Research Policy 52 (2023) 104670

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Linguistic metrics for patent disclosure: Evidence from university versus corporate patents☆

Nancy Kong [a,*], Uwe Dulleck [b], Adam B. Jaffe [c], Shupeng Sun [d], Sowmya Vajjala [e]

[a] Queensland University of Technology, The University of Sydney and IZA, 2 George Street, Brisbane City, QLD, 4000, Australia
[b] Centre of Behavioural Economics, Society and Technology (BEST), Queensland University of Technology, Crawford School of Public Policy, Australian National University, and CESifo, LMU Munich, Australia
[c] Brandeis University, Motu Research, and Queensland University of Technology, USA
[d] Queensland Treasury, Australia
[e] National Research Council, Canada

ARTICLE INFO

JEL classification:
K11
O31
O34

Keywords:
Patent disclosure
Computational linguistic analysis
Readability
University patents
Corporate patents

ABSTRACT

Encouraging disclosure is important for the patent system, yet the technical information in patent applications is often inadequate. We use algorithms from computational linguistics to quantify the effectiveness of disclosure in patent applications. Relying on the expectation that universities have more ability and incentive to disclose their inventions than corporations, we analyze 64 linguistic measures of patent applications, and show that university patents are more readable by 0.4 SD of a synthetic measure of readability. Results are robust to controlling for non-disclosure-related invention heterogeneity. The linguistic metrics are evaluated by a panel of "expert" student engineers and further examined by USPTO 112(a) – lack of disclosure – rejection. The ability to quantify disclosure opens new research paths and potentially facilitates improvement of disclosure.

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 resources wasted on duplicate inventions (Hegde et al., 2022); 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 receives limited attention. This raises concerns; 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. 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 or information accessibility. We propose and demonstrate that extant metrics developed in computational linguistics can help to fill this gap.

☆ We thank Andrew A. Toole (Chief Economist of The US Patent and Trademark Office), Nicholas Pairolero (USPTO), Lisa Larrimore Ouellette, Heidi Williams, Monika Schnitzer, Mike Teodorescu, Lesley Millar-Nicholson (Director of The Technology Licensing Office, MIT), and Timothy Oyer (President of Wolf Greenfield Intellectual Property Law), as well as participants at the Innovation Information Initiative Technical Working Group, Chief Economist Speaker Series at USPTO, Technology & Policy Research Initiative IP day at Boston University, and Queensland University of Technology. We appreciate Hamish Macintosh and Yi Wang for their technical support. 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. This project is funded by Australian Research Council Discovery Grant DP180103856.

* Corresponding author.
E-mail address: nancy.kong@sydney.edu.au (N. Kong).

[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).

https://doi.org/10.1016/j.respol.2022.104670
Received 30 March 2021; Received in revised form 26 September 2022; Accepted 11 November 2022
Available online 5 December 2022

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 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. Our proposed linguistic measures are likely to serve as an informative proxy for the explicitly or implicitly chosen level of disclosure. The goal of this article is simply to demonstrate that these measures do appear to capture meaningful differences in accessibility or disclosure, and thereby opening up the possibility of research on the causes and effects of variations in disclosure.

Our strategy for demonstrating the relevance of linguistic readability metrics is to identify a situation in which 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 see this as an indication to treat them as potentially useful. We compare patent applications from universities with those of corporations. Both strategic reasons and the costs of revealing information inform our expectations. From a strategic perspective, universities, with their focus on licensing of patents have an interest in making their patents more accessible. In contrast, corporations (particularly practicing corporations) may benefit from limiting the accessibility of information. From a cost perspective, drafting patents is usually informed by documentation of the relevant research or process of innovation. Given university researchers' primary interest in accessible publications and the relevant standards of documentation, the source material available to an attorney drafting a patent may be much better than in the case of the same attorney drafting a patent for a corporation, in which the need for such documentation is much less. The literature also supports this expectation (Trajtenberg et al., 1997; Henderson et al., 1998; Cockburn et al., 2002).

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,[3] which could further affect the level of disclosure.

Corporations, particularly those with a focus on in-house production, on the other hand, have a greater incentive to obfuscate crucial technical information to deter competitors from understanding, using, and building on their inventions. The profit-maximizing motive, as well as a lack of incentive to thoroughly document the invention, 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 corporations for a particular reason. 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, which indicates the value of further research and the need to reconsider patent examination with respect to the accessibility and disclosure of information contained in patents.

Similar to the finance literature, we use a computational linguistic program designed to assess the reading difficulty of texts using 64 measures from second language acquisition research. The indicators cover the 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; thus, the textual communication required for corporate inventions could differ. To address this concern, our identification strategy employs the following. First, to account for the unobserved heterogeneity in linguistic characteristics intrinsic to technical fields, our econometric method controls for U.S. patent subclass fixed effects. This enables us to measure disclosure as the degree of readability relative to other technologically similar patents. Second, we use patent attorney fixed effects to control for systematic disclosure effects from the drafting agents. This compares the university and corporate patents drafted by the same patent attorney. Third, we employ cited-patent fixed effects with a data compression technique, least absolute shrinkage and selection operator (LASSO), to further control for the nature of inventions. This is because 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. Fourth, to deal with any selection bias from observables, we use a doubly robust estimation that combines propensity score matching and regression adjustment. This enables us to compare university and corporate patents with similar attributes.

Our results show that corporate patents are 0.4 SD more difficult to read and require 1.1–1.6 years more education to comprehend than university patents. We find that the difference is more prominent for more experienced patent applicants, and that licensing corporate patents disclose more than other corporate patents, 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 lies in the provision of many examples in order to conceal the "best mode" of inventions.

This paper is one of the first to specifically use textual analysis to examine patent disclosure (with exception of Dyer et al. (2020) who focus on patent examiners' leniency) and validate the measure. We obtain the whole set of full text patent applications in categories related to nanotechnology, batteries, and electricity from 2000 to 2019, totaling 40,949, and apply our linguistic analysis model to the technical descriptions of these patents. We expand readability studies in related literature that rely heavily on traditional readability indices such as Gunning Fog, Kincaid, and Flesch Reading Ease by including lexical richness, syntactic complexity, and discourse features. We use the best non-commercial readability software (Vajjala and Meurers, 2014b) to capture the multidimensional linguistic features of 64 indicators, and perform a more in-depth linguistic analysis (Loughran and McDonald, 2016) than previous studies. We also use principal components analysis to construct synthetic overall measures of readability.

Having developed this rich set of readability measures, we validate them as indicators of effective patent disclosure by testing whether the lexical measures show patents to be more readable in several real-world

[3] University patents are typically drafted by patent attorneys based on patent disclosures filed by the university inventors. These disclosures often contain large blocks of text copied and pasted from the associated academic articles. It is thus not uncommon for university patents to contain text that originated in a scientific paper.

contexts. Our primary comparison is between university and corporate patents. The licensing aims of universities and absence of market driven competitive motives mean that they have greater incentive to disclose – less incentive to conceal – key information relative to corporations. Through analyses that control for sources of variation in readability, we find that university patents are, indeed, more readable. We support this main analysis with several other comparisons. Intellectual Ventures – a corporation that, akin to universities, seeks to license its patents over competing in the market – also holds patents with above average readability. Several large corporations known to be active patent licensors (IBM, Qualcomm, and HP) similarly exhibit higher readability. Additionally, a set of patents that can be presumed to have been reassigned also exhibit higher readability than otherwise similar patents. Finally, we compared the computational readability measures to subjective evaluations of readability and disclosure for a small number of patents, and assessed the readability of patents rejected by the USPTO for reasons that include failure to adequately disclose the technology.

We see the role of this paper as analogous to Trajtenberg et al. (1997), who first introduced metrics of patent "importance", "generality" and "originality" based on patent citation data. We imitate their strategy to test whether our proposed new measures reveal the contrast we expect between university and corporate patents, and argue that the finding – that they display the predicted pattern – can be taken as initial evidence that they capture meaningful variation in unobservable patent disclosure quality. The introduction and initial validation of these measures open up the possibility of quantitative treatment of extent of disclosure in patents, both for social science research on the sources and effects of better or worse disclosure, and potentially for use in more systematic treatment of the disclosure obligation in the patent examination process.

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. We examine attorney fixed effects and cited-patent fixed effects in Section 6, and one channel that corporations could use to obscure patent applications in Section 7. We show heterogeneous effects in Section 8 and usefulness tests in Section 9, and conclude in Section 10.

2. Literature review

2.1. Textual analysis

Textual analysis has only recently been used in the economic literature. For example, Gentzkow et al. (2019) propose a practical overview of textual analysis and statistical analysis using text as data, and Hansen et al. (2018) examine the effects of transparency in central banks on monetary policies using a statistical model for content analysis.[4]

Similarly, computational linguistics has only recently been used in patent research. For example, Younge and Kuhn (2016), Arts et al. (2018), Whalen et al. (2020) and Helmers et al. (2019) use textual analysis to examine patent similarity. 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.[5] However, most of these studies focus only on the technologies 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.

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". That is, the technical description must meet the requirements for (1) written description, (2) enablement, and (3) best mode. Outside the US, the World Trade Organization states that "members shall require that an applicant for a patent shall disclose the invention in a manner sufficiently clear and complete for the invention to be carried out by a person skilled in the art and may require the applicant to indicate the best mode for carrying out the invention known to the inventor at the filing date or, where priority is claimed, at the priority date of the application" (see Article 29 of the Agreement on Trade-Related Aspects of Intellectual Property Rights). The European Patent Convention also has a requirement that "the European patent application must disclose the invention in a manner sufficiently clear and complete for it to be carried out by a person skilled in the art" (see Article 83).

A number of papers investigate the effects of disclosure in patents by studying the effects of the publication of patent applications. For example, Baruffaldi and Simeth (2020) show that early publication increases forward citation counts, and Hegde et al. (2022) find that speedy publication reduces duplicative R&D. These papers demonstrate that some kind of consequential information is revealed by the publication of a patent application, but do not explore how the degree of disclosure varies in different patent texts. Indeed, in these papers it is impossible to know whether the consequences of publication flow solely from the revelation of the mere existence of an application, or are also conditioned by the nature and extent of specifics about the invention that are revealed. The development and validation of metrics for the effectiveness of disclosure in the patent text would allow for a much richer exploration of these questions.

Coming from the other direction, Dyer et al. (2020) calculate metrics for the effectiveness of disclosure in patent texts that are related to those we propose here. They use these disclosure metrics to identify "lenient" examiners, which they take to be examiners who allow patents with low disclosure levels. In doing this, they assume that the metrics capture poor disclosure, but do not attempt to demonstrate

[4] A limited number of studies employ textual analysis to examine gender discrimination in the publication and job market process. For instance, Hengel (2022); Card et al. (2020); and Wu (2018).

[5] Also, see Teodorescu (2017) for a comprehensive survey of the natural language processing method used in strategic research.

the validity of the metrics for this purpose. We extend the metrics they use to 64 linguistic measures using state-of-the-art computational linguistic algorithms, and then validate the meaningfulness of the measures as proxies for disclosure by showing that they conform to a priori expectations about the difference in disclosure effectiveness between universities and corporations, while carefully controlling for other differences between the patents.

