The Scientization of Literary Study

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

Andrew Piper
Languages, Literatures, and Cultures
McGill University
Montreal, Canada
andrew.piper@mcgill.ca

Abstract

Scholarly practices within the humanities have historically been perceived as distinct from the natural sciences. We look at literary studies, a discipline strongly anchored in the humanities, and hypothesize that over the past half-century literary studies has instead undergone a process of "scientization", adopting linguistic behavior similar to the sciences. We test this using methods based on information theory, comparing a corpus of literary studies articles (around 63,400) with a corpus of standard English and scientific English respectively. We show evidence for "scientization" effects in literary studies, though at a more muted level than scientific English, suggesting that literary studies occupies a middle ground with respect to standard English in the larger space of academic disciplines. More generally, our methodology can be applied to investigate the social positioning and development of language use across different domains (e.g. scientific disciplines, language varieties, registers).

1 Introduction

The study of literature has historically been seen as a scholarly practice that is distinct from the natural sciences (Wellmon, 2017; Rickman, 1976). This view became particularly pronounced in the twentieth century with the growth of scientific disciplines within universities and the expansion of government funding for such initiatives. Today, it remains a commonplace to argue that literary studies, as a subset of the humanities more generally, has a distinctive set of methods, concepts, and practices that produce a unique form of knowledge (Nussbaum, 1997; Kramnick, 2018).

Our aim in this paper is to test the opposing view to this consensus, namely, that literary studies has over the past half-century become more "scientific". By this we do not mean that literary studies has gradually come to share similar vocabulary or concepts to other scientific disciplines. To be "like science" in this sense does not mean the adoption of a distinctly scientific language. Rather, we define the process of *scientization* as a set of three interlocking linguistic practices, which we set out to test here: social differentiation, diachronic specialization, and phrasal standardization.

By social differentiation we mean the extent to which the language of a scholarly discipline distinguishes itself from standard linguistic practices within a given language or culture. The more distinctive a field is with respect to "common language use" the more socially differentiated that field is (Ure, 1982). As Degaetano-Ortlieb and Teich (2016) have shown, scientific language in English has gradually become increasingly divergent from standard representations of English over time. This is the first hypothesis of scientization: that literary studies should look increasingly different from standard English over time (H1).

Specialization on the other hand refers to a process of *self*-differentiation over time. Teich et al. (2016) and Degaetano-Ortlieb et al. (2019) have shown that as a scientific field develops, it will become increasingly specialized and expert-oriented. As a field specializes, it develops more technical and differentiated vocabulary (cf. Halliday (1988); Teich et al. (2016)), while retaining some past linguistic practices and frameworks. A growing aspect of its vocabulary will thus not be accounted for by its own past vocabulary. Past and present will become asymmetrically different from one another. Specialization thus captures the effect of directional linguistic change over time. To reflect increasing specialization, we hypothesize greater linguistic divergence between past and present than vice versa (H2).

Finally, we hypothesize that scientific language is partially defined by a growth of phrasal (i.e.

Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 18–28
Minneapolis, MN, USA, June 7, 2019.

lexico-grammatical) standardization (H3). Less surprise at the local contextual level of linguistic phrases, i.e. more predictable word sequences, allows for more efficient communication – arguably important for the building of scientific knowledge (Harris, 2002; Halliday, 2006). For a discipline to become more scientific it should show evidence of greater standardization at the level of linguistic phrases.

Taken together, our model allows us to test the extent to which a particular field, in this case literary studies, indicates a process of linguistic scientization over time. As we will show, there is evidence that this has been the case, although with important caveats. While literary studies appears to remain more linguistically similar to standard English than scientific language, over time it has shown increased levels of all three dimensions of scientization we measure here: it has become more socially differentiated, diachronically specialized, and phrasally standardized. Our findings suggest that literary studies remains distinctive within the linguistic landscape of "science" in terms of its proximity to standard English, but has simultaneously undergone trends of scientization that point towards its allegiance to the larger project of scientific inquiry. Such conflicting points of view have important implications for any future meta-reflections on the place of literary studies within the university. We see this as a potential indicator of literary studies' bridge-like nature within the academic landscape, a hybrid undertaking that mediates between more fully specialized and differentiated disciplines and common public discourse.

2 Related work

Disciplinary self-knowledge has been integral to the study of literature for well over two-thousand years. As scholars have long demonstrated, the reproduction and reception of literary works was traditionally accompanied by prior critical voices, either in the form of marginal gloss or printed commentary (Reynolds and Wilson, 1991; Tribble, 1993). The "state of the field", as we might now refer to it, was part of the circulation of the field's objects of study. With the institutionalization of literary studies as an academic discipline in the twentieth century, there have been numerous meta-studies of different national and historical contexts of literary study (Kennedy, 1989; Fohrmann and Vosskamp, 1991; Graff, 2007).

