Does Machine Translation Affect International Trade? Evidence from a Large Digital Platform*

Erik Brynjolfsson
MIT and NBER

Xiang Hui
Washington University in St. Louis and MIT

Meng Liu†
Washington University in St. Louis and MIT

December 3, 2018

Abstract

Artificial intelligence (AI) is surpassing human performance in a growing number of domains. However, there is limited evidence of its economic effects. Using data from a digital platform, we study a key application of AI: machine translation. We find that the introduction of an upgraded machine translation system has significantly increased international trade on this platform, increasing exports by 10.9%. Furthermore, heterogeneous treatment effects are consistent with a substantial reduction in translation costs. Our results provide causal evidence that language barriers significantly hinder trade and that AI has already begun to improve economic efficiency in at least one domain.

Keywords: Artificial Intelligence, International Trade, Machine Translation, Machine Learning, Digital Platforms

*We thank David Atkin, Andrey Fradkin, Avi Gannamaneni, Avi Goldfarb, and Alejandro Molnar for helpful discussions. We also thank eBay, especially Brian Bieron, for providing the data for this project. We acknowledge the support of the MIT Initiative on the Digital Economy (http://ide.mit.edu/).

†erikb@mit.edu, hui@wustl.edu, mengl@wustl.edu

1 Introduction

Artificial intelligence (AI) is one of the most important technological advances of our era. Recent progress of AI and, in particular, machine learning (ML), has dramatically increased predictive power in many areas such as speech recognition, image recognition, and credit scoring (Agrawal et al. [2016] and Mullainathan and Spiess [2017]). Unlike the last generation of information technology that required humans to codify tasks explicitly, ML is designed to learn the patterns automatically from examples (Brynjolfsson and Mitchell [2017]). This has opened a broad new frontier of applications and economic implications that are, as yet, largely undeveloped.

Does AI adoption affect economic activities? This is an important question because extensive resources are being allocated to AI research and implementation.[1] Surprisingly, no matter how obvious the answer may *a priori* be when thinking of AI's capability and general applicability, there is virtually no direct casual evidence between the use of AI and changes in economic activities. In particular, contributions from AI are not found in aggregate productivity measures. Brynjolfsson et al. [2017] argue that the gap between expectations and statistics is likely due to lags in complementary innovations. Noting this, the best domains to assess AI's impact are settings where AI applications are seamlessly embedded in systems with established complementary innovations. In particular, digital platforms are pioneers of AI adoption, providing ideal opportunities for early assessment of AI's economic effects.

In this paper, we document a causal effect of AI adoption on economic activities by analyzing the introduction of eBay Machine Translation (eMT) *for product listing titles* on international trade on eBay. This platform mediated more than 14 billion dollars of global trade among more than 200 countries in 2014. The focal AI technology, eMT, is an in-house ML system that statistically learns how to translate among different languages. It replaces and improves translation quality over the previous generation of translation technology used by eBay (BLEU score changed from 41.01 to 45.24). We exploit the introduction of eMT to Spanish-speaking Latin American countries as a natural experiment, and study its consequence on U.S. exports to these countries on eBay.

A standard approach of evaluating the effectiveness of machine translation (MT) would exploit across-country variation in the availability or quality of MT, both in this setting and broader settings. There are two key challenges with this approach. First, country pairs with better trade

[1] For example, overall investments in AI startups increased 150% globally year-over-year in 2017, growing from $4 billion in 2016 to $10.7 billion last year. Sources: https://www.abiresearch.com/market-research/product/1030415-artificial-intelligence-investment-monitor/

prospects likely select into introducing or improving MT. Hence what appears to be MT's effect among countries with higher translation quality may simply be self-selection. Second, contemporaneous events around the improvement of MT may bias the estimation, such as changes in marketing activities or macroeconomic conditions.

In this paper, we adopt a *within-country*, continuous difference-in-difference (DiD) approach that exploits variation in the number of words across listing titles. Essentially, the approach compares the post-policy change in U.S. exports to Spanish-speaking Latin American countries for listings with longer titles, against those for listings with shorter titles. This approach assumes that the reduction in the cost of translating listing titles is larger for titles with higher ex-ante translation costs, which are proxied by title lengths. For example, one could imagine that eMT reduces per-word translation cost by x, and a listing with n words experiences a reduction of nx.

