers deemed not virtuous enough for inclusion in the canon. What question does the inclusion of Samuel Beckett's Waiting for Godot or James Joyce's Finnegan's Wake in the literary canon raise? D. Should artistic merit be the sole criterion for inclusion in the canon? The law of large numbers just tells us that my sample mean will approach my expected value of the random variable. Or I could also write it as my sample mean will approach my population mean for n approaching infinity. And I'll be a little informal with what does approach or what does convergence mean? But I think you have the general intuitive sense that if I take a large enough sample here that I'm going to end up getting the expected value of the population as a whole. And I think to a lot of us that's kind of intuitive. That if I do enough trials that over large samples, the Canonical Traditions in Comparative Perspective Human societies establish collections of classics and canons within various segments of life.1 In this setting, a canon would provisionally mean â€œany 1. See variously Hallberg, Canons, 1984; Gorak, Making of the Modern Canon, 1991; Hjort, Rules and Conventions, 1992; Guillory, Cultural Capital, 1993; Heydebrand, Kanon â€“ Macht â€“ Kultur, 1996; Baehr, Founders, Classics, Canons, 2002; Reed, New.Â The composition, which possibly emerged in the fifth or fourth century BCE, has one brief chapter professing to have been written by the master, and nine chapters of commentary on the masterâ€™s text.