Patents and knowledge diffusion

Paper to be presented at DRUID21
Copenhagen Business School, Copenhagen, Denmark
October 18-20, 2021

Patents and knowledge diffusion - The impact of Machine Translation

Benjamin Buettner
Eindhoven University of Technology
Technology, Innovation & Society
b.buettner@tue.nl

Emilio Raiteri
Eindhoven University of Technology
Technology, Innovation & Society
e.raiteri@tue.nl

Abstract

One of the main rationales for the patent system's existence is to encourage knowledge diffusion and follow-on innovation through the full disclosure of the technical knowledge embodied in a patented invention. Yet, economists and legal scholars cast doubts on the validity of the disclosure theory and stress that inventors do not learn their science from reading patents. The empirical evidence on the actual benefits of the disclosure function is, indeed, limited. The present paper aims at expanding our understanding of how information spreads via patent disclosure and exploits recent improvements in machine translation (MT) to identify the effect of broader access to patented knowledge. More specifically, the paper uses a unique natural experiment. In September 2013, Google launched a major upgrade of its Google Patents service and added patent applications from the China National Intellectual Property Agency(CNIPA) to its searchable patent database. To do so, Google used its own neural machine translation service to translate patent documents previously available to the general public only in Chinese. Using a difference-in-differences approach, we show that the translation of the Chinese patents into English resulted in an increase in citations received from patents filed by U.S. inventors, compared to a suitable control group composed of patents which Google translated only in 2016. Our results suggest that increased access to patented knowledge promotes technological progress. This finding seems to confirm the beneficial effect of patent disclosure.

Patents and knowledge diffusion
The impact of Machine Translation

immediate

August 11, 2021

Abstract

One of the main rationales for the patent system's existence is to encourage knowledge diffusion and follow-on innovation through the full disclosure of the technical knowledge embodied in a patented invention. Yet, economists and legal scholars cast doubts on the validity of the disclosure theory. The empirical evidence on the actual benefits of the disclosure function is still limited. The present paper aims at expanding our understanding of how information spreads via patent disclosure and exploits recent improvements in machine translation (MT) to identify the effect of broader access to patented knowledge. More specifically, the paper uses a unique natural experiment. In September 2013, Google launched a major upgrade of its Google Patents service and added patent applications from the China National Intellectual Property Agency (CNIPA) to its searchable patent database. Using a difference-in-differences approach, we show that the translation of the Chinese patents into English resulted in an increase in citations received from patents filed by US inventors, compared to a suitable control group composed of patents which Google translated only in 2016. Our results suggest that improved access to patented knowledge fosters cumulative innovation.

Keywords: patent disclosure, knowledge diffusion, machine translation

1 Introduction

"English's emergence as the global language, along with *the rapid progress in machine translation* [...] make it less clear that the substantial investment necessary to speak a foreign tongue is universally worthwhile. While there is no

> gainsaying the insights that come from mastering a language, it will over time become *less essential in doing business* [...].“
>
> (Lawrence H. Summers — NY Times, 2012)

One of the rationales behind the existence of the patent system is to foster innovation by facilitating knowledge diffusion. Patents ‘‘promote the progress of science’’ (US Constitution art. I, § 8, cl. 8) by granting the inventor a temporary monopoly for his invention while assuring the disclosure of the technical information on which the invention bases on. For this reason, the patent system is often characterized as *quid pro quo* system, in which the inventor trades the public release of new technical knowledge to society against the rights to exclusive use. However, the scholarly debate is far from reaching agreement on a positive effect of patent disclosure on the pace of technical progress. Several scholars question the legitimacy of disclosure theory as a justification for the existence of the patent system and argue that the disclosure requirement is not really working as planned for society (Lemley, 2012; Risch, 2007; Devlin, 2010). This paper aims to contribute to this discussion and to shed light on the mechanism of knowledge diffusion through the patent system, focusing on cross-borders and cross-language knowledge flows.

Historically, language differences represented one of the main barriers to the diffusion of scientific and technical knowledge. In the the last three centuries, scientists, researchers, and inventors resorted to different solutions to facilitate communications between diverse language communities. The use of Latin as *lingua franca*, the development of international auxiliary languages such as Ido and Esperanto, the publication of journal abstracts that systematically translated the abstracts of the contributions appeared in the most important scientific journals are only a few well-known examples of attempts to overcome language barriers to the progress of science (Gordin, 2015).

Even though it was not its main objective, the progressive harmonization of patent law across countries was probably one of the main element that fostered the diffusion of novel technical knowledge across language borders. In the current system, the legal protection granted by patented invention could be extended to other jurisdictions within twelve months from the original application.[1] Most jurisdictions require the foreign application to be translated in one of the official languages of the country of the receiving patent authority. Therefore, a Japanese inventor that would like to obtain the right to exclude others from using her invention in the United States needs to translate her patent application in English and file it to the United States Trademark and Patent Office (USPTO). In such a case, the existence of the patent system not only pushed the inventor to publicly disclose the technical knowledge that should allow to reproduce her invention, but the prospect of protection

[1]Patent application filed via the international route provided by the Patent Cooperation Treaty can be filed to a national office up to thirty months from the original application.

beyond the national context gave her the incentive to translate said knowledge, allowing for a potentially much larger interested audience.

In spite of its potential relevance, the importance of mandated translation in fulfilling the objectives of the disclosure requirement has received very limited attention in the scholarly debate (Ouellette, 2017). This is especially surprising in an era in which increasingly sophisticated machine translation tools are lowering the costs of access to knowledge behind a language barriers in way that have no precedent in human history. This paper aims to fill this gap in the literature and, more specifically, to assess the effect of disclosure through automated translation of technical knowledge on follow-on innovation.

To do so, we exploit a recent natural experiment. In 2013, Google launched a major improvement to Google Patents, adding documents from the Chinese Patent Office, the China National Intellectual Property Administration (CNIPA), and using its own neural machine translation service to translate patents previously available only in Chinese. Thanks to the automatic translations, Chinese patents were made available and searchable in both their original Chinese version and in the English language.[2] This event provides an ideal research setting for evaluating whether disclosure of technical knowledge through patents and their translation fosters knowledge diffusion. We exploit this sudden change in the availability of translation of Chinese patents to implement a difference-in-differences analysis, in which we evaluate whether US inventors rely more on knowledge embodied in Chinese patents after the automated translation provided by Google, relatively to a set of suitable control groups. Our results show a positive effect of machine translation of Chinese patents on follow-on innovation in the US, an increase of about 7.2 percent in patent citations received by Chinese patents after the translation. The effect is more pronounced in technological areas in which China is at the technological frontier, such as computing- and data processing-related technologies, with up to 14.5 percent increase in citations received by translated patents. Excluding citations from patents filed by US residents of Chinese origins strengthens the aforementioned results. Furthermore, we find that inventors are more likely to cite translated patents that are considered as more readable. All in all, our results suggest that inventors do use the knowledge disclosed in patent documents as an input in the invention process. Policy that improve the disclosure function of the patent system are likely to increase the benefits the society can reap from having a patent system in place.

The rest of the paper is organized as follows. Section 2 provides the theoretical framework on the disclosure function of the patent system. Section 3 illustrates our identification strategy, whereas section 4 gives an overview of the data and method used. Section 5 discusses the econometric results and section 5.4 describes a few robustness checks. Section 7 concludes.

[2]One of the declared objectives of Google's effort was to increase the discoverability of foreign inventions to *improve the quality of patents in the US and worldwide.*

2 Theoretical Background

.

2.1 Patent disclosure and cumulative innovation

The promotion of innovation through the disclosure of technical knowledge is one of the pillars on which the patent system bases on (Ouellette, 2012; Lemley, 2012; Furman et al., 2018). The patentee obtains exclusive rights to the commercial exploitation of her invention and has the possibility of earning quasi-rents for a limited time (Sampat and Williams, 2019). In return, the inventor has to disclose her invention to society so that science and technology can progress by building on the divulged knowledge (Ouellette, 2012; Fromer, 2009). To fulfill this function, in most jurisdictions, a patent application needs to include a precise written description of the invention, that would allow a person trained in the relevant field to reproduce it. Indeed, disclosure is one of the requirement for which the TRIPS Agreement impose a minimum standard for protection. Article 29.1 of the TRIPS states 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''.

However, several legal scholars and economists have argued that the information disclosed in patent applications is of little use to inventors, as they do not learn their science from patents (Lemley, 2012; Boldrin and Levine, 2013). In particular, the works in this strand of literature suggest that patents are generally drafted using a rather obscure and vague jargon aimed at hiding strategic information or at broadening the scope of the patent's claims (Risch, 2007; Devlin, 2010). If these claims were correct, at least a part of the patent bargain would not be satisfied. The society would be paying the costs that exclusivity entails, i.e., higher prices for innovated products and services and the consequential under consumption, without receiving the benefits that patent disclosure should bring about in terms of additional cumulative innovations. At present, the empirical evidence on the positive effect of patent disclosure on innovation is still limited and does not fully rule out this possibility.

On the one hand, empirical studies based on survey administered to inventors provide mixed results. Jaffe et al. (2000) asked US patent inventors to list the most important inputs for the conception of their patented invention. Only five percent of the respondents identified the patent literature as having significant influence in the invention process. A similar a survey administered to R&D managers by Cohen et al. (2000) asked them to rank the relevance of different information channels for the completion of a recent R&D project. Results showed that patents were considered less important than other information sources like scientific publications and informal exchanges. More recent survey-based studies draw a

more optimistic picture for disclosure theory. Based on a survey administered to researchers active in the nanotechnology field, Ouellette (2012) shows that 64 percent of the respondents have read at least a patent for a research purpose. In a follow-up study based on a survey administered to a more diverse group of researchers, Ouellette (2017) finds that only a minority of respondents never read a patent, and that sixty percent of the patent readers found useful scientific information in the most recent patent they read.

On the other hand, a recent strand of research based on quasi-experimental approaches has produced growing empirical evidence of a positive effect of patent disclosure on innovation (Furman et al., 2018; Baruffaldi and Simeth, 2020; de Rassenfosse et al., 2020).

Furman et al. (2018) investigate the impact of information disclosure through patents on subsequent innovation, and exploit the opening of new patent libraries in the pre-internet era for identification. They find that the improved access to the patent literature made possible by the opening of a library in a given region lead to a 17 percent increase of local patenting relative to suitable control regions. Furman et al. (2018) firmly suggest that the increase in patenting is driven by the disclosure of technical information in the patent documents.