This within-country identification directly addresses the selection issue at the country level, because it holds the set of countries fixed. The approach could also address the concern of contemporaneous events, if they are orthogonal to title lengths, which we believe is a reasonable assumption in our setting. For example, the effect of increased advertising spending in Latin American countries should similarly affect exports of listings with different title lengths.

Based on this approach, we estimate that exports increase is 1.06% larger for items with one more word in the listing title after the introduction of eMT on title translation. This estimate translates into a 10.9% increase in overall exports to Spanish-speaking Latin American countries. Furthermore, we identify heterogeneous treatment effects that confirm eMT's mechanism in reducing translation costs for market participants: The export increase is more pronounced for differentiated products, cheaper products, and less experienced buyers.

We should caution that this estimation approach could suffer from omitted variable bias. In particular, listings with different title lengths could be correlated with unobserved product characteristics that affect exports. This confounding correlation could be (1) time-invariant, (2) time-variant and serially-correlated, or (3) time-variant and serially-uncorrelated. To deal with (1), we control for title length-specific fixed effects. To mitigate the concern of (2), we perform various placebo tests using different months before the introduction of the eMT. For (3), we include many time-varying, title length-specific market characteristics in the regressions. The results are consistent with the validity of the exclusion restriction assumption in the continuous DiD model.

An additional concern is that sellers may strategically change the title lengths of their listings to take advantage of the eMT. To mitigate this concern, we study items that were listed *before* the

policy change and have not been modified afterwards. We find consistent policy effect for this set of listings, suggesting that sellers' strategic behavior is not large.

Lastly, we exploit eMT's rollouts in the European Union and Russia, and estimate comparable eMT effects for other language pairs: English–French, English–Italian, and English–Russian. We estimate that the export promotion effect of eMT is equivalent to reducing bilateral geographical distance by 26.1%.

1.1 Related Literature and Contribution

1.1.1 Language Barriers in Trade

Empirical studies using gravity models, as specified in Anderson and Van Wincoop [2003], have established the existence of a robust negative correlation between bilateral trade and language barriers. Typically, researchers regress bilateral trade on a dummy variable for whether the two countries share the same official language, and find that this coefficient is strongly positive (e.g., Melitz [2008], Egger and Lassmann [2012], and Melitz and Toubal [2014]). However, these cross-sectional regressions are vulnerable to endogeneity biases, even after controlling for the usual set of variables in the gravity equation. For example, two countries with the same official language can be similar in preference, which also affects trade.

A key contribution of our paper, therefore, is that it exploits a natural experiment on eBay to identify the effect of changing language barriers on international trade. The online marketplace provides us with a uniquely powerful laboratory to study the consequences on bilateral trade after an exogenous decrease in language barriers *for a given language pair* . Our finding that even a quality upgrade of machine translation could increase exports by 5.6%–10.9% is consistent with Lohmann [2011] and Molnar [2013], who argue that language barriers may be far more trade-hindering than suggested by previous literature.

1.1.2 AI and Economic Welfare

The current generation of AI represents a revolution of prediction capabilities (e.g., Brynjolfsson and McAfee [2017]). Recent breakthroughs in ML, especially supervised learning systems using deep neural networks, have made possible substantial improvements in many technical capabilities. Machines have surpassed humans at tasks as diverse as playing the game Go (Silver et al. [2016]) and recognizing cancer from medical images (Esteva et al. [2017]). There is active work converting these

breakthroughs into practical applications such as self-driving cars, substitutes for human-powered call-centers, and new roles for radiologists and pathologists, but the complementary innovations required are often costly (Brynjolfsson et al. [2017]).

