NBER WORKING PAPER SERIES

LINGUISTIC METRICS FOR PATENT DISCLOSURE: EVIDENCE FROM UNIVERSITY VERSUS CORPORATE PATENTS

Nancy Kong
Uwe Dulleck
Adam B. Jaffe
Shupeng Sun
Sowmya Vajjala

Working Paper 27803
http://www.nber.org/papers/w27803

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2020

This research uses data from the Lens (https://www.lens.org/) and we are grateful to Richard Jefferson and Aaron Ballagh for their constructive comments and data support. For thoughtful discussions, we thank Lesley Millar-Nicholson (Director at Technology Licensing Office, MIT), and Timothy Oyer (President at Wolf Greenfield Intellectual Property Law), as well as seminar participants at Queensland University of Technology. We appreciate Hamish Macintosh and Yi Wang for their technical support. This research is funded by Australian Research Council Discovery Grant DP180103856. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Linguistic Metrics for Patent Disclosure: Evidence from University Versus Corporate Patents
Nancy Kong, Uwe Dulleck, Adam B. Jaffe, Shupeng Sun, and Sowmya Vajjala
NBER Working Paper No. 27803
September 2020
JEL No. O31,O34

ABSTRACT

This paper proposes a novel approach to measure disclosure in patent applications using algorithms from computational linguistics. Borrowing methods from the literature on second language acquisition, we analyze core linguistic features of 40,949 U.S. applications in three patent categories related to nanotechnology, batteries, and electricity from 2000 to 2019. Relying on the expectation that universities have more incentives to disclose their inventions than corporations for either incentive reasons or for different source documents that patent attorneys can draw on, we confirm the relevance and usefulness of the linguistic measures by showing that university patents are more readable. Combining the multiple measures using principal component analysis, we find that the gap in disclosure is 0.4 SD, with a wider gap between top applicants. Our results do not change after accounting for the heterogeneity of inventions by controlling for cited-patent fixed effects. We also explore whether one pathway by which corporate patents become less readable is use of multiple examples to mask the “best mode” of inventions. By confirming that computational linguistic measures are useful indicators of readability of patents, we suggest that the disclosure function of patents can be explored empirically in a way that has not previously been feasible.

Nancy Kong
Queensland University of Technology (QUT)
nancy.kong@qut.edu.au

Uwe Dulleck
Queensland University of Technology (QUT)
2 George St
Brisbane, QLD 4000
Australia
uwe.dulleck@qut.edu.au

Adam B. Jaffe
188 Brookline Avenue
Apartment 26A
Boston, MA 02215
and Brandeis University
and Queensland University of Technology
and also NBER
adam.jaffe@motu.org.nz

Shupeng Sun
Queensland Treasury
1 George St
Brisbane, QLD 4000
Australia
shupeng.sun@treasury.qld.gov.au

Sowmya Vajjala
National Research Council
1200 Montreal Road
Ottawa, Ontario K1K 2E1
Canada
sowmya.vajjala@nrc-cnrc.gc.ca

1 Introduction

The patent system serves two purposes: "encouraging new inventions" and "adding knowledge to the public domain."[1] The former incentivizes creation, development, and commercialization by protecting inventors' exclusive ownership for a limited period of time. The latter encourages disclosure of new technologies by requiring "full, clear, concise, and exact terms" in describing inventions.[2] Sufficient disclosure in patents has three major benefits: (1) fostering later inventions (Jaffe and Trajtenberg, 2002; Scotchmer and Green, 1990; Denicolò and Franzoni, 2003); (2) reducing waste resources wasted on duplicate inventions; and (3) inducing more informed investment in innovation (Roin, 2005).

Despite a large body of literature on the patent incentivizing function (Cornelli and Schankerman, 1999; Kitch, 1977; Tauman and Weng, 2012; Cohen et al., 2002), patent disclosure has not been studied systematically. This raises concern; as Roin (2005), Devlin (2009), Sampat (2018), Arinas (2012) and Ouellette (2011) document, the technical information contained in patent documents is often inadequate and unclear. To date, little empirical research has been conducted on patent disclosure. Important questions, such as how to measure disclosure, potential incentives behind disclosure, heterogeneous levels of disclosure by entities, and the tactic of avoiding the disclosure requirement, have not been directly investigated. A major barrier to such empirical research has been the lack of broadly applicable, reproducible quantitative measures of the extent of disclosure. We propose that extant metrics developed in computational linguistics can fill this gap.

In using computational linguistic metrics to compare the readability of documents, we follow researchers in the finance and accounting literature, who have used readability metrics to gauge whether readers are able to extract information efficiently from financial reports (Li, 2008; Miller, 2010; You and Zhang, 2009; Lawrence, 2013). This

[1] See Eldred v. Ashcroft, 537 U.S. 186, 226-27 (2003) and Pfaff v. Wells Elecs., Inc., 525 U.S. 55, 63 (1998).

[2] See 35 U.S.C. §112 (2000).

literature posits that more complex texts increase the information processing cost for investors (Grossman and Stiglitz, 1980; Bloomfield, 2002) and finds, for example, that companies are likely to hide negative performance in complicated text to obfuscate that information (You and Zhang, 2009).

Although patent applications differ from corporate annual reports, the research question regarding strategic obfuscation is similar: Documents are created subject to regulation, in which the purpose of the regulation is to compel disclosure, but the party completing the document may have incentives to obscure information. We propose that the linguistic measures we use are likely to serve as an informative proxy for the explicitly or implicitly chosen level of disclosure. The goal of this paper is simply to demonstrate that these measures do appear to capture meaningful differences in disclosure, thereby opening up the possibility of research into the causes and effects of variations in disclosure.

Our strategy for demonstrating the relevance of the linguistic readability metrics is to identify a situation where we have a strong a priori expectation of a systematic difference in disclosure across two groups of patents. If the proposed metrics show the expected difference, we propose to treat them as potentially useful. We use university and corporate patents to exploit the expectation that universities disclose more information in patent documents (Trajtenberg et al., 1997; Henderson et al., 1998). Universities and corporations follow different business models for patenting: technology transfer versus in-house commercialization. Patents applied for by universities, with a focus on generating income from the licensing of inventions, should have a higher level of disclosure because transparent information makes it easier to signal the technology contained in the patent and attract potential investors. As a result, they are more readable than corporate patents. The readability difference could be further magnified by the moral requirements of university research as well as the rigor of academic writing.

Corporations, on the other hand, focus on in-house production, and are therefore have a greater incentive to obfuscate crucial technical information to deter competitors from understanding, using, and building on their inventions. The profit-maximization motive, as well as a lack of incentive to document the invention thoroughly could also contribute to the low level of disclosure. Together, it is reasonable to assume that universities may strategically (or unconsciously) choose a higher disclosure level in patent applications than corporations. We emphasize that we do not see this analysis as testing the hypothesis that universities engage in more disclosure than firms. Rather, we take this as a maintained hypothesis and show—conditional on that maintained hypothesis—that the linguistic measures meaningfully capture differences in disclosure across patents.

Similar to finance literature, we use a computational linguistic program designed to assess reading difficulty of texts using 64 measures from second language acquisition research. The indicators cover lexical, syntactic, and discourse aspects of language along with traditional readability formulae. We apply them to a full set of U.S. patent application texts in three cutting-edge industries from the past 20 years. Our baseline OLS estimations reveal significant differences between university and corporate patents.

Using principal component analysis (PCA) to combine the 64 indicators and create synthetic readability measures, we show that composite indices detect strong differences between university and corporate patents, which lends support to the validity of our measures.

The key empirical challenge is that the nature of corporate and university inventions might differ, and thus the textual communication required for corporate inventions could differ. To address this concern, our identification strategy employs cited-patent fixed effects; this assumes that university and corporate patents that cite the same previous patents build on the same prior knowledge, and are therefore likely to be technologically similar inventions.

Statistical analysis of cited-patent fixed effects requires that any given patent can be in more than one fixed-effect group. We use high-degree fixed effects estimations to overcome this challenge and employ a data compression technique, least absolute shrinkage and selection operator (LASSO), to handle the large number of fixed-effect groups.

Our results show that corporate patents are 0.4 SD more difficult to read and require 1.2 years more education to comprehend than university patents. We find that the difference is more prominent in more experienced patent applicants, which we believe supports the idea that the differences in readability are at least somewhat intentional. We also show that a potential channel for obfuscation is to provide many examples in order to conceal the "best mode" of inventions.

Our main contributions are the following. First, we construct a set of measures as a proxy for otherwise unobserved disclosure levels in patent documents, and apply this approach to university and corporate patents. Our results confirm that our proxies capture legitimate differences between university and corporate patent documents, and prove our hypothesis that university patents have a higher disclosure level.

Second, this paper is the first study to apply textual analysis of patent applications on a large scale. We obtain the whole set of full text patent applications in categories related to nanotechnology, batteries, and electricity from 2000 to 2019, totalling 40,949, and apply our linguistic analysis model to the technical descriptions of these patents.

Third, we expand readability studies in related literature which relied heavily on traditional readability indices, such as Gunning Fog, Kincaid, and Flesch Reading Ease, to include lexical richness, syntactic complexity, and discourse features. We use the best noncommercial readability software (Vajjala and Meurers, 2014b) to capture the multidimensional linguistic features of 64 indicators, and perform a much more in-depth linguistic analysis (Loughran and McDonald, 2016) than previous studies. Employing Romano and Wolf's (2005) stepdown multiple hypothesis testing, we show that corporate patents vary significantly on readability in 38 of the indicators.

Fourth, to handle the complicated data structure, our statistical analysis employs big data techniques such as PCA and LASSO. We show that our results are robust across different specifications.

Fifth, this is the first study that documents empirically different levels of patent disclosure across business entities, and extends the "incomplete revelation hypothesis"

(Bloomfield, 2002; Schrand and Walther, 2000) from financial reports to patent applications.

The rest of the paper proceeds as follows. Section 3 explains the linguistic measures used in the study. In Section 2, we review the relevant literature and lay out our hypothesis of differences in disclosure between university and corporate patent applications. Section 4 presents our data and baseline estimation, followed by our main results in Section 5. Synthetic measures are proposed in Section 6. We examine the cited-patent fixed effects in Section 7 and one channel that corporations could use to obscure patent application in Section 8. We show heterogeneous effects in Section 9 and conclude in Section 10.