2.3. University and corporate patents

We choose to compare patents filed by universities and corporations because they have different business models for patenting. As Cockburn et al. (2002) suggested, university or public sector patents are written less strategically than those of private corporations, and this may partially explain why university patents are more highly cited than those by private corporations (Henderson et al., 1998). 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.

By 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., to 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 than 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 standardized rules across grant-making agencies and accelerated changes to make patenting by universities easier and more routine (Sampat, 2006). It explicitly renders 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".[6]

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, written to communicate the results, which may have been subject to review and editing designed to increase its readability. Corporate patents, in contrast, are typically drafted based on a disclosure written by the inventors. The availability of a previously written scholarly paper may provide a basis for patent drafting that intrinsically leads to greater readability.

Despite the aforementioned incentives that render universities likely to disclose more effectively than corporations, there may be situations in which universities have incentive to obscure information; for instance, their diminished legal capacity for monitoring and acting on infringement may lead to less disclosure. Conversely, some corporations may have an incentive to offer more transparent disclosure in licensing patents. Of exclusive importance to our validity test is that, on average, universities have a greater incentive to disclose. To the extent that this difference is diminished as a result of particular case-by-case counter-incentives, it would be more difficult to find significant differences between universities and corporations with our metrics. We do believe, however, that this does not undermine the validity of the conclusions if the expected difference is found.

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

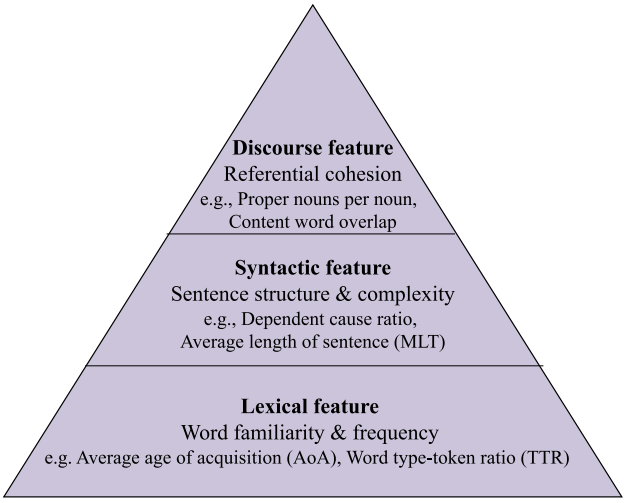

Fig. 1. Hierarchy of linguistic analysis.
Note: This hierarchy of linguistic analysis is derived from "Key aspects of text readability" from Collins-Thompson (2014) and Loughran and McDonald (2016). We selected the relevant and feasible levels of analysis in a patent context.

3. Linguistic measures

We use the readability assessment program developed by Vajjala and Meurers (2014b), which has been shown to be the best non-commercial readability assessment approach for English (Vajjala and Meurers, 2014b) and 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 traditional measures, such as Gunning Fog and Flesch–Kincaid grade level, the rest of the measures from this program are divided into three categories: lexical features, syntactic features, and discourse features (see Fig. 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.[7] 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 1 presents the definitions, interpretation, implications, and sources of representative 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

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

Table 1
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 2 to 6. ⊕ indicates a positive relationship with "hard to read" in the linguistic literature, and ⊖ indicates a negative relationship. See the Appendix B 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).

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 average length of phrases. We use *dependent clauses to total clauses ratio* and the mean length of T-unit (*MLT*)[8] as the representative measures for this category.

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

Table 2 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*.[9]

4. Empirical strategy

4.1. Data

Using the Lens database,[10] 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. These areas have been marked by a high degree of innovation over the past two decades, and have been studied by Ouellette (2017, 2011).[11] We exclude headers and claims,[12] and strip technical description text files 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, and NPL resolved citation count.[13]

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, 21,234

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

[9] For further information on the computational linguistic program, see Vajjala and Meurers (2014b); The code is openly available at https://bitbucket.org/nishkalavallabhi/readinglevelpredictor.

[10] 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.

[11] We also conduct the same analysis on a set of biomedical patents (see Table B.1 in Appendix B). The results show that university patents are easier to read than corporate patents, with a slightly smaller magnitude.

[12] We were advised by practitioners that claims are usually written in standardized legal terms, and would therefore be less likely to reveal differences in clarity of exposition. We have, however, reproduced the results reported below including claims text along with the descriptions in the analysis. The results (not reported) are qualitatively very similar to those reported.

[13] The NPL resolved citation count is a metric constructed by the Lens that purges the NPL citation count of references that cannot be identified as scholarly articles. It is generated by matching NPLs to PubMed and Crossref metadata (see Jefferson et al., 2018, for more information).

Table 2
Summary statistics of representative variables.

Categories	Variables	(1) Universities'		(2) Corporations'		(3) Differences	
		Mean	SD	Mean	SD	Mean	t-stat
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_Resolved_Citation_Count	0.82	1.82	0.19	0.83	0.63***	(19.82)
	Number_Inventors	3.42	1.80	3.01	1.93	0.41***	(12.36)
	ClaimCounts	27.86	23.38	22.27	20.57	5.59***	(13.17)
	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. ⊕ indicates a positive relationship with "hard to read" in the linguistic literature, and ⊖ indicates a negative relationship. The full sample summary statistics, including patents jointly filed by universities and corporations as well as by other entities, are presented in Table B.2 in Appendix B. Detailed information on the sample is provided in Section 4.1. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

from corporations, 1,644 jointly filed by universities and corporations, and 14,657 filed by other entities, such as individuals and government organizations.

We then apply Vajjala and Meurers' 2014b. computational linguistic model to the 40,949 full-text patent applications using a high performance computing platform,[14] and apply the 64 linguistic measures to each application. Table 2 presents summary statistics for the metadata and linguistic measures of patent applications filed by universities and corporations. All nine representative indicators differ significantly for universities and corporations.[15] It also shows that most characteristics, such as citation counts, family size, number of inventors, and number of claims are significantly different. We control for these observed differences between universities and corporations.

We use PCA to consolidate our 64 linguistic measures. PCA is a non-parametric statistical technique primarily used to reduce dimensions; it explores the highest variability in variables.[16] For easy interpretation, PCA components are standardized.

4.2. Baseline estimation

We estimate the following OLS regression:

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

where Y_{ij} is the PCA measure or one of the 64 linguistic indicators of application i in subclassification j; $Corp = 1$ if the patent application is filed by a corporation, and 0 otherwise; $Joint$ is a dummy variable if a patent is jointly filed by a corporate and a university; $Other$ captures the rest of the patents (i.e., patents by government agencies, and individuals); and university patents are the base.

Hsu et al. (2021) noted that university patents differ from corporate patents in forward-citation counts and international family sizes. We therefore explicitly control for these variables: $\boldsymbol{X_{ij}}$ is a vector of forward-citation counts, simple and extended family sizes, sequence count, NPL resolved citation count, number of inventors, and claim counts; δ_j is U.S. patent subclassification fixed effects; η_t is the application year fixed effects; and ϵ_{ij} is the error term clustered at 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 enable us to estimate the difference within the finest possible technical area, and in effect account for area-specific competition.[17] We include the number of claims in the patents as a proxy for the breadth of the knowledge contained in the patent applications, and the number of inventors as a proxy for the depth and/or complexity of the knowledge embedded in the underlying invention. Application year fixed effects are included to control for any time-specific effects.[18] The hypothesis is that β_1 is significant and positively correlated with "hard to read" indices compared with university patents.

[14] 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.

[15] Summary statistics of the 64 variables in the full sample include joint patents and other patents; see Table B.2 in Appendix B.

[16] We present components of the linguistic variables in Table B.3 in Appendix B. Fig. B.1 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). Table B.4 shows the estimates of corporate patents using components 2 to 4. 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.

[17] According to our PCA measures, the readability ranks in the order of nanotechnology, photoelectric, and batteries. Specifically, nanotechnology is 1.90 SD harder to read than photoelectric, which is 1.92 SD harder to read than batteries. These technical area effects are captured by the sub-USPC fixed effects in our estimation.

[18] We also add a control for domestic/foreign patents, and the results are similar.

Table 3
Baseline estimates from representative indicators.

Features	Synthetic	Traditional			Lexical		Syntactic		Discourse	
Variables	PCA	Fog ⊕	Flesch ⊖	Kincaid ⊕	AoA_Kup	Word_TTR	DependentClauseR	MLT	ProperNounsPerNoun	ContentWordOverlap
Raw Y										
Corporate	1.429***	1.418***	−4.410**	1.622***	0.0246*	−0.0152**	0.0171***	0.422***	−0.0186***	123.6**
	(0.00362)	(0.132)	(0.557)	(0.126)	(0.00698)	(0.00216)	(0.00109)	(0.0298)	(0.00124)	(21.30)
Standradized Y										
Corporate	0.420***	0.211***	−0.226**	0.249***	0.0867*	−0.333**	0.205***	0.213***	−0.471***	0.245**
	(0.00106)	(0.0197)	(0.0285)	(0.0193)	(0.0246)	(0.0472)	(0.0130)	(0.0151)	(0.0313)	(0.0422)
Control variables from standardized Y										
Joint	0.394**	0.00469	0.0856	0.00603	0.215	−0.314***	−0.0693	−0.0341	−0.440**	0.159***
	(0.0406)	(0.0916)	(0.109)	(0.0996)	(0.0829)	(0.0180)	(0.115)	(0.0661)	(0.0586)	(0.00578)
Others	0.306***	0.203***	−0.222***	0.236***	−0.0107	−0.112*	0.256**	0.175**	−0.393***	0.137*
	(0.0254)	(0.00875)	(0.00924)	(0.0111)	(0.0622)	(0.0285)	(0.0481)	(0.0290)	(0.00174)	(0.0320)
Forward Citations[a]	−0.184*	−0.224	0.127	−0.197	0.169*	−0.340	−0.0979	−0.0727	0.154*	0.340**
	(0.0477)	(0.103)	(0.0951)	(0.0951)	(0.0525)	(0.128)	(0.0471)	(0.0300)	(0.0505)	(0.0735)
Simple Family Size[a]	0.843	1.034*	−0.810*	1.005*	−0.164	−0.896*	0.805*	0.914**	−0.207***	0.872**
	(0.331)	(0.281)	(0.196)	(0.271)	(0.292)	(0.287)	(0.203)	(0.191)	(0.0196)	(0.144)
Extended Family Size[a]	−0.142*	−0.191***	0.125***	−0.181***	−0.0890*	−0.155	−0.134**	−0.140*	0.196***	0.289**
	(0.0422)	(0.0173)	(0.00306)	(0.0127)	(0.0297)	(0.0658)	(0.0299)	(0.0333)	(0.0156)	(0.0295)
Sequence Count[b]	−1.749***	−1.085***	0.849***	−0.927***	−1.213***	0.946**	0.349*	−2.536***	2.248***	−0.393**
	(0.0768)	(0.0296)	(0.0486)	(0.0366)	(0.0411)	(0.131)	(0.0840)	(0.0995)	(0.0383)	(0.0625)
NPL Resolved Citations[a]	−4.275**	−0.390**	−0.0452	−0.671***	−0.203	2.004	−1.674***	−0.202*	3.961**	−0.424
	(0.837)	(0.0687)	(0.284)	(0.0365)	(0.693)	(0.893)	(0.0557)	(0.0497)	(0.809)	(1.229)
Number of Inventors[a]	1.381**	1.928**	−1.417	1.695*	1.420	−4.026**	−0.262	3.448**	1.224**	3.421**
	(0.192)	(0.443)	(0.513)	(0.433)	(0.528)	(0.598)	(0.372)	(0.461)	(0.266)	(0.556)
	(0.00192)	(0.00443)	(0.00513)	(0.00433)	(0.00528)	(0.00598)	(0.00372)	(0.00461)	(0.00266)	(0.00556)
Claim Counts[a]	−0.275	−0.288	0.111	−0.261	−0.120	−0.577***	−0.201	−0.143	0.189**	0.714*
	(0.135)	(0.115)	(0.0732)	(0.110)	(0.0539)	(0.0341)	(0.121)	(0.150)	(0.0217)	(0.199)
Observations	40,949	40,949	40,949	40,949	40,949	40,949	40,949	40,949	40,949	40,949
R-squared	0.295	0.058	0.077	0.056	0.152	0.176	0.178	0.069	0.183	0.115
Sub-USPC & Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS estimates of corporate patents are obtained from Eq. (1), using university patents as the base. Estimations control for application year fixed effects and U.S. patent subclassification fixed effects. Both raw linguistic measures and standardized linguistic measures (mean = 0, $SD = 1$) are used as dependent variables. ⊕ indicates a positive relationship with "hard to read" in the linguistic literature, and ⊖ indicates a negative relationship. Standard errors are clustered at U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