More recently, a number of studies have argued for the distinctive nature of literary studies with respect to the social and natural sciences (Nussbaum, 1997; Lamont, 2009; Biber and Gray, 2016; Kramnick, 2018). This work draws on an older tradition that emerged at the start of the twentieth century in response to the era known as "big science" (Rickman, 1976; Wellmon, 2017). The study of creative writing was seen, then as now, as an important protection against the "rationalization" and "standardization" of scientific knowledge. While different hypotheses have been posited as to the unique contribution of literary study as a form of knowledge (whether it makes us more empathetic or critical minded for example), what is consistent throughout this work is the assumption that literary studies is distinct from the broader endeavor known as "science."

All of this work is importantly qualitative in nature. With one exception (Goldstone and Underwood, 2014), no studies have attempted to understand the field of literary studies from a quantitative perspective. In this respect we see our work as part of a growing body of research concerned with the data-driven study of academic disciplines, known as "metaknowledge" or the "science of science" (Evans and Foster, 2011; Fortunato et al., 2018). Researchers have examined the discursive evolution of scientific disciplines (Shi, 2004; Chavalarias and Cointet, 2013; Goldstone and Underwood, 2014), as well as the relationship between tradition and innovation within particular scientific fields (Foster et al., 2015) and the role that highly productive researchers play (Azoulay et al., 2014). Biber and Gray (2010, 2011, 2016) (a.o.) have studied the evolution of scientific writing towards increased linguistic complexity. Degaetano-Ortlieb and Teich (2018) have analyzed the development of scientific writing from the mid 17th to the 19th century towards an optimal code for scientific communication. Vilhena et al. (2014) have examined the linguistic relationships between disciplines and Teich et al. (2016) the linguistic development of interdisciplinary disciplines. Recent work has also studied the notion of paradigmaticness with respect to linguistic behavior within disciplines (Evans et al., 2016). Based on the idea of the productivity of scientific "paradigms" inherited from the work of Thomas Kuhn (Kuhn, 1962), Evans et al. (2016) observe distinctions between disciplines based on

the extent of linguistic consensus and marginal innovation.

Our work fits within this line of research and extends it in novel ways. Similar to prior work, we use an information-theoretic notion of entropy and surprisal to model linguistic relationships (Hughes et al., 2012; Bochkarev et al., 2014; Fankhauser et al., 2014; Vilhena et al., 2014; Evans et al., 2016; Degaetano-Ortlieb, 2018; Degaetano-Ortlieb and Teich, 2018). The consideration of analyzing language change and the development of sublanguages from an information-theoretic perspective goes back to Harris (1991): in striving for successful communication, distinctive codes develop which facilitate communication – over time and within subgroups. However, where prior work has focused on relationships between disciplines or the evolution of individual disciplines with respect to notions of innovation or paradigmaticness, our interest is in developing a more general linguistic understanding of the process of scientization itself. Degaetano-Ortlieb and Teich (2016), e.g., have shown how scientific language and common language become increasingly distinct over time. In the same vain, we ask how disciplines evolve with respect to common language (extra-scientific meaning) and with respect to their own language in terms of specialization and standardization (intra-scientific meaning). Thus, adopting their methodology, we similarly add a further dimension to theories of scientific consensus-building, while also working on developing a theory of scientization more generally.

Finally, our work is important because all of the above mentioned quantitative work has focused on the natural and social sciences rather than the humanities. There is a paucity of large-scale understanding about the behavior of fields like literary studies. Given the commitment to a particular world-view as a means of disciplinary self-understanding and given the larger institutional importance of the field, it is vital that more empirical evidence is provided to justify, refute, or nuance beliefs about the field. We see our work and the data set we are introducing as initiating the means to do so.

3 Methodology

3.1 Data

Literary Research Article Corpus (LRA) The LRA corpus consists of 63,397 articles published between 1950 and 2010 drawn from 60 academic journals with approx. 285 million tokens. The data is provided by the JSTOR Data for Research platform which provides metadata and ngrams using their own methods of parsing and cleaning. Journals represent different dimensions of the discipline, including leading generalist journals (PMLA, New Literary History, Critical Inquiry, MLN), genre or period-specific journals (Studies in Romanticism, Studies in the Novel, Shakespeare Quarterly, Science Fiction Studies), language- or culture-specific journals (Yale French Studies, New German Critique, African American Review, Journal of Arabic Literature), as well as more theoretically oriented journals (boundary 2, Social Text, Transition).

Royal Society Corpus (RSC) The RSC corpus consists of journal publications of the Proceedings and Transactions of the Royal Society of London, the first and longest-running English periodical of scientific writing (Kermes et al., 2016). The full version of the RSC spans from 1665 to 1996 amounting at approx. 300 million tokens. Here, we only use texts from 1950 to 1996, containing approx. 170 million tokens, to match the LRA corpus. Metadata of the RSC contain text type (article, abstract), author, title, date of publication, and time periods (decades and fifty years). The corpus provides linguistic annotation at the level of tokens (with normalized and original forms), lemmas, and parts of speech using TreeTagger (Schmid, 1995). The current release of the RSC (version 4.0) is freely available as a vertical text format (vrt) on the CLARIN-D repository[1].