2 Literature review

2.1 Textual analysis

Textual analysis is introduced to the economic literature only in the last few years. A few studies employ computational linguistic analysis. For example, Gentzkow et al. (2019) propose a practical overview of textual analysis and statistical analysis using text as data. Hansen et al. (2018) examine the effects of transparency in central bank on monetary policies using a statistical model for content analysis. A limited number of studies employ textual analysis to examine gender discrimination in the publication and job market process. For instance, Hengel (2017) examines articles in the peer-review process using traditional linguistic measures such as the Flesch Reading Ease, Flesch-Kincaid, and Gunning Fog indices and finds that female-authored papers are 1%–6% better written, but tougher editorial standards and/or biased referee assignment are consistently used for female researchers. Card et al. (2020) use Gunning Fog and the Coleman-Liau index to examine whether the complexity of the abstract is gender-dependent. They find that female-authored papers receive about 25% more citations, but there is a 7 percentage points lower probability of a revise and resubmit verdict for female-authored papers. Wu (2018) uses text scraped from Economics Job Market Rumors Forum and apply a LASSO-logistic model to extract the words with the strongest predictive power for each gender. Wu finds that comments about female academics are mainly concern physical appearances and posts about male academics are more relevant to their academic abilities.

Computational linguistics has not been used much in research on patents with a few exceptions. For example, Younge and Kuhn (2016) and Arts et al. (2018) use textual analysis to examine patent similarity.[3] De Clercq et al. (2019) use natural language processing tools on electric vehicle patent information extraction and dynamic visualization. To examine which type of invention ("new idea-based" or "old idea-based") is more likely to stimulate follow-up innovation, Packalen and Bhattacharya (2015) investigate words and word sequences related to a certain technical term as the concept, and count the number of patents that use these concepts. They find that inventions based on new ideas are more likely to stimulate follow-up inventions than those based on old ideas. Kelly et al. (2018) employ similar methods to measure the novelty of patented inventions by searching for new words. However, these studies focus only on the technologies that patents contain, and linguistic methods are used to extract technical terms rather than measure the disclosure level.

The use of readability measures in accounting and finance provides us with a precedent for our own use of readability measures with patent documents. Loughran and McDonald (2016) show that the readability of financial documents determines whether readers can reasonably extract the information. Other studies show that the readability of financial reports (usually annual or 10-K reports) may affect investors' behavior, or be affected by the firm's performance (Li, 2008; Miller, 2010; You and Zhang, 2009; Lawrence, 2013). We therefore base our study on previous finance literature, but expand it to patent documents and apply a series of computational linguistic measures as proxies for disclosure.

[3] See Teodorescu (2017) for a comprehensive survey on natural language processing method used in strategic research.

2.2 Patent disclosure

It is a legal requirement that an adequate description of the invention be stated in the patent application. According to 35 U.S. Code §112, the patent specification "shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same, and shall set forth the best mode contemplated by the inventor or joint inventors of carrying out the invention." In addition to the US, the European Patent Convention and the World Trade Organization also have similar requirements.[4] That is, the technical description must meet the requirements for (1) written description, (2) enablement, and (3) best mode.

Despite the legal obligations, several studies document a lack of transparency in patent documents; however most of these claims are based on anecdotal evidence. For example, Devlin (2009) uses several references to support his argument without direct evidence. Roin (2005) does not have empirical results to support his claims, which is based on a speech by the CEO of a consulting firm specializing in intellectual property licensing. Cohen et al. (2002) provide evidence on the disclosure function of patents based on surveys, in which inventors are asked whether they prefer patents or other sources to obtain technical information in the U.S. and Japan, and the results show that patents play a greater role in knowledge diffusion in Japan. Walsh and Nagaoka (2009) also survey patentees in the two countries and find that patent literature is more important as an information source for Japanese firms than American firms. Ouellette (2011) studies how patents can be used as a source of technical information in the nanotechnology industry, and finds that 70% of respondents are looking for useful technical information. This indicates that the patent system is an important channel for obtaining information by researchers.

These studies are mostly based on small-scale surveys in a specific field. In con-

[4] See Article 29 of the Agreement on Trade-Related Aspects of Intellectual Property Rights, and Article 83 of the European Patent Convention.

trast, the initial drafting process for patents from inventors' perspective has not been examined. Our paper aims to close this gap by directly examining large-scale multi-discipline patent texts.

2.3 University and corporate patents

We choose to compare patents filed by universities and corporations because they have different business models for patenting. Universities' main purposes are teaching and research, and the dominant business model for university technology transfer is licensing patents (Valdivia, 2013). In order to attract potential investors, universities would describe their inventions more clearly, in relative terms, because this can signal the technical information contained in patents and facilitate technology transfers.

In contrast, corporations typically seek to self-commercialize their R&D results and maximize profits. They are likely to regard patent disclosure as "a limitation on the monopoly power" of their inventions (Landes and Posner, 2009). Baker and Mezzetti (2005) show that in reality, corporations may only disclose technical information for defensive purposes; for example, by disclosing some key information to the public (i.e., enlarge the prior art) to make it more difficult for competitors to apply for patents in a related area. Therefore, we propose that corporations are more reluctant to clearly disclose technical information compared with universities,

Several additional aspects of the institutional environment reinforce the underlying difference between universities and corporations in disclosure incentives. First, the 1980 Bayh-Dole Act created the legal framework in the U.S. for universities to own patents on publicly funded research (which includes almost all of their patents). It explicitly renders the realization of the economic and social benefits of the invention a goal of the law and enables universities to foster the diffusion of their patents (Henderson et al., 1998).

Second, universities' fundamental purpose is to promote knowledge flows. In 2007, a group of universities, including Caltech, Stanford, MIT, and Harvard, signed a statement in which they promised to be mindful of public interest and declared that "exclusive licenses should be structured in a manner that encourages technology development and use."[5]

Finally, the process of patent drafting frequently differs for university patents. In many cases, the patent is drafted on the basis of a scientific paper, which is written to communicate the results, and may have been subject to review and editing designed to increase its readability. Corporate patents are typically drafted based on a disclosure written by the inventors. The availability of a previously written scholarly paper may provide a base for patent drafting that intrinsically leads to greater readability.

3 Linguistic measures

We use the readability assessment program developed by Vajjala and Meurers (2014b), which is shown to be the best non-commercial readability assessment approach for English (Vajjala and Meurers, 2014b) and is demonstrated to be useful in other experimental settings (Vajjala and Meurers, 2012, 2013, 2014a). We use 64 measures from this program, which were previously used in text readability research. In addition to traditionally used measures, such as Gunning Fog and Flesh-Kincaid grade level, the rest of the measures from this program are divided into three categories: lexical features, syntactic features, and discourse features (see Figure 1, the hierarchy of linguistic analysis). This classification is a customized combination of those of Loughran and McDonald (2016) and Collins-Thompson (2014) that is relevant to patent documents.[6] The caveat is that lexical, syntactic, and discourse measures have not been tested on patent documents, and thus we report the differences for those, but only interpret traditional measures in the direction of readability.

Table I presents the definitions, interpretation, implications and sources of rep-

[5] See https://otl.stanford.edu/documents/whitepaper-10.pdf.

[6] Loughran and McDonald (2016) propose the following hierarchy of analysis: lexical, collocation, syntactic, semantic, pragmatic, and discourse. Since semantics, pragmatics are both in general open problems in the computational modeling of language, we don't have software that can extract such features yet.

resentative variables in each category. These variables are chosen according to their high frequency of use in the literature, and because they are easily understood by non-linguists. For example, traditional measures, such as Gunning Fog and Flesch Reading ease scores, are the most widely used readability measures. *Fog*, or Gunning Fog, combines average sentence length in words and the ratio of words with more than three syllables to all words. It describes how many years of formal education are needed to understand the text on first reading. *Kincaid*, or Flesch-Kincaid, combines the average word length in syllables and average sentence length. The result is a number that corresponds with a U.S. grade level. *Flesch*, or the Flesch Reading Ease score, combines average word length and average sentence length, ranging from 0 to 100. Unlike Fog and Kincaid, a low score is associated with a "hard to read" text.

The lexical features describe word complexity and diversity and examine the building blocks of readability. We use the average age of acquisition of words (*AoA*) from the language acquisition literature, and the word type-token ratio (*TTR*), which is the ratio of unique words to total words, to represent the lexical feature.

The syntactic features focus on the structure of sentences, such as the average length of various syntactic units, number of phrases of various categories, and the average length of phrases. We use *dependent clauses to total clause ratio* and mean length of T-unit (*MLT*),[7] as the representative measures for this category.

The discourse features examine textual cohesion. It refers to the process of linking the different parts of the text together to achieve overall coherence. One way to achieve this is by the use of appropriate connective words between sentences. We use referring expressions (Todirascu et al., 2013) and word overlap features implemented based on the Coh-Metrix tool (McNamara et al., 2002) for our analysis. In this category, the representative indicators are *ratio of proper nouns to nouns* and *global content word overlap* between all pairs of sentences as the representative measures.

[7] A T-Unit is the "shortest grammatically allowable sentences into which (writing can be split) or minimally terminable unit" (Hunt, 1965). It is linguistically defined as "one main clause plus any subordinate clause or nonclausal structure that is attached to or embedded in it" (Lu, 2010).

Table II presents the means, standard deviations, and *t*-statistics for universities and corporations. *Fog* shows that it takes 22.1 years of education to understand university patents, whereas for corporate patents, it takes 23.6. It also suggests that corporate patents have higher values for *AoA, dependent clauses ratio, content word overlaps*, and *MLT*, and lower values for *proper noun ratio* and *TTR*.

4 Empirical strategy

4.1 Data

Using the Lens database,[8] we obtain the full text of U.S. patent applications in three classes—Nanotechnology (977); Batteries: Thermoelectric and Photoelectric (136); and Electricity: Battery or Capacitor Charging or Discharging (320)—from January 1, 2000 to July 8, 2019. We choose these three research areas because both universities and corporations invest heavily in these fields; therefore, we can gather enough patent samples from these patent classes. We strip technical description text files, excluding headers and claims, from the full-text files, and obtain 40,949 patent applications. We also acquire patent metadata, such as application date, priority numbers, applicants, inventors, forward-citation counts, simple and extended family sizes, sequence count, NPL citation count, NPL resolved citation count, etc.

To identify universities and corporations, we manually researched the top 100 applicants to determine which were universities. On this basis, we identified text strings such as "univ," "inst," and "college," and then classified all applicants whose name contains these strings as universities. Similarly, corporations are identified as applicants containing strings such as "INC," "LTD," "CORP," "LLC," and "CO."

Our sample consists of 3,414 patent applications from universities, and 21,234 from corporations, 1,644 jointly filed by universities and corporations, and 14,657 filed by other entities, such as individuals and government organizations (see Table 10 in the

[8] The Lens is a public benefit project of the global non-profit Cambia. See `https://www.lens.org/` and Jefferson et al. (2018) for more information.

Appendix for detailed summary statistics).