[a]Scaled by multiplying by 0.01.

[b]Scaled by multiplying by 0.000001.

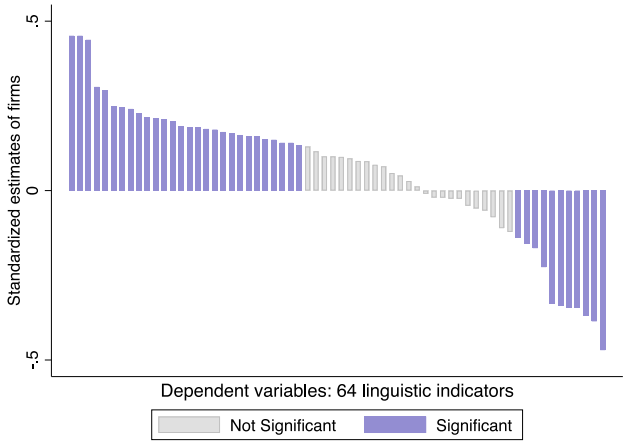

Fig. 2. Baseline estimates of corporate patents plotted with significance using multiple hypothesis testing.
Note: The Y-axis indicates the estimates from Table B.5 using Eq. (1). Each bar represents one linguistic measure. Significance is defined as $p < 0.1$. Multiple hypothesis testing uses Romano 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.

5. Results

Table 3 presents estimates for corporate patents on representative individual linguistic measures. We show both the estimates using raw linguistic scores for magnitude interpretations (first row) and standardized linguistic scores for easy comparison across different measures (the rest of rows).

The PCA measure shows that corporate patents are 0.42 SD different from their university counterparts. The Fog and Kincaid measures both correspond with the years of education required to understand the text; their estimates are 1.4 and 1.6, respectively. This 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 estimate of −4.4 indicates harder to read texts for corporate patents.

In terms of lexical feature, corporate patents have words with higher age of acquisition (9% SD) and lower unique word ratio (33% SD). For syntactic feature, corporate patents have more dependent clauses (20% SD) and longer T-units (21% SD). For discourse feature, corporate patents have fewer proper nouns (47% SD) and more content word overlap (25% SD). From these representative linguistic measures, discourse feature seems to suggest the largest difference between corporate and university patents.

We also present the estimates from the full set of 64 outcome indicators in Table B.5 in Appendix B, applying multiple hypothesis testing (Romano and Wolf, 2005) with stepdown adjusted p-values that enable strong control of the family-wise error rate. Fig. 2 shows the 64 estimates by significance, and 39 linguistic indicators are significant. This means that the linguistic measures effectively capture the differences in patent applications between universities and corporations.

Additionally, to test for potential selection bias from observables, we follow Wooldridge (2010) and Imbens and Rubin (2015) and conduct doubly robust estimations that combine propensity score matching and regression adjustment. This method offers a desirable property—as long as either PSM or the regression is correctly specified, we

Table 4
Robustness checks: Panel A: high-degree patent attorney fixed effects; Panel B: high-degree cited-patent fixed effects using LASSO.

Features	Synthetic	Traditional			Lexical		Syntactic		Discourse	
Variables	PCA	Fog ⊕	Flesch ⊖	Kincaid ⊕	AoA_Kup	Word_TTR	DependentClauseR	MLT	ProperNounsPerNoun	ContentWordOverlap
Panel A: Patent attorney fixed effects										
Raw Y										
Corporate		1.017**	−2.623**	1.164***	0.0255***	−0.0136**	0.0125	0.314***	−0.0159***	109.9**
		(0.105)	(0.424)	(0.101)	(0.00245)	(0.00246)	(0.00443)	(0.0256)	(0.00117)	(18.57)
Standardized Y										
Corporate	0.394***	0.152**	−0.134**	0.179***	0.0900***	−0.296**	0.15	0.159***	−0.403***	0.218**
	−0.0381	−0.0157	−0.0217	−0.0155	−0.00864	−0.0537	−0.0532	−0.0129	−0.0298	−0.0368
Observations	15,275	15,275	15,275	15,275	15,275	15,275	15,275	15,275	15,275	15,275
R-squared	0.510	0.263	0.247	0.255	0.238	0.370	0.390	0.270	0.322	0.270
Panel B: Cited patent fixed effects										
Raw Y										
Corporate		1.109***	−3.290***	1.261***	0.0269***	−0.0167***	0.0182***	0.372***	−0.0194***	131.2***
		(0.142)	(0.449)	(0.139)	(0.00777)	(0.00114)	(0.00203)	(0.0522)	(0.00151)	(10.55)
Standardized Y										
Corporate	0.445***	0.165***	−0.169***	0.194***	0.0948***	−0.364***	0.218***	0.188***	−0.491***	0.260***
	(0.0235)	(0.0212)	(0.0230)	(0.0213)	(0.0274)	(0.0248)	(0.0244)	(0.0264)	(0.0382)	(0.0209)
Observations	20,571	20,571	20,571	20,571	20,571	20,571	20,571	20,571	20,571	20,571

Note: Estimations control for joint patents and other patents (using university patents as the base), various citation counts, simple and extended family size, number of inventors, claim counts, application year fixed effects and U.S. patent subclassification fixed effects. Panel A adds patent attorney fixed effects and Panel B uses cited patent fixed effects (LASSO). Both raw linguistic measures and standardized linguistic measures (mean = 0, $SD = 1$) are used as dependent variables. ⊕ indicates a positive relationship with "hard to read" in the linguistic literature, and ⊖ indicates a negative relationship. Standard errors are clustered at U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

can obtain correct estimates of treatment effect (i.e., corporate patent estimate). The results are highly consistent to the baseline estimation (see Table B.6).

6. Robustness checks

6.1. Patent attorney fixed effects

Patent applications for both corporations and universities are generally drafted by outside patent attorneys. We would expect that some patent attorneys systematically draft more-, but also, some draft less-readable patents. Of course, the choice of patent attorney is not random, so patenting entities that wish to disclose could choose attorneys who write readable patents, and vice versa. The readability is also affected by applicant behavior that the attorneys do not control, such as the information provided to attorneys. But other factors affect which attorneys are hired; legal budgets vary between universities and corporations, and thus attorney quality could contribute to the differences. In this section, we control for patent attorney fixed effects by linking full-text patent applications to patent attorney data from *PatentsView*.[19]

Our linked data are at the patent application level and have attorney dummies. To perform attorney fixed effects and to ensure that we have enough observations in each fixed-effect group, we include only "frequently hired attorneys" in the data. We present results using the cut-off of attorneys who have drafted 10 or more patents,[20] which reduces the number of fixed-effect groups from 4,927 to 450.

Table 4 Panel A shows the results: After controlling for attorney fixed effects, the effect size is reduced slightly but significance remains. Compared with standardized outcomes in baseline estimation (averaging 0.25 SD), after controlling for patent attorney fixed effects the standardized scores have an average effect size of 0.20 SD.[21] Specifically, corporate estimate for Fog is about 1.0 year, and for PCA is about 0.39 SD. The magnitudes are slightly smaller than those of the baseline, which are about 1.4 years and 0.42 SD. Thus, at least for this subsample of the overall data, a portion of the difference between corporations and universities can be associated with the identity of the attorneys chosen; however, most of the effect is not explained by the choice of attorney.

6.2. Cited-patent fixed effects

Even though we are comparing patents in the same patent classes, there may be a residual concern that the observed systematic difference between university and corporate patents stems from unobserved differences in the nature of the inventions rather than different choices regarding expression. To address this concern, we propose that patent applications that cite the same previous patent tend to be similar inventions. If one is filed by a university and the other by a corporation, the difference is more likely to arise from the entity rather than the invention. Indeed, most patents cite many previous patents, and we can allow for a fixed effect for each cited patent. Two patents that both cite multiple overlapping previous patents are likely to cover very similar inventions. Therefore, we ask whether the estimated differences between university and corporate patents change materially after controlling for the intrinsic nature of inventions in this way.

The empirical challenges 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, and one patent could belong to multiple citation fixed-effect groups; and (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

[19] Specifically, we obtain the application dataset (information on applications for granted patents) and patent lawyer dataset (metadata table for many-to-many relationships) from https://www.patentsview.org/download/. We matched Lens data using USPTO ID to obtain patent ID (24,836 matched, since the PatentsView download version does not include non-granted patent applications in the current release), then using patent ID to match patent attorney data (m:m) and obtain attorney ID (23,303 matched). By keeping only patents written by patent attorneys who have filed at least 10 patents, 9,766 observations are dropped. We reshaped the data to patent application level and matched 15,275 patent applications.

[20] Other cut-offs produce consistent results, which are available upon request.

[21] We calculate the average effects in Table 3, Panel B, second row, and Table 4, Panel B, second row, excluding PCA. Absolute values are used.

Table 5
Estimates of example frequencies.