Corpus of Historical American English (COHA) The COHA corpus is the largest structured corpus of historical English spanning from the 1810s to the 2000s. It contains more than 400 million words of text in more than 100,000 individual texts, balanced by genre across decades. It covers the major genres of fiction, magazine, newspaper and non-fiction. A detailed description of each genre and genre size is available at https://corpus.byu.edu/coha/. Fiction

[1] https://fedora.clarin-d.uni-saarland.de/rsc

is the largest genre with 48-55% of the total in each decade, followed by magazine with around 23-30%, news with 11-15% and non-fiction with 11-13%. We use the COHA corpus to represent standard English.

3.2 Methods

Our methodology is based on two information-theoretic measures. First, to investigate how much LRAs diverge from standard English and scientific language and to investigate specialization processes (H1 and H2) we use *Kullback-Leibler Divergence* (KLD; cf. Kullback and Leibler (1951)). Second, for the analysis of diachronic trends of standardization (H3) we use *Surprisal* to calculate the amount of information linguistic units transmit in text.

3.3 Divergence

Kullback-Leibler Divergence is an asymmetric measure of divergence calculating the additional bits of information needed between two models A and B:

$$D(A||B) = \sum_i p(item_i|A) log_2 \frac{p(item_i|A)}{p(item_i|B)} \quad (1)$$

Here, $p(item_i|A)$ is the probability of the ith item (in our case a word) in corpus A and $p(item_i|B)$ of that item in corpus B. Thus, divergence D between A and B, $D(A||B)$, is the sum of the probabilities of all items in A by the log_2 probability of the item in A divided by the probability of the item in B. This allows us to measure the amount of additional bits needed to encode words distributed according to a corpus A by the words' distribution in corpus B. The higher the amounts of bits, the more the two corpora diverge according to the probability distributions of their words. Difference in vocabulary size is controlled for by using ngram language models with Jelinek-Mercer smoothing (lambda at 0.05; cf. Zhai and Lafferty (2004); Fankhauser et al. (2014)). In our case, we compare *language* models between the language of literary research articles (LRAs), standard English, and scientific language.

For the investigation of H1 (LRAs vs. standard English and scientific language), we build yearly models and compare each year model across LRAs, standard English and scientific language, determining the degree of divergence between the models. The models are based on a vocabulary of 3,000 top occurring words of each corpus (LRA, COHA, RSC), excluding punctuation, stop words, and words shorter than three characters. The vocabulary lists are manually evaluated to ensure omission of possible noise in the data. For H2 (specialization of LRAs over time), we build KLD models on decades to investigate the degree of divergence of LRAs over time. Comparison is done between each decade (e.g. 1950 vs. 1960, 1950 vs. 1970, etc.). The inherent asymmetry of KLD allows us to inspect changes from past to present by $D(2000||1950)$, i.e. how well can the present be modeled by the past, and from present to past by $D(1950||2000)$, i.e. how well can the past be modeled by the present.

3.4 Surprisal

Surprisal is a measure of informativity and can be thought of as the amount of information a word transmits in a message (Shannon, 1948). In online-comprehension, surprisal is used to estimate how probable a unit (e.g. a word) is in a particular context (see Equation 2).

$$S(unit) = -log_2 p(unit|context) \quad (2)$$

Surprisal has two fundamental properties: (1) linguistic units with low probability convey more information than those with high probability, and (2) not only the unit itself but crucially the context in which a unit occurs determines the information a unit conveys. The intuition behind this is that linguistic units that are highly predictable in a given context convey less information than those that are less predictable and thus surprising (see Hale (2001); Levy (2008) for psycholinguistic accounts and Crocker et al. (2016) for surprisal and linguistic encoding across levels of linguistic representation (e.g. phonetic, psycholinguistic, discourse, register)).

We use surprisal to observe possible phrasal standardization of literary research articles over time (H3). As the LRA corpus comes in an ngram version (uni- to trigrams), we use surprisal on trigrams calculating surprisal of the last word, w_i, in the trigram based on its preceding context consisting of two previous words, w_{i-1} and w_{i-2} (a trigram model, see Equation 3).

$$S(w_i) = -log_2 p(w_i|w_{i-1}w_{i-2}) \quad (3)$$

Training is done on the COHA corpus, confining the data to span the same time period as the

LRA corpus (i.e. using texts from 1950 onwards), converting the corpus to lower-case and excluding sentence markers. In addition, we exclude from the training data sentences with a sequence of @ signs, which are part of COHA due to copyright. In addition we confine our selection of trigrams per document by matching the last word in a trigram with a dictionary consisting of the 3,000 most often occurring words in LRA, COHA and RSC plus function words. To test our hypothesis of phrasal standardization over time in LRA, we compare surprisal values of documents across years and decades. Assuming an increased phrasal standardization, the proportion of low surprisal per document will increase over time.