Baruffaldi and Simeth (2020) focus instead on the effect of early disclosure of patent applications and investigate the effect of a policy change that affected their publication time in the US. In 2000 the American Inventors Protection Act reduced the default publication time of patent application filed at the United States Patent and Trademark Office (USPTO) to eighteen months. Their findings confirm the importance of early disclosure in facilitating information diffusion, which also highlights that technological knowledge may become obsolete quickly. In addition, they find the effect mainly unaffected by geographical distance. The results of Baruffaldi and Simeth (2020) imply that knowledge diffusion increases if the time frame between the invention and its disclosure decreases, highlighting the importance of timing of disclosure for the usefulness of the disclosed information for future inventions.

de Rassenfosse et al. (2020) evaluate the extent to which knowledge flows are disrupted when a patent application to the USPTO is temporarily kept secret because of national security concerns.[3] Their analysis shows that patented inventions cited by a patent that is subject to a secrecy order receive on average between 30 and 50 percent less forward citations than a group of suitable control patents during the period in which the secrecy order is in place. These findings suggest that secrecy orders hinder cumulative effects in knowledge production, highlighting the importance of patent disclosure to enable follow-on inventions.

All in all, these recent studies appear to confirm that the disclosure function of the patent system is working as planned. Inventors seem to be using the knowledge disclosed in patent documents as an input for their inventive efforts. However, the works in this strand mainly

[3]Secrecy orders are imposed by the Commissioner of Patents in accordance with the Invention Secrecy Act of 1951.

focus on the flows of knowledge that patent disclosure allows within a single jurisdiction. Ouellette (2017) suggests instead that patent disclosure might be particularly beneficial for improving access to knowledge that would otherwise be inaccessible to local inventors. In particular, patents might be especially useful in facilitating information flows across language boarders through mandatory translations of knowledge that would be otherwise trapped behind a language barrier. Anecdotal evidence supports this hypothesis. For instance, the adoption of one of the most fundamental innovations in coronary angioplasty treatment, the balloon catheter developed by Andreas Grüntzig, was delayed because the authors described the new method in an article that was published exclusively in German. The publication could be read only by individuals proficient in the German language, and was restricted to German-speaking countries (Husmann and Barton, 2014). The global adoption of this technique followed the filing of several patent applications for the balloon angioplasty device in Switzerland, Germany, France, the United Kingdom, the United States, and Japan. In a recent paper, Choudhury and Kim (2019) exploit a natural experiment to find that ethnic migrant inventors are instrumental in transferring contextual knowledge across borders. They constructed a dataset of herbal patents to evaluate whether knowledge of Chinese and Indian herbal medicine is transferred to the west by first-generation migrant Chinese and Indian scientists. Choudhury and Kim (2019) find that an increase in the supply of first-generation ethnic migrant inventors increases the rate of codification of herbal knowledge by US patent assignees by 4.5 percent. Their results indicate that knowledge locked within specific cultural regions becomes accessible only through migrants capable of speaking the language, confirming the importance of language as a barrier to knowledge flows.

The benefits of patent disclosure may not exclusively derive from the requirements imposed within a single jurisdiction, but also from the existence of an international patent system that builds on a broadly harmonized patent law. The Paris Convention for the Protection of Industrial Property, one of the first international treaties on intellectual property matters, establishes the right to priority. This right means that, on the basis of a regular first application filed at the patent authority of one of the Contracting States, the applicant may apply for protection in any of the other Contracting States within twelve months from the first filing and that these subsequent applications will be examined as if they were filed on the same date as the original application.[4] The Patent Cooperation Treaty extends the time window for filing subsequent applications in different jurisdictions to thirty months.[5] Each jurisdiction requires the extended application to be filed in accordance to the local filing rules and, where applicable, this entails the translation of the original application in (one of) the official language(s) of the receiving patent authority. The additional time that

[4]The Convention was originally signed in 1883 and it is still in force today with 177 signatory members. See `https://www.wipo.int/treaties/en/ip/paris/`

[5]The PCT is an international treaty making it possible to seek patent protection for an invention simultaneously in a large number of countries. The granting of patents remains under the control of the national Offices. See `https://www.wipo.int/pct/en/texts/articles/atoc.html`

the treaties grant to applicants for extending their applications abroad is exactly intended to provide enough time for the applicants to translate and adjust their patent applications to the requirements of a foreign patent office. In addition, all the main international IP treaties in force today, the Paris Convention, the PCT, and the TRIPS, impose the national treatment principle, which states that within each jurisdiction, foreign applicants must receive treatment equal to that accorded to domestic applicants.[6] On the one hand, the progressive harmonization of patent law ensures that in most countries applicants should comply with the disclosure requirements. On the other, the existence of an international patent system provides a framework in which inventors can safely extend the protection granted to their inventions beyond their domestic jurisdiction.

All in all, the role of the patent system in fostering knowledge diffusion across countries might have been underappreciated so far, and much of this effect might be driven by the translation of knowledge that would be otherwise trapped behind a language barrier. This lack of attention towards the role of patent translation is even more surprising in an era in which advances in computer science made machine translation tools highly sophisticated and lowered the costs of access to knowledge codified in a foreign language. While research exists on the precision and improvements of machine translation in general, its reliability for translations of patent documents or highly technological information has been only recently assessed.[7] Research done by Zulfiqar et al. (2018) evaluates the accuracy of Google Translate among others, with an emphasis on German scientific literature.[8] Their study shows that the service is reliable and an instantaneous tool for translating ''not only short phrases but even large passages'' into English. Focusing on the disclosure of patent information, research done by Larroyed (2018), shows the effectiveness of machine translation. Larroyed (2018) estimated the level of disclosure at almost eighty percent of the patent content between Western languages and almost 70 percent for Chinese to English translations. The estimate is based on a selection of 100 patents related to clean technologies, applied for between 2013 and 2015. Based on the LISA (Localization Industry Standards Association) Quality Assessment, the machine translation of these patents was evaluated against the manual translation done by native speakers. In addition, the translated documents were blind reviewed by persons skilled in the art. The translation was evaluated based on multiple aspects, such as Accuracy, Terminology, Syntax and Style, followed by classification of

[6]TRIPS Article 3, See `https://www.wto.org/english/docs_e/legal_e/27-trips_03_e.htm`; Paris Convention Article 2: `https://wipolex.wipo.int/en/text/288514`; PCT (not sure which article/paragraph), see: `https://www.wipo.int/edocs/pubdocs/en/wipo_pub_274_2020.pdf`

[7]Research shows, that conventional phrase-based machine translation is able to translate even complex languages into grammatically correct English, yet (Groves and Mundt, 2015) demonstrate it is still error-prone. However, recent improvements, the implementation of Google's neural machine translation further bridges the gap between human and computer-aided translation (Junczys-Dowmunt et al., 2016). The modern approach reduces translation errors by an average of 60 percent compared to a human side-by-side evaluation on a set of isolated simple sentences (Wu et al., 2016).

[8]While Google Patents is the database storing the patent information, Google Translate is the backbone, providing the translation of the content.

errors as minor, major or critical. The average number of errors from both evaluations was 22.75 for Western languages, resulting in a disclosure of 80 percent of information. Although the author states that machine translation still needs improvement, it ''clearly discloses patent information'' and represents one of the main tools of communication within the patent system. Indeed, machine translation is increasingly used by patent specialists, companies, and even patent authorities. The examination guidelines of European Patent Office currently states that ''in order to overcome the language barrier constituted by a document [...], it might be appropriate for the [patent] examiner to rely on a machine translation of said document'' (EPO, 2021).[9]

In short, one of the channels through which patent disclosure positively affect follow-on innovation might be the translation of otherwise unavailable technical knowledge. If that is the case, we would expect that the automatic translation of large body of patent literature previously unreachable for a substantial share of the global population of inventors should have a relevant and positive impact on cumulative innovation. The aim of this paper is exactly to contribute to the discussion on the role of patent disclosure in fostering cumulative innovation and to provide evidence about its importance for international knowledge flows. The next section discusses the strategy we adopt to identify the effect of automated translation of patents on cumulative innovation.

3 Identification strategy

To investigate the impact of machine translation of patent documents on follow-on innovation, it is necessary to identify a well-defined study population, a set of treatment conditions, where it is easy to distinguish between a treatment and control group and between pre-treatment and post-treatment time periods. Ideally, we would run an experiment on a sample of patents searchable on a publicly available patent search website that are written in a specific language, for which no translation, automated or otherwise, is available. We would then randomly split the group in a treated and a control set and make available a machine translation for the patents in the treatment group. We would then assess whether the patents in the treatment group lead to more follow-on innovation than patents in the control group after the treatment. Even though it would be hard to implement such an experiment, in this paper, we exploit a natural experiment that could be considered a close substitute of the ideal setting described above. In 2013, Google launched a major improvement to Google Patents, Google's search engine that allows to search and read the full-text of patents is-

[9]The USPTO is also highlighting the importance of machine translation in the examination guideline by stating that examiners are ''encouraged to use [...] machine translations where possible in the early phases of examination''. See `https://mpep.uspto.gov/RDMS/MPEP/e8r9#/current/d0e113207.html`

sued and filed at several different patent authorities.[10] In September 2013, Google patents added the text of the patent documents from the China National Intellectual Property Administration (CNIPA), the Chinese patent office, using Google's neural machine translation service Google Translate to automatically translate patents, available only in Chinese till that moment (Orwant, 2013). From September 17, 2013, onward the text of all Chinese patents was available on Google patents in both its original language and in English. The observational data from this natural experiment allows us to construct an ideal treatment group: patents published in China before September 17, 2013.

Choosing patented inventions realized in China as a treatment group is highly interesting for multiple reasons that go beyond their publication on the Google patent platform. In global comparison, China's patent activity experienced unprecedented growth both in terms of quantity, CNIPA received the highest number of patent applications in recent years (WIPO, 2019; WIPO, 2020), and quality. Dominguez Lacasa et al. (2019) show China's growing relevance in the innovation and intellectual property landscape in the context of technological catch-up by the BRICS economies. Their findings show that China is unique among BRICS in terms of rapid improvements of the technological intensity, fast structural change in the direction of dynamic frontier activities, and technology diversification, which is also expressed in the diversification of technological knowledge. China's rapid scientific progress is also evident in the production of scientific papers. Zhou and Leydesdorff (2006) provide evidence that its contribution to world science shows exponential growth not only in quantity but also in quality. Along with the exponential increase of scientific publications, the citation rates of Chinese publications are increasing exponentially. However, knowledge produced in China seems to be more relevant in certain technological areas, such as renewable energies (Trancik, 2014) and physical sciences and electronics (Zhou and Leydesdorff, 2006). Especially in the fast-growing field of data processing China evolved into a technological leader. In 2017 China's global share of research papers in the field of Artificial Intelligence has vaulted to 27.68 percent. Also the number of companies working in this domain grew to 1,189, second only to the US (Li et al., 2021).

The increasing centrality of China in the global knowledge economy makes Chinese patents the ideal treatment group to assess the impact of improved access to knowledge through automated translations. To identify a suitable control group, we searched for jurisdictions that were not affected by the Google Patent improvement of 2013 that have patenting activities comparable to China. South Korea proved to be a good candidate. Korean patents are locked behind a similar language barrier as their Chinese counterparts. Without being translated, documents published in Korean are, to a large extent, as inaccessible for English-speaking inventors as patents written in the Chinese language. Unlike for the CNIPA, patents filed at the Korean Intellectual Property Office (KIPO) were added to

[10] Google Patents currently indexes more than 87 million patents and patent applications. See `https://patents.google.com/`

the Google Patent platform only on August 30th 2016, about three years after the inclusion of the Chinese documents. Therefore, while English speaking inventors were able to access the knowledge disclosed in Chinese documents through machine translations available on Google Patents, the content of Korean patents remained behind a language barrier for a longer period (Wetherbee, 2016). In addition, South Korea is also an important global competitor in the IP domain, responsible for the most patent filings after China, the U.S., and Japan in 2019 (WIPO, 2020) and most filings in the field of big data technologies and AI after China and the US.