Features	Synthetic	Traditional			Lexical		Syntactic		Discourse	
Variables	PCA	Fog ⊕	Flesch ⊖	Kincaid ⊕	AoA_Kup	Word_TTR	DependentClauseR	MLT	ProperNounsPerNoun	ContentWordOverlap
Num_examples[a]	10.50**	6.072	−107.9*	10.32*	0.544	−1.715**	−0.0679	36.50**	−0.0180	40,560**
	(2.203)	(0)	(26.27)	(3.169)	(0.734)	(0.375)	(0)	(6.740)	(0.0190)	(7,733)
Corporate	0.411***	1.413	−4.324**	1.614***	0.0242*	−0.0139**	0.0172	0.393***	−0.0186***	91.47**
	(0.00194)	(0)	(0.562)	(0.126)	(0.00748)	(0.00190)	(0)	(0.0245)	(0.00122)	(17.14)
Observations	40,857	40,857	40,857	40,857	40,857	40,857	40,857	40,857	40,857	40,857
R-squared	0.295	0.056	0.076	0.054	0.150	0.210	0.177	0.076	0.181	0.280

Notes: *Num_examples* is the frequency of "for example" and "e.g." in technical descriptions in patent applications. Estimates of corporate patents are presented with university patents as the base. All estimations control for joint patents and other patents (using university patents as the base), various citation counts, simple and extended family size, number of inventors, claim counts, application year fixed effects, and U.S. patent subclassification fixed effects. ⊕ indicates a positive relationship with "hard to read" in the linguistic literature, and ⊖ indicates a negative relationship. Standard errors are clustered at U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
[a]Scaled by multiplying by 0.0001.

variables. LASSO performs the following estimation:

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

We limit the sample to patents that cite "highly cited patents" (≥ 10 citations), and there are 20,571 patent applications with 6,163 fixed-effect groups.[22] We perform LASSO linear "post-double-selection" inference model (Belloni et al., 2014):

$$Y_{ijk} = \alpha + \beta_1 Corp_{ijk} + \beta_2 Joint_{ijk} + \beta_3 Other_{ijk} + \lambda \boldsymbol{X_{ijk}} + \eta_t + \delta_j + \sum_{k=1}^{6163} \gamma_k + \epsilon_{ijk}, \quad (3)$$

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

Table 4 Panel B 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 shows a very similar magnitude (0.45 compared with 0.42), and the traditional measures have slightly reduced magnitudes relative to the baseline estimation: The raw Fog measure decreased from 1.4 to 1.1; the Kincaid from 1.6 to 1.3; and the Flesch from −4.4 to −3.3. 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 1.3 more years of education to comprehend.

7. A possible channel

In this section, we explore a strategy that corporations may use that could partially explain the differences in readability and disclosure.[23] As a matter of patent law, the so-called "best mode" rule specifies that if there are multiple 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 to minimize disclosure is by burying 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.[24] We extract *Num_examples*, the occurrences of "for example" 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 $t-stats = -2.36$. When we look at the overall distribution of the number of examples for the two groups, university patents are systematically under-represented among patents with 10 or fewer examples, and systematically over-represented among patents with 10–60 examples.

We add the *Num_examples* to Eq. (1) as an independent variable.[25] Table 5 shows that the number of examples is positively correlated with the synthetic variable, Kincaid, MLT, and content word overlap, and negatively correlated with Flesch and *TTR*. In general, *Num_examples* mostly corresponds 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 this is reflected in our linguistic measures.[26]

8. Heterogeneous effects

8.1. Top applicants

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 top applicants (10.1% are university patents), and the rest are in the "other" category (7.4% are university patents).

We estimate Eq. (1) separately for the top 100 applicants and the rest. The results are presented in Table 6, Panel A. 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 0.66 SD harder to read than university patents among top applicants, compared with 0.25 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

[22] 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-effect estimation without LASSO, and the results are highly consistent.

[23] We thank an anonymous practicing patent attorney who identified such strategy. We do not claim our results suggest a causal relationship, however.

[24] The legal status of the "best mode" rule was changed in 2011, which eliminated the threat of invalidity for failure to disclose the best mode. There are, however, still reasons to disclose the best mode in a patent, including to ensure that the patent's claims are construed to cover the best mode. The benefits of burying the best mode in a bunch of examples remain the same. For these reasons, we believe that the best-mode effect we discuss in this section was likely present throughout our study period.

[25] We also log transform *Num_examples*; results are presented in Table B.7.

[26] Other jurisdictions do not have a "best mode" requirement like the one in the U.S. To the extent that the differences in readability we find are, in fact, due to this requirement, this difference might therefore be smaller for patents granted in other jurisdictions.

Table 6
Heterogeneous effects by top applicants, corporate patent estimates (Panel A); licensing corporate patent estimates (Panel B); and Intellectual Ventures matching estimates (Panel C).

Features	Synthetic	Traditional			Lexical		Syntactic		Discourse	
Variables	PCA	Fog ⊕	Flesch ⊖	Kincaid ⊕	AoA_Kup	Word_TTR	DependentClauseR	MLT	ProperNounsPerNoun	ContentWordOverlap
Panel A: Top applicants										
Corporate	0.657***	2.226**	−6.006**	2.408**	0.0166	−0.0202***	0.0300*	0.714**	−0.0262***	171.4**
	(0.0231)	(0.293)	(0.895)	(0.326)	(0.0123)	(0.00125)	(0.00737)	(0.109)	(0.00244)	(24.76)
Observations	11,844	11,844	11,844	11,844	11,844	11,844	11,844	11,844	11,844	11,844
R-squared	0.378	0.111	0.120	0.111	0.215	0.225	0.277	0.121	0.255	0.144
Panel A: Others										
Corporate	0.254***	0.947*	−3.675*	1.190**	0.0260**	−0.0104**	0.00900	0.212**	−0.0140***	86.56**
	(0.0121)	(0.269)	(1.049)	(0.270)	(0.00449)	(0.00215)	(0.00359)	(0.0377)	(0.000930)	(19.32)
Observations	29,105	29,105	29,105	29,105	29,105	29,105	29,105	29,105	29,105	29,105
R-squared	0.274	0.056	0.078	0.053	0.134	0.165	0.160	0.069	0.180	0.117
Panel B: Licensing corporate patents (corporate patent sample)										
LicenseCorp	−0.278	−2.565***	4.994**	−2.400***	0.0133	0.00462**	−0.00978	−0.595	−0.00434	−126.1***
	(0.184)	(0.238)	(1.063)	(0.232)	(0.0495)	(0.000683)	(0.00785)	(0.256)	(0.00223)	(7.435)
Observations	21,452	21,452	21,452	21,452	21,452	21,452	21,452	21,452	21,452	21,452
R-squared	0.243	0.054	0.080	0.051	0.161	0.104	0.174	0.062	0.160	0.049
Panel C: Intellectual Ventures and matching patents										
Intellectual Ventures	−0.373***	−0.499**	−0.658	−0.327	−0.0753***	0.0158***	−0.0115*	−0.458***	0.00267***	−128.0**
	(0.0827)	(0.207)	(0.991)	(0.209)	(0.0215)	(0.00337)	(0.00626)	(0.136)	(0.000843)	(50.10)
Observations	2,022	2,022	2,022	2,022	2,022	2,022	2,022	2,022	2,022	2,022
R-squared	0.010	0.003	0.000	0.001	0.006	0.011	0.002	0.006	0.005	0.003

Notes: Panel A presents corporate patent estimates. The top 100 applicants are defined by patent application counts in the sample. Row 1 shows estimates from the top applicants sample and Row 2 uses the rest of the sample. Estimations follows Eq. (1) and controls for joint patents and other patents (using university patents as the base), various citation counts, simple and extended family size, number of inventors, claim counts, application year fixed effects, and U.S. patent subclassification fixed effects. Panel B presents licensing corporate patent estimates (with "other corporate patents" as the base) and estimations use controls above. Panel C presents propensity score matching estimates of Intellectual Ventures patents and their matching patents using entity types, various citation counts, simple and extended family size, number of inventors, claim counts, application year, and U.S. patent subclassification. ⊕ indicates a positive relationship with "hard to read" in the linguistic literature, and ⊖ indicates a negative relationship. Standard errors are clustered at U.S. patent classification level in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

to read than university patents for top applicants (compared with 0.9 and 1.2 for other applicants), which means that the readability gap is 2 to 2.4 times between top applicants and 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 for top corporations versus 23.3 for other corporations). In general, we would expect that corporations 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 corporations for the most experienced applicants as reinforcing the interpretation that the measured differences are indicative of the different strategic objectives of universities and corporations. Another explanation is that patent experience is correlated with the size of the entity. New corporations may act more like universities because their goal is often to be bought out. Therefore, we observe a smaller gap between less experienced corporations and universities.[27]

8.2. Licensing corporations

Our a priori expectation of better disclosure by universities is based on the fact that they seek to license their patents rather than practice them. By the same reasoning, corporations that seek to license rather than practice their patents ought to exhibit better disclosure than other corporations. While a systematic characterization of corporations in terms of their licensing stance is beyond the scope of this paper, as a rough test we identified Qualcomm, HP and IBM as corporations that are known to license a large proportion of patents[28]; these corporations hold 4.3% of the corporate patents in our sample. We regress the linguistic measures on a dummy variable for *licensing corporates* in the corporate patent sample, and the results are presented in Table 6, Panel B. It shows that the patents of these licensing corporations are indeed more readable than those of other corporations; for example, 0.28 SD lower in the synthetic hard-to-read measure, which aligns with expectations.

We also conduct another robustness check on licensing corporations. We select Intellectual Ventures, a company whose business model is to license patents (Hagiu et al., 2009). The hypothesis is that companies whose mission is to monetize intellectual properties would invest in creating and purchasing more readable patents. We obtain full-text patents owned by Intellectual Ventures as well as their matching patents identified by their application month and sub-USPC. We conduct propensity score matching using all patent attributes presented in the baseline estimation, and results strongly support our hypothesis

[27] We conduct two tests to examine whether the top applicants' results are driven by a small number of entities. First, we exclude Qualcomm, HP and IBM from the top applicants; the corporate estimates among the rest of the top applicants is 0.68 and is significant at 1% level. Second, we conduct a simulation exercise by dropping each top applicant at a time and repeating the process 100 times. The distribution of point estimates for PCA and Fog are plotted in Fig. B.2 which shows results are concentrated in the OLS point estimate indicated by the red vertical line.

[28] According to the 10-K reports of Qualcomm, HP and IBM in 2019, licensing of intellectual property generates at least $400 million annually for each company.

Table 7
Estimates of expert-evaluated readability and disclosure on selected linguistic measures based on experts' comments.