4 Analysis

In the analysis, we test our three hypotheses of scientization reflected in the process of social differentiation (H1, Section 4.1), diachronic specialization (H2, Section 4.2), and phrasal standardization (H3, Section 4.3).

4.1 Social Differentiation

As a humanistic discipline literary studies is often claimed to be more unique than other scientific disciplines (especially those from the 'hard' sciences) and to have a lower degree of scientificness. We thus hypothesize that literary studies should (1) diverge less from standard English than scientific disciplines and (2) diverge less from standard English than from scientific disciplines. To test this, we use three corpora: Literary Research articles (LRAs), COHA as a standard American English corpus to be comparable with LRAs, and the Royal Society Corpus (RSC) as a diachronic corpus of science. As a measure of divergence we use Kullback-Leibler Divergence D (see Section 3.2) comparing years between LRA vs. COHA, RSC vs. COHA, and LRA vs. COHA, assuming the following:

(1) LRAs will diverge less from standard English than scientific language from standard English: $D(lra||coha) < D(rsc||coha)$

(2) LRAs will diverge less from standard English than LRAs from scientific language: $D(lra||coha) < D(lra||rsc)$

For our first assumption, Figure 1 shows KLD over time from the 1950s to the early 2000s on a 5-year basis[2]. In general, LRAs diverge less from standard English than scientific language diverges from standard English, confirming our first assumption.

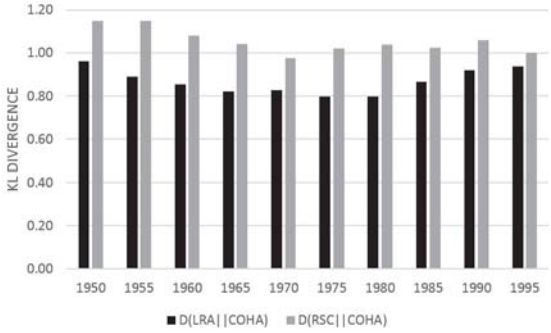


Figure 1: KLD over time for the comparisons of LRAs vs. COHA and RSC vs. COHA.

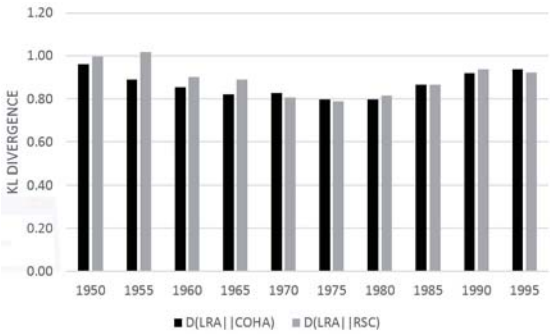
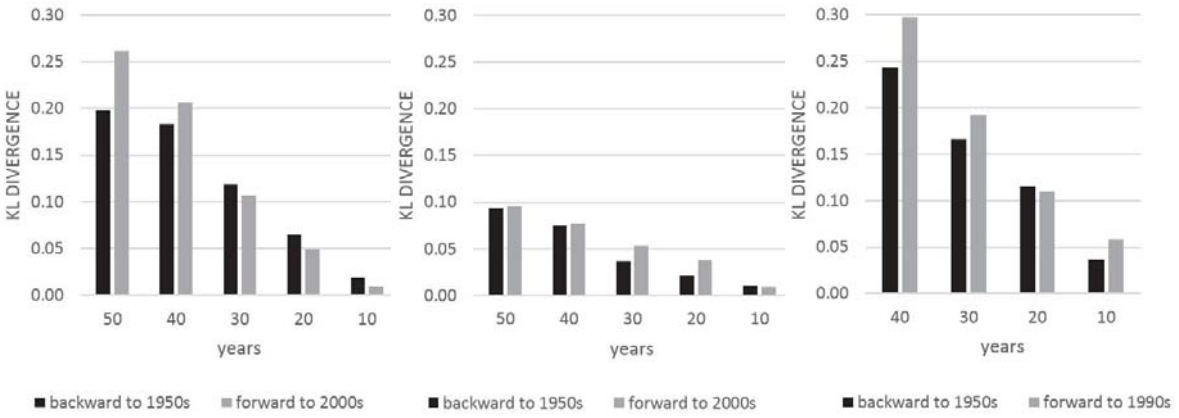


Figure 2: KLD over time for the comparisons of LRAs vs. COHA and LRAs vs. RSC.