We then apply Vajjala and Meurers' (2014c) computational linguistic model to the 40,949 full-text patent applications using a high performance computing platform,[9] and apply the 64 linguistic measures to each application.

Table II presents summary statistics for the metadata and linguistic measures of patent applications filed by universities and corporations. All nine representative indicators are significantly different for universities and corporations.[10] It also shows that all characteristics, such as citation counts and family size, are significantly different. We control for these observed differences between universities and corporations.

4.2 Baseline estimation

We estimate the following OLS regression:

$$Y_{ij} = \alpha + \beta_1 Corp_{ij} + \beta_2 Joint_{ij} + \beta_3 Other_{ij} + \boldsymbol{\lambda X_{ij}} + \delta_j + \varepsilon_{ij}, \qquad (1)$$

where Y_{ij} is one of the 64 linguistic indicators of application i in subclassification j; $Corp_{ij} = 1$ if the patent application is filed by a firm using Uni as the base; $\boldsymbol{X_{ij}}$ is a vector of forward-citation counts, simple and extended family sizes, sequence count, NPL citation count, and NPL resolved citation count; δ_j is U.S. patent subclassification fixed effects; and ε_{ij} is the error term clustered at the U.S. patent classification level.

The baseline estimation controls for forward citation counts, which is a strong indicator of patent quality. We also control for the 574 subclassification fixed effects, which in effect account for the area-specific competition.

The hypothesis is that β_1 is significant and positively correlated with "hard to read"

[9] We use a high performance computing platform at Queensland University of Technology that employs a heterogeneous cluster consisting of several different architectures of CPUs, GPUs, and node configurations. It uses PBSPro to schedule jobs on the cluster and SLES 12 for its operating system. The linguistic software is run parallel by the cluster.

[10] The summary statistics of 64 variables in the full sample include joint patents and other patents; see Table 10 in the Appendix.

indices compared with university patents.

5 Results

Table III presents estimates for corporate patents on representative individual linguistic measures, applying multiple hypothesis testing (Romano and Wolf, 2005) with step-down adjusted p-values enabling strong control of the familywise error rate. We show first the difference between the university and firm patents without controlling for patent attributes such as citations received, family size and non-patent literature citations made (see Panel A). It is not clear whether ideally one would compare the two groups with or without these controls. For example, it is possible that greater disclosure in fact facilitates subsequent citation, so controlling for citations received might inappropriately capture some of the underlying variation in disclosure. However, to be conservative in trying to ensure that we have controlled for patent differences other than readability, in our preferred specification we include these control variables (see Panel B). We show both the estimates using raw linguistic scores for magnitude interpretations and standardized linguistic scores for easy comparison across different measures.

The Fog and Kincaid measures both correspond with the years of education required to understand the text, their estimates are 1.5 and 1.7 without controls, and 1.4 and 1.6 with controls, respectively, which means that corporate patents require 1.4 to 1.6 more years of education to comprehend than university patents. Since the Flesch score is reversely correlated with "hard to read," the point estimates of -4.6 (without controls) and -4.3 (with controls) indicate harder to read texts for corporate patents.

On average, corporate patents have words with higher age of acquisition, more dependent clauses, longer t-units, fewer proper nouns, and more content word overlap. The standardized magnitudes for most linguistic measures are approximately 0.2 to 0.3 SD (except for *AoA* proper noun ratio).

We also present the full estimates in Table A.2 in Appendix. Figure 2 shows the 64 estimates by significance, and 38 linguistic indicators are significant. This means that the linguistic measures effectively capture the differences in patent applications between universities and corporations.

6 Synthetic indicators

To consolidate our results, we use PCA to combine the 64 linguistic measures. PCA is a nonparametric statistical technique primarily used to reduce dimensions. The mechanism is described as follows:

$$\arg\max_{w} \{||\boldsymbol{Yw}||^2\} \qquad \text{s.t. } \boldsymbol{w}^2 = 1,$$

where Y is a vector of the outcome variables and w is the weight assigned to Y. The PCA explores the highest variability in variables, and rotates the coordinates so that data points become orthogonal.

Figure 3 presents a scree plot of eigenvalues after PCA, and the largest distances between the first four components show that they are the most relevant (Onatski, 2010). We thus use those synthetic indicators as the dependent variables and re-estimate Equation 1.[11]

Table IV shows the estimates of corporate patents using components 1 to 4. For easy interpretation, all components are standardized. Component 1 shows that corporate patents are 0.43 SD different from their university counterparts. Components 2 to 4 indicate significant differences between corporate and university patents. For the rest of the paper, we will use component 1, which captures the most explanatory power of the linguistic indicators, as the PCA index.

Based on these results, it is clear that PCA captures a significant difference in the implicit structure of linguistic measures. This lends support to the underlying disclosure gap in patents from universities and corporations.

[11] We present components of the linguistic variables in Table 10 in the Appendix.

7 Cited-patent fixed effects

Are the differences we find are in fact driven by the different natures of corporate and university inventions? To address this concern, we propose that patent applications that cite the same previous patent would be somewhat similar inventions. If one is filed by a university and the other by a corporation, the difference more likely arises from the entity than the invention. Therefore, we ask whether the estimated differences between university and corporate patents change materially after controlling for the intrinsic nature of inventions.

The empirical challenges are that there are (1) multiple group identifiers for the citation fixed effects, which we address by using a high degree of fixed effects with each dummy variable to represent every previous patent cited; (2) a large number of dummy variables, which we address by using LASSO to perform variable selection and shrinkage (Tibshirani, 2011) to reduce the dimensionality of the right-hand-side variables. LASSO performs the following estimation:

$$min \sum_{i=1}^{N} (y_i - \sum x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

We limit the sample to patents that cite "highly cited patents" ($\geq$ 10 citations), and there are 20,571 patent applications with 6,163 fixed-effect groups.[12] We perform LASSO linear "post-double-selection" inference model (Belloni et al., 2014):

$$Y_{ijk} = \alpha + \beta_1 Corp_{ijk} + \beta_2 Both_{ijk} + \beta_3 Other_{ijk} + \boldsymbol{\lambda X_{ijk}} + \delta_j + \sum_{k=1}^{6163} \gamma_k + \varepsilon_{ijk},$$

where $\sum_{k=1}^{6163} \gamma_k$ is cited patent k fixed effects, and $Corp_{ijk}$, $Joint_{ijk}$, $Other_{ijk}$, $\mathbf{X}_{ijk}$ and δ_j are always included; LASSO chooses whether to include or exclude terms in $\sum_{k=1}^{6163} \gamma_k$.

[12] We also conduct a sensitivity analysis of different thresholds for "highly cited patents": more than 100 cites with 945 patents and 9 fixed effect groups, and 50 cites with 5,082 patents and 168 fixed effect groups. Those sensitivity tests are also done using high-degree fixed effects estimation without LASSO. The results are highly consistent.

Table V presents cited-patent fixed effect estimations. Synthetic, traditional, lexical, syntactic, and discourse features are all highly significant for corporate patents. Compared with the previous estimations, the PCA (component 1) shows a very similar magnitude (0.44 compared with 0.43), and the traditional measures have slightly reduced magnitudes relative to the baseline estimation: The Fog measure decreased from 1.4 to 1.1; the Kincaid from 1.6 to 1.3; and the Flesch from -4.3 to -3.1. This suggests a small proportion of the estimated difference is absorbed by the nature of inventions, but the estimates do not change materially: Corporate patent applications are still significantly different from university patents in readability, and require 1.1 to 1.6 more years of education to comprehend.

8 A Possible Channel

In this section, we explore a potential strategy that corporations may use that could partially explain the differences in readability and disclosure. As a matter of patent law, the so-called "best mode" rule specifies that if there are multiple different ways of implementing the patented technology, and one of these "modes"is known to be better than the others, this "best mode" must be disclosed. There is, however, no requirement that it be identified as such. This means that one way of minimizing disclosure is to bury the revelation of the best mode within a list of other (less effective or satisfactory) implementations of the invention. This means that long lists of examples may be evidence of obfuscation.

We extract *Num_examples*, the occurrences of "for examples" and "e.g.," in the patent document. The average number of examples in university patents is 24 and in corporate patents is 26; the difference is significant with $t\text{-}stats = -2.36$.

We add the *Num_examples* to Equation 1 as an independent variable. Table VI shows that the number of examples is positively correlated with the synthetic variable, sentence length, and content word overlap, and negatively correlated with Flesch and *TTR*. In general, *Num_examples* mostly correspond to hard to read. This lends support to our hypothesis that corporations and universities have different levels of disclosure, as evidenced by the number of examples, and is reflected in our linguistic measures.

9 Heterogeneous effects

Lastly, we test whether the gap between university and corporate patents is more prominent in more experienced applicants. We select the top 100 applicants in our sample. The number of patent applications filed by those applicants ranges from 51 to 835. Of the 40,949 applications, 11,844 are filed by the top applicants (10.1% are university patents), and the rest are in "other" category (7.4% are university patents).

We estimate Equation 1 separately for the top 100 applicants and the rest. Results are presented in Table VII. We find that across all measures (with the exception of *AoA*), the top 100 applicants have a significantly higher gap between universities and corporations. The PCA variable shows that corporate patents are 0.68 SD harder to read than university patents among top applicants, compared with 0.26 SD in other applicants. This means that top applicants have a 2.6 times higher difference relative to others. The Fog and Kincaid measures indicate that corporate patents require 2.2 to 2.4 more years of education to read than university patents for top applicants (compared with 0.9 and 1.1 for other applicants), which means that the readability gap is 2.4 times wider between top applicants than other applicants.

According to the Fog measures, we find that this widened gap arises from both the increased readability of top university applications (21.6 for top universities versus 22.3 for other universities), and the decreased readability for top corporate applications (24.0 top corporations versus 23.3 other corporations). In general, we would expect that firms or universities get better at achieving their own objectives (whatever those objectives may be) the more patents they have filed. Thus we interpret the modestly wider gap between universities and firms among the most experienced applicants to reinforce the interpretation that the measured differences are indicative of the different strategic objectives of universities and firms.

10 Conclusion

This paper proposes a novel approach that uses computational linguistic measures to study patent disclosure by examining large-scale patent text data, and combines high-degree statistical techniques. Based on the maintained hypothesis that universities and corporations have different business models for patenting inventions (Trajtenberg et al., 1997), and universities have incentives to disclose more in their patent documents (Henderson et al., 1998), we find evidence that our proposed measures capture significant differences in the applications' wording, sentence structure, and referential coherence. Compared with university patents, corporate patents require 1.1 to 1.6 more years of education to read using the Fog and Kincaid measures, and are 0.4 SD harder to comprehend using a composite index. We show that such a gap is 2.2 to 2.6 times larger between the top 100 applicants, which further supports our hypothesis that this difference may stem from a strategic motive whereby corporations intentionally obscure their inventions to deter competitors from adopting the innovation. We also find evidence that our measures are negatively correlated with the number of examples, which could suggest that corporations use many examples to hide the "best mode" of the invention in patent applications. In general, the robust results from statistical models and tests suggest that our proposed measures are effective and stable in capturing linguistic differences in patent documents, and shed light by quantifying the level of disclosure in patent applications.