Variables	Correlation with readability	(1)		(2)	
		Evaluated disclosure (1-5)	Evaluation supports measure	Evaluated readability (1-5)	Evaluation supports measure
Fog	-	−0.0325		−0.175***	Supportive
		(0.0297)		(0.0367)	
Flesch	+	0.0134		0.172***	Supportive
		(0.0341)		(0.0426)	
Kincaid	-	−0.0218		−0.163***	Supportive
		(0.0303)		(0.0377)	
Mean length of clauses	-	0.0937**	Not supportive	0.0892*	Not supportive
		(0.0375)		(0.0495)	
Dependent clause ratio	-	0.0407		−0.239***	Supportive
		(0.0405)		(0.0499)	
MRC age of acquisition	-	−0.0769		−0.350***	Supportive
		(0.0514)		(0.0621)	
MRC word concreteness	+	0.0396		−0.0734	
		(0.0449)		(0.0586)	
Global argument overlap count	+	0.210**	Supportive	0.211*	Supportive
		(0.0843)		(0.111)	
Global content word overlap	+	−0.0478		−0.204**	Not supportive
		(0.0642)		(0.0827)	
Global noun overlap count	+	0.236***	Supportive	0.222**	Supportive
		(0.0795)		(0.105)	
Global stem overlap count	+	0.207**	Supportive	0.218*	Supportive
		(0.0846)		(0.111)	
Local argument overlap count	+	0.311***	Supportive	0.0310	
		(0.0444)		(0.0659)	
Local content word overlap	+	−0.0464		−0.316***	Not supportive
		(0.0759)		(0.0964)	
Local noun overlap count	+	0.311***	Supportive	0.0307	
		(0.0439)		(0.0653)	
Local stem overlap count	+	0.335***	Supportive	0.0531	
		(0.0445)		(0.0671)	

Notes: N = 258. The linguistic measures are selected according to experts' comments. The dependent variables are expert-evaluated disclosure and readability on 5-point scales. The linguistic measures are standardized and selected according to experts' comments. Whether the linguistic measure supports rejections is given only for significant measures. All estimation follows Eq. (1) and controls for joint patents and other patents (using university patents as the base), various citation counts, simple and extended family size, number of inventors, claim counts, application year fixed effects, and U.S. patent subclassification fixed effects. Standard errors are clustered at U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

(see Table 6, Panel C). Intellectual Ventures patents are 0.37 SD easier to read and require 0.5 years less education to comprehend than their matching patents.[29]

9. Testing the usefulness of linguistic measures

Our results show that university patents are more readable, as estimated using standard computational linguistic measures, than corporate patents, and that corporate patents that follow a licensing strategy are more readable than other corporate patents. Given our maintained hypothesis that corporations that do not license have more incentive to restrict disclosure than do entities that intend to license, this provides inferential evidence that linguistic measures of readability are a (noisy) indicator for effective disclosure. We recognize, however, that the linguistic measures were developed for a different purpose. Disclosure in the patent context means effective conveyance of the key technical aspects of the patent, such that skilled practitioners of the relevant technology are able to understand and build upon the patented invention. We believe that the university/corporate contrast we have demonstrated suggests that linguistic readability is useful as an indicator of disclosure, but we also undertook to directly test whether it is correlated with disclosure in the specific patent sense.

First, we recruited high-degree research engineering students who are experts in the relevant disciplines to read patents and provide two subjective Likert scale ratings: one of how readable the patent is, and one of how effectively the patent discloses the technical information regarding the invention.[30] We had 18 students on the expert panel, each of whom reviewed 16 patents, so we were able to have 96 patents reviewed by 3 different experts. We found that the students' subjective evaluations of "technical disclosure" and "readability" were significantly correlated (Pearson correlation=0.48). We then regressed both the expert-evaluated readability and expert-evaluated effectiveness of disclosure on our linguistic measures, controlling for expert fixed effects and patent attributes. We found that some measures (Fog, Flesch, dependent clause ratio, MRC age of acquisition, global argument, noun and stem overlap) exhibit the predicted relationship with expert-evaluated readability. Somewhat fewer measures (global and local argument, noun and stem overlap), exhibit the predicted relationship with expert-evaluated disclosure (see Table 7). Overall, the linguistic readability measures are modestly predictive of students' assessments of readability and disclosure. This survey has its limitations, such as the small sample size, the fact that most of the students (like engineering students at many universities) were not native English-speakers, and our inability to control how much care or effort the students put into their evaluations. Nonetheless, it does suggest that our

[29] We also find reassigned patents are easier to read than the others (see Table B.8). To do so, we match our main sample of analysis to USPTO assignment data https://www.uspto.gov/ip-policy/economic-research/research-datasets/patent-assignment-dataset. We identify reassigned patent = 1 if "conveyance type" is "assignment", excluding (1) within firm transfers from inventing employees to their employer assignees or (2) transaction date is the same date as application filing date. We add the reassigned dummy to baseline estimation (1), and the results show that reassigned patents are 0.12 SD easier to read and require 0.49 years less education to comprehend than the rest of the patents in the sample.

[30] More details on this exercise and the first-action rejection exercise described below is provided in Appendix A. Appendix C provides the survey questionnaire.

Table 8
Linear probability estimates of first-action 112 (a) rejections on selected linguistic measures based on experts' comments.

Variables	Correlation with readability	FirstRejection112a	Linguistic measure supports rejection
Fog	-	−0.00161	
		(0.00154)	
Flesch	+	−1.83e−05	
		(0.000625)	
Kincaid	-	−0.00166	
		(0.00129)	
Mean length of clauses	-	0.00254**	Supportive
		(0.000399)	
Dependent clause ratio	-	−0.000434	
		(0.000905)	
MRC age of acquisition	-	−0.00554	
		(0.00502)	
MRC word concreteness	+	−0.00894	
		(0.00515)	
Global argument overlap count	+	0.00104	
		(0.000526)	
Global content word overlap	+	0.00165	
		(0.000917)	
Global noun overlap count	+	0.000257	
		(0.000552)	
Global stem overlap count	+	0.00122	
		(0.000588)	
Local argument overlap count	+	−0.0121**	Supportive
		(0.00206)	
Local content word overlap	+	−0.00119*	Supportive
		(0.000314)	
Local noun overlap count	+	−0.0125**	Supportive
		(0.00289)	
Local stem overlap count	+	−0.0121**	Supportive
		(0.00206)	

Notes: N = 26,070. Each linguistic measure represents a separate regression. The dependent variable is the likelihood of first-action 112 (a) rejections. The linguistic measures are standardized and selected according to experts' comments. Whether the linguistic measure supports rejection is given only for significant measures. All estimation follows Eq. (1) and controls for joint patents and other patents (using university patents as the base), various citation counts, simple and extended family size, number of inventors, claim counts, application year fixed effects, and U.S. patent subclassification fixed effects. Standard errors are clustered at U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

linguistic measures, at least in some respects, align with experts' evaluations, and that the question of the relationship between readability in the linguistic sense and effective disclosure in the patent sense merits further study in future work.

Our other test was to examine patents rejected by the examiner in a first action under 112 (a) of the patent statute. In effect, these are patent applications that were initially rejected because the examiner determined that they did not adequately disclose the invention.[31] We were able to find about 60% of our sample in a data source that provides information on these first actions; of these, about 6% had 112(a) first-action rejections. We regressed the rejection outcome on the linguistic measures, controlling for patent characteristics. Once again, we found modest support for the meaningfulness of the linguistic measures in this context. Lower readability, as indicated by mean length of clauses, local argument, and content, noun, and stem overlaps are significantly correlated with rejection (see Table 8). Note that it is unclear, at a conceptual level, how strongly we would expect these measures to "work" at this stage of the patent process. The policy issue of inadequate disclosure relates, of course, to the degree of disclosure in granted patents. Even if linguistic measures are useful for scoring the degree of disclosure in granted patents, it is not clear that they would necessarily capture well the wholly inadequate disclosure that generates a first action 112(a) rejection.[32] This underscores the need for more work to understand the true relationship between linguistic readability and patent disclosure.

10. Conclusion

This paper proposes a novel approach that uses computational linguistic measures to study patent disclosure – the accessibility of the information contained in patent documents – by examining large-scale patent text data and applying 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; Cockburn et al., 2002), 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 1.4 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 to 2.6 times larger between the top 100 applicants and that licensing corporate patents disclose more than other corporate patents. This 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 find evidence that our measures are negatively correlated with the number of examples; another argument made by patent professionals who suggest that corporations use many examples to hide the "best mode" of the invention in patent applications. Last, we show that our linguistic measures modestly predict the expert evaluations, and the first-action 112 (a) rejections. 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 on quantifying the level of disclosure in patent applications.

In summary, we see that our various linguistic measures show different degrees of success in explaining and predicting outcomes that are related to disclosure. For example, Gunning Fog captures the difference between universities and corporations, and successfully predicts expert-evaluated readability in patents. Various "global overlap" measures in discourse features are better for predicting expert-evaluated disclosure in patents, whereas the average length of clause and several "local overlap" measures are better for predicting 112 (a) rejections. Therefore, our findings demonstrate the complexity of disclosure and highlight the need to study how different linguistic measures capture disclosure in different contexts.

University and corporate patents differ in many other ways besides readability. One concern is 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. Future research could further address this issue by using other identification strategies, such as disclosure law changes or instrumental variables. It would also be useful to identify other situations in which a strong prior expectation about differences in readability could be used

[31] 112 (a) rejections are described as "not being supported by an enabling disclosure because the person skilled in the art would not know how to make and use the invention without a description of elements to perform the function" or "as lacking adequate written description because the specification does not describe the claimed invention in sufficient detail that one skilled in the art can reasonably conclude that the inventor had possession of the claimed invention" (USPTO).

[32] Multiple specific statutory requirements can be the basis of a 112(a) rejection, which relate to the generic concept of disclosure to varying degrees; see Ashtor (2022).

to test the validity of the measures. Besides examining the differences between university patents and corporate patents, future research could compare patents with a paired paper versus those without. Another approach is to use time of publication as an indicator for an incentive to obscure the information.

The second-language acquisition indicators used in the paper (which are widely used in the linguistic 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 available proxy for patent readability. We believe 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 and information accessibility in patents. If further research confirms our preliminary results, or if the linguistic measures are applied with appropriate caution, we see two applications. The first informs patent practice. Readability scores, particularly where the developed algorithm captures readability, disclosure, and accessibility, provide a valuable tool that enables innovators and their agents to write better patents. It also could help improve the performance of the patent examination process. The second application is derived from a research and policy evaluation perspective. Such measures could help take into account the effectiveness of policy choices where the goals are not only to increase patenting activities and to foster innovation (direct benefits of the patent system), but also to provide insights into the readability, disclosure, and accessibility this innovation has (indirect benefits of a successful patent system). Given that measures of readability capture systematic differences across business entities, it should be of particular interest from a policy perspective to investigate the discrepancies in knowledge disclosure and thus improve the performance of the patent system.

The code for calculating all of the linguistic metrics is shared on *Bitbucket.org,* and the STATA program for making comparisons across groups while controlling econometrically for technology field and other characteristics is shared on the Harvard Dataverse, in the hope that others will further test and use these metrics. With respect to the examination process, having quantitative measures of readability should allow more systematic analysis of the effectiveness of the statutory disclosure standard, including both whether the standard is high enough overall and whether there is unacceptable variation in how it is applied in different contexts. Being able to address these questions quantitatively should assist the patent office in potentially increasing the extent to which the knowledge dissemination goal of the system is realized. And for innovation research, disclosure metrics allow empirical analysis of what competitive, legal, cultural, and institutional factors affect the level of readability in granted patents, and how differences in readability play out in the marketplace and in technology evolution. Such analysis can advance our understanding of the role of information disclosure in the innovation process.