Based on Figure 2, our second assumption is only partially confirmed: from 1950 until the mid-1970s, LRAs are indeed more similar to standard English than they are to scientific language. However, the diachronic trend is a decreasing one. After 1965, LRAs tend to be equally distinct from standard English and scientific language, with an increasing divergence from both over time (from approx. 0.8 to 0.9 bits). By contrast, divergence between scientific language and standard English during that period remains relatively stable (around 1.05 bits). Thus, in the 1950s and 1960s, LRAs seem to have a lower degree of scientificness, being more similar to standard En-

[2]Note that COHA is genre-balanced by decades only. Thus, a yearly representation would be strongly biased by the change in genre distribution in COHA across years. We have chosen to use a 5-year scale, as the distribution across genres is relatively stable. An inspection of our word lists does not suggest that the differences we are seeing are due to differences in British and American spelling.



(a) LRA corpus (b) COHA corpus (c) RSC corpus

Figure 3: KLD over time for LRA, COHA, and RSC. (KLD models are built for the 1950s in comparison to the other decades (e.g., 10 years: $D(1950||1960)$, 20 years: $D(1950||1970)$, etc.). The same applies for the 2000s and 1990s.)

glish than scientific language. The 1970s seem to mark a transition point, where LRAs equally diverge from both standard English and scientific language. From the 1980s onwards, LRAs increasingly diverge from standard and scientific English possibly undergoing a process of specialization as their language use diverges both from scientific language and from common language.

4.2 Specialization of LRAs

We inspect a possible process of specialization by considering divergence between different time periods of the LRA corpus. The evolution of disciplines is inherently accompanied by periods of lexical expansion due to new discoveries, which are paralleled by processes of terminology formation as well as periods of lexical consolidation (cf. Degaetano-Ortlieb and Teich (2018)). Thus, as a discipline evolves, its vocabulary typically changes over time. In information-theoretic terms this would imply, first, that a language model of an earlier time period will match a more contemporary time period less well and vice versa. Second, we expect this process to be gradual, where more adjacent time periods will diverge less from each other than periods that are further apart. Finally, while vocabulary changes over time, we expect that it will keep elements from the past while developing new terminology. If a process of specialization is at work, more contemporary articles will be modeled less well by earlier time periods than vice versa because the present will enclose the vocabulary of the past in ways that the past cannot enclose the present. Past and present become asymmetrically different from one another. Thus, for the LRA corpus, we hypothesize the following:

(1) LRAs of the 1950s will be better modeled by LRAs of the 2000s than vice versa, reflected in a lower divergence: $D(lra1950||lra2000) < D(lra2000||lra1950)$

(2) The closer the time periods, the lower their divergence: $D(lra1950||lra1960) < D(lra1950||lra1970)$

To test this, we build *forward* KLD models, i.e. models of the 2000s (or 1990s for the RSC) using past decades, e.g. $D(2000||1990)$, as well as *backward* models, i.e. models of the 1950s using future decades, e.g. $D(1950||1960)$. Figure 3a shows each model performance – the higher the KLD value the less well the models perform. As expected, the more adjacent the periods (e.g. only 10 years apart), the better the model in either direction, i.e. the forward model $D(1950||1960)$ performs quite well in modeling texts of 1950 when using 1960 texts (and vice versa). We also see our hypothesis about the assymetry in diachronic modeling confirmed, as the forward models show considerably higher divergence than the backward models for the longest time spans for both LRAs and the RSC (i.e. models 50 years apart).

A comparison to COHA (see Figure 3b) shows that the process of specialization (as defined here) does not adhere to standard English: KLD across comparisons is much lower than for LRAs, and the 50 year comparison $D(1950||2000)$ is almost equal to $D(2000||1950)$. In other words, we do not see the same directionality at work in general language use.

The growth in divergence over time and overall asymmetry between forward and backward models provide evidence to support our assumption of LRAs undergoing a process of specialization over time, similar to other disciplines (compare Figure 3a and 3c).

4.3 Standardization of Literary Research articles over time

At the level of linguistic phrases, we hypothesize a growth of phrasal standardization over time, i.e. a diachronic increase of standardized phrases in LRAs. While we have seen evidence above for the growing divergence from past linguistic practices in the field, our question here is whether there are higher levels of within-text standardization over time.

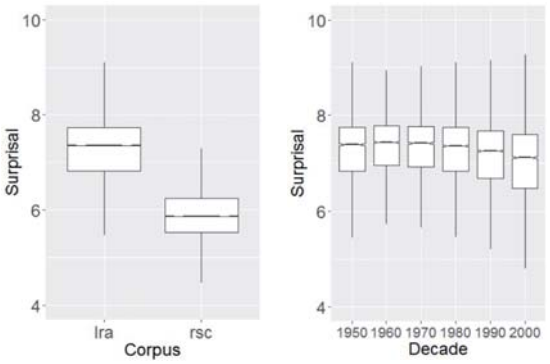
Surprisal is a well suited method for this kind of analysis, as it measures predictability of words in context. High predictability of words in phrases is reflected in low surprisal of these words and indicates standardized language use. To test this, we use a trigram version of the LRA corpus, approximating linguistic phrases by trigrams. We calculate surprisal of the last word in each trigram (see Section 3.4) to estimate predictability of possible phrases. In addition, we compare results to the RSC corpus to assess diachronic trends of standardization.