University and corporate patents differ in many ways other than readability. One has to be concerned that these differences might somehow lead to systematic differences in the scores on these particular metrics, without actually being reflected in true readability. We have employed several strategies to minimize this issue, including both very fine subclass-level technological area controls, and cited-patent fixed effects. However, it is possible that future research could address this issue using better identification strategies, such as disclosure law changes or instrumental variables. It would also be useful to identify other situations where a strong prior expectation about differences in readability could be used to test the validity of the measures. In addition, we plan in future work to compare the linguistic metrics to subjective evaluations of the extent of disclosure given by subject domain experts based on their reading the patents.

The linguistic indicators used in the paper (which are widely used in the literature) are not specifically designed for patent texts. Many of the measures were developed in the context of second language acquisition, and some readability results may not necessarily reflect the same direction of readability in patent data. For example, “solar” might require a higher age of acquisition in standard contexts, but it is a standard word in the photoelectric patent category. Since we do not have a field-specific dictionary available, this is the best proxy available for patent readability. We believe that we have demonstrated that these measures pass a threshold of providing a useful set of metrics for patent readability, but it is likely that they could be refined to capture readability more precisely in the patent context.

We view this analysis as proof of concept for the use of computational metrics of readability as proxies for disclosure in patents. While further development and validation of the metrics is certainly warranted, the real payoff will come in the use of these metrics to begin to establish empirically what competitive, legal, cultural and institutional factors affect the level of readability in patents, and how differences in readability play out in the market place and in technology evolution.

References

Arinas, I. (2012, July). How Vague Can Your Patent Be? Vagueness Strategies in U.S. Patents. SSRN Scholarly Paper ID 2117827, Social Science Research Network, Rochester, NY.

Arts, S., B. Cassiman, and J. C. Gomez (2018). Text matching to measure patent similarity. *Strategic Management Journal 39*(1), 62–84. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.2699.

Baker, S. and C. Mezzetti (2005). Disclosure as a Strategy in the Patent Race. *The Journal of Law and Economics 48*(1), 173–194. Publisher: The University of Chicago Press.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650. Publisher: Oxford University Press.

Bloomfield, R. J. (2002, September). The "Incomplete Revelation Hypothesis" and Financial Reporting. *Accounting Horizons 16*(3), 233–243. Publisher: American Accounting Association.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2020). Are Referees and Editors in Economics Gender Neutral? *The Quarterly Journal of Economics 135*(1), 269–327. Publisher: Oxford University Press.

Cohen, W. M., A. Goto, A. Nagata, R. R. Nelson, and J. P. Walsh (2002). R&D spillovers, patents and the incentives to innovate in Japan and the United States. *Research policy 31*(8-9), 1349–1367. Publisher: Elsevier.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics 165*(2), 97–135. Publisher: John Benjamins.

Cornelli, F. and M. Schankerman (1999). Patent Renewals and R&D Incentives. *The RAND Journal of Economics 30*(2), 197–213. Publisher: [RAND Corporation, Wiley].

De Clercq, D., N.-F. Diop, D. Jain, B. Tan, and Z. Wen (2019, September). Multi-label classification and interactive NLP-based visualization of electric vehicle patent data. *World Patent Information 58*, 101903.

Denicolò, V. and L. A. Franzoni (2003). The contract theory of patents. *International Review of Law and Economics 23*(4), 365–380. Publisher: Elsevier.

Devlin, A. (2009). The misunderstood function of disclosure in patent law. *Harvard Journal of Law and Technology 23*, 401. Publisher: HeinOnline.

Gentzkow, M., B. Kelly, and M. Taddy (2019, September). Text as Data. *Journal of Economic Literature 57*(3), 535–574.

Grossman, S. J. and J. E. Stiglitz (1980). On the impossibility of informationally efficient markets. *The American economic review 70*(3), 393–408. Publisher: JSTOR.

Hansen, S., M. McMahon, and A. Prat (2018, May). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics 133*(2), 801–870. Publisher: Oxford Academic.

Henderson, R., A. B. Jaffe, and M. Trajtenberg (1998). Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. *Review of Economics and statistics 80*(1), 119–127. Publisher: MIT Press.

Hengel, E. (2017). Publishing while Female. Are women held to higher standards? Evidence from peer review. Publisher: University of Cambridge.

Hunt, K. W. (1965). Grammatical Structures Written at Three Grade Levels. NCTE Research Report No. 3. Publisher: ERIC.

Jaffe, A. B. and M. Trajtenberg (2002). *Patents, citations, and innovations: A window on the knowledge economy.* MIT press.

Jefferson, O. A., A. Jaffe, D. Ashton, B. Warren, D. Koellhofer, U. Dulleck, A. Ballagh, J. Moe, M. DiCuccio, and K. Ward (2018). Mapping the global influence of published research on industry and innovation. *Nature biotechnology 36*(1), 31–39. Publisher: Nature Publishing Group.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2018). Measuring technological innovation over the long run. Technical report, National Bureau of Economic Research.

Kitch, E. W. (1977). The Nature and Function of the Patent System. *The Journal of Law and Economics 20*(2), 265–290. Publisher: The University of Chicago Press.

Kuperman, V., H. Stadthagen-Gonzalez, and M. Brysbaert (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods 4*(44), 978–990.

Landes, W. M. and R. A. Posner (2009). *The economic structure of intellectual property law*. Harvard University Press.

Lawrence, A. (2013). Individual investors and financial disclosure. *Journal of Accounting and Economics 56*(1), 130–147. Publisher: Elsevier.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics 45*(2), 221–247.

Loughran, T. and B. McDonald (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research 54*(4), 1187–1230. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12123.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics 15*(4), 474–496. Publisher: John Benjamins.

McNamara, D. S., M. M. Louwerse, and A. C. Graesser (2002). Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Technical report, Technical report, Institute for Intelligent Systems, University of Memphis ....

Miller, B. P. (2010). The effects of reporting complexity on small and large investor trading. *The Accounting Review 85*(6), 2107–2143.

Onatski, A. (2010, November). Determining the Number of Factors from Empirical Distribution of Eigenvalues. *Review of Economics and Statistics 92*(4), 1004–1016.

Ouellette, L. L. (2011). Do patents disclose useful information. *Harvard Journal of Law and Technology 25*, 545. Publisher: HeinOnline.

Packalen, M. and J. Bhattacharya (2015). New ideas in invention. Technical report, National Bureau of Economic Research.

Roin, B. N. (2005). The disclosure function of the patent system (or lack thereof). *Harvard Law Review*. Publisher: Harvard University, Harvard Law School.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282. Publisher: Wiley Online Library.

Sampat, B. (2018, December). A Survey of Empirical Evidence on Patents and Innovation. Technical Report w25383, National Bureau of Economic Research, Cambridge, MA.

Schrand, C. M. and B. R. Walther (2000, April). Strategic Benchmarks in Earnings Announcements: The Selective Disclosure of Prior-Period Earnings Components. *The Accounting Review 75*(2), 151–177. Publisher: Allen Press.

Scotchmer, S. and J. Green (1990). Novelty and disclosure in patent law. *The RAND Journal of Economics*, 131–146. Publisher: JSTOR.

Tauman, Y. and M.-H. Weng (2012, March). Selling patent rights and the incentive to innovate. *Economics Letters 114*(3), 241–244.

Teodorescu, M. (2017). Machine learning methods for strategy research. *Harvard Business School Research Paper Series* (18-011).

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(3), 273–282. Publisher: Wiley Online Library.

Todirascu, A., T. François, N. Gala, C. Fairon, A.-L. Ligozat, and D. Bernhard (2013). Coherence and cohesion for the assessment of text readability. *Natural Language Processing and Cognitive Science 11*, 11–19.

Trajtenberg, M., R. Henderson, and A. Jaffe (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and new technology 5*(1), 19–50. Publisher: Taylor & Francis.

Vajjala, S. and D. Meurers (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pp. 163–173.

Vajjala, S. and D. Meurers (2013). On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pp. 59–68.

Vajjala, S. and D. Meurers (2014a). Exploring measures of "readability" for spoken language: Analyzing linguistic features of subtitles to identify age-specific TV programs. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pp. 21–29.

Vajjala, S. and D. Meurers (2014b). Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics 165*(2), 194–222. Publisher: John Benjamins.

Valdivia, W. D. (2013). University start-ups: Critical for improving technology transfer. *Center for Technology Innovation at Brookings. Washington, DC: Brookings Institution.*

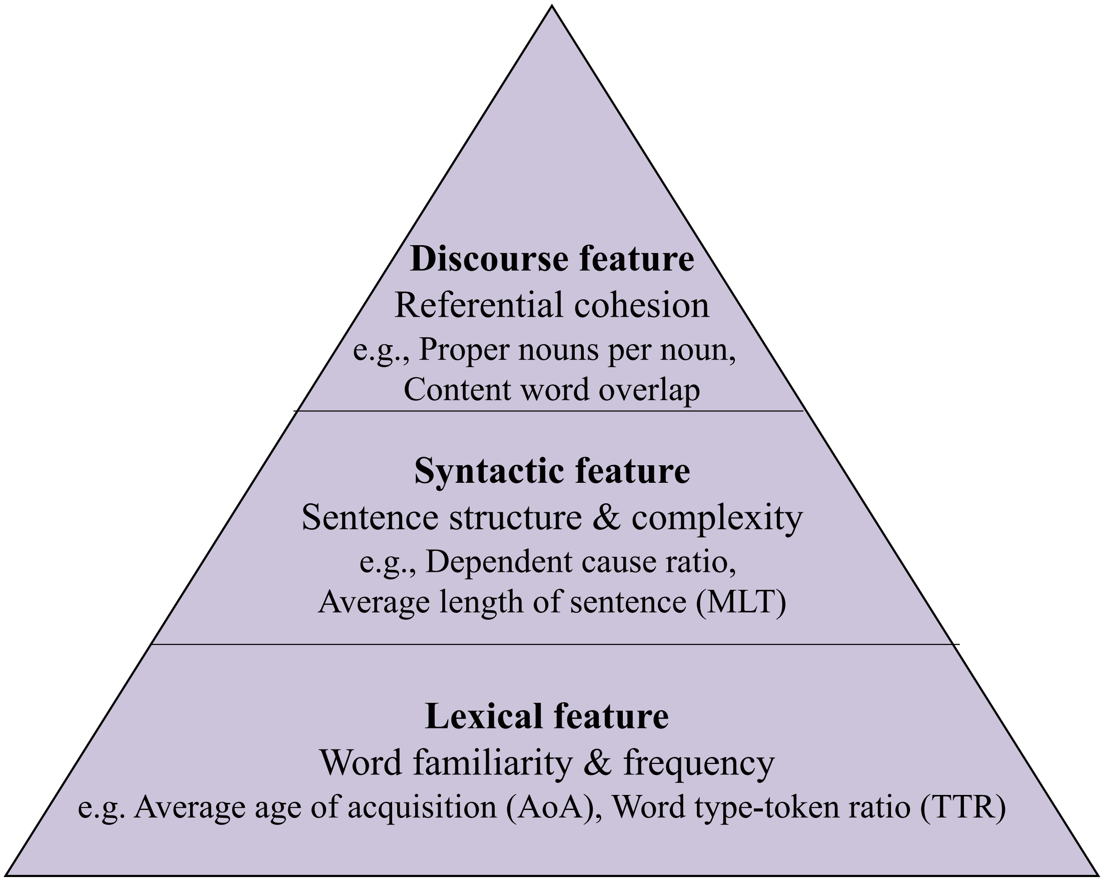
Walsh, J. P. and S. Nagaoka (2009). Who invents? Evidence from the Japan & US inventor survey. *RIETI Discussion papers.*