CRediT authorship contribution statement

Nancy Kong: Conceptualization, Data curation, Methodology, Formal analysis, Software, Writing – original draft, Writing – review & editing, Investigation. **Uwe Dulleck:** Conceptualization, Methodology, Writing – review & editing, Funding acquisition, Supervision, Investigation. **Adam B. Jaffe:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition, Investigation. **Shupeng Sun:** Conceptualization, Methodology. **Sowmya Vajjala:** Software, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the data and do files to Harvard Dataverse. It is indicated in the conclusion.

Appendix A

Survey of patent readability and disclosure

We conducted expert panel evaluation of patent readability and disclosure. The purpose of this exercise is twofold: First, to validate whether readability in patents is a good proxy for disclosure; second, to explore what linguistic metrics are correlated with expert panel evaluation. We recruited 18 experts (one is a postdoctoral researcher, 15 are Ph.D. students, and two are honors students who have extensive research experience) who are specialized in the respective patent fields as research assistants. We assigned 16 patents to each research assistant,[33] for a total of 96 patents, each evaluated by 3 students. Patent selection criteria are having a mixture of university and corporate patents, and high and low readability levels, as determined by our linguistic measure in each technological field.[34] STATA randomly selected the patents within these criteria. Research assistants read patents in their respective fields and rated readability and disclosure levels. Specifically, we asked "How is the disclosure level of the technical information in the patent? 1 is very low and 5 is very high", and "How is the disclosure level of the technical information in the patent? 1 is very low and 5 is very high" (see Appendix C for the questionnaire). We calculated the average of the three ratings as the indicator for evaluated readability and disclosure. We also assigned patent pairs within the same USPC that ask the expert panel to give a pairwise comparison on their readability and disclosure.

At the end of the evaluations, we also asked open-ended questions about what their evaluations were based on. The disclosure evaluation was often based on "clarity", "accuracy", "specificity", and "adequacy" and being "concise", "well-structured", and "easy to understand". Readability evaluation was based on "simple words", "length and the simplicity of the sentences", and "well-organized structure", and reversely related to "jargon", and "long names" (see Appendix C

[33] We recruited the expert panel from the engineering faculty in two universities in Australia. Ten post-graduate students are from the University of Queensland and eight from Queensland University of Technology (Ethics Approval number: 3520). The research assistants (RAs) have 20 years of education on average. Among the 18 RAs, 15 are applied researchers and 3 are theoretical researchers, with 6 evaluators specialized in batteries, 3 in photovoltaics, and 5 in nanotechnology; the rest have crossover in these areas. Regarding familiarity with patenting, 33% of evaluators have read patents before, 17% have filed patents before, and 39% have considered filing patents. There are 44% female evaluators and 56% male evaluators. In terms of first languages, 2 speak English, 9 Chinese, 3 Persian, 2 Sinhala, 1 Japanese, and 1 Marathi.

[34] Specifically, each RA evaluated 8 pairs of patents: 2 pairs for a university patent (easy to read) against a university patent (hard to read); 2 pairs for a university patent (easy to read) against a corporate patent (hard to read); 2 pairs for a university patent (hard to read) against corporate patent (easy to read); and 2 pairs for corporate patent (easy to read) against a corporate patent (hard to read). Easy to read is defined as the Fog and PCA measures jointly in the bottom 20th percentile, whereas hard to read is defined as the Fog and PCA jointly in the top 20th percentile.

Table A.1
Evaluated results by their evaluated pairwise readability.

Evaluated results	(1) Evaluated hard to read		(2) Evaluated easy to read		(3) Difference	
	Mean	SD	Mean	SD	Mean	t-stat
Disclosure (5pt)	3.19	0.69	3.85	0.60	−0.66***	(−7.80)
Readability (5pt)	3.29	0.71	4.23	0.50	−0.95***	(−11.84)
Pairwise: more readable	0.33	0.47	0.69	0.46	−0.37***	(−5.94)
Can understand (5pt)	3.54	0.69	4.31	0.51	−0.77***	(−9.74)
Can utilize (5pt)	3.32	0.67	3.85	0.64	−0.53***	(−6.21)
Can create (5pt)	2.92	0.72	3.62	0.77	−0.69***	(−7.13)
University patent	0.43	0.50	0.48	0.50	−0.05	(−0.79)
Observations	118		120		238	

Notes: N = 258. Each linguistic measure represents a separate regression. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

for their comments). We therefore focus on linguistic measures that capture those aspects of the text.

To examine whether readability is significantly correlated with disclosure, in A.1 we divide patents according to the pairwise comparison for being more readable. We combine the 3-expert evaluation by using the mode rating for analysis (if the mode does not exist, then the average rating is used).[35] It shows that more readable patents are significantly higher on the 5-point disclosure, and pairwise disclosure, and more likely to agree with statements that the patent is "worded in such a way that you or another researcher in the field could understand/utilize/recreate the invention without additional information" (5-point scale). We therefore view readability as a good proxy for patent disclosure. To study how our linguistic measures contribute to understanding the expert evaluation, we regress expert-evaluated disclosure and readability (5-point) on the linguistic measures selected to fit the expert comment. Table 7 shows that the Fog index and average length of clauses (both higher values mean harder to read) are negatively associated with the evaluated readability. Argument, noun, and stem overlaps are positively correlated with disclosure. Overall, we view the results as moderately supportive of our proposed metrics.[36]

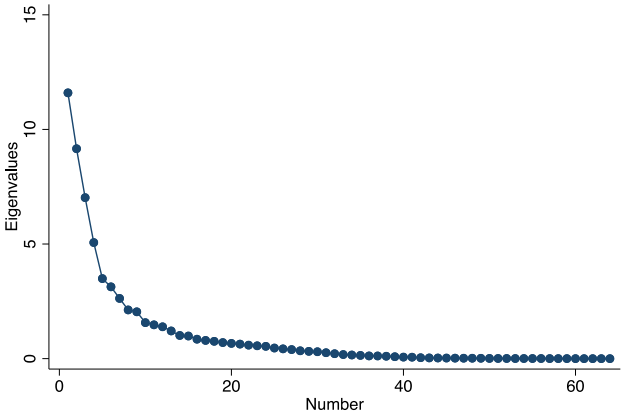

Fig. B.1. Scree plot of eigenvalues after PCA.
Note: The figure presents the scree plot of the eigenvalues of correlation metrics after PCA, which combines 64 linguistic indicators into synthetic variables, as described in Section 4.1. According to the largest distance rule of Onatski (2010), we present estimates of components 1–4.

USPTO first-action 112 (a) rejections

We further explore the USPTO examination outcomes—particularly first-action 112 (a) rejections, which are based on a lack of written description and enablement. We obtain the first actions using the PatEx dataset and limiting transaction types to being non-final rejections, allowances, and final rejections. We then sort the dates and use the earliest transaction date as the first action, before merging the first-actions file with the Office Action Datasets (OAD),[37] in which we identify the 112 rejections. The rejection data further determine the subtypes of the 112 rejections. We are able to merge 26,070 applications (60% of our sample). Of the merged applications, 5.5% have 112 (a) first-action rejections (N=1,444). When we break these down by entity type, 7.5% of university patents (SD=0.26) are first-action rejected on the basis of 112 (a), and 4.5% for corporate patents (SD=0.21).

We test whether our linguistic measures are correlated with first-action rejections, and regress first-action rejections on various linguistic measures separately. We find that first-action 112 (a) rejections are significantly negatively correlated with PCA component 3, local argument overlap, local content word overlap, local noun overlap, and local stem overlap; and significantly positively correlated with PCA component 4, mean length of clauses, global noun overlap, and global stem overlap.

We discuss these robustness results with caution. According to Ashtor (2022), 112 (a) rejections may contain multiple statutory requirements that are not disambiguated in the dataset. Therefore, our limited matching of the USPTO data due to their availability, as well as the multiple requirements in 112 (a), are likely to confound our results. Additionally, previous research has documented that patent examiners exercise different degree of leniency (Dyer et al., 2020; Farre-Mensa et al., 2020; Feng and Jaravel, 2020; Gaule, 2018; Tabakovic and Wollmann, 2018). Therefore, we argue that these results are inconclusive, and cannot indicate a floor level.

Appendix B

See Tables B.1–B.8 and Figs. B.1 and B.2

[35] We implemented quality control in the survey by asking the same questions at two separate time points, and kept those that have the same answer. We also checked whether pairwise preference is reflected in the 5-point evaluation. After quality control, 258 out of 288 (90%) of the evaluations were used.

[36] A few caveats about the survey: First, the evaluations are by postgraduate students (with one exception of a postdoctoral researcher); therefore, the evaluations may differ from those of typical inventors. Second, most of the RAs are non-native English speakers, which could affect the evaluation of readability and disclosure. Third, due to COVID, the evaluations were done online. While we assigned half an hour to each patent, we did not observe the true efforts made by the RAs to read and evaluate patents, especially since most of the evaluations are multiple-choice questions. Fourth, 18 RAs evaluated a total of 96 patents, which means that our evaluation sample size is not particularly large, and the results may not be representative. Hence, we do not claim that these results are conclusive.

[37] The OAD datafile only contains applications that start with 12, 13, 14, or 15 in the application number.

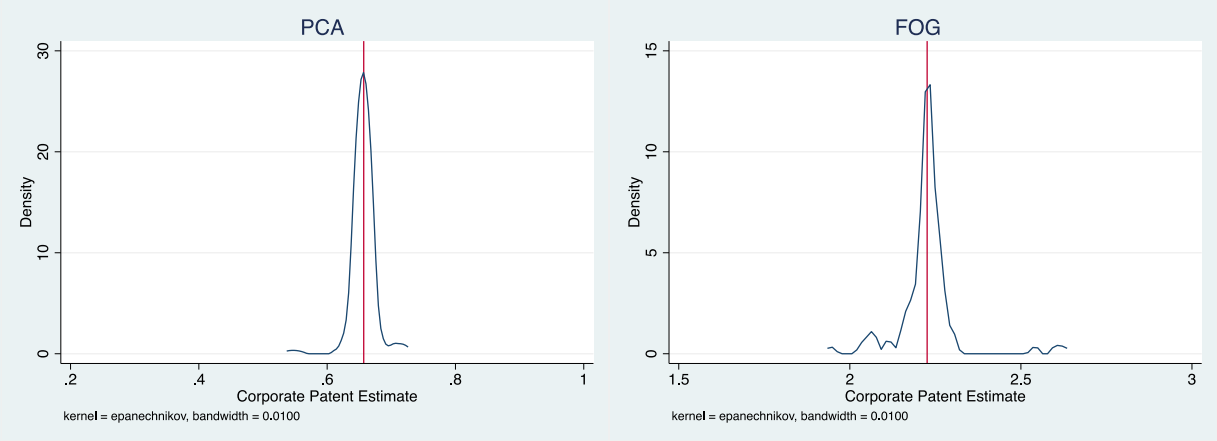

Fig. B.2. Simulations for top applicants.
Note: The figure presents the distribution of estimated coefficients of corporate patent from the baseline regression. We drop each top applicant at a time and repeat 100 times. The vertical line represents the estimate obtained from the baseline estimation in Table 6 Panel A Row 1.

Table B.1
Biomedical patents estimations.