Comparing follow-on innovations generated by patents filed at the Korean and at the Chinese patent office before and after the machine translation of the Chinese documents operated by Google is the corner stone of our identification strategy. However, the identification of the effect of machine translation confront us with a few additional challenges.

First, we have to ensure the comparability of the treated and the control group. In particular, patents filed at CNIPA and KIPO, may have a different propensity to be extended to a foreign jurisdiction and, therefore, to be officially translated to another language to seek patent protection abroad. To mitigate the risk of such a relevant confounding factor, we decided to focus exclusively on granted patents filed at CNIPA (KIPO) and never extended (or applied for) to another jurisdiction, i.e., we constrain the sample to single-child patent families. This choice substantially reduces the chance of the invention being translated and publicly disclosed to the non-Chinese (Korean) speaking world through a mechanism other than Google's machine translation. Limiting the data to single-child families, we isolate a body of knowledge that is only available at the national level and that remained likely trapped behind the language barrier for Western inventors. To further reduce the possibility of access to this knowledge before the machine translation implemented by Google in 2013, we remove from our working sample any patent that lists an inventor with a residence address outside of China (South Korea) and any patent whose applicant is a non-Chinese (non-Korean) entity, based on the residence address of the applicant. The rationale for this choice is that non-Chinese-speaking (non-Korean-speaking) or mixed teams could already share knowledge in a different language using communication channels beyond the one offered by patent disclosure.

Second, we have to determine a way to measure follow-on innovations to our treated and control patents and, more specifically, a measure of follow-on innovation able to capture the impact of a reduced cost of access to information codified in a foreign language. Here we follow an extensive literature that use patent citations as quantifiable trails of knowledge flows (e.g., Belenzon, 2012; Galasso and Schankerman, 2014; Moser et al., 2018). As explained by Hall et al. (2005), if patent B cites patent A, it implies that patent A represents a piece of previously existing knowledge upon which patent B builds and over which B cannot have a claim. Following the works in this strand, we use the count of forward citations as a measure of follow-on invention. To better capture the effect of machine trans-

lation of documents written in Chinese, we count the number of citations to the patents in our treated and control group arriving from patents granted at the USPTO and filed by US-based inventors.[11]

Therefore, we will adopt a difference-in-differences approach and compare the number of forward citations made by US inventors to our treated patents—patents filed at CNIPA by Chinese applicants that were never extended to another jurisdiction—before and after the introduction of machine translations by Google Patents, with the number of US citations received by the patents in our control group— patents filed by Korean applicants, never extended to a foreign patent authority and translated by Google patents only in the last part of 2016. We would interpret a relative increase for the treated groups after the translation as an indicator of the positive influence of machine translation on knowledge diffusion. This would strongly suggest that US inventors read the translated documents and that the information disclosed in the patent affected their invention process.[12] The next section presents the data and the methodology we use to estimate the effect of machine translation.

4 Data and methods

4.1 Empirical model

As discussed in section 3, we use a difference-in-differences approach to assess the effect of machine translations on follow-on innovation by contrasting a group of Chinese patents, our treated group, and a group of Korean patents, our control group. To do so we estimate an econometric model that follows from our identification strategy:

$$y_{it} = \beta_1 Chinese_i + \beta_2 Translated_{it} + \beta_3 Chinese \times Translated_{it} + \delta_{ipc} + \delta_{year} + \gamma X_i + \epsilon_{it}$$

where the dependent variable y_{it} is the number of citations received from US patents to Chinese or Korean patent i, divided into the time t before and after the introduction of the automatic translation by Google. *Chinese* reports whether the focal patent i is Chinese or not (South-Korean) and *Translated* is our treatment indicator, which takes the value 1 in the period after the automated translation and value 0 before the translation. Our main variable of interest is $Chinese \times Translated$: a positive and significant coefficient for this

[11] Yet, the limitations of citation data need to be highlighted as well. Their use to track follow-on inventions has been criticized, especially the role of examiners (Sampat, 2010) and citations added by examiners (Alcácer et al., 2008). Since we estimate differences between trends and view examiner citations separately, these limitations do not affect the effect that drives our results, which is ensured by implementing several checks to account for their robustness, discussed in subsubsection 5.4.2 and subsubsection 5.3.1

[12] Other scholars discuss a more fundamental flaw of citations, arguing that a small minority of patent applications generates a large majority of patent citations (Kuhn et al., 2019). To address this issue, we remove citations arriving from patents with more than 250 backward citations. Additionally, we re-estimated the regression removing patents with more than 400 and more than 500 backward citations, as well as without removing any of these patents, which still showed significant and positive results.

interaction term would confirm a positive effect of the machine translation of Chinese patents on follow-on invention in the US δ_{ipc} accounts for International Patent Classification (IPC) fixed effects and δ_{year} includes publication year fixed effects of the patents in our sample.

The vector X_i captures features of the focal patent, including number of applicants and the number of inventors, listed on the patent document, the count of different IPC sections assigned to the patent, the number of citations received from patents published within the same country, the number of backward citations and the number of independent claims included in the patent. In subsection 4.2 we further discuss the relevance of these features. ϵ is the error term.

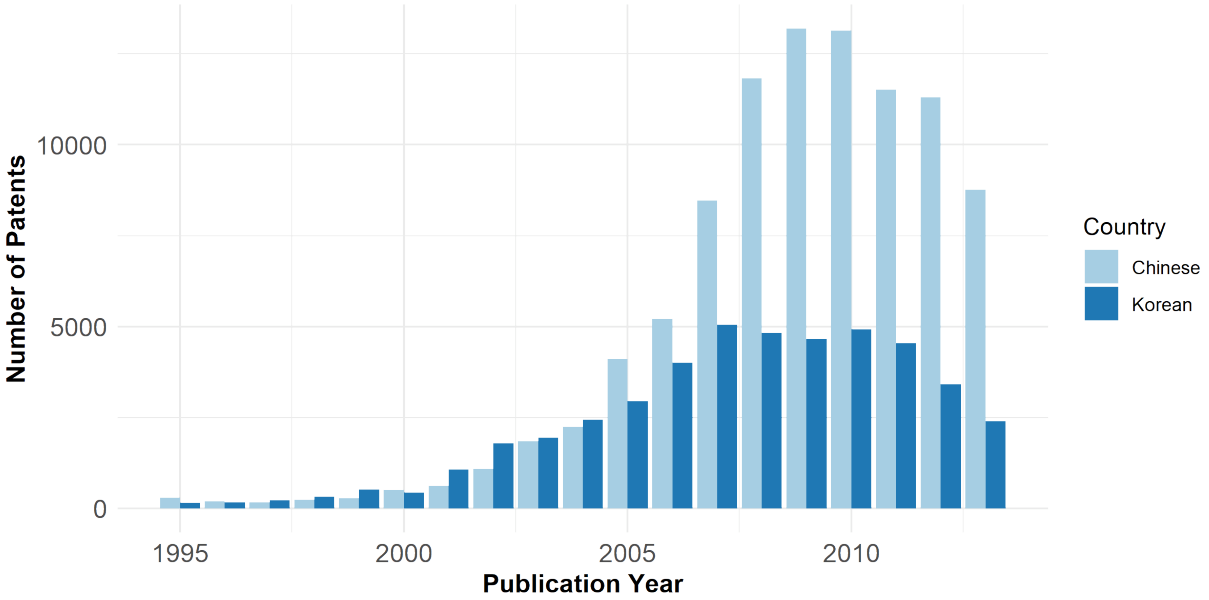
We estimate our baseline model using an OLS and, to account for the count nature of the citation data, a Poisson regression.It is important to note, that the validity of the difference-in-difference analysis is dependent on two factors: the absence of any pre-event trends within the data and the assumption, that both groups follow a similar trend previous to the treatment. We discuss potential dissimilarities between the treated and the control group and the presence of possible pre-trends in subsection 5.4.

4.2 Data preparation and descriptive statistics

Our main data source is the EPO'S Worldwide Patent Database PATSTAT (April 2020 release). PATSTAT is one of the most comprehensive patent databases and contains data on more than 100 million patents and patent applications from 90 patent authorities. To populate our treatment and control group, we exploit PATSTAT to identify patents granted by the Chinese and Korean patent authorities and first published between 1995 and September 17, 2013. To ensure that they were never extended, and hence potentially translated, outside the country of origin, we limit our samples to patents filed by Chinese (Korean) residents that belong to patent families of size one, i.e. they are not claimed as priority filing by any other patent application within or outside the country of first filing. Additionally, we removed Utility models from the set due to the different standard for protection required by the Chinese and the Korean patent law. Based on the person country code provided in PATSTAT, we also excluded all patents that listed among its inventors or applicants an individual or an entity that reported a foreign residence address. The final set contains 49,004 Chinese and 19,758 Korean patents. Figure 1 displays the distribution of patents by country of filing and year of filing. Both countries follow a similar trend, with a peak between 2008 and 2010, though the number of Chinese patents grew drastically since 2005. Figure 2 shows the distribution by IPC sections.[13] Previous contributions agree on the importance of controlling for technology classification and the focus on specific industrial sectors (Baruffaldi and Simeth, 2020; Berkes and Nencka, 2020; Furman et al., 2018). Pre-

[13]The International Patent Classification (IPC) provides a hierarchical system of letters for the classification of patents and utility models according to the different areas of technology to which they pertain. The IPC divides technology into eight sections, classified into approx. 70000 classes (WIPO, 2020).

Figure 1: Frequency distribution of patents by Country of filing.



liminary descriptive statistics shown in Figure 2 already indicate a significant difference. The technological sections Physics (G) and Electricity (H) account for more than 50 percent of all patents. China's catching up to the global frontier in computing and data processing can also be observed in our data. If we split Physics into classes, the majority of Chinese patents are classified as 'G06', which refers to 'computing' including the 'processing of information and the structure of the database.' Figure 3 confirms that the majority of the patents in our sample belong to a technological domain in which China supposedly leads.

We exploit the PATSTAT database to construct our dependent variable by counting forward citations arriving from US patents and divide the citation count into pre-treatment and post-treatment time period. Our treatment being the machine translation of the Chinese patents in September 17, 2013. The pre-treatment time period includes data from January 1, 2008 to September 17, 2013 and the post-treatment time period a similar timespan from September 18, 2013 to December 31st, 2017. Forward citations to the USPTO were chosen due to the stricter rules of the US patent office concerning citations. Because of the incompleteness of non granted patent information in the original data set, only forward citations from granted patents were taken into account. Following the critique discussed by Sampat (2010), we focus exclusively on forward citations made by the applicants.[14] Arguing that citations added by others, e.g., the patent office examiner, were not relevant in tracking the diffusion of technical knowledge from the original inventor of the Chinese (Korean) invention to the US However, in section 5.3.1 we use examiner citations to control for the robustness of our approach.