Wu, A. H. (2018). Gendered language on the economics job market rumors forum. In *AEA Papers and Proceedings*, Volume 108, pp. 175–79.

You, H. and X.-j. Zhang (2009). Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting studies 14*(4), 559–586. Publisher: Springer.

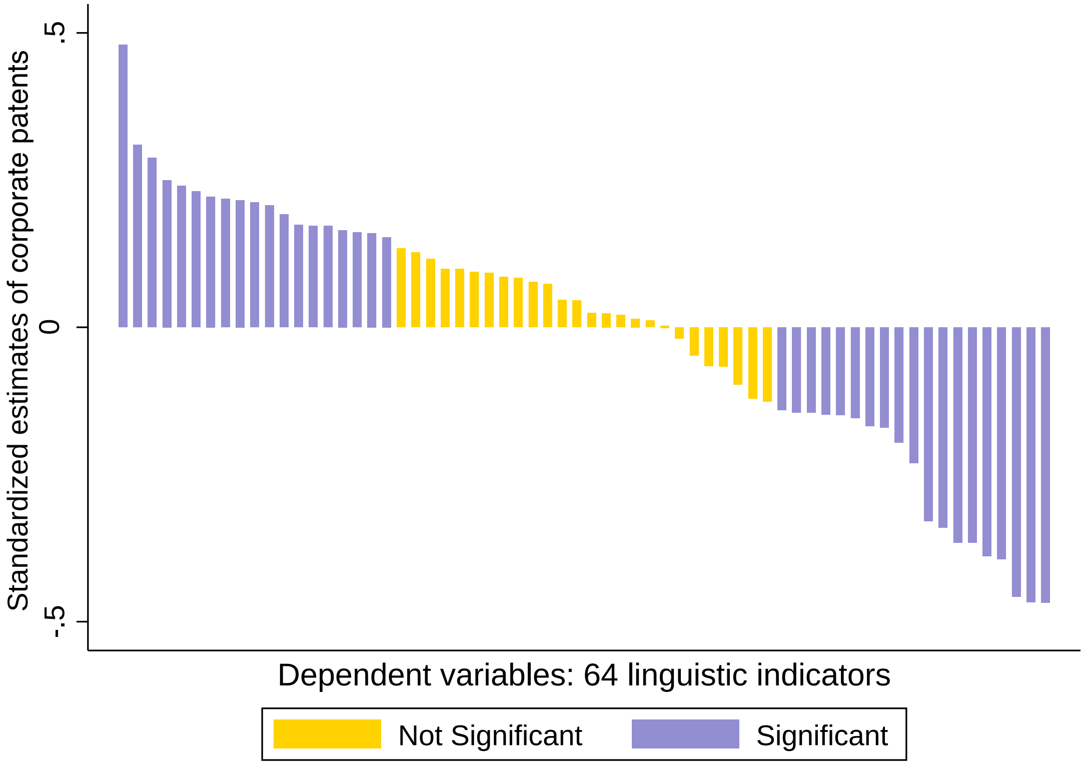
Younge, K. A. and J. M. Kuhn (2016, July). Patent-to-Patent Similarity: A Vector Space Model. SSRN Scholarly Paper ID 2709238, Social Science Research Network, Rochester, NY.

Figure 1. Hierarchy of linguistic analysis


Discourse feature
Referential cohesion
e.g., Proper nouns per noun,
Content word overlap
Syntactic feature
Sentence structure & complexity
e.g., Dependent cause ratio,
Average length of sentence (MLT)
Lexical feature
Word familiarity & frequency
e.g. Average age of acquisition (AoA), Word type-token ratio (TTR)

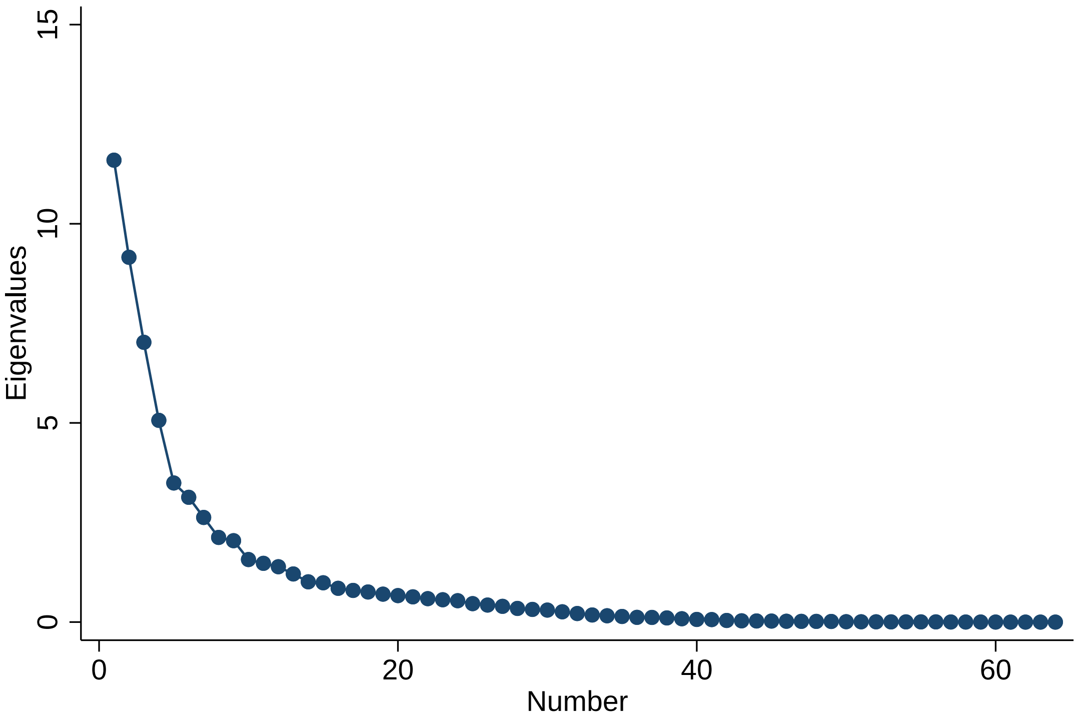
Note: This hierarchy of linguistic analysis is derived by "Key aspects of text readability" from Collins-Thompson (2014) and Loughran and McDonald (2016). We selected the relevant and feasible levels of analysis in a patent context.

Figure 2. Baseline estimates of corporate patents plotted with significance using multiple hypothesis testing



Note: Y-axis indicates the estimates from Table A.2 using Equation 1. Each bar represents one linguistic measure. Significance is defined as $p < 0.1$. Multiple hypothesis testing uses Romano and Wolf (2005) stepdown adjusted p-values with 250 bootstrap replications. The sample is 40,949 patent applications in three patent categories related to nanotechnology, batteries, and electricity in the U.S. from 2000 to 2019, as described in Section 4.1.

Figure 3. Scree plot of eigenvalues after PCA



Note: The figure presents the scree plot of the eigenvalues of correlation metrics after PCA that combines 64 linguistic indicators into synthetic variables, as described in Section 6. According to the largest distance rule from Onatski (2010), we present estimates of components 1-4.

TABLE I. Linguistic measures for patent applications

LINGUISTIC OUTCOMES	**Traditional measures**	**Formula**	**Notes**
Traditional	Fog	0.4 (ASL + 100 RHW)[a]	Corresponds to years of formal education to understand the text on first reading
	Flesch	06.356 -84.6 AWL -1.015 ASL[b]	Ranging from 0 (professional reading level) to 100 (5th grade reading level).
	Kincaid	-15.59 + 11.8 AWL + 0.39 ASL	Corresponds with a U.S. grade level. It is relevant when the number is greater than 10, with no upper bound.
Levels of Linguistic features		**Definition**	
Lexical	AoA_Kup	Age of acquisition of words[c]	
	Word_TTR	# unique words / # total words[d]	
Syntactic	DependentClauseR	# dependent clauses/ # total clauses[e]	
	MLT	Average length of a t-unit: # of words / # of T-units[f]	
Discourse	ProperNounsPerNoun	Ratio of proper nouns to nouns[g]	
	ContentWordOverlap	# content word overlap between all pairs of sentences / # total sentences	

Note: Representative linguistic measures used in Tables II to VII. See the Appendix for the full list of linguistic measures. Vajjala and Meurers' (2014c) computational linguistic model is used to calculate all linguistic measures.

[a] ASL is average sentence length and RHW is the ratio of hard words to all words. Hard words are defined as words of more than three syllable.

[b] AWL is average word length in syllables.

[c] Compiled from Kuperman et al. (2012) Psycholinguistic database.

[d] Total number of different words occurring in a text divided by the total number of words.

[e] A dependent clause has a subject and verb but does not express a complete thought. A dependent clause cannot be a sentence, as opposed to an independent clause (a sentence).

[f] T units are the shortest grammatically allowable sentences. See Lu (2010).

[g] A proper noun is a specific (i.e., not generic) name for a particular person, place, or thing. See Todirascu et al. (2013).