Variables	(1) PCA	(2) Fog	(3) Flesch	(4) Kincaid	(5) AoA_Kup
Corporate	0.0227	0.605***	−2.627***	0.749***	0.0233***
	(0.0435)	(0.0810)	(0.197)	(0.0494)	(0.00724)
Observations	31,034	31,034	31,034	31,034	31,034
R-squared	0.093	0.123	0.092	0.066	0.137
Variables	(6) Word_TTR	(7) DependentClauseR	(8) MLT	(9) ProperNounsPerNoun	(10) ContentWordOverlap
Corporate	−0.0111***	0.0124***	0.452***	−0.000932***	305.2***
	(0.000939)	(0.00279)	(0.123)	(0.000201)	(74.55)
Observations	31,034	31,034	31,034	31,034	31,034
R-squared	0.127	0.080	0.113	0.026	0.393

Note: Estimates are corporate patents from Eq. (1). We search "biomedical" as the keyword in patents from January 1, 2000, to July 8, 2019, and obtain and match 31,034 biomedical patents over 67 USPCs (USPCs with ≥ 30 patents in the sample). We conduct baseline estimations as outlined in Equation 1. We show that corporate patents are more difficult to read. Estimations control for joint patents and other patents (using university patents as the base), various citation counts, simple and extended family size, number of inventors, claim counts, application year fixed effects, and U.S. patent subclassification fixed effects. Estimates of corporate patents are presented with university patents as the base. Standard errors are clustered at U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.2
Summary statistics of 64 individual linguistic measures and controls by business model.

	(1) Corporations		(2) Universities		(3) Joint		(4) 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_Resolved_Citation_Count	0.19	0.83	0.82	1.82	0.64	1.71	0.25	0.95
NumInventors	3.01	1.93	3.42	1.80	3.84	1.95	2.69	1.79
ClaimCounts	22.27	20.57	27.86	23.38	24.18	29.58	23.56	22.48
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

(*continued on next page*)

Table B.2 (*continued*).

	(1) Corporations		(2) Universities		(3) Joint		(4) Others	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
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
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
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
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 the sample is provided in Section 4.1. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.3
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
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 Fig. B.1 for eigenvalues.

Table B.4
Estimates of principal component analysis as synthetic linguistic indicators.

Variables	(1) Component 2	(2) Component 3	(3) Component 4
Corporations	0.104** (0.0152)	−0.132 (0.0529)	−0.137 (0.0537)
R-squared	0.161	0.157	0.128

Notes: N = 40,949. The dependent variable is the PCA generated by 64 linguistic measures. See Table B3 for detailed compositions and Fig. B.1 for eigenvalues. OLS estimates of corporate patents are obtained from Eq. (1), using university patents as the base. All estimations control for joint patents and other patents (using university patents as the base), various citation counts, simple and extended family size, number of inventors, claim counts, application year fixed effects, and U.S. patent subclassification fixed effects. Standard errors are clustered at U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.5
OLS estimates of 64 individual readability measures and adjusted p-value using Romano–Wolf stepdown multiple hypothesis testing.

	Variables	Standardized		R-squared
(1)	AoA_Bird_Lem	0.0874	(0.0244)	0.097
(2)	AoA_Bristol_Lem	−0.0241	(0.0336)	0.185
(3)	AoA_Cort_Lem	−0.0528	(0.0760)	0.101
(4)	AoA_Kup	0.0867	(0.0246)	0.152
(5)	AoA_Kup_Lem	−0.122	(0.0109)	0.214
(6)	DISC_RefExprDefArtPerSen	0.134*	(0.0110)	0.208
(7)	DISC_RefExprDefArtPerWord	0.0518	(0.0348)	0.239
(8)	DISC_RefExprPerProPerWord	−0.0210	(0.0225)	0.027
(9)	DISC_RefExprPerPronounsPerSen	−0.0208	(0.0392)	0.056
(10)	DISC_RefExprPossProPerSen	−0.157**	(0.0144)	0.053
(11)	DISC_RefExprPossProPerWord	−0.0242	(0.0346)	0.007
(12)	DISC_RefExprPronounsPerNoun	−0.111	(0.0289)	0.067
(13)	DISC_RefExprPronounsPerSen	−0.0789	(0.0355)	0.058
(14)	DISC_RefExprPronounsPerWord	−0.0448	(0.0386)	0.033
(15)	DISC_RefExprProperNounsPerNoun	−0.471***	(0.0313)	0.183
(16)	DISC_globalArgumentOverlapCount	0.164***	(0.0419)	0.191
(17)	DISC_globalContentWordOverlapCount	0.245***	(0.0422)	0.115
(18)	DISC_globalNounOverlapCount	0.186***	(0.0436)	0.175
(19)	DISC_globalStemOverlapCount	0.160***	(0.0422)	0.194
(20)	DISC_localArgumentOverlapCount	0.457***	(0.0213)	0.271
(21)	DISC_localContentWordOverlapCount	0.0952	(0.0130)	0.019
(22)	DISC_localNounOverlapCount	0.445***	(0.0150)	0.287
(23)	DISC_localStemOverlapCount	0.457***	(0.0224)	0.276
(24)	MRCAoA	0.0716	(0.0281)	0.135
(25)	MRCColMeaningfulness	0.150**	(0.0448)	0.146
(26)	MRCConcreteness	0.186***	(0.0273)	0.143
(27)	MRCFamiliarity	0.116	(0.0319)	0.072
(28)	MRCImageability	0.141*	(0.0284)	0.107
(29)	MRCPavioMeaningfulness	−0.00950	(0.0296)	0.168
(30)	POS_adjVar	0.0446	(0.0675)	0.238
(31)	POS_advVar	0.182***	(0.0305)	0.061
(32)	POS_correctedVV1	0.180***	(0.0636)	0.167
(33)	POS_modVar	0.101	(0.0720)	0.236
(34)	POS_nounVar	−0.170***	(0.0668)	0.195
(35)	POS_squaredVerbVar1	0.151**	(0.0270)	0.099
(36)	POS_verbVar1	0.296***	(0.0512)	0.094
(37)	POS_verbVar2	0.0123	(0.0189)	0.220
(38)	SYN_CNPerClause	0.0995	(0.0125)	0.067
(39)	SYN_CNPerTunit	0.130	(0.0104)	0.056
(40)	SYN_ComplexTunitRatio	0.169***	(0.00759)	0.150
(41)	SYN_CoordPerClause	0.101	(0.0285)	0.108
(42)	SYN_CoordPerTunit	0.173***	(0.0244)	0.081
(43)	SYN_DependentClauseRatio	0.205***	(0.0130)	0.178
(44)	SYN_DependentClausesPerTunit	0.189***	(0.0163)	0.163
(45)	SYN_MLC	0.0764	(0.0290)	0.095
(46)	SYN_MLT	0.213***	(0.0151)	0.069
(47)	SYN_TunitComplexityRatio	0.160***	(0.0173)	0.106
(48)	SYN_VPPerTunit	0.306***	(0.0138)	0.134
(49)	TRAD_ARI	0.228***	(0.0127)	0.055
(50)	TRAD_Coleman	−0.0591	(0.0120)	0.168
(51)	TRAD_FOG	0.211***	(0.0197)	0.058
(52)	TRAD_FORCAST	0.141**	(0.0462)	0.224
(53)	TRAD_Flesch	−0.226***	(0.0285)	0.077
(54)	TRAD_Kincaid	0.249***	(0.0193)	0.056

(*continued on next page*)

Table B.5 (*continued*).

	Variables	Standardized		R-squared
(55)	TRAD_LIX	0.217***	(0.0152)	0.061
(56)	TRAD_SMOG	0.241***	(0.0339)	0.086
(57)	TRAD_numChars	−0.139**	(0.00805)	0.186
(58)	TRAD_numSyll	0.0282	(0.0373)	0.205
(59)	Word_BilogTTR	−0.386***	(0.0211)	0.261
(60)	Word_CTTR	−0.345***	(0.0503)	0.437
(61)	Word_MTLD	−0.340***	(0.0444)	0.082
(62)	Word_RTTR	−0.345***	(0.0503)	0.437
(63)	Word_TTR	−0.333***	(0.0472)	0.176
(64)	Word_UberIndex	−0.370***	(0.0372)	0.407

Note: Each row represents one estimation. Estimates are from Eq. (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 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 the base), various citation counts, simple and extended family size, number of inventors, claim counts, application year fixed effects, and U.S. patent subclassification fixed effects. Estimates of corporate patents are presented with university patents as the base. Standard errors are clustered at U.S. patent classification level in parentheses.

Table B.6
Doubly robust propensity score matching.

Variables	(1) PCA	(2) Fog	(3) Flesch	(4) Kincaid	(5) AoA_Kup
Corporate	0.453***	1.351**	−3.611*	1.548**	0.00152
	(0.0345)	(0.246)	(1.138)	(0.249)	(0.0151)
Observations	22,869	22,869	22,869	22,869	22,869
R-squared	0.309	0.072	0.075	0.071	0.168
Variables	(10) Word_TTR	(12) DependentClauseR	(14) MLT	(16) ProperNounsPerNoun	(18) ContentWordOverlap
Corporate	−0.0149***	0.0209**	0.447***	−0.0161**	141.7**
	(0.00106)	(0.00279)	(0.0448)	(0.00258)	(19.05)
Observations	22,869	22,869	22,869	22,869	22,869
R-squared	0.194	0.163	0.073	0.199	0.146

Note: The estimates are generated from the doubly robust estimation procedure. We first use propensity score to match corporate patents and university patents based on all patent attributes identified in the baseline estimation. Then we re-weight university patents using the weight generated from the matching to create a counterpart corporate patent group based on observed characteristics. Therefore, the re-weighted university patents would have observed characteristics similar to those in the corporate patents. Last, we use the matched corporate and university patents to conduct the baseline estimation. This method produces consistent estimates as long as either propensity score matching or the regression adjustment is correctly specified (Wooldridge, 2010; Kantarevic and Kralj, 2013). Standard errors are clustered at U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.7
Log transformed examples.

Variables	(1) PCA	(2) Fog	(3) Flesch	(4) Kincaid	(5) AoA_Kup
log examples	0.00467	−0.133	−0.428	−0.0951	0.00322
	(0.0223)	(0.119)	(0.339)	(0.108)	(0.00982)
Corporate	0.418***	1.463***	−4.263**	1.655***	0.0235
	(0.00786)	(0.135)	(0.554)	(0.119)	(0.0102)
Observations	40,949	40,949	40,949	40,949	40,949
R-squared	0.295	0.059	0.078	0.056	0.152
VARIABLES	(6) Word_TTR	(7) DependentClauseR	(8) MLT	(9) ProperNounsPerNoun	(10) ContentWordOverlap
log examples	−0.0140***	0.00240*	0.211**	0.000722	121.8***
	(0.000318)	(0.000625)	(0.0375)	(0.000295)	(6.611)
Corporate	−0.0105**	0.0163***	0.350***	−0.0188***	81.89**
	(0.00157)	(0.000866)	(0.0254)	(0.00132)	(15.49)
Observations	40,949	40,949	40,949	40,949	40,949
R-squared	0.290	0.179	0.083	0.184	0.186

Note: *log examples* is log transformed frequency of "for example" and "e.g." in technical descriptions of patent applications. Estimates of corporate patents are presented with university patents as the base. All estimations control for joint patents and other patents (using university patents as the base), various citation counts, simple and extended family size, number of inventors, claim counts, application year fixed effects, and U.S. patent subclassification fixed effects. Standard errors are clustered at U.S. patent classification level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.8
Reassigned patents estimations.