[14]By design all Chinese and Korean patents used in the analysis were cited at least once by a US patent.

Figure 2: Percentage frequency distribution of patents by Country of filing and IPC sections they are assigned to.

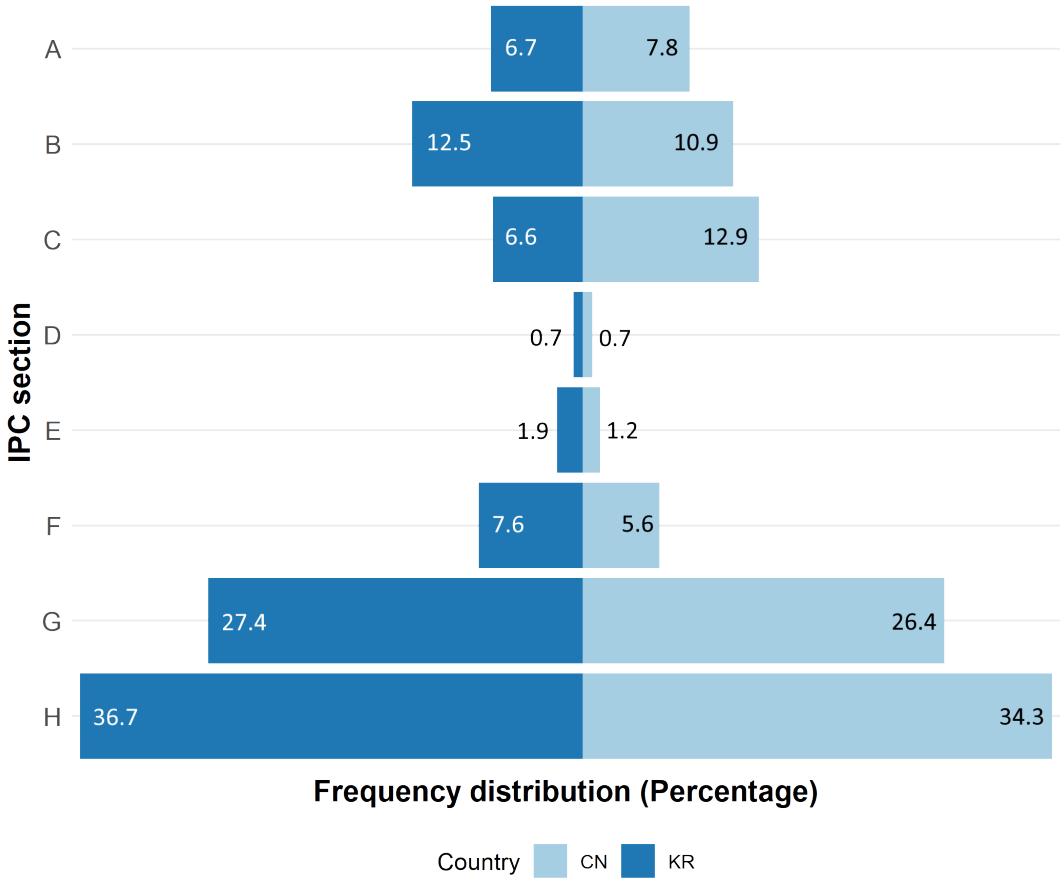
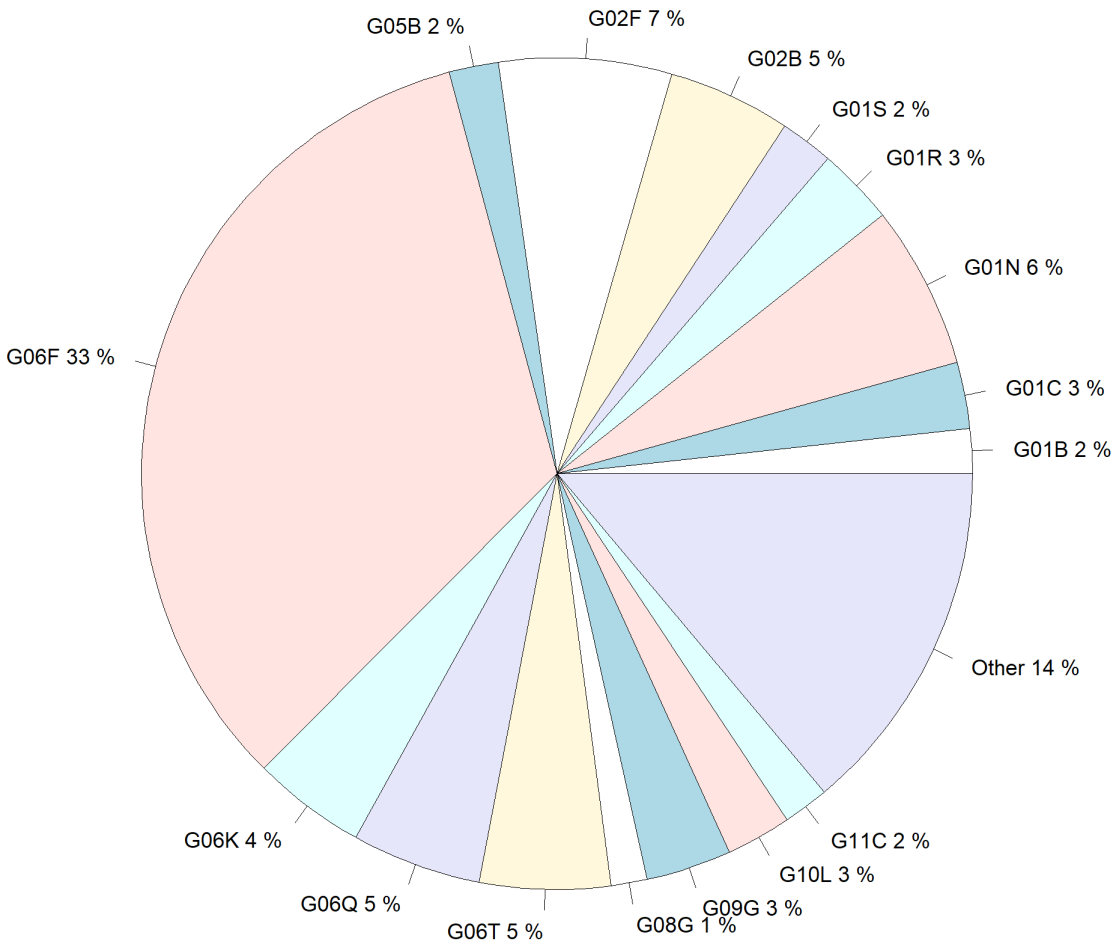


Figure 3: IPC classes assigned to Chinese patents, only considering patents related to Physics (IPC section G).



To limit the probability of a native speaker being involved with the US patent application process, all citations from US patents filed by Chinese (Korean) inventors and companies were excluded from the count. The distinction was made based on the residence address reported in PATSTAT, leaving 128,292 citations to the patents in our full sample. However, this approach does not fully rule out the possibility that some of these citations may come from Chinese (Korean) inventors able to read Chinese (Korean) patent applications independently of the machine translation offered in Google patents. Patent documents, and hence the PATSTAT database, do not report the nationality of the inventors but merely their registered address of residence. A Chinese inventor living in the US would be labeled as US inventor following the method discussed above, yet she would still be able to fully understand the Chinese language and would not need a translation to access information disclosed in Chinese patents. Following the approach of Breschi and Lissoni (2009); Breschi et al. (2015), to have a more fine-grained distinction, in subsection 5.2 we made use of two different natural language processing libraries to construct two additional dependent variables that exclude citations from two types of patents. First, we exclude citations coming from patents filed by inventors that have a name that the algorithm identifies as Chinese or Korean. Second, to mitigate issues that may arise from names that are used for both Chinese and Korean individuals, we exclude from the citation count all the citations coming from patents with inventors that have names the algorithm identifies as Asian names.

After constructing the outcome variable, we also construct several control variables capturing characteristics of the focal patent that may potentially influence the relationship of interest. From PATSTAT, we extract the number of applicants and inventors as listed in the patent document, the number of IPC classes a patent is assigned to, the number of backward citations received, and the number of citations received from other Chinese or Korean patents.

The number of IPC classes reports the total number of four-digit IPC classes which are assigned to a patent. Patent documents covering many IPC classes are commonly used in literature as proxy for the technological scope of the invention (Harhoff et al., 2003).

Both, the number of backward citations received, and the number of citations received from other Chinese or Korean patents, are used as a proxy for the value of patents. Later, represents the patent's significance within the Chinese or Korean language environment and therefore relates to the importance for US inventors to access its content.

Additionally, we used Google Patents to create the variable number of claims, which reports the number of independent claims listed in the patent application. Independent claims describe the essential features of the invention and this variable could then be interpreted as a proxy for the scope of the invention.

Table 2 shows summary statistics of the key variables used in our analysis for the full sample. Table 1 considers instead the total sample and the two sub-samples of treated and

control focal patents. On average, the treated patents appear to have a smaller number of received citations. Moreover, Chinese patents make on average less backward citations and have a significantly smaller number of claims.

Table 1: Average of the dependent variable and the control variables computed over control group, Korean patents, and treated group, Chinese patents.

	Control	Treated	Mean-Diff
Received Citations	2.10	1.84	0.26
No. of applicants	1.10	1.09	0.01
No. of inventors	2.61	3.43	-0.81
No. of IPC classes	1.33	1.60	-0.27
No. of CN/KR citations	1.38	1.31	0.06
No. of backward citations	0.94	0.79	0.15
No. of claims	3.55	1.49	2.06

Notes: The sample consists of 19,758 Korean patents and 49,004 Chinese patents.

Table 2: Summary Statistics; treated and control focal patents

	Treated				Control			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Received Citations	1.848	2.679	1	144	2.123	4.119	0	194
No. of applicants	1.092	0.346	1	9	1.109	0.394	1	6
No. of inventors	3.439	2.378	1	29	2.637	2.009	1	39
No. of IPC classes	1.605	0.911	1	30	1.339	0.614	1	8
No. of CN/KR citations	1.319	3.945	0	196	1.390	3.721	0	129
No. of backward citations	0.791	1.809	0	32	0.950	1.644	0	15
No. of claims	1.499	1.016	0	24	3.583	3.589	0	92

Notes: Total No. of observations 68,762, Korean patents (control group) 19,758, Chinese patents (treated group) 49,004.