In Figure 4a, we see surprisal averaged by documents for the LRA and RSC corpora, showing significantly higher surprisal for LRAs (tested with a Wilcoxon rank sum test; p-value <2e-16). Inspecting the diachronic tendency of surprisal for LRAs, we can see how it significantly decreases over time, especially for the later time periods (see Figure 4b and Table 1). Thus, while LRAs use less standardized phrases than scientific language, over time surprisal of phrases in LRAs decreases, indicating an increase of standardized phrases.

	1950	1960	1970	1980	1990
1960	0.00019	-	-	-	-
1970	0.21622	0.00130	-	-	-
1980	0.04975	2.1e-12	3.0e-05	-	-
1990	1.9e-08	< 2e-16	< 2e-16	2.9e-07	-
2000	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16

Table 1: Pairwise comparisons of surprisal levels in LRAs by decade using Wilcoxon rank sum test and p-value adjustment with Benjamini-Hochberg method.

When inspecting the data more closely, we posit that a surprisal value <=0.5 bits appears to indicate standardized phrases in the LRA corpus (see first five examples in Table 2). These phrases transmit low informational content, indicated both by their surprisal value and their qualitative content. As we move up the surprisal scale, the information content transmitted appears to increase (compare *in order to* with *high cultural poetics*). This is in line with studies showing surprisal to be an indicator of processing effort, i.e. longer, low frequency words show higher surprisal, while shorter, high frequency words lower surprisal (cf. Hale (2001); Levy (2008)). In fact, phrases on the high surprisal end in Table 2 are lexical phrases (encompassing lower frequency words but high in information content), while phrases on the low surprisal end are grammatical phrases (encom-



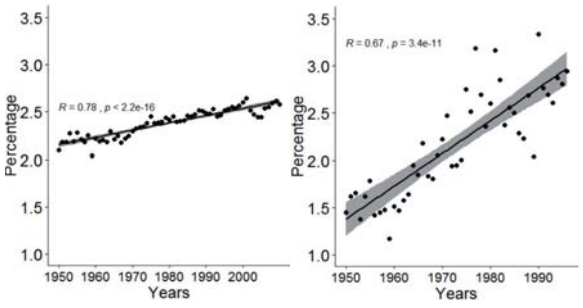
(a) LRA and RSC corpora (b) LRA corpus over time

Figure 4: Surprisal for LRA and RSC.

phrase	surprisal
on behalf of	0.0116
be able to	0.0144
the nineteenth century	0.1710
in order to	0.2934
been forced to	0.4128
writings from the	1.2075
elaboration of the	2.0679
he complained of	3.1327
have suggested the	4.0291
his works of	5.0548
posits women as	6.9722
full of hope	7.7751
wrote two novels	7.8494
movement protesting on	8.0463
starving child like	9.3617
eighteenth century rhetoric	17.9100
high cultural romanticism	18.7972
a democratic poem	19.0587
a critical anti	19.0712
high cultural poetics	21.4387

Table 2: Examples of phrases from very low to high surprisal (LRA corpus).

passing high frequency words with lower information content). If we consider only phrases that fall below our 0.5 threshold, i.e. highly standardized phrases, we see how their percentage grows over time (Figure 5a), though modestly when compared to the science corpus (Figure 5b). In other words, the LRA corpus indicates a similar process of standardization as the scientific corpus, but it does so less strongly. It lends support to the scientization hypothesis, that the field engages in more standardized language now than in the past, but also the differentiation theory, that LRAs are still less "scientific" than science articles.



(a) LRA corpus (b) RSC corpus

Figure 5: Percentage of standardized phrases (surprisal <=0.5 bits) over time.

5 Conclusion

We have investigated the evolution of literary studies with respect to two different kinds of language use: standard English on the one hand and scientific English on the other. In particular, we have tested three hypotheses with respect to a process of what we term scientization: social differentiation (H1, Section 4.1), diachronic specialization (H2, Section 4.2), and phrasal standardization (H3, Section 4.3). Methodologically, we used the information-theoretic measures of relative entropy (Kullback-Leibler Divergence) and surprisal. Kullback-Leibler Divergence is used to determine diverging trends between corpora/time periods. Surprisal is used to model the amount of information of words in context, providing us with a measure of phrasal standardization (the lower the surprisal, the more standardized a phrase is).