Variables	(1) PCA	(2) Fog	(3) Flesch	(4) Kincaid	(5) AoA_Kup
Reassigned	−0.118***	−0.494**	0.694**	−0.433**	−0.00801
	(0.00744)	(0.0818)	(0.0759)	(0.0634)	(0.00337)
Observations	39,020	39,020	39,020	39,020	39,020
R-squared	0.300	0.061	0.078	0.058	0.154
Controls	Yes	Yes	Yes	Yes	Yes
Sub-USPC & Year FE	Yes	Yes	Yes	Yes	Yes
Variables	(6) Word_TTR	(7) DependentClauseR	(8) MLT	(9) ProperNounsPerNoun	(10) ContentWordOverlap
Reassigned	0.00395**	−0.00320	−0.136**	0.00160**	−29.93*
	(0.000811)	(0.00175)	(0.0250)	(0.000309)	(9.217)
Observations	39,020	39,020	39,020	39,020	39,020
R-squared	0.178	0.178	0.071	0.186	0.120
Controls	Yes	Yes	Yes	Yes	Yes
Sub-USPC & Year FE	Yes	Yes	Yes	Yes	Yes

Note: Reassigned patents are identified by matching with USPTO assignment dataset https://www.uspto.gov/ip-policy/economic-research/research-datasets/patent-assignment-dataset. We define reassigned patent = 1 if "conveyance type" is "assignment", excluding (1) within firm transfers from inventing employees to their employer assignees, or (2) transaction date is the same date as application filing date. We add the reassigned dummy to baseline estimation (1). Estimations control for entity types, various citation counts, simple and extended family size, number of inventors, claim counts, application year fixed effects, and U.S. patent subclassification fixed effects. Standard errors are clustered at U.S. patent classification level in parentheses.

References

Arinas, I., 2012. How vague can your patent be? Vagueness strategies in U.S. patents. SSRN Scholarly Paper ID 2117827, Social Science Research Network.

Arts, S., Cassiman, B., Gomez, J.C., 2018. Text matching to measure patent similarity. Strateg. Manag. J. 39 (1), 62–84.

Ashtor, J.H., 2022. Modeling patent clarity. Res. Policy 51 (2), 104415.

Baker, S., Mezzetti, C., 2005. Disclosure as a strategy in the patent race. J. Law Econ. 48 (1), 173–194.

Baruffaldi, S.H., Simeth, M., 2020. Patents and knowledge diffusion: The effect of early disclosure. Res. Policy 49 (4), 103927.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection among high-dimensional controls. Rev. Econom. Stud. 81 (2), 608–650.

Bloomfield, R.J., 2002. The "incomplete revelation hypothesis" and financial reporting. Account. Horiz. 16 (3), 233–243.

Card, D., DellaVigna, S., Funk, P., Iriberri, N., 2020. Are referees and editors in economics gender neutral? Q. J. Econ. 135 (1), 269–327.

Cockburn, I.M., Kortum, S., Stern, S., 2002. Are All Patent Examiners Equal? The Impact of Examiner Characteristics. Working Paper 8980, National Bureau of Economic Research, Series: Working Paper Series.

Cohen, W.M., Goto, A., Nagata, A., Nelson, R.R., Walsh, J.P., 2002. R&D spillovers, patents and the incentives to innovate in Japan and the United States. Res. Policy 31 (8–9), 1349–1367.

Collins-Thompson, K., 2014. Computational assessment of text readability: A survey of current and future research. ITL-Int. J. Appl. Linguist. 165 (2), 97–135.

Cornelli, F., Schankerman, M., 1999. Patent renewals and R&D incentives. Rand J. Econ. 30 (2), 197–213.

De Clercq, D., Diop, N.-F., Jain, D., Tan, B., Wen, Z., 2019. Multi-label classification and interactive NLP-based visualization of electric vehicle patent data. World Pat. Inf. 58, 101903.

Denicolò, V., Franzoni, L.A., 2003. The contract theory of patents. Int. Rev. Law Econ. 23 (4), 365–380.

Devlin, A., 2009. The misunderstood function of disclosure in patent law. Harv. J. Law Technol. 23, 401.

Dyer, T., Glaeser, S., Lang, M.H., Sprecher, C., 2020. The Effect of Patent Disclosure Quality on Innovation. SSRN Scholarly Paper 3711128, Social Science Research Network, Rochester, NY.

Farre-Mensa, J., Hegde, D., Ljungqvist, A., 2020. What is a patent worth? Evidence from the US patent "lottery". J. Finance 75 (2), 639–682.

Feng, J., Jaravel, X., 2020. Crafting intellectual property rights: Implications for patent assertion entities, litigation, and innovation. Am. Econ. J. Appl. Econ. 12 (1), 140–181.

Gaule, P., 2018. Patents and the success of venture-capital backed startups: Using examiner assignment to estimate causal effects. J. Ind. Econ. 66 (2), 350–376.

Gentzkow, M., Kelly, B., Taddy, M., 2019. Text as Data. J. Econ. Lit. 57 (3), 535–574.

Grossman, S.J., Stiglitz, J.E., 1980. On the impossibility of informationally efficient markets. Amer. Econ. Rev. 70 (3), 393–408.

Hagiu, A., Yoffie, D., Wagonfeld, A.B., 2009. Intellectual ventures. HBS Case (710–423).

Hansen, S., McMahon, M., Prat, A., 2018. Transparency and deliberation within the FOMC: A computational linguistics approach. Q. J. Econ. 133 (2), 801–870.

Hegde, D., Herkenhoff, K.F., Zhu, C., 2022. Patent publication and innovation. Working Paper 29770, National Bureau of Economic Research, Series: Working Paper Series.

Helmers, L., Horn, F., Biegler, F., Oppermann, T., Müller, K.-R., 2019. Automating the search for a patent's prior art with a full text similarity search. PLOS ONE 14 (3), e0212103.

Henderson, R., Jaffe, A.B., Trajtenberg, M., 1998. Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. Rev. Econ. Stat. 80 (1), 119–127.

Hengel, E., 2022. Publishing while female: are women held to higher standards? evidence from peer review. The Economic Journal 132 (648), 2951–2991.

Hsu, D.H., Hsu, P.-H., Zhou, T., Ziedonis, A.A., 2021. Benchmarking U.S. university patent value and commercialization efforts: A new approach. Res. Policy 50 (1), 104076.

Hunt, K.W., 1965. Grammatical structures written at three grade levels. NCTE Research Report No. 3.

Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.

Jaffe, A.B., Trajtenberg, M., 2002. Patents, Citations, and Innovations: A Window on the Knowledge Economy. MIT Press.

Jefferson, O.A., Jaffe, A., Ashton, D., Warren, B., Koellhofer, D., Dulleck, U., Ballagh, A., Moe, J., DiCuccio, M., Ward, K., Bilder, G., Dolby, K., Jefferson, R.A., 2018. Mapping the global influence of published research on industry and innovation. Nature Biotechnol. 36 (1), 31–39.

Kantarevic, J., Kralj, B., 2013. Link between pay for performance incentives and physician payment mechanisms: Evidence from the diabetes management incentive in ontario. Health Econ. 22 (12), 1417–1439.

Kelly, B., Papanikolaou, D., Seru, A., Taddy, M., 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. J. Law Econ. 20 (2), 265–290.

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

Landes, W.M., Posner, R.A., 2009. The Economic Structure of Intellectual Property Law. Harvard University Press.

Lawrence, A., 2013. Individual investors and financial disclosure. J. Account. Econ. 56 (1), 130–147.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. J. Account. Econ. 45 (2), 221–247.

Loughran, T., McDonald, B., 2016. Textual analysis in accounting and finance: A Survey. J. Account. Res. 54 (4), 1187–1230.

Lu, X., 2010. Automatic analysis of syntactic complexity in second language writing. Int. J. Corpus Linguist 15 (4), 474–496.

McNamara, D.S., Louwerse, M.M., Graesser, A.C., 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. Account. Rev. 85 (6), 2107–2143.

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. Rev. Econ. Stat. 92 (4), 1004–1016.
Ouellette, L.L., 2011. Do patents disclose useful information? Harv. J. Law Technol. 25, 545.
Ouellette, L.L., 2017. Who reads patents? Nature Biotechnol. 35 (5), 421–424.
Packalen, M., Bhattacharya, J., 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). Harv. Law Rev..
Romano, J.P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73 (4), 1237–1282.
Sampat, B.N., 2006. Patenting and US academic research in the 20th century: The world before and after Bayh-Dole. Res. Policy 35 (6), 772–789.
Sampat, B., 2018. A survey of empirical evidence on patents and innovation. Technical Report w25383, National Bureau of Economic Research, Cambridge, MA, p. w25383.
Scotchmer, S., Green, J., 1990. Novelty and disclosure in patent law. Rand J. Econ. 131–146.
Tabakovic, H., Wollmann, T.G., 2018. From revolving doors to regulatory capture? Evidence from patent examiners. Technical report, National Bureau of Economic Research.
Tauman, Y., Weng, M.-H., 2012. Selling patent rights and the incentive to innovate. Econom. Lett. 114 (3), 241–244.
Teodorescu, M., 2017. Machine learning methods for strategy research. Harv. Bus. School Res. Pap. Series (18–011).
Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: A retrospective. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (3), 273–282.
Todirascu, A., François, T., Gala, N., Fairon, C., Ligozat, A.-L., Bernhard, D., 2013. Coherence and cohesion for the assessment of text readability. Nat. Lang. Proc. Cogn. Sci. 11, 11–19.
Trajtenberg, M., Henderson, R., Jaffe, A., 1997. University versus corporate patents: A window on the basicness of invention. Econ. Innov. New Technol. 5 (1), 19–50.
Vajjala, S., Meurers, D., 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., Meurers, D., 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., Meurers, D., 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., Meurers, D., 2014b. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. ITL-Int. J. Appl. Linguist. 165 (2), 194–222.
Valdivia, W.D., 2013. University start-ups: Critical for improving technology transfer. Center for Technology Innovation At Brookings. Washington, DC: Brookings Institution.
Whalen, R., Lungeanu, A., DeChurch, L., Contractor, N., 2020. Patent similarity data and innovation metrics. J. Empir. Leg. Stud. 17 (3), 615–639.
Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT Press.
Wu, A.H., 2018. Gendered language on the economics job market rumors forum. In: AEA Papers and Proceedings, Vol. 108. pp. 175–179.
You, H., Zhang, X., 2009. Financial reporting complexity and investor underreaction to 10-K information. Rev. Account. Stud. 14 (4), 559–586.
Younge, K.A., Kuhn, J.M., 2016. Patent-to-patent similarity: A vector space model. SSRN Scholarly Paper ID 2709238, Social Science Research Network.