5 Results

5.1 Baseline result

Table 3 presents the results of the baseline OLS estimates for the main dependent variable. As previously mentioned, to better identify the effect of knowledge diffusion, we restrict the analysis exclusively to the sample of treated focal patents and their correspondent control patents that have been cited at least once by a US patent filed between 2008 and 2018. The results, reported in columns 1, show a positive and highly statistically significant coefficient for the interaction term $Chinese \times Translated$, confirming a positive relationship between

machine translation and follow-on invention. The OLS results suggests that Chinese patents receives on average about 0.186 more citations than Korean patents for which no automated translation existed at the time on Google Patents. Compared to an average of 1.93 citations received by the patents in our sample, this implies a 9.7 percent increase in forward citations after the treatment period.

Table 3: The effect of machine translation on forward citations, OLS and Poisson estimates.

	OLS	Poisson
Chinese x Translated	.186***	.138***
	(.025)	(.012)
Chinese	−.217***	−.163***
	(.019)	(.008)
Translated	−.772***	−.685***
	(.021)	(.010)
No. of applicants	−.022	−.024***
	(.017)	(.007)
No. of inventors	.004	.005***
	(.003)	(.001)
No. of ICP classes	.019***	.019***
	(.007)	(.003)
No. of CN/KR citations	.062***	.019***
	(.001)	(.000)
No. of backward citations	−.073***	−.020***
	(.004)	(.001)
No. of claims	.011***	.010***
	(.003)	(.001)
Class fixed effects	Yes	Yes
Publn. year fixed effects	Yes	Yes
Mean of dep. var.	1.927	1.927
Num. obs.	68,762	68,762

Notes: Significant at ***1%, **5% and 10%*. Robust Standard errors in parentheses. The dependent variable records the cumulative forward citations that the focal patent i has received from U.S. patents earliest filed between 2008 and 2019. Not-granted patent citations and foreign (not-U.S.) citations are excluded from the count of the dependent variable. The variable Chinese x Translated is a binary indicator that identifies those focal patents that was treated – added to Google Patent database and automatically translated in 2013. To address the skewness of citation data, that a small minority of patent applications is generating a large majority of patent citations (Kuhn et al., 2019), citations arriving from patents with more than 250 backward citations were removed.

To account for the fact that citations are count data, we also estimate the baseline model using a Poisson regression. Table 3 reports the estimates in column 2, which yields a slightly smaller marginal effect than OLS. The Poisson estimates an increase of 0.138, implying an increase of 7.2 percent.

While the baseline results show a positive and statistically significant effect, the effect

Table 4: The effect of machine translation on forward citations divided into technological areas – Poisson estimates.

	Chinese x Translated	Chinese	Translated	No. Obs.	Avg.
A (Human Necessities)	$-.363^{***}$	$.152^{***}$	$-.338^{***}$	4,810	2.200
	(.041)	(.030)	(.038)		
B (Performing Ops.)	$.128^{***}$	$-.058^{**}$	$-.314^{***}$	7,003	1.713
	(.035)	(.024)	(.027)		
C (Chemistry)	$.228^{***}$	$-.169^{***}$	$-.799^{***}$	9,868	1.941
	(.040)	(.026)	(.037)		
D (Textiles)	.033	.079	$-.609^{***}$	721	1.778
	(.108)	(.069)	(.093)		
E (Constructions)	.044	.058	$-.449^{***}$	1,011	1.653
	(.084)	(.057)	(.065)		
F (Engineering)	.008	$-.098^{***}$	$-.380^{***}$	3,691	1.729
	(.043)	(.031)	(.034)		
G (Physics)	$.273^{***}$	$-.208^{***}$	$-.755^{***}$	16,817	1.889
	(.025)	(.016)	(.021)		
H (Electricity)	$.181^{***}$	$-.266^{***}$	$-.875^{***}$	24,474	2.000
	(.020)	(.012)	(.017)		

Notes: Significant at ***1%, **5% and 10%*. Robust Standard errors in parentheses. OLS estimates are not reported in the table, however, their estimates follow along with the reported results. The focal patents were split into technological areas, based on their International Patent Classification. This table shows only the regressor of interest. Regressors not listed are those of Table 3, see text for details. The dependent variable records the cumulative forward citations that the focal patent i has received from U.S. patents earliest filed between 2008 and 2019. Not-granted patent citations and foreign (not-U.S.) citations are excluded from the count of the dependent variable. The variable Chinese x Translated is a binary indicator that identifies those focal patents that was treated – added to Google Patent database and automatically translated in 2013. To address the skewness of citation data, that a small minority of patent applications is generating a large majority of patent citations (Kuhn et al., 2019), citations arriving from patents with more than 250 backward citations were removed.

of translation on knowledge diffusion becomes more evident when we split the data by technological sections. The Poisson estimates reported in column 1 of Table 4 point to a strong effect especially in Chemistry, Physics, and Electricity classes. With an increase of 14.5 percent in forward citations after the translation, patents from the Physics domain seem to drive the results. This result appear to fit well with our interpretation of the result: improvements in the accessibility to an otherwise unavailable body of knowledge fostered cumulative innovation. As previously discussed, Physics is the IPC section under which technologies related to computing are categorized. Information technology, and computer science in particular, is one of the domain in which China has rapidly evolved to a global leader in the past ten to fifteen years (O'Meara, 2019; Li et al., 2021). The fact that the benefits of machine translation are especially evident in fields in which China is closer to the technological frontier seems to suggest that the improved access to knowledge allows US inventors to identify and use relevant technological contributions made by Chinese inventors. In addition, the different effects by IPC categories are coherent with the evidence provided in Galasso and Schankerman (2014). As expected, we find a stronger effect of the translation in technological areas such as ICT where innovation is more cumulative in nature. Instead, in the area of Human Necessities, which includes pharmaceutical products and medical drugs, disclosure happening through translation appears to have a negative effect on subsequent innovation. In those fields innovation happens in a less cumulative fashion and the discovery of previously unknown patented knowledge may reduce the likelihood of obtaining a patent for an invention that builds on a closely related body of knowledge.

5.2 Citations origin

As discussed in subsection 4.2, one major limitation of the data that might potentially threaten our interpretation of our baseline estimates, lies in the way patent documents reports the country of origin of inventors and applicants: PATSTAT reports their registered residence address and not their actual nationality. For this reason, a Chinese inventor who lives in the United States would simply be registered as US residents in PATSTAT. However, a Chinese inventor living in the US obviously does not face the same language barrier as her non-Chinese speaking colleagues when it comes to searching and reading the Chinese-only patent literature. In such a scenario, we cannot entirely rule out the hypothesis that the effect we recover in the baseline estimates could be driven by confounding factors such as sudden increase in high-skilled Chinese migration happening contemporaneously to our treatment. Indeed, the number of high-skilled workers from China is steadily increasing and since 2013 they represent the second largest group receiving H1-B visas (USCIS, 2020).

To rule out the concern that this migratory phenomenon is driving our result, we use two natural-language-processing algorithms to infer the nationality of the inventors based on their name and surname. First, we use the library *name2nat* (Park, 2020), which was trained

on names and nationalities extracted from Wikipedia and assign a specific nationality to a combination of first name and surname. Second, we use *Ethnicolr*, a machine-learning-based classifier trained on a specific dataset and implemented in Python to impute the ethnic background (Laohaprapanon and Sood, 2021). This algorithm assigns persons based on their first and last names to categories that combine ethnic backgrounds. The algorithm was trained on Florida voter registration data and Wikipedia data from 2000 and 2010. It assigns to a given name a specific probability of belonging to one of four classes, 'white', 'black', 'asian' or 'hispanic', and then assign the imputed ethnic background by selecting the one associated with the highest probability.

Using the *name2nat* library on the names of the US patent inventors citing our focal patents, we identified patents that listed inventors considered as Chinese or Korean by the algorithm. We then constructed an alternative outcome variable for which we remove these patents from the citation count. Using this approach, we excluded 43,356 citations from the original count, which left us with 84,936 citations from US patents without any Chinese or Korean inventors involved.

We instead use the *Ethnicolr* classifier to create a citation count without any patents that list inventors with a combination of name and surname that the algorithm identifies as Asian. We remove patents with inventors that have probability of more than fifty percent of being Asian – which results in the removal of 79,235 citations.[15]

We then re-estimate the baseline model using the newly created dependent variables Table 5 reports the marginal effects recovered through the Poisson regression, together with the OLS coefficient as a benchmark.

All results show a positive and statistically significant effect, which is overall stronger than the baseline effect. Compared with the average of 1.28 citations received by focal patents, the Poisson estimate implies (columns 2) a 9.5 percent increase in forward citations from patents without any inventor with a Chinese or Korean name involved, after the translation. If we consider citations arriving from patents without any inventors with an Asian name, the estimator yields a stronger marginal effect, implying a 13.1 percent increase of forward citations.

The overall difference with the baseline estimate appears to be rather small, but if we look into the different technological areas, in some areas we observe substantial changes. Table 6 reports the results of the Poisson regression of the dependent variable without Chinese and Korean inventors involved in column 1 and without any Asian inventors in column 4.

While the tendencies of the results are similar to the baseline estimates, the effect is once again mainly driven by fields in which China became is on the technological frontier, the effect in the Physics domain almost doubles after the removal of patent citations from

[15]To account for the accuracy of name2nat and Ethnicolr, we randomly choose 1000 Chinese and 1000 Korean names and manually verified the prediction. We calculated an accuracy of 88.3 percent for name2nat and 96 percent for Ethnicolr's predictions.

Table 5: The effect of machine translation on citations arriving from patents without Chinese and Korean inventors and without Asian inventors involved. OLS and Poisson estimates.

	Baseline	CN/KR removed	Asian removed
OLS Estimations			
Chinese x Translated	.186***	.232***	.200***
	(.025)	(.023)	(.018)
Chinese	−.217***	−.347***	−.299***
	(.019)	(.017)	(.014)
Translated	−.772***	−.606***	−.418***
	(.021)	(.019)	(.015)
Poisson Estimations			
Chinese x Translated	.138***	.122***	.097***
	(.012)	(.009)	(.007)
Chinese	−.163***	−.264***	−.226***
	(.008)	(.007)	(.005)
Translated	−.685***	−.476***	−.301***
	(.010)	(.008)	(.006)
Num. obs.	68,762	68,762	68,762
Mean of dep. var.	1.927	1.278	0.738

Notes: Significant at ***1%, **5% and 10%*. Robust Standard errors in parentheses. This table shows only the regressor of interest. Regressors not listed are those of Table 3, see text for details. The dependent variable records the cumulative forward citations that the focal patent i has received from U.S. patents earliest filed between 2008 and 2019. Not-granted patent citations and foreign (not-U.S.) citations are excluded from the count of all dependent variables. The variable Chinese x Translated is a binary indicator that identifies those focal patents that was treated – added to Google Patent database and automatically translated in 2013.

Table 6: The effect of machine translation on citations arriving from patents without Chinese and Korean inventors and without Asian inventors involved divided into technological areas – Poisson estimates.