Doing so has indicated for us a complex portrait of the field, offering evidence to support two competing theories of disciplinary identity. On the one hand, we see evidence to support the idea that literary studies has indeed undergone a process of "scientization", which we define as the increased divergence from standard English, the increased divergence from past linguistic practices, and the increased use of standardized phrases. On the other hand, we see evidence to suggest that literary studies continues to occupy a middle-ground between science and common language. Literary research articles have remained consistently more similar to standard English than scientific articles, though the level of the difference of divergence has declined over time. Similarly, the divergence with past practices is considerably higher in LRAs than in standard English though somewhat lower than scientific articles. Language from the most recent decade is less well modeled by language from the past than the other way around, suggesting the emergence of field-specific vocabulary, even if not quite as strongly as in the RSC corpus. Finally, we see the uptick of standardized phrases, though once again with less overall strength than scientific articles.

These insights are important benchmarks for understanding the position of literary studies within the larger space of academic disciplines. They challenge the idea of literary studies' absolute distinctiveness from other disciplines and suggest that the field is gradually moving closer to the linguistic behavior of scientific domains. On the other hand, they indicate that this process is potentially not as distinctive for literary studies, as the field still maintains a closer approximation to common language than scientific fields. It suggests that one of the distinctive identities of literary studies might be its ability to mediate between scientific language practices on the one hand and common language practices on the other.

Our study could be expanded in various ways. Our collection of LRAs is limited to an Anglo-Saxon context and thus cannot account for disciplinary practices specific to other national contexts. Exploring further national frameworks within the discipline would reveal useful points of comparison. Second, as the title of our collection indicates, our results are only valid for articles, not monographs. While monographs play an important role in the field, articles are an equally central genre of scholarly discourse within literary studies. It would indeed be of interest to learn whether monographs behave differently with respect to the linguistic practices we uncover here. In terms of our language models used, one could test whether a broader vocabulary or the integration of function

words and punctuation could lead to more insights on changing practices of grammatical consolidation (see e.g. Rubino et al. (2016); Degaetano-Ortlieb and Teich (2018)). And while we capture semantic context using trigrams, one could explore the effect of using word embeddings that capture broader contextual windows.

Finally, it is also important to point out that our definition of scientization does not encapsulate the full range of practices that belong to the linguistic or methodological behavior of academic disciplines. Citation practices and evidentiary norms are two obvious ways that disciplines communicate knowledge that are not captured by our models. It could be that these practices follow our trends or diverge in telling ways. Future research will have to decide. Similarly, our models cannot explain what is driving this process of scientization, which we see as the subject of future work. What mechanisms are at work that contribute to these movements toward scientization, such as editorial behavior of journals, administrative pressures of institutions, or demographic changes in the profession? Are different effects occurring at different points in time? While we cannot yet answer these questions they are essential for understanding the logic through which disciplines constitute themselves and produce new knowledge.

Acknowledgments

Funding for this project was provided by the Social Sciences and Humanities Research Council of Canada and by the German Research Foundation (Deutsche Forschungsgemeinschaft) under the grant SFB1102: Information Density and Linguistic Encoding (www.sfb1102.uni-saarland.de). We are also indebted to Stefan Fischer for support in corpus processing and Elke Teich for her comments on a previous version of this paper. Also, we thank the anonymous reviewers for their constructive and valuable comments.

References

Pierre Azoulay, Toby Stuart, and Yanbo Wang. 2014. Matthew: Effect or Fable? *Management Science*, 60(1):92–109.

Douglas Biber and Bethany Gray. 2010. Challenging Stereotypes about Academic Writing: Complexity, Elaboration, Explicitness. *Journal of English for Academic Purposes*, 9:2–20.

Douglas Biber and Bethany Gray. 2011. The Historical Shift of Scientific Academic Prose in English towards Less Explicit Styles of Expression: Writing without Verbs. In Vijay Bathia, Purificación Sánchez, and Pascual Pérez-Paredes, editors, *Researching Specialized Languages*, pages 11–24. John Benjamins, Amsterdam.

Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English Language. Cambridge University Press, Cambridge, UK.

Vladimir Bochkarev, Valery D. Solovyev, and Soren Wichmann. 2014. Universals versus Historical Contingencies in Lexical Evolution. *Journal of The Royal Society Interface*, 11(101).

David Chavalarias and Jean-Philippe Cointet. 2013. Phylomemetic Patterns in Science Evolution - The Rise and Fall of Scientific Fields. *PloS one*, 8(2):e54847.

Matthew W. Crocker, Vera Demberg, and Elke Teich. 2016. Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz*, 30(1):77–81.

Stefania Degaetano-Ortlieb. 2018. Stylistic Variation over 200 Years of Court Proceedings according to Gender and Social Class. In *Proceedings of the 2nd Workshop on Stylistic Variation at NAACL*, pages 1–10, New Orleans, USA. ACL.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2019. An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. In Carla Suhr, Terttu Nevalainen, and Irma Taavitsainen, editors, *From Data to Evidence in English Language Research*, Language and Computers, pages 258–281. Brill, Leiden.