Machine translation has also experienced significant improvement due to advances in ML. For instance, the best score at the Workshop on Machine Translation for translating English into German improved from 23.5 in 2011 to 49.9 in 2018, according to a widely used BLEU score.[2] Much of the recent progress in MT has been a shift from symbolic approaches towards statistical and deep neural network approaches. For our study, an important characteristic of eMT is that replacing human translators with MT or upgrading MT is typically relatively seamless. For instance, for product listings on eBay, users simply consume the output of the translation system, but otherwise need not change their buying or selling process. While users care about the quality of translation, it makes no difference whether it was produced by a human or machine. Thus, adoption of MT can be very fast and its economic effects, especially on digital platforms, immediate. While so far much of the work on the economic effects of AI has been theoretical (Acemoglu and Restrepo [2018], Aghion et al. [2017], Korinek and Stiglitz [2017], Sachs and Kotlikoff [2012]), and notably Goldfarb and Trefler [2018] in the case of global trade, the introduction of improved MT on eBay gives us an early opportunity to assess the economic effects of AI using plausible natural experiments.

1.1.3 Peer-to-Peer Platforms and Matching Frictions

Einav et al. [2016] and Goldfarb and Tucker [2017] provide great surveys on how digital technology has reduced matching frictions and improved market efficiency. Reduced matching frictions affect price dispersion, as evidenced in Brynjolfsson and Smith [2000], Brown and Goolsbee [2002], and Cavallo [2017]. These reduced frictions also mitigate geographic inequality in economic activities in the case of ride-sharing platforms (Lam and Liu [2017] and Liu et al. [2018]), short-term lodging platforms (Farronato and Fradkin [2018]), crowdfunding platforms (Catalini and Hui [2017]), and e-commerce platforms (Blum and Goldfarb [2006], Lendle et al. [2016], and Hui [2018]). We contribute to this literature by documenting the significant matching frictions between consumers and sellers who speak different languages. Specifically, we find that efforts to remove language barriers increase market efficiency substantially.

[2]Source: http://matrix.statmt.org/matrix

2 Background

The primary goal of eMT is to support international trade by making it easier for buyers to search for and understand the features of items not listed in their language. In particular, eMT is a set of statistical translation models that output probabilistic results generated from vast amounts of parallel language data. These ML models are trained on both eBay data and other data scraped from the Web. Some hand-crafted rules were applied, such as preserving named entities, to make eMT more suited for the eBay environment.

eBay rolled out eMT for query translation in May 2014 and eMT for title translation in July 2014 between the U.S. and Spanish-speaking Latin American countries. We discuss both introductions in this section to comprehensively illustrate buyers' interaction with eMT, although we the main focus of this paper is to identify the effect of eMT for *item title translation*. To shop on eBay, buyers in Spanish-speaking Latin American countries visit www.ebay.com, and see items from sellers who sell to buyers' countries. eBay recognizes buyers' IP addresses from Spanish-speaking countries in Latin America, and shows buyers the website in Spanish. Note that the translation of non-user-generated content on the website, such as product categories, existed before and was not affected by the introduction of eMT. Instead, the introduction of eMT affected translation quality of only *search queries* and *listing titles* (not item descriptions) in the period we study.[3] In particular, when buyers enter search keywords in Spanish, eMT translates them into English and the search engine retrieves listings in the search results page based on the translated query. Next, for this set of listings, eMT translates the titles from English into Spanish.

Prior to eMT, eBay used Bing Translator for query and item title translation. Therefore, the policy treatment here is *an improvement in translation quality*. To understand the magnitude of quality improvement, we follow the MT evaluation literature and report qualities based on both the BLEU score and human evaluation. The BLEU score measures how "close" the MT translation output is to one or more reference translations by linguistic experts (for details, see Papineni et al. [2002]). It is an automated measure that has been shown to highly correlate with human judgment of quality (Callison-Burch et al. [2006]). However, BLEU scores are not easily interpretable and should not be compared across languages (Denkowski and Lavie [2014]). Generally, scores over 30 reflect understandable translations, and scores over 50 reflect good translation (Lavie [2010]. On

[3] eBay prioritized search queries and item titles because searching for a product and viewing search results in buyers' own language allows consumers to make informed decisions on which listings to open, which is very important for sales.

the other hand, although human evaluations are highly interpretable, they are very costly and can be less consistent.

Comparing Bing and eMT translation for item titles from English to Spanish, the BLEU score increased from 41.01 to 45.24, and human acceptance rate (HAR) increased from 82.4% to 90.2%. To compute HAR, three linguistic experts vote either yes or no for translations based on adequacy only (whether the translation is acceptable for minimum understanding), and the majority vote is then used to determine the translation quality. In comparison, the BLEU score is rated based on both adequacy and fluency, because it compares the MT output with the translation by someone fluent in the language. As a result, in cases where grammar and style of translation is not of first-order importance, such as translating listing titles, one might prefer the use of HAR over the BLEU score for measuring translation quality.