TABLE II. Summary statistics of representative variables

		(1) Universities'		(2) Corporations'		(3) Difference	
Categories	Variables	Mean	SD	Mean	SD	b	t
LINGUISTIC OUTCOMES							
Traditional	Fog	22.10	4.61	23.59	6.90	-1.49***	(-16.18)
	Flesch	40.72	15.10	38.55	20.01	2.16***	(7.39)
	Kincaid	15.23	4.56	16.92	6.68	-1.69***	(-18.65)
Lexical	AoA_Kup	5.19	0.26	5.21	0.29	-0.02***	(-3.65)
	Word_TTR	0.16	0.04	0.13	0.04	0.03***	(34.58)
Syntactic	DependentClauseR	0.33	0.07	0.37	0.08	-0.04***	(-32.46)
	MLT	12.10	1.72	12.59	2.00	-0.49***	(-15.21)
Discourse	ProperNounsPerNoun	0.08	0.06	0.05	0.03	0.03***	(32.56)
	ContentWordOverlap	367.43	263.34	551.29	546.77	-183.86***	(-31.35)
CONTROLS							
	Cited_by_Patent_Count	10.93	17.86	13.93	26.53	-3.00***	(-8.42)
	Simple_Family_Size	5.49	5.27	7.43	9.47	-1.95***	(-17.49)
	Extended_Family_Size	6.78	10.20	11.72	28.52	-4.94***	(-18.84)
	Sequence_Count	7.01	261.33	39.91	5114.95	-32.90	(-0.93)
	NPL_Citation_Count	1.12	2.28	0.38	1.25	0.74***	(18.47)
	NPL_Resolved_Citation_Count	0.82	1.82	0.19	0.83	0.63***	(19.82)
	uspc136	0.24	0.43	0.35	0.48	-0.11***	(-13.90)
	uspc320	0.03	0.18	0.32	0.47	-0.29***	(-64.01)
	uspc977	0.76	0.43	0.34	0.47	0.42***	(52.03)
	Observations	3,414		21,234		24,648	

Note: The sample is patent applications filed by universities and corporations in three patent categories related to nanotechnology, batteries, and electricity in the U.S. from 2000 to 2019. The full sample, including patents jointly filed by universities and corporations as well as by other entities, are presented in Table 10 in the Appendix. Detailed information on the sample is provided in Section 4.1. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

TABLE III. Baseline estimates of corporate patents

		Panel A: Without controlling for X		**Panel B: Controlling for X**		
Features	VARIABLES	Raw	R-squared	Raw	Standardized	R-squared
Traditional	Fog	1.529***	0.044	1.424***	0.212***	0.054
		(0.0750)		(0.116)	(0.0173)	
	Flesch	-4.592***	0.071	-4.330***	-0.222***	0.076
		(0.503)		(0.543)	(0.0278)	
	Kincaid	1.735***	0.043	1.626***	0.250***	0.052
		(0.0796)		(0.112)	(0.0172)	
Lexical	AoA_Kup	0.0235	0.147	0.0261	0.0921	0.149
		(0.00892)		(0.00792)	(0.0279)	
	Word_TTR	-0.0171***	0.120	-0.0151***	-0.329***	0.143
		(0.00173)		(0.00165)	(0.0360)	
Syntactic	DependentClauseR	0.0187***	0.171	0.0172***	0.207***	0.175
		(0.00152)		(0.00139)	(0.0166)	
	MLT	0.458***	0.057	0.427***	0.216***	0.064
		(0.0330)		(0.0253)	(0.0128)	
Discourse	ProperNounsPerNoun	-0.0192	0.173	-0.019	-0.48	0.180
		(0.00117)		(0.00128)	(0.0324)	
	ContentWordOverlap	141.0***	0.051	116.9***	0.231***	0.084
		(17.30)		(16.48)	(0.0326)	

Note: N=40,949. Rows represent dependent variables. Estimations in Panel A control for the U.S. patent subclassification fixed effects; and regressions in Panel B control for joint patents and other patents (using university patents as base), various citation counts, simple and extended family size, and the U.S. patent subclassification fixed effects. Both raw linguistic measures and standardized linguistic measures (mean=0, SD=1) are used as dependent variables in Panel B. OLS estimates of corporate patents are obtained from Equation 1, using university patents as the base. Standard errors are clustered at the U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ for Romano-Wolf adjusted p-values using 64 linguistic measures.

TABLE IV. Estimates of principal component analysis as synthetic linguistic indicators

	(1)	(2)	(3)	(4)
VARIABLES	Component 1	Component 2	Component 3	Component 4
Corporations	0.429***	0.0929**	-0.123*	-0.157*
	(0.00171)	(0.0183)	(0.0386)	(0.0529)
R-squared	0.290	0.155	0.103	0.113

Notes: N=40,949. The dependent variable is the PCA generated by 64 linguistic measures. See Table 10 for detailed compositions and Figure 3 for eigenvalues. OLS estimates of corporate patents are obtained from Equation 1, using university patents as the base. All estimations control for joint patents and other patents (using university patents as base), various citation counts, simple and extended family size, and the U.S. patent subclassification fixed effects. Standard errors are clustered at the U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

TABLE V. Corporate patent estimates with high-degree cited-patent fixed effects using LASSO

Features	VARIABLES	Raw	Standardized
Synthetic	PCA	0.435***	0.430***
		(0.0240)	(0.0242)
Traditional	Fog	1.072***	0.159***
		(0.151)	(0.0218)
	Flesch	-3.096***	-0.161***
		(0.472)	(0.0236)
	Kincaid	1.265***	0.196***
		(0.148)	(0.0227)
Lexical	AoA_Kup	0.0241***	0.0686**
		(0.00793)	(0.0282)
	Word_TTR	-0.0154***	-0.336***
		(0.00121)	(0.0268)
Syntactic	DependentClauseR	0.0183***	0.214***
		(0.00215)	(0.0252)
	MLT	0.394***	0.188***
		(0.0532)	(0.0275)
Discourse	ProperNounsPerNoun	-0.0191***	-0.505***
		(0.00153)	(0.0387)
	ContentWordOverlap	114.5***	0.208***
		(10.80)	(0.0184)

Notes: N=20,571. Rows represents dependent variables. Both raw linguistic measures and standardized linguistic measures (mean=0, SD=1) are used. Sample includes patents that have cited “highly cited patents” (citation$\geq$ 10). All estimations control for joint patents and other patents (using university patents as base), various citation counts, simple and extended family size, and the U.S. patent subclassification fixed effects. Estimates of corporate patents are presented with university patents as the base. Standard errors are in parentheses. [para,flushleft] *** p <0.01, ** p <0.05, * p <0.1.

TABLE VI. Estimates of example frequencies

Features	Variables	Num_examples	Firms	R-squared
Synthetic	PCA	0.000790***	0.423***	0.289
		(7.54e-05)	(0.00171)	
Traditional	Fog	-0.00109	1.432***	0.052
		(0.000508)	(0.113)	
	Flesch	-0.00810**	-4.275**	0.075
		(0.00120)	(0.535)	
	Kincaid	-0.000507	1.630***	0.050
		(0.000573)	(0.108)	
Lexical	AoA_Kup	3.45e-05	0.0259*	0.148
		(7.65e-05)	(0.00830)	
	Word_TTR	-0.000195**	-0.0138**	0.190
		(4.12e-05)	(0.00147)	
Syntactic	DependentClauseR	-2.47e-05	0.0174***	0.174
		(1.23e-05)	(0.00141)	
	MLT	0.00337**	0.404***	0.070
		(0.000503)	(0.0243)	
Discourse	ProperNounsPerNoun	5.62e-06	-0.0190***	0.178
		(3.75e-06)	(0.00126)	
	ContentWordOverlap	4.223**	88.24**	0.274
		(0.807)	(15.03)	

Notes: N=40,949. Rows represent dependent variables. *Num_examples* is the frequency of “for example” and “e.g.” in technical descriptions in patent applications. All estimations control for joint patents and other patents (using university patents as base), various citation counts, simple and extended family size, and the U.S. patent subclassification fixed effects. Standard errors are clustered at the U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

TABLE VII. Heterogeneous effects by top applicants

		Panel A: Top 100 applicants		**Panel B: Other**	
Features	Variables	Corporations	R-squared	Corporations	R-squared
Synthetic	PCA	0.675***	0.373	0.255***	0.270
		(0.0150)		(0.0135)	
Traditional	Fog	2.228**	0.105	0.912*	0.052
		(0.257)		(0.271)	
	Flesch	-5.741**	0.115	-3.583*	0.077
		(0.810)		(1.052)	
	Kincaid	2.398**	0.106	1.159*	0.049
		(0.296)		(0.273)	
Lexical	AoA_Kup	0.0196	0.214	0.0265**	0.132
		(0.0107)		(0.00528)	
	Word_TTR	-0.0191***	0.199	-0.0111**	0.128
		(0.000881)		(0.00186)	
Syntactic	DependentClauseR	0.0303**	0.274	0.00858	0.156
		(0.00685)		(0.00344)	
	MLT	0.704**	0.116	0.215**	0.064
		(0.111)		(0.0432)	
Discourse	ProperNounsPerNoun	-0.0268***	0.251	-0.0141***	0.177
		(0.00261)		(0.000885)	
	ContentWordOverlap	145.9**	0.119	92.09**	0.083
		(25.79)		(16.73)	
	Observations	11,844		29,105	

Notes: The top 100 applicants are defined by patent application counts in the sample. Rows represent dependent variables. Panel A shows estimates from the top applicants sample and Panel B uses the rest of the sample. All estimation follows Equation 1 and control for joint patents and other patents (using university patents as base), various citation counts, simple and extended family size, and the U.S. patent subclassification fixed effects. Standard errors are clustered at the U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

APPENDIX

Table A.1. Summary statistics of 64 individual linguistic measures and controls by business model