Stefania Degaetano-Ortlieb and Elke Teich. 2016. Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In *Proceedings of the 10th LaTeCH Workshop*, pages 165–173, Berlin. ACL.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using Relative Entropy for Detection and Analysis of Periods of Diachronic Linguistic Change. In *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at COLING2018*, pages 22–33, Santa Fe, NM, USA. ACL.

Eliza D. Evans, Charles J. Gomez, and Daniel A. McFarland. 2016. Measuring Paradigmaticness of Disciplines Using Text. *Sociological Science*, 3(32):757–778.

James A. Evans and Jacob G. Foster. 2011. Metaknowledge. *Science*, 331(6018):721–725.

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and Visualizing Variation in Language Resources. In *Proceedings of the 9th LREC*, pages 4125–4128, Reykjavik. ELRA.

Jürgen Fohrmann and Wilhelm Vosskamp, editors. 1991. *Wissenschaft und Nation: Studien zur Entstehungsgeschichte der Deutschen Literaturwissenschaft*. Fink.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. 2018. Science of Science. *Science*, 359(6379).

Jacob G. Foster, Andrey Rzhetsky, and James A. Evans. 2015. Tradition and Innovation in Scientists Research Strategies. *American Sociological Review*, 80(5):875–908.

Andrew Goldstone and Ted Underwood. 2014. The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History*, 45(3):359–384.

Gerald Graff, editor. 2007. *Professing Literature: An Institutional History*, twentieth anniversary edition. University of Chicago Press, Chicago.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pages 1–8. Association for Computational Linguistics.

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

M.A.K. Halliday. 2006. *Language of Science*, volume 5. Bloomsbury Publishing, Continuum, London.

Zellig Harris. 1991. *A Theory of Language and Information. A Mathematical Approach*. Clarendon Pess, Oxford.

Zellig S. Harris. 2002. The Structure of Science Information. *Journal of Biomedical Informatics*, 35(4):215 – 221.

James M. Hughes, Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore. 2012. Quantitative Patterns of Stylistic Influence in the Evolution of Literature. *Proceedings of the National Academy of Sciences*, 109(20):7682–7686.

George A. Kennedy, editor. 1989. *The Cambridge History of Literary Criticism*. Cambridge University Press, Cambridge.

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

Jonathan Kramnick. 2018. *Paper Minds: Literature and the Ecology of Consciousness*. University of Chicago Press, Chicago.

Thomas S Kuhn. 1962. *The Structure of Scientific Revolutions*, 3rd edition. University of Chicago Press, Chicago.

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Michèle Lamont, editor. 2009. *How Professors Think: Inside the Curious World of Academic Judgment*. Harvard University Press, Cambridge.

R. Levy. 2008. Expectation-based Syntactic Comprehension. *Cognition*, 106(3):1126–1177.

Martha C. Nussbaum, editor. 1997. *Cultivating Humanity: A Classical Defense of Reform in Liberal Education*. Harvard University Press, Cambridge.

Leighton Durham Reynolds and Nigel Guy Wilson. 1991. *Scribes and Scholars - A Guide to the Transmission of Greek and Latin Literature*. Oxford University Press, Oxford.

H.P. Rickman, editor. 1976. *W. Dilthey Selected Writings*. Cambridge University Press, Cambridge.

Raphael Rubino, Stefania Degaetano-Ortlieb, Elke Teich, and Joseph van Genabith. 2016. Modeling Diachronic Change in Scientific Writing with Information Density. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 750–761, Osaka, Japan. ACL.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Kyoto, Japan.

Claude E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Tian Shi. 2004. Ecological Economics as a Policy Science: Rhetoric or Commitment towards an Improved Decision-making Process on Sustainability. *Ecological Economics*, 48(1):23–36.

Elke Teich, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2016. The Linguistic Construal of Disciplinarity: A Data Mining Approach Using Register Features. *Journal of the Association for Information Science and Technology (JASIST)*, 67(7):1668–1678.

Evelyn B. Tribble, editor. 1993. *Margins and Marginality: The Printed Page in Early Modern England*. University Press of Virginia, Charlottesville.

Jean Ure. 1982. Introduction: Approaches to the Study of Register Range. *International Journal of the Sociology of Language*, 35:5–23.

Daril A. Vilhena, Jacob G. Foster, Martin Rosvall, Jevin D. West, James Evans, and Carl T. Bergstrom. 2014. Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication. *Sociological Science*, 1:221.

Chad Wellmon. 2017. Loyal Workers and Distinguished Scholars: Big Humanities and the Ethics of Knowledge. *Modern Intellectual History*, pages 1–39.

Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, 22(2):179–214.