We should note that, when eBay first introduced eMT for query translation in May 2014, it also made other localization changes. In particular, it catered local deals to buyers' home country and allowed buyers to see prices in local currencies.[4] Also, eBay may have increased their advertising spending in Latin American countries. As will be discussed in Section 3, our identification addresses confounding effects that are orthogonal to title lengths. Additionally, the eMT for title translation was introduced two months after introducing the eMT for query translation and other localization effects, so we also shrink the estimation window to exclude this period.

In 2014, eBay also rolled out eMT in Russia (January, English–Russian) and the European Union (July, English–French, English–Italian, English–Spanish). In our main analyses we focus on the rollout in Latin America for two reasons: (1) the rollout in Russia was followed by Russia's annexation of Crimea, which prompted international sanctions; (2) the rollout of eMT for query translation (May) and for item title translations (July) were two-month apart in Latin America, but they happened in the same month for the EU. Therefore, studying the policy impact in July in the Latin American rollout allows us to separate the treatment effect of improving translation quality of listing titles from improving query translation. Nonetheless, we replicate our analyses in the rollouts for EU and Russia as robustness checks.

[4]Source: https://www.ebayinc.com/stories/news/ebay-delivers-localized-shopping-experiences-latin-america/.

3 Data and Empirical Strategy

This paper uses administrative data from eBay, which include detailed product, listing, and buyer characteristics. Importantly, we observe the number of words in the title of a listing, providing information on the exact number of words that are translated from English to Spanish.[5] We restrict the reporting of summary statistics to comply with eBay's data policy.

Our identification exploits the fact that a better translation system was implemented across listings with differential translation costs to begin with, and assumes that the reduction in translation costs (treatment intensity) is larger for listings with higher ex-ante translation costs. For example, one could imagine that eMT reduces per-word translation cost by x, and a listing with n words experiences a reduction of nx. Since we do not directly obverse translation costs, we use number of words in listing titles as a proxy. This implicitly assumes that translation costs are non-decreasing in number of words in the title. While this assumption is innocuous for any given listing, we should be cautious of selection of different types of products and sellers into longer titles, as will be discussed soon.

We aim to create treatment and control groups using variations in title lengths across different listings. Figure 1 plots the share of exports across listings with 5–16 words in the title (95% of U.S. exports to affected countries). Note that exports refer to the export quantity throughout the paper, unless otherwise mentioned, to purge away eMT's effect on price, because we are mainly interested in its effect on short-run exporting activities. From the figure, we see a decent amount of variation in title lengths, which we use as the treatment intensity in our continuous difference-in-difference (DiD) estimation, similar to Mian and Sufi [2012] and Hui et al. [2018]:

$$log(Y_{clt}) = \beta Num\_Words_{clt} \times Post_t + \gamma XR_{ct} + \eta_c + Num\_Words_g + \xi_t + \epsilon_{clt}. \quad (1)$$

The regression is performed at the country–title length–time period level. Y_{clt} is the exports to country c of title length l in period t; Num_Words_{clt} is the title length; $Post_t$ is the dummy for the introduction of eMT; XR_{ct} is the average daily bilateral exchange rate at t; η_c are importing country fixed effects; Num_Words_g are title length fixed effects; and ξ_t are time fixed effects. The coefficient β represents marginal policy effect on listings with one more word in the title. Throughout the paper, the standard errors are clustered at the country level to account for serial

[5] Note that a good translation should preserve brand names.

correlation of exports.

In nutshell, we estimate the policy effect by comparing the post-policy change in U.S. exports to Spanish-speaking Latin American countries for listings with longer titles, against the change in U.S. exports to the same countries for listings with shorter titles. This comparison can handle the existence of contemporaneous events, if they are orthogonal to title lengths. For example, displaying local deals, showing prices in local currencies, and increasing advertising spending should affect exports similarly across title lengths.