	(1)		(2)		(3)		(4)	
	Corporations		Universities		Joint		Others	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Controls								
Cited_by_Patent_Count	13.93	26.53	10.93	17.86	10.60	18.34	14.39	25.62
Simple_Family_Size	7.43	9.47	5.49	5.27	7.02	5.91	6.89	6.79
Extended_Family_Size	11.72	28.52	6.78	10.20	14.14	30.29	10.63	24.50
Sequence_Count	39.91	5114.95	7.01	261.33	1.62	33.83	3.79	168.31
NPL_Citation_Count	0.38	1.25	1.12	2.28	0.92	2.22	0.42	1.30
NPL_Resolved_Citation_Count	0.19	0.83	0.82	1.82	0.64	1.71	0.25	0.95
uspc136	0.35	0.48	0.24	0.43	0.19	0.39	0.32	0.46
uspc320	0.32	0.47	0.03	0.18	0.08	0.27	0.28	0.45
uspc977	0.34	0.47	0.76	0.43	0.75	0.43	0.42	0.49
Linguistic Measures								
AoA_Bird_Lem	3.20	0.20	3.19	0.17	3.17	0.19	3.20	0.19
AoA_Bristol_Lem	1.52	0.27	1.45	0.21	1.56	0.30	1.51	0.26
AoA_Cort_Lem	2.22	0.20	2.21	0.15	2.22	0.16	2.23	0.19
AoA_Kup	5.21	0.29	5.19	0.26	5.26	0.26	5.17	0.28
AoA_Kup_Lem	6.51	0.25	6.61	0.22	6.54	0.22	6.51	0.26
DISC_RefExprDefArtPerSen	2.56	1.11	2.00	0.74	2.33	0.77	2.40	1.05
DISC_RefExprDefArtPerWord	0.08	0.02	0.07	0.02	0.08	0.02	0.08	0.02
DISC_RefExprPerProPerWord	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DISC_RefExprPerPronounsPerSen	0.08	0.07	0.08	0.06	0.07	0.06	0.09	0.08
DISC_RefExprPossProPerSen	0.02	0.03	0.03	0.03	0.02	0.03	0.03	0.04
DISC_RefExprPossProPerWord	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DISC_RefExprPronounsPerNoun	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table A.1. Summary statistics of 64 individual linguistic measures and controls by business model

	(1)		(2)		(3)		(4)	
	Corporations		Universities		Joint		Others	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
DISC_RefExprPronounsPerSen	0.11	0.08	0.11	0.07	0.09	0.07	0.13	0.10
DISC_RefExprPronounsPerWord	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DISC_RefExprProperNounsPerNoun	0.05	0.03	0.08	0.06	0.06	0.04	0.05	0.04
DISC_globalArgumentOverlapCount	57.67	46.67	49.02	32.31	54.05	36.85	51.41	43.20
DISC_globalContentWordOverlapCount	551.29	546.77	367.43	263.34	469.41	374.69	455.02	487.39
DISC_globalNounOverlapCount	51.22	40.98	41.48	26.88	47.96	32.44	44.82	37.54
DISC_globalStemOverlapCount	59.64	48.15	51.07	33.70	55.70	38.12	53.27	44.76
DISC_localArgumentOverlapCount	0.79	0.09	0.70	0.12	0.77	0.10	0.76	0.11
DISC_localContentWordOverlapCount	9.37	17.78	6.29	4.86	7.96	5.69	8.24	13.99
DISC_localNounOverlapCount	0.76	0.10	0.66	0.12	0.74	0.11	0.72	0.12
DISC_localStemOverlapCount	0.80	0.09	0.71	0.11	0.78	0.10	0.77	0.10
MRCAoA	0.33	0.09	0.31	0.07	0.32	0.08	0.33	0.08
MRCColMeaningfulness	1.76	0.12	1.75	0.10	1.80	0.14	1.75	0.12
MRCConcreteness	1.80	0.13	1.77	0.12	1.85	0.17	1.79	0.13
MRCFamiliarity	3.93	0.19	3.90	0.17	3.97	0.19	3.93	0.19
MRCImageability	1.97	0.13	1.93	0.11	2.00	0.15	1.96	0.13
MRCPavioMeaningfulness	0.26	0.11	0.23	0.09	0.22	0.09	0.25	0.11
POS_adjVar	0.17	0.04	0.16	0.03	0.16	0.03	0.17	0.04
POS_advVar	0.05	0.01	0.04	0.01	0.04	0.01	0.05	0.01
POS_correctedVV1	44.61	25.99	41.33	19.93	41.19	21.34	40.71	23.58
POS_modVar	0.22	0.04	0.21	0.04	0.21	0.04	0.22	0.04
POS_nounVar	0.54	0.05	0.56	0.04	0.56	0.04	0.54	0.05
POS_squaredVerbVar1	5330.99	10561.22	4210.79	5486.37	4304.39	5913.33	4427.59	8681.50
POS_verbVar1	4.33	1.93	3.65	1.21	3.95	1.45	3.88	1.65
POS_verbVar2	0.21	0.03	0.20	0.03	0.20	0.03	0.21	0.03

Table A.1. Summary statistics of 64 individual linguistic measures and controls by business model

	(1)		(2)		(3)		(4)	
	Corporations		Universities		Joint		Others	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
SYN_CNPerClause	8.36	4.98	8.69	6.94	8.09	6.31	8.37	6.49
SYN_CNPerTunit	5.20	3.04	5.14	4.27	4.86	3.82	5.19	3.88
SYN_ComplexTunitRatio	0.25	0.10	0.21	0.07	0.21	0.09	0.25	0.10
SYN_CoordPerClause	0.46	0.16	0.48	0.14	0.43	0.13	0.47	0.16
SYN_CoordPerTunit	0.28	0.09	0.28	0.08	0.26	0.08	0.28	0.09
SYN_DependentClauseRatio	0.37	0.08	0.33	0.07	0.33	0.09	0.37	0.08
SYN_DependentClausesPerTunit	0.24	0.08	0.20	0.06	0.20	0.08	0.24	0.08
SYN_MLC	20.24	2.89	20.49	2.73	20.06	2.55	20.06	2.97
SYN_MLT	12.59	2.00	12.10	1.72	12.00	1.77	12.43	2.00
SYN_TunitComplexityRatio	0.62	0.09	0.59	0.07	0.60	0.08	0.62	0.09
SYN_VPPerTunit	1.70	0.24	1.56	0.21	1.58	0.23	1.68	0.24
TRAD_ARI	21.26	8.43	19.00	5.59	19.26	7.69	20.95	8.36
TRAD_Coleman	13.86	1.40	14.40	1.34	13.88	1.24	13.97	1.45
TRAD_FOG	23.59	6.90	22.10	4.61	22.08	6.28	23.45	6.83
TRAD_FORCAST	16.61	0.54	16.33	0.47	16.54	0.47	16.54	0.55
TRAD_Flesch	38.55	20.01	40.72	15.10	43.56	18.33	38.16	19.73
TRAD_Kincaid	16.92	6.68	15.23	4.56	15.17	6.14	16.77	6.62
TRAD_LIX	69.35	17.14	64.88	11.45	65.06	15.82	68.77	16.97
TRAD_SMOG	19.20	3.29	18.59	2.68	18.38	3.08	19.16	3.31
TRAD_numChars	5.19	0.23	5.31	0.22	5.22	0.20	5.22	0.24
TRAD_numSyll	1.55	0.12	1.59	0.11	1.54	0.11	1.56	0.12
Word_BilogTTR	0.77	0.03	0.79	0.02	0.77	0.03	0.78	0.03
Word_CTTR	8.18	2.26	10.30	2.80	8.72	2.61	8.79	2.53
Word_MTLD	6.75	0.46	6.92	0.50	6.76	0.44	6.85	0.47
Word_RTTR	11.56	3.20	14.57	3.96	12.33	3.69	12.44	3.58

Table A.1. Summary statistics of 64 individual linguistic measures and controls by business model

	(1)		(2)		(3)		(4)	
	Corporations		Universities		Joint		Others	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Word_TTR	0.13	0.04	0.16	0.04	0.14	0.04	0.15	0.05
Word_UberIndex	39.47	4.94	44.22	5.54	40.65	5.33	41.05	5.46
Observations	21,234		3,414		1,644		14,657	

Note: The sample is patent applications filed by all entities in three patent categories related to nanotechnology, batteries, and electricity in the U.S. from 2000 to 2019. Detailed information on sample is provided in Section 4.1. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.2. OLS estimates 64 individual readability measures and adjusted p-value using Romano-Wolf stepdown multiple hypothesis testing

	VARIABLES	Raw		Standardized		R-squared
(1)	AoA_Bird_Lem	0.0165	(0.00475)	0.0854	(0.0245)	0.096
(2)	AoA_Bristol_Lem	-0.00506	(0.00922)	-0.0191	(0.0349)	0.182
(3)	AoA_Cort_Lem	-0.00931	(0.0155)	-0.0481	(0.0801)	0.097
(4)	AoA_Kup	0.0261	(0.00792)	0.0921	(0.0279)	0.149
(5)	AoA_Kup_Lem	-0.0319	(0.00358)	-0.126	(0.0142)	0.211
(6)	DISC_RefExprDefArtPerSen	0.154**	(0.00931)	0.145**	(0.00877)	0.199
(7)	DISC_RefExprDefArtPerWord	0.00143	(0.000741)	0.0655	(0.0339)	0.229
(8)	DISC_RefExprPerProPerWord	-2.20e-05	(1.97e-05)	-0.0209	(0.0188)	0.025
(9)	DISC_RefExprPerPronounsPerSen	-0.00174	(0.00218)	-0.0245	(0.0306)	0.044
(10)	DISC_RefExprPossProPerSen	-0.00537**	(0.000399)	-0.161**	(0.0120)	0.050
(11)	DISC_RefExprPossProPerWord	-5.88e-06	(8.39e-06)	-0.0238	(0.0340)	0.007
(12)	DISC_RefExprPronounsPerNoun	-0.00111	(0.000196)	-0.116	(0.0206)	0.057
(13)	DISC_RefExprPronounsPerSen	-0.00738	(0.00240)	-0.0837	(0.0272)	0.048
(14)	DISC_RefExprPronounsPerWord	-8.58e-05	(6.29e-05)	-0.0457	(0.0335)	0.030
(15)	DISC_RefExprProperNounsPerNoun	-0.0190***	(0.00128)	-0.480***	(0.0324)	0.180
(16)	DISC_globalArgumentOverlapCount	6.416*	(1.336)	0.145*	(0.0302)	0.126
(17)	DISC_globalContentWordOverlapCount	116.9***	(16.48)	0.231***	(0.0326)	0.084
(18)	DISC_globalNounOverlapCount	6.546***	(1.216)	0.170***	(0.0315)	0.117
(19)	DISC_globalStemOverlapCount	6.435*	(1.397)	0.141*	(0.0306)	0.129
(20)	DISC_localArgumentOverlapCount	0.0479***	(0.00291)	0.468***	(0.0285)	0.265
(21)	DISC_localContentWordOverlapCount	1.499	(0.188)	0.0972	(0.0122)	0.019
(22)	DISC_localNounOverlapCount	0.0513***	(0.00247)	0.458***	(0.0220)	0.281
(23)	DISC_localStemOverlapCount	0.0468***	(0.00298)	0.467***	(0.0297)	0.271
(24)	MRCAoA	0.00648	(0.00286)	0.0769	(0.0339)	0.131
(25)	MRCColMeaningfulness	0.0187**	(0.00551)	0.154**	(0.0454)	0.144
(26)	MRCConcreteness	0.0264***	(0.00416)	0.196***	(0.0309)	0.138
(27)	MRCFamiliarity	0.0226	(0.00708)	0.121	(0.0381)	0.066
(28)	MRCImageability	0.0193**	(0.00420)	0.149**	(0.0325)	0.103
(29)	MRCPavioMeaningfulness	-0.000128	(0.00313)	-0.00117	(0.0286)	0.163
(30)	POS_adjVar	0.00172	(0.00255)	0.0463	(0.0688)	0.234