	CN/KR removed	No. Obs.	Avg.	Asian removed	No. Obs.	Avg.
A (Human Necessities)	$-.311^{***}$	4917	1.747	$-.133^{***}$	4,810	1.158
B (Performing Ops.)	$-.006$	7073	1.302	.017	7,003	0.814
C (Chemistry)	$.178^{***}$	9942	1.399	$.098^{***}$	9,868	0.754
D (Textiles)	.111	724	1.202	.094	721	0.778
E (Constructions)	.016	1020	1.372	$-.003$	1,011	1.084
F (Engineering)	.002	3735	1.356	$.071^{**}$	3,691	0.921
G (Physics)	$.301^{***}$	16984	1.253	$.175^{***}$	16,817	0.699
H (Electricity)	$.183^{***}$	24682	1.134	$.167^{***}$	24,474	0.611

Notes: Significant at ***1%, **5% and 10%*. Robust Standard errors in parentheses. OLS estimates are not reported in the table, however, their estimates follow along with the reported results. The focal patents were split into technological areas, based on their International Patent Classification. This table shows only the regressor of interest. Regressors not listed are those of Table 3, see text for details. The dependent variable record the cumulative forward citations that the focal patent i has received from U.S. patents earliest filed between 2008 and 2019 and with Chinese and Korean inventors removed (columns 1-3) and Asian inventors removed (columns 4-6). Not-granted patent citations and foreign (not-U.S.) citations are excluded from the count of the dependent variable. The variable Chinese x Translated is a binary indicator that identifies those focal patents that was treated – added to Google Patent database and automatically translated in 2013.

Chinese and Korean inventors. Based on an average of 1.253 citations per patent, Chinese patents related to Physics show a 24 percent increase in forward citations after the treatment period in comparison to the control group. Column 4, the citation count without Asian inventors involved, confirms the results, indicating a 25 percent increase in the number of citations after the implementation of the machine translation by Google.

These estimates indicate that our initial assumption discussed in the identification strategy in section 3 is highly relevant. The effect is not driven by citations coming from patents with inventors already capable of speaking the language but from inventors who likely got access through translation.[16] As in the baseline, this is especially evident in areas in which China has become a technological leader, such as computing and data processing.

5.3 Small Entities

Even though Google Translate is arguably the most popular machine translation service, it is fair to assume that specific kind of organizations such as large multinational enterprises may have in-house professional translation and machine translation services already in place in house and they might use these services also to parse and examine the patent literature relevant to them. Clearly, we should expect the benefits of a generalized machine translation

[16]In the rest of the paper we report the results only for these narrowly defined outcome variables as they allow us to mitigate potential confounding factors. Nevertheless, we obtain similar results in terms of significance and sign of the effect using our original depend variable.

service to be quite limited in such a case. Yet, these professional translation services or ad hoc machine translation software are pricey, and their cost can be an obstacle especially for small entities or independent inventors. Being open access and free of charge, Google Patents sets itself apart from paid solutions, therefore lowering the barrier for its usage and enabling small entities to make use of the service. For this reason, we would expect a stronger effect of the treatment for citations arriving from patents which were filed by smaller entities and single inventors.

Since there is a high chance of big companies having already internal machine or professional translation services in place, we constructed an additional alternative outcome variable that exclusively counts citations arriving from US patents filed by small entities. To identify small entities, we follow the definition used by the USPTO (section 3 of the United States Small Business Act) to grant a fifty percent reduction in application fees. This definition includes independent inventors, small businesses, and nonprofit organizations. Using the USPTO Patent Examination Research Dataset (PatEx) (Graham et al., 2015), we identify the citations arriving from small entities and create the new outcome variable.[17]

Table 7: The effect of machine translation on citations from patents filed by small entities: counting all citations from U.S. patents without Chinese and Korean inventors – reporting OLS and Poisson estimates.

	OLS CN/KR removed	Poisson CN/KR removed
Chinese x Translated	.054***	.040***
	(.007)	(.004)
Class fixed effects	Yes	Yes
Publn. year fixed effects	Yes	Yes
Mean of dep. var.	0.214	0.214
Num. obs.	137524	137524

Notes: Significant at ***1%, **5% and 10%*. Robust Standard errors in parentheses. This table shows only the regressor of interest. Regressors not listed are those of Table 3, see text for details. The dependent variable records the cumulative forward citations that the focal patent i has received from U.S. patents earliest filed between 2008 and 2019 and filed by small entities without Chinese/Korean inventors. To identify small entities we follow the definition used by the USPTO (section 3 of the United States Small Business Act) to grant a fifty per cent reduction in application fees. In addition, not-granted patent citations and foreign (not-U.S.) citations are excluded from the count of the dependent variable. The variable Chinese x Translated is a binary indicator that identifies those focal patents that was treated – added to Google Patent database and automatically translated in 2013.

Table Table 7 reports the results. Both the OLS and Poisson estimates are positive and statistically significant. The non-linear estimator yields a slightly smaller marginal effect

[17]To flag small entities, we used the USPTO Patent Examination Research Dataset (PatEx). The PatEx dataset provides access to the bulk data collected by the Public Patent Application Information Retrieval system (Public PAIR) and is available at `https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-examination-research-dataset-public-pair`

than OLS. Yet, the magnitude of the effect is much stronger than the baseline results if we only consider citations from small entities. Based on an average of 0.21, the model estimates an increase of 18.6 percent (25.2 percent for the OLS) in citations after the treatment.

These results confirms that the benefits of the improved access to knowledge behind a language barrier are especially important for those categories, small businesses, nonprofit organizations and researchers, for which the costs of access to this knowledge through alternative means are higher and need a free of charge and open access service.

5.3.1 Examiner citations

As alluded to in section 3, using patent citations as an indicator for knowledge flows has limitations that need to be addressed. One important aspect to consider is whether citations are added by the applicant or by the patent examiner that review the patent application. Previous literature acknowledges that examiner citations add measurement error and not reporting them separately adds unknown noise to the data (Alcácer and Gittelman, 2006). In our identification strategy, we only consider citations added by the applicants, even though patent examiners – government agents who approve patent applications – are also involved in drafting the content of patents. However, their citations are unlikely to reflect knowledge flows since they are not involved in the innovation and innovation process itself (Jaffe and Trajtenberg, 1999). Yet, Alcácer et al. (2008) show that examiner citations account for 63 percent of all citations on the average USPTO patent, making them a relevant factor to consider for testing our model's robustness.

Besides not being involved with the initial knowledge flow, patent offices and examiners already had translation services in place previous to Google's patent translation. They were already able to search and access the Chinese patent literature before 2013. Hence, citations added by examiners should not be affected by the machine translation services offered by Google. To verify this assumption, we recompute our dependent variable by counting exclusively citations coming from patent examiners.[18]

Table 8 reports the results of the OLS and the Poisson regression. All coefficients associated are not significantly different from zero, showing no change of received citations after the translation in 2013. This result appear to confirm that the treatment effect in the focal analysis is not driven by confounding factors that would also affect the examiner behavior, but by the inventors actually benefiting from improved access to the Chinese patent literature.

[18] An interesting observation is that the examiner citations in our data, only account for roughly 4.2 percent of overall citations.

Table 8: The effect of machine translation on citations added by examiners: counting citations from U.S. patents without Chinese and Korean inventors based on core and matched samples.

	Not matched	1:1 matched	2:1 matched
OLS Estimations			
Chinese x Translated	.038*	.061	.002
	(.023)	(.067)	(.058)
Poisson Estimations			
Chinese x Translated	.010	−.032	−.068
	(.026)	(.071)	(.058)
Num. obs.	69,102	30,4391	41,348
Mean of dep. var.	0.023	0.025	0.025

Notes: Significant at ***1%, **5% and 10%*. Robust Standard errors in parentheses. This table shows only the regressor of interest. Regressors not listed are those of Table 3, see text for details. The dependent variable records the cumulative forward citations added by the examiner that the focal patent i has received from U.S. patents earliest filed between 2008 and 2019 (column 1), the 1:1 matched sample (column 2) and the 2:1 matched sample (column 3). Not-granted patent citations and foreign (not-U.S.) citations are excluded from the count of the dependent variable. The variable Chinese x Translated is a binary indicator that identifies those focal patents that was treated – added to Google Patent database and automatically translated in 2013.

5.4 Robustness Checks

As discussed in section 3, there is a general threat to the validity of our approach, that we discuss in this section. It mainly relates to the comparability of the treated and the control group. As the descriptive statistics in 4.2 show, Chinese and Korean patents do present some differences in their observable characteristics and we might have the concerns that these differences could be also correlated to unobservable time-variant factors that are not adequately addressed by the difference-in-differences approach we adopted in our analyses.

To rule out these concerns, we perform two additional robustness checks that confirm the validity of the results presented in the previous sections.

5.4.1 A conditional Difference-in-Difference approach

To better account for potential differences between the treated and the control group, we show that the results are robust to the implementation of an exact matching approach combined with the difference-in-differences method used in the focal analysis, i.e., to the adoption of a conditional Difference-in-Difference approach. We perform an exact matching on the IPC class[19] and the application year of our focal patents. Therefore, this procedure creates bins containing treated and control inventions that belong to the same technology

[19]There are in total 573 unique IPC classes in the data set.

field and were developed around the same time. Considering the imbalance of our core set, having three times as much Chinese patents as Korean patents, we accounted for the robustness of our baseline results by creating two different matching sets. In the first one, each treated patent is matched to one perfect twin in the control group, whereas in the second, we match using a two-to-one ratio, keeping up to two Chinese patents in each bin. In both cases, we removed patents without a suitable twin.

The one-to-one exact matching removes roughly half of the patents, resulting in a loss of 33,864 Chinese and 4,610 Korean patents. The overall sample size of the matched set is 30,288. Exact matching with a ratio of two-to-one gives a sample size of 41,142, removing 24,266 treated patents.

Table 9: The effect of machine translation on forward citations on two different matched samples, OLS and Poisson estimates

	Not matched	Matched	
		1:1	2:1
OLS Estimations			
Chinese x Translated	.232***	.229***	.235***
	(.023)	(.034)	(.028)
Poisson Estimations			
Chinese x Translated	.122***	.070***	.093***
	(.009)	(.014)	(.012)
Num. obs.	68,762	30,288	41,142
Mean of dep. var.	1.278	1.340	1.301

Notes: Significant at ***1%, **5% and 10%*. Robust Standard errors in parentheses. This table shows only the regressor of interest. Regressors not listed are those of Table 3, see text for details. The dependent variable records the cumulative forward citations that the matched focal patent i has received from U.S. patents earliest filed between 2008 and 2019 The matched samples were also re-estimated with citations arriving from non-Asian patents, which yielded statistically significant and positive effects, but not reported in the table. In addition, not-granted patent citations and foreign (not-U.S.) citations are excluded from the count of the dependent variable. The variable Chinese x Translated is a binary indicator that identifies those focal patents that was treated – added to Google Patent database and automatically translated in 2013.

We re-estimate the baseline model from patents without Chinese and Korean inventors involved using the Poisson and OLS estimators based on the two matched samples. Table 9 reports the estimates together with the previously-estimated marginal effects for the full sample as a benchmark. The Poisson regression yields smaller marginal effects than the OLS, yet overall all effects are statistically significant and positive. If we consider citations to the one-to-one matched sample, the OLS estimate implies a 17.1 percent and the Poisson a 5.2 percent increase. The estimators for the two-to-one matched sample yield similar positive and significant effects: an increase of 18 percent (OLS estimate) and 7.2 percent (Poisson estimate) in the citations arriving from US patents after the machine translation.