The identification assumption is an exclusion restriction: the length of titles does not correlate with product or seller characteristics that affect exports. Note that we control for title length-specific fixed effects, which should take care of time-invariant omitted variables. To deal with time-variant and serially-correlated omitted variables, we perform various placebo tests using different months before the introduction of the eMT. Furthermore, to deal with time-variant and serially-uncorrelated omitted variables, we include many time-varying, title length-specific market characteristics in the regressions. The results are consistent with the validity of the exclusion restriction assumption in the continuous DiD model. Lastly, we repeat the analyses on listings whose titles were not changed after the policy change to control for sellers' strategic behavior, and estimate similar treatment effects.

As a robustness check for our within-country continuous DiD specification, we also adopt an across-country DiD specification:

$$log(Y_{ct}) = \beta T_c \times Post_t + \gamma XR_{ct} + \eta_c + \xi_t + \epsilon_{ct}, \tag{2}$$

where Y_{ct} is the exports to country c at time t, and T_c is the dummy for the treatment status of country c. The identification comes from comparing the intertemporal change in exports in the treatment group (countries that become eligible for eMT) against the baseline change in exports in the control group (countries that remain ineligible for eMT). The coefficient β represents the average treatment effect of eMT on exports across all treated countries.

4 Results

4.1 Overall Policy Effect

We first visually inspect if the parallel trend assumption holds for our continuous DiD specification. Figure 2a plots the average monthly U.S. exports to Spanish-speaking Latin American countries by number of words in the listing titles. The two vertical lines at t=0 and t=2 correspond to the introduction of eMT for query (May 2014) and title translation (July 2014), respectively. Exports are normalized based on the value at t=-1. From the figure, we see that the four series were close to each other in the six months from t=-6 to t=-1, and stayed close in the two months after the introduction of eMT for query translation (t=0 and t=1). However, in the six months after the introduction of eMT for title translation at t=2, export change is larger for listings with more words in the title, suggesting that this introduction reduces translation costs in understanding item titles.

We estimate the policy effect with equation (1) using the same data as in Figure 2a. Specifically, the sample includes 14 months $\times$ 18 Spanish-speaking Latin American countries $\times$ 16 categories of title lengths (e.g. 5 words, 6 words, ..., 16 words). In Table 1, "Post" dummy turns to 1 at t=2 when eMT title translation is introduced. Column (1) in Panel A shows that export increase is 1.06% larger for listings with one more word in the listing title. This implies a 10.9% overall export increase, given that the average title contains 10.26 words.

In column (2), we control for product characteristics (average sales price, number of distinct products, share of used items) and market characteristics (average seller size, share of Top Rated Sellers, share of buyer disputes) for different title lengths in each time period. We see that the inclusion of these characteristics does not reduce the estimated policy effect. The estimated coefficients of these characteristics are also sensible: an increase in number of distinct products, average seller size, and share of Top Rated Sellers is positively correlated with exports, while an increase in average sales price, share of used items, and share of buyer disputes is negatively correlated with exports.

Next, we shorten the estimation window to six weeks before and after the introduction of eMT for title translation (July 2014). The rationale is to exclude the period containing the introduction of eMT for query translation (May 2014) among other advertising and localization efforts in the estimation. We estimate a slightly smaller, yet statistically significant, policy effect of 0.8% (Panel A column (3)). Column (4) shows that this effect is robust to the inclusion of title length-specific market characteristics.

We have estimated policy effect on a per-word basis. Alternatively, we estimate policy effect by comparing U.S. exports across treated and non-treated countries. In Figure 2b, we plot monthly U.S. exports to the 18 Spanish-speaking Latin American countries and to the 86 non-Latin American countries. Exports are normalized by the value in the month before the introduction of improved query translation (t=-1). The two series moved along closely in the six months before the first eMT introduction (from t=-6 to t=-1), providing evidence for the parallel trends assumption. After the introduction of improved query translation (t=0), the U.S. exports to the treated countries increased relative to those in the non-treated country. The gap between the two series became even larger exports after the introduction of better title translation (t=2). The observations suggest that both introductions increased exports, although the introduction for title translation seems to have a bigger effect.