Table A.2. OLS estimates 64 individual readability measures and adjusted p-value using Romano-Wolf stepdown multiple hypothesis testing

	VARIABLES	Raw		Standardized		R-squared
(31)	POS_advVar	0.00238***	(0.000377)	0.172***	(0.0273)	0.055
(32)	POS_correctedVV1	3.929**	(1.311)	0.160**	(0.0533)	0.121
(33)	POS_modVar	0.00403	(0.00292)	0.0993	(0.0720)	0.232
(34)	POS_nounVar	-0.00778***	(0.00290)	-0.168***	(0.0627)	0.190
(35)	POS_squaredVerbVar1	1,264	(185.0)	0.134	(0.0196)	0.071
(36)	POS_verbVar1	0.513***	(0.0786)	0.288***	(0.0441)	0.080
(37)	POS_verbVar2	0.000377	(0.000779)	0.0115	(0.0238)	0.204
(38)	SYN_CNPerClause	0.546	(0.0776)	0.0943	(0.0134)	0.065
(39)	SYN_CNPerTunit	0.446	(0.0356)	0.127	(0.0102)	0.055
(40)	SYN_ComplexTunitRatio	0.0162***	(0.00117)	0.172***	(0.0124)	0.144
(41)	SYN_CoordPerClause	0.0156	(0.00440)	0.0991	(0.0280)	0.099
(42)	SYN_CoordPerTunit	0.0157***	(0.00210)	0.174***	(0.0232)	0.076
(43)	SYN_DependentClauseRatio	0.0172***	(0.00139)	0.207***	(0.0166)	0.175
(44)	SYN_DependentClausesPerTunit	0.0159***	(0.00167)	0.192***	(0.0201)	0.157
(45)	SYN_MLC	0.213	(0.0754)	0.0736	(0.0260)	0.088
(46)	SYN_MLT	0.427***	(0.0253)	0.216***	(0.0128)	0.064
(47)	SYN_TunitComplexityRatio	0.0149***	(0.00161)	0.165***	(0.0179)	0.099
(48)	SYN_VPPerTunit	0.0734***	(0.00306)	0.310***	(0.0129)	0.131
(49)	TRAD_ARI	1.893***	(0.0806)	0.231***	(0.00982)	0.050
(50)	TRAD_Coleman	-0.0940	(0.0186)	-0.0665	(0.0132)	0.167
(51)	TRAD_FOG	1.424***	(0.116)	0.212***	(0.0173)	0.054
(52)	TRAD_FORCAST	0.0831***	(0.0258)	0.153***	(0.0476)	0.219
(53)	TRAD_Flesch	-4.330***	(0.543)	-0.222***	(0.0278)	0.076
(54)	TRAD_Kincaid	1.626***	(0.112)	0.250***	(0.0172)	0.052
(55)	TRAD_LIX	3.642***	(0.216)	0.218***	(0.0130)	0.056
(56)	TRAD_SMOG	0.781***	(0.106)	0.240***	(0.0327)	0.081
(57)	TRAD_numChars	-0.0349**	(0.00220)	-0.148**	(0.00935)	0.183
(58)	TRAD_numSyll	0.00179	(0.00490)	0.0146	(0.0399)	0.198
(59)	Word_BilogTTR	-0.0112***	(0.000443)	-0.394***	(0.0155)	0.252
(60)	Word_CTTR	-0.912***	(0.119)	-0.366***	(0.0477)	0.413

Table A.2. OLS estimates 64 individual readability measures and adjusted p-value using Romano-Wolf stepdown multiple hypothesis testing

	VARIABLES	Raw		Standardized		R-squared
(61)	Word_MTLD	-0.159***	(0.0216)	-0.340***	(0.0461)	0.080
(62)	Word_RTTR	-1.290***	(0.169)	-0.366***	(0.0478)	0.413
(63)	Word_TTR	-0.0151***	(0.00165)	-0.329***	(0.0360)	0.143
(64)	Word_UberIndex	-2.087***	(0.190)	-0.389***	(0.0354)	0.394

Note: Each row represents one estimation. Estimates are from Equation 1. Multiple hypothesis testing uses Romano and Wolf (2005) stepdown adjusted p-values with 250 bootstrap replications: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ (adjusted). The sample consists of the 40,949 patent applications in three patent categories related to nanotechnology, batteries, and electricity in the U.S. from 2000 to 2019, as described in Section 4.1. Estimations control for joint patents and other patents (using university patents as base), various citation counts, simple and extended family size, and the U.S. patent subclassification fixed effects. Standard errors are clustered at the U.S. patent classification level in parentheses.

Table III. Synthetic readability composition

Variable	Comp1	Comp2	Comp3	Comp4
AoA_Bird_Lem	0.04	-0.0552	0.0545	0.132
AoA_Bristol_Lem	0.0772	-0.0676	-0.0149	-0.0898
AoA_Cort_Lem	0.0462	-0.1279	0.044	0.1544
AoA_Kup	-0.0097	0.0941	-0.0743	-0.1843
AoA_Kup_Lem	-0.1337	0.1564	-0.0683	-0.1207
DISC_RefExprDefArtPerSen	0.233	0.0203	0.0472	-0.0612
DISC_RefExprDefArtPerWord	0.1693	-0.1221	0.0097	-0.0891
DISC_RefExprPerProPerWord	0.0149	-0.0107	0.0984	0.1281
DISC_RefExprPerPronounsPerSen	0.0942	0.0437	0.2067	0.2093
DISC_RefExprPossProPerSen	-0.0131	0.028	0.1529	0.1696
DISC_RefExprPossProPerWord	0.0042	-0.0047	0.0379	0.0479
DISC_RefExprPronounsPerNoun	0.0144	-0.0231	0.217	0.2434
DISC_RefExprPronounsPerSen	0.0697	0.0463	0.2247	0.2347
DISC_RefExprPronounsPerWord	0.0161	-0.0143	0.1416	0.1727
DISC_RefExprProperNounsPerNoun	-0.1443	0.0259	-0.0318	0.1387
DISC_globalArgumentOverlapCount	0.0691	0.0628	-0.2859	0.1996
DISC_globalContentWordOverlapCount	0.1479	0.0983	-0.2255	0.1581
DISC_globalNounOverlapCount	0.0866	0.0575	-0.2887	0.1833
DISC_globalStemOverlapCount	0.0667	0.0633	-0.2855	0.2037
DISC_localArgumentOverlapCount	0.2006	-0.0255	-0.0726	-0.1578
DISC_localContentWordOverlapCount	0.1142	0.0966	0.029	-0.0204
DISC_localNounOverlapCount	0.2104	-0.0278	-0.0826	-0.1698
DISC_localStemOverlapCount	0.2001	-0.0266	-0.0712	-0.1546
MRCAoA	0.108	-0.0519	-0.0058	-0.0869
MRCColMeaningfulness	0.0567	-0.1143	0.0324	0.112
MRCConcreteness	0.1012	-0.1261	0.0173	0.0147
MRCFamiliarity	0.1337	-0.1333	0.0577	0.0869
MRCImageability	0.1049	-0.1342	0.0299	0.0376
MRCPavioMeaningfulness	0.0919	-0.0297	-0.0096	-0.083
POS_adjVar	-0.0018	0.0376	0.0324	-0.1011
POS_advVar	-0.004	-0.0119	0.132	0.0941
POS_correctedVV1	0.0802	0.0953	-0.2608	0.2338
POS_modVar	-0.0031	0.0301	0.0747	-0.0609
POS_nounVar	-0.0587	0.0617	-0.1259	0.0203
POS_squaredVerbVar1	0.0634	0.0782	-0.2343	0.22
POS_verbVar1	0.1465	0.0671	-0.2297	0.1234
POS_verbVar2	0.0972	-0.1001	0.0928	0.0611
SYN_CNPerClause	-0.0056	0.1648	-0.0297	0
SYN_CNPerTunit	0.0316	0.1699	-0.0078	0.0187
SYN_ComplexTunitRatio	0.1991	0.055	0.124	0.1089

Table III. Synthetic readability composition

Variable	Comp1	Comp2	Comp3	Comp4
SYN_CoordPerClause	-0.0739	0.1516	-0.0417	-0.0456
SYN_CoordPerTunit	-0.0071	0.1685	-0.0034	-0.0075
SYN_DependentClauseRatio	0.1834	0.0568	0.128	0.1047
SYN_DependentClausesPerTunit	0.1967	0.0437	0.1224	0.1101
SYN_MLC	-0.0178	0.183	-0.078	-0.0805
SYN_MLT	0.14	0.183	0.0155	0.0156
SYN_TunitComplexityRatio	0.1754	0.0179	0.0928	0.1001
SYN_VPPerTunit	0.1892	0.0653	0.0982	0.0399
TRAD_ARI	0.1535	0.222	0.1151	-0.0303
TRAD_Coleman	-0.0872	0.2384	0.0161	-0.0946
TRAD_FOG	0.141	0.2354	0.1151	-0.031
TRAD_FORCAST	0.1195	-0.2127	-0.0103	0.0626
TRAD_Flesch	-0.0725	-0.2818	-0.1035	0.0457
TRAD_Kincaid	0.137	0.2391	0.1155	-0.0303
TRAD_LIX	0.1488	0.2287	0.1133	-0.0405
TRAD_SMOG	0.1172	0.2588	0.103	-0.0444
TRAD_numChars	-0.1247	0.2098	-0.0012	-0.0916
TRAD_numSyll	-0.137	0.2189	0.0097	-0.0571
Word_BilogTTR	-0.2228	0.0322	0.1738	0.0796
Word_CTTR	-0.1956	0.1161	-0.015	0.219
Word_MTLD	-0.0742	-0.058	0.0841	0.0439
Word_RTTR	-0.1957	0.1161	-0.015	0.219
Word_TTR	-0.1723	-0.0239	0.2326	-0.0213
Word_UberIndex	-0.2109	0.1036	0.0299	0.2073

Notes: The first four principal components (eigenvectors) from principal component analysis of 64 linguistic measures are presented. Eigenvectors are orthonormal, which is uncorrelated and normalized. See Figure 3 for eigenvalues.