Comparing the results for the matched samples to the one obtained in the focal analysis, the matched samples show a slightly weaker effect but, overall, the results are qualitatively very similar to the one obtained for the full sample, confirming the validity of the findings discussed above.

5.4.2 Pre-event trends

Another potential threat to our identification strategy comes from the possible presence of pre-trends driving our results. To mitigate this concern, we perform a placebo test. To do so, we construct the data in the same way as introduced in subsection 4.2, but pretend that the machine translation took place in 2011, two years prior to the real treatment. In addition, we removed patents published after September 17, 2011, and adjusted the time period for the collection of citations coming from US patents accordingly.

Table 10 presents the estimates of the fake shock for the core set in columns 1 and 2 and, additionally, only for Physics-related patents in columns 3 and 4.[20]

Table 10: Effect of a robustness test to investigate possible pre-event trends for citations without Chinese and Korean inventors on the core and matched sets - OLS and Poisson estimates.

	Not matched	1:1 matched	2:1 matched
OLS Estimations			
Chinese x Translated	−.016	.008	.023
	(.019)	(.021)	(.018)
Poisson Estimations			
Chinese x Translated	−.029**	−.013	.001
	(.012)	(.013)	(.011)
Num. obs.	52,651	25,481	33,608
Mean of dep. var.	1.829	1.189	1.189

Notes: Significant at ***1%, **5% and 10%*. Robust Standard errors in parentheses. This table shows only the regressor of interest. Regressors not listed are those of Table 3, see text for details. To investigate possible pre-event trends, we removed patents published after 2011. The dependent variable records the cumulative forward citations that the focal patent i has received from U.S. patents earliest filed between 2006 and 2017 (column 1), on the 1:1 matched sample (column 2) and 2:1 matched sample (column 3). Not-granted patent citations and foreign (not-U.S.) citations are excluded from the count of the dependent variable. The variable Chinese x Translated is a binary indicator that identifies those focal patents that received the fake treatment.

The OLS estimator for the core set is not significantly different from zero (column 1, row 1). The Poisson estimate shows a weak statistically significant negative effect. The

[20]The results for the Physics class is reported because, as discussed in section 5, citations to the patents belonging to this class appear to be driving the results.

treatment group was cited less after the fictitious treatment compared to the control patents.

Re-estimating the model using the previously introduced matched samples, the one-to-one (column 1, row 1 and 2) matched and the two-to-one matched sample (column 2, row 1 and 2), shows no effect of the placebo treatment, both the Poisson and OLS coefficient for the interaction term are not significantly different from zero.

The test confirms that there is a weak pre-event trend when we consider the full sample. Yet, it is a weakly negative trend, therefore, if anything, it should have lead to a reduction in the magnitude of the actual effect and does not seem to affect the validity of our results. In addition, the effect disappears when we increase the similarity between the treated and the control group by matching on the application year and the technology class.

6 Do inventors read machine translations?

In section 5.2, we discussed a few factors that may have confounded our results and rule them out by using alternative methodologies to construct our outcome variables. However, these additional analyses only provide indirect evidence that the main channel through which machine translations of Chinese patents by Google patents lead to an increase in citations by inventors located in the US is an improvement in the accessibility to knowledge that was previously available only in Chinese. This section aims to provide more direct evidence of a direct link between machine translations and knowledge flows. To do so, we run two additional analyses in which we adopt a triple-differences approach.

6.1 University-owned patents

In a recent article Kong et al. (2020) investigate the difference between university and corporate patents. They use linguistic measures to show that university patents need 1.1 to 1.6 years of education less to read. Considering corporate patents, the gap is 2.2 to 2.6 times larger between the top 100 applicants, which further supports the hypothesis that this difference may stem from a strategic motive whereby corporations intentionally obscure their inventions to deter competitors from adopting the innovation. The relevance for machine translation is evident. Considering the finding of Larroyed (2018)—patents with easier sentence structure result in higher quality translations—we would expect a stronger effect of machine translation on patents owned by universities that have a higher readability than corporate patents. Clearly, if patents with a higher readability received more citations as result of the machine translation also in comparison with patents with a lower readability, this would strongly suggests that US-based inventors actually read those patents and used the knowledge codified in the patent documents. A first descriptive analysis of our data seems to confirm this hypothesis. In total 22,843 patents are owned by universities. Five out of ten of the treated patents with the highest increase of citations after the treatment

are University-owned patents, whereas no university-owned patents appear in the top ten for the control group.

To investigate the role of university patents, we expand our initial model and use a difference-in-difference-in-differences estimator, which distinguishes between patents filed by universities and corporations. We estimate the following model:

$$y_{it} = \beta_1 Chinese_i + \beta_2 Translated_{it} + \beta_3 Chinese \times Translated_{it}$$
$$+\beta_3 University_i + Chinese_i \times University_i + Translated_{it} \times University_i$$
$$+\mathbf{Chinese_i} \times \mathbf{Translated_{it}} \times \mathbf{University_i} + \delta_{ipc} + \delta_{year} + \gamma X_i + \epsilon_{it}$$

The baseline model discussed in subsection 4.1 is extended by the variable *University*, which reports whether the focal patent i is filed by a University.[21] In the triple-differences model, the main variable of interest is the variable $Chinese \times Translated \times University$. This is a binary variable taking value 1 if the patent is Chinese, a University patent, and translated after 2013. Therefore, a positive and significant coefficient will provide indications on the effect of machine translation of Chinese university patents on follow-on invention in the US. Table 11 reports the marginal effects of the Poisson estimators for citations arriving from patents without Chinese and Korean inventors involved (Columns 1), and citations without Asians (Column 2).

The analyses show a highly significant and positive effect of the triple-interaction term for both the dependent variables. These results imply that more readable patents are more frequently cited by US-based inventors as a consequence of the machine translation. This, in turn, appears to confirm that US inventors learn new technical knowledge by reading machine translated patents.

6.2 Illustrations in patents

In the previous section, we show that university-owned patents that are likely to be more readable than corporate patents are cited more often after the machine translation. We interpret this finding as evidence that US-based inventors read and learn from Chinese patents once the cost of accessing the knowledge is lowered. Another characteristic of a patent that can make it more easily understandable and readable is the presence of illustrations in its claims. The availability of a graphical figure in the patent claims could straighten potential error in the interpretation deriving from an inaccurate or flawed machine translation. In addition, illustrations increase the overall readability of the patent claim. If this is the case, we should expect that patents with illustrations in the claims receive more citations on average after the introduction of the treatment in 2013.

[21] Patents classified as University patents include all patents filed by educational and academic entities. To identify such patents, we use the PATSTAT reported applicant name. Every patent containing the following keywords is classified as a University patent: 'university', 'school', 'academy', 'college', and 'institute'. The Chinese and Korean translation of these terms is used accordingly.

Table 11: The effect of machine translation on forward citations to University patents, Poisson estimates.

	CN/KR removed	Asian removed
Chinese x Translated x University	.133***	.045**
	(.026)	(.018)
Chinese x Translated	.023**	.049***
	(.011)	(.009)
Translated x University	.007	.030**
	(.020)	(.015)
Chinese x University	.027*	.005
	(.014)	(.010)
Chinese	−.304***	−.245***
	(.008)	(.006)
Translated	−.432***	−.300***
	(.008)	(.006)
University	−.022*	.012
	(.012)	(.009)
Class fixed effects	Yes	Yes
Publn. year fixed effects	Yes	Yes
Mean of dep. var.	1.245	0.697
Num. obs.	136790	136790

Notes: Significant at ***1%, **5% and 10%*. Robust Standard errors in parentheses. This table shows only the regressor of interest. Regressors not listed are those of Table 3, see text for details. The dependent variable records the cumulative forward citations that the focal patent i has received from U.S. patents earliest filed between 2008 and 2019 and without Chinese/Korean inventors involved (column 1) and without Asian inventors (column 2). In addition, not-granted patent citations and foreign (not-U.S.) citations are excluded from the count of the dependent variable. Patents were classified as University patents using the in PATSTAT reported applicant name. In total 22,843 patents in the data set are University-owned. The variable Chinese x Translated X University is a binary indicator that identifies those focal patents that were treated – filed by a University, added to Google Patent database and automatically translated in 2013.

In order to test this hypothesis, we estimate another triple-difference model:

$$y_{it} = \beta_1 Chinese_i + \beta_2 Translated_{it} + \beta_3 Chinese \times Translated_{it}$$
$$+\beta_3 Figure_i + Chinese_i \times Figure_i + Translated_{it} \times Figure_i$$
$$+\mathbf{Chinese_i} \times \mathbf{Translated_{it}} \times \mathbf{Figure_i} + \delta_{ipc} + \delta_{year} + \gamma X_i + \epsilon_{it}$$

The variable *Figure* reports whether the focal patent i contains a figure in the claims.[22] In this case, our main variable of interest is $Chinese \times Translated \times Figure$, which takes value 1 for Chinese patents with at least one figure in their claims, after the translation. In total 3,714 patents in the data set have at least one illustration in their claims.

Table 12: The effect of machine translation on forward citations to patent documents containing figures, Poisson estimates.

	CN/KR removed	Asian removed
Chinese x Translated x Figures	.601***	.396***
	(.100)	(.095)
Chinese x Translated	.103***	.087***
	(.010)	(.007)
Translated x Figures	−.225***	−.138***
	(.030)	(.026)
Chinese x Figures	−.119***	.043*
	(.022)	(.024)
Chinese	−.257***	−.225***
	(.007)	(.005)
Translated	−.468***	−.297***
	(.008)	(.006)
Figures	.080***	−.076***
	(.025)	(.015)
Class fixed effects	Yes	Yes
Publn. year fixed effects	Yes	Yes
Mean of dep. var.	1.245	0.697
Num. obs.	136790	136790

Notes: Significant at ***1%, **5% and 10%*. Robust Standard errors in parentheses. This table shows only the regressor of interest. Regressors not listed are those of Table 3, see text for details. The dependent variable records the cumulative forward citations that the focal patent i has received from U.S. patents earliest filed between 2008 and 2019 and without Chinese/Korean inventors involved (column 1) and without Asian inventors (column 2). In addition, not-granted patent citations and foreign (not-U.S.) citations are excluded from the count of the dependent variable. The dummy coded variable 'Figures' indicates, that a patent has at least one figure in the claim – based on the data retrieved from Google Patents. In total 3,714 patents in the data set have at least one illustration in their claims. The variable Chinese x Translated X Figures is a binary indicator that identifies those focal patents that were treated – containing at least one figure in the claims, added to Google Patent database and automatically translated in 2013.