When comparing Figure 2b with Figure 2a, we see that even though improved query translation seems to increase U.S. exports when comparing across countries, this change does not differ by number of words in listings. On the other hand, the effect of improved title translation on exports is detected in both the within-country and across-country comparisons. This result makes sense: While improved query translation may differentially affect search query of different lengths, it should not have differential effect on title with different lengths. Therefore, this comparison provides a falsification test for our continuous DiD strategy exploiting different title lengths, which we discuss in detail in Section 5.

Interestingly, although improved query translation appears to increase U.S. exports when we compare across countries (t=0 in Figure 2b), it does not lead to differential export changes across title lengths within affected countries (t=0 in Figure 2a). In contrast, improved title translation (t=2) is associated with export changes in both across-country and across-title length comparisons. This contrast essentially provides a falsification test for our continuous DiD specification that exploits title lengths as treatment intensities, because the initial eMT query translation should not have differential effects across title lengths (more details in Section 5).

4.2 Heterogeneous Policy Effects

If the observed export changes estimated from Equation (1) are caused by the eMT for title translation, we would expect to see more pronounced effects in categories with higher translation costs (see a theoretical framework in the appendix). We leverage the data richness to explore the heterogeneous effects of the policy change.

We begin by comparing eMT's effect between homogeneous products (e.g., cellphones and books, which have standard identifiers) and differentiated products (e.g., antiques and clothing, which have more variation in product attributes). Since the language requirement of translating the specifics of differentiated products is likely higher, eMT's effect should also be larger for these products. We distinguish the two types of product based on whether a product is assigned a "Product ID" on eBay. Product IDs are the most fine-grained catalogs on eBay defined for homogeneous products. For instance, an "Apple iPhone 8-256 GB-Space Gray-AT&T-GSM" has a different Product ID from other versions of iPhones. For books or CDs, Product IDs are ISBN codes. Conversely, Product IDs are rarely defined for products in fashion, clothing, art, and jewelry categories.

In Panel B of Table 1, we perform the continuous DiD regression for the two types of products using exports aggregated at the country–title length–product type–time period level. We control for product type fixed effects in addition to the controls in Equation (1). Column (1) shows that export increase for each additional word is 1.85% for differentiated produts, but is only 0.65% for homogeneous products. This estimated difference is robust to the inclusion of title length characteristics and using shorter estimation window. This heterogeneity is consistent with the theoretical prediction.

Next, we explore how the policy effect differs by product value. Since the translation cost as a fraction of item value is higher for cheaper items, we expect a larger export increase for cheaper items than for more expensive items. For example, imagine there are two items with same title length, one worth \$5 and the other one worth \$500. Although translation costs are the same for both items, before eMT a buyer presumably was more likely to incur the translation cost for the \$500 item, assuming the utility of consuming more expensive items is higher. Therefore, the policy effect should be smaller for expensive items.

To test this hypothesis, we divide items into four value bins: $[0, \$10)$, $[\$10, \$50)$, $[\$50, \$200)$, and \$200 and above. Following Einav et al. [2015], product value is defined as the average sales price in posted price format in the 6 months preceding the policy change. Column (1) in Panel C of Table 1 shows that export increase for cheap products is 1.44% for each additional word in listing titles, but decreases to 9.2% for expensive products, and the difference is statistically significant. This finding is consistent across columns (2)–(4) as we include more controls and shrink estimation window, and suggest that the policy effect for items worth more than \$200 is around 2/3 of that for products worth less than \$10.

Besides heterogeneous translation costs across product types, buyers may also be subject to

heterogeneous language barriers. Although we do not directly observe buyers' translation cost, we follow Hui et al. [2016] and consider buyers' experience on eBay as a proxy for translation cost: experienced buyers are deal seekers and spend more time on eBay. This suggests that they have smaller search costs, which in turn suggests that they were more likely to incur the hassle cost of using translation tools. Therefore, improved translation quality should mainly affect inexperienced buyers who may value their time more.

For this analysis, we define experienced buyers as those who spent more than $2,500 in the previous year on eBay, which roughly corresponds to eBay's definition for buyer experience. Column (1) in Panel D of Table 1 shows that the increase in exports is 1.24% for each additional word in listing titles for inexperienced buyers, but is only 0.69% for experienced ones. The qualitative finding is persistent when we add title length-specific controls and narrowing estimation windows.