The coefficients of the variable of interest reported in Table 12 are always positive and highly significant in all models. The Poisson estimates that a Chinese patent with a figure

[22]The data was obtained from Google Patents. We only considered figures in the claims, to ensure the relevance of the figure for the innovation itself.

in its claims, once translated, receives on average about 0.601 citations more. Compared to an average of 1.25 citations, this implies a 48.3 percent increase in forward citation after the treatment period. The effect is even stronger if we re-estimate the model for the dependent variable counting citations from patents without Asian inventors. The result of the Poisson shows a 56.8 percent increase.

To our knowledge, no research has yet investigated the relationship of translation, knowledge diffusion, and graphical devices in patent documents. Our results strongly suggest the existence of such a relationship. Patents with pictures in their claims receive more citations after being translated, implying that inventors have an easier access to this knowledge once it is translated into English.

7 Discussion and conclusion

One of the main rationale for the existence of the patent system is the promotion of innovation through the disclosure of technical knowledge. However, the scholarly debate about the actual impact of patent disclosure on cumulative innovation is still open. Our paper seeks to contribute to the discussion by investigating the influence of machine translation.

Our identification strategy exploits an interesting natural experiment, the automatic translation of patent documents issued by the Chinese Patent Office implemented by the Google Patent service in 2013. This improvement in the patent search service allowed inventors to read patents previously only available in Chinese in English.

Our results suggest that translation facilitates knowledge diffusion. We use citations from US patents to 49,217 prior Chinese patents that were never filed, and hence translated, outside the country. We use a group of 19,860 Korean patents as control group, exploiting the fact that patents issued by the Korean patent office were included and automatically translated in English by Google Patents three years after the patents issued by the Chinese patent office. We find a positive and statistically significant relationship between the machine translation and the number of received citations. Our estimates indicate that treated patents receive on average 7.2 percent more forward citations than the control group after the translation. The effect is much stronger in technical domains in which China is a technological leader, such as computing and data processing related technologies, showing an increase of 14.5 percent. The main results are robust to a broad range of alternative specifications, like assuming a fake treatment period to account for potential pre-event trends. We find no effect for examiner added citations, further providing evidence supporting the disclosure argument. The effect is robust to considering exclusively citations arriving from non-Chinese or non-Asian inventors, as identified by natural language processing algorithms, indicating that the effect is driven by inventors not capable of speaking the language prior to the translation.

In addition, we investigate the heterogeneous effect of translation for university-owned

and corporate-owned patents. University patents, due to their easier readability, have a higher likelihood to result in an accurate translation. Our findings support this claim and suggest that US-based inventors read and use the knowledge disclosed in Chinese documents. We reach a similar conclusion by focusing on patents that include illustrations in their claims.

These findings come with several policy implications. A first, immediate consideration is that the patent system produces the intended effects at least when it comes to the disclosure function. The increase of cumulative innovation produced by lowering the costs of access to knowledge disclosed in patent documents shows that inventors are reading and learning from patents during the invention process. This is especially evident in areas where innovation happens in a cumulative fashion.

Second, our findings show that one of the main channels through which the benefits of patent disclosure are realized is the translation of knowledge that would be otherwise trapped behind a language barrier. The existence of an international and harmonized patent system makes relevant knowledge available to a wide and interested audience and facilitates knowledge flows between geographically and culturally distant areas of the world.

Third, our results also show that there are still substantial innovation-related gains that could be obtained by improving the quality and the availability of the knowledge disclosed in patents. As shown by the differential impact of translation between university and corporate patents, it is clear that the usual critique made by scholars that are skeptical about disclosure theory is not far-fetched: corporate patents often hide, rather than disclose, relevant information. The adoption of policies to enhance the enforcement of the disclosure requirements during the patent prosecution process could be a step in the right direction to enjoy the full benefits of patent disclosure.

References

Alcácer, J. and M. Gittelman (2006): "Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations," *The Review of Economics and Statistics*, 88, 774–779.

Alcácer, J., M. Gittelman, and B. N. Sampat (2008): "Applicant and Examiner Citations in US Patents: An Overview and Analysis," Working Paper 09-016, Harvard Business School.

Baruffaldi, S. H. and M. Simeth (2020): "Patents and Knowledge Diffusion: The Effect of Early Disclosure," *Research Policy*, 49, 103927.

Belenzon, S. (2012): "Cumulative Innovation and Market Value: Evidence from Patent Citations," *The Economic Journal*, 122, 265–285.

Berkes, E. and P. Nencka (2020): "Knowledge Access: The E Ects of Carnegie Libraries on Innovation," 58.

Boldrin, M. and D. K. Levine (2013): "The Case Against Patents," *Journal of Economic Perspectives*, 27, 3–22.

Breschi, S. and F. Lissoni (2009): "Mobility of Skilled Workers and Co-Invention Networks: An Anatomy of Localized Knowledge Flows," *Journal of Economic Geography*, 9, 439–468.

Breschi, S., F. Lissoni, and E. Miguelez (2015): "Foreign Inventors in the US: Testing for Diaspora and Brain Gain Effects," 66.

Choudhury, P. and D. Y. Kim (2019): "The Ethnic Migrant Inventor Effect: Codification and Recombination of Knowledge across Borders," *Strategic Management Journal*, 40, 203–229.

Cohen, W. M., R. R. Nelson, and J. P. Walsh (2000): "Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)," Working Paper 7552, National Bureau of Economic Research.

de Rassenfosse, G., G. Pellegrino, and E. Raiteri (2020): "Do Patents Enable Disclosure? Evidence from the Invention Secrecy Act," *SSRN Electronic Journal*.

Devlin, A. J. (2010): "The Misunderstood Function of Disclosure in Patent Law," *Harvard Journal of Law & Technology*, 23, 46.

Dominguez Lacasa, I., B. Jindra, S. Radosevic, and M. Shubbak (2019): "Paths of Technology Upgrading in the BRICS Economies," *Research Policy*, 48, 262–280.

EPO (2021): "Guidelines for Examination," Tech. Rep. T 0991/01.

Fromer, J. C. (2009): "Patent Disclosure," SSRN Scholarly Paper ID 1116020, Social Science Research Network, Rochester, NY.

Furman, J. L., M. Nagler, and M. Watzinger (2018): "Disclosure and Subsequent Innovation: Evidence from the Patent Depository Library Program," Working Paper 24660, National Bureau of Economic Research.

Galasso, A. and M. Schankerman (2014): "Patents and Cumulative Innovation: Causal Evidence from the Courts," Working Paper 20269, National Bureau of Economic Research.

Gordin, M. D. (2015): *Scientific Babel: How science was done before and after global English*, University of Chicago Press.

Graham, S., A. Marco, and U. S. Patent (2015): "The USPTO Patent Examination Research Dataset: A Window on the Process of Patent Examination," 119.

Groves, M. and K. Mundt (2015): "Friend or Foe? Google Translate in Language for Academic Purposes," *English for Specific Purposes*, 37, 112–121.

Hall, B. H., A. Jaffe, and M. Trajtenberg (2005): "Market Value and Patent Citations," *The RAND Journal of Economics*, 36, 16–38.

Harhoff, D., F. M. Scherer, and K. Vopel (2003): "Citations, Family Size, Opposition and the Value of Patent Rights," *Research Policy*, 32, 1343–1363.

Husmann, M. and M. Barton (2014): "Advancing and Translating Knowledge in Vascular Medicine," *Frontiers in Cardiovascular Medicine*, 1.

Jaffe, A. B. and M. Trajtenberg (1999): "International Knowledge Flows: Evidence From Patent Citations," *Economics of Innovation and New Technology*, 8, 105–136.

Jaffe, A. B., M. Trajtenberg, and M. S. Fogarty (2000): "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors," *American Economic Review*, 90, 215–218.

Junczys-Dowmunt, M., T. Dwojak, and H. Hoang (2016): "Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions," *arXiv:1610.01108 [cs]*.

Kong, N., U. Dulleck, A. Jaffe, S. Sun, and S. Vajjala (2020): "Linguistic Metrics for Patent Disclosure: Evidence from University Versus Corporate Patents," Tech. Rep. w27803, National Bureau of Economic Research, Cambridge, MA.

Kuhn, J. M., K. A. Younge, and A. C. Marco (2019): "Patent Citations Reexamined," SSRN Scholarly Paper ID 2714954, Social Science Research Network, Rochester, NY.

Laohaprapanon, S. and G. Sood (2021): "Appeler/Ethnicolr," appeler.

Larroyed, A. A. (2018): "Machine Translation and Disclosure of Patent Information," *IIC - International Review of Intellectual Property and Competition Law*, 49, 763–786.

Lemley, M. (2012): "The Myth of the Sole Inventor," *Michigan Law Review*, 110, 709–760.

Li, D., T. W. Tong, and Y. Xiao (2021): "Is China Emerging as the Global Leader in AI?" *Harvard Business Review*.

Moser, P., J. Ohmstedt, and P. W. Rhode (2018): "Patent Citations—An Analysis of Quality Differences and Citing Practices in Hybrid Corn," *Management Science*, 64, 1926–1940.

O'Meara, S. (2019): "Will China Lead the World in AI by 2030?" *Nature*, 572, 427–428.

Orwant, J. (2013): "Broadening Google Patents," .

Ouellette, L. L. (2012): "Do Patents Disclose Useful Information?" *Harvard Journal of Law & Technology*, 25, 64.

——— (2017): "Who Reads Patents?" *Nature Biotechnology*, 35, 421–424.

Park, K. (2020): "Name2nat: A Python Package for Nationality Prediction from a Name," *GitHub repository*.

Risch, M. (2007): "Why Do We Have Trade Secrets?" *Marquette Intellectual Property Law Review*, 11, 76.

Sampat, B. and H. L. Williams (2019): "How do patents affect follow-on innovation? Evidence from the human genome," *American Economic Review*, 109, 203–36.

Sampat, B. N. (2010): "When Do Applicants Search for Prior Art?" *The Journal of Law and Economics*, 53, 399–416.

Trancik, J. E. (2014): "Renewable Energy: Back the Renewables Boom," *Nature News*, 507, 300.

USCIS (2020): "H-1B Petitions by Gender and Country," https://travel.state.gov/content/travel/en/legal/visa-law0/visa-statistics.html.

Wetherbee, I. (2016): "11 New Countries Available in Google Patents," .

WIPO (2019): "World Intellectual Property Indicators: Filings for Patents, Trademarks, Industrial Designs Reach Record Heights in 2018," https://www.wipo.int/pressroom/en/articles/2019/article_0012.html.

——— (2020): "Guides de Recherche · Research Guides: Patents: Patent Information & Classification," Tech. rep., World Intellectual Property Organization.

WIPO (2020): *World Intellectual Property Indicators 2020.*

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean (2016): "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *arXiv:1609.08144 [cs].*

Zhou, P. and L. Leydesdorff (2006): "The Emergence of China as a Leading Nation in Science," *Research Policy*, 35, 83–104.

Zulfiqar, S., M. F. Wahab, M. I. Sarwar, and I. Lieberwirth (2018): "Is Machine Translation a Reliable Tool for Reading German Scientific Databases and Research Articles?" *Journal of Chemical Information and Modeling*, 58, 2214–2223.