5 Placebo Tests

The key identification assumption in Equation (1) is that there are no time-varying, title length-specific characteristics that correlate with exports conditional on the specified set of fixed effects. Otherwise, our identification would suffer from omitted variable bias. Note that we have controlled for some product and market characteristics in Table 1, and the estimated policy effects do not change dramatically. However, it is still possible that the model does not control for all relevant, time-varying, and title length-specific characteristics. To mitigate this concern, we run a series of placebo tests to see whether there were differential change in exports related to title lengths even before the policy change. The idea is as follows: if there exists time-varying characteristics that correlates simultaneously with title length and exports, then this confounding relationship should spuriously drove differential exports change even before the policy change, when no actual treatment took place. This test assumes that the confounding correlation has some persistence over time.

Recall that the actual treatment is the introduction of eMT for title translation (t=2). Our first placebo treatment is two months before the actual treatment, i.e. t=0, when eMT for query translation was introduced. We have seen in Figures 2a and 2b that exports increased both at t=0 and t=2, but the effects differed across title length only after the actual treatment. To test this formally, we added another interaction, "No. Words*Placebo", in Equation (1), where "Placebo" equals 1 on or after t=0. The coefficient of this interaction captures potential policy effect for

improved query translation.

Results analogous to Figure 2a are reported in columns (1) and (2) in Panel A of Table 2. We see that improved query translation does not increase trade more for longer titles, with or without controlling for title length-specific characteristics. The corresponding coefficients for Figure 2b are reported in columns (1) and (2) in Panel B, estimated using Equation (2). The results suggest that the improved query translation improves exports, and the magnitude is half of that for improved title translation. However, we should be cautious of the causal interpretation here, because the across-country DiD is vulnerable to unobserved contemporaneous events, such as increased advertising or other localization effort.

Our second placebo treatment is six months before the actual treatment, namely t= -4. In columns (3) and (4) in Panel A, "Placebo" equals 1 on or after t=-4, and we use data from the six months before and six months after to estimate the placebo effect. We find no increase in exports that differ in title lengths. In Panel B, we estimate this placebo treatment using the cross-country Equation (2), and again did not find any effect on exports.

Lastly, our third placebo treatment is twelve months before the actual treatment, namely t= -10. In this case, "Placebo" equals 1 on or after t=-10. Using data from the six months before and after this placebo month, we did not detect any increase in exports, as indicated in columns (5) and (6). This is the case both according to Equation (1) as indicated in Panel A, as well as according to Equation (2) as shown in Panel B.

To sum up, our preferred continuous DiD specification (Equation 1) estimates a 10.9% overall export increase after the introduction of eMT for title translation. Interestingly, the across-country DiD specification (Equation 2) estimates the policy effect of improved title translation to be 11.9%, which is similar to the first specification. On the other hand, although export increased by 6.1% after the introduction of eMT for query translation according to Equation (2), this effect does not differ by title lengths. Similarly, we did not find any spurious relationship between title lengths and exports in the six and twelve months before the introduction of eMT for title translation. These null results based on Equation (1) are reassuring evidence of the validity of our main identification strategy.

6 Other Robustness Checks

Besides the omitted variable biases discussed in the previous section, one might worry that some sellers could endogenously change their listing attributes, such as title lengths, to take advantage of the improved translation system. This would cause bias because some of the estimated export increase would come from changes in seller behavior, rather than reduction in translation cost.

To deal with this concern, we focus on listings that were listed in the four weeks before the policy change and were not modified in the four weeks before and after the policy change. For this set of listings, the listing attributes were set ex-ante and therefore are less subjective to strategic behaviors. The results reported in the Panel A of Table 3 show policy effect of similar magnitude, suggesting that bias from sellers' strategic behavior is small.

Next, we study how number of photos in a listing moderates the effect of improved title translation. Since both photos and titles provide information on item characteristics, there might be a substitutability between the two instruments. We therefore further interact "No. Words*Post" with "No. Photos" in Equation (1), and control for all the standalone and two-way interaction dummies.

Results are reported in Panel B of Table 3. We estimate the moderating effect both using listings with 0 to 8 photos (95% of listings), and with 1 to 4 photos (80% of listings). We find that having one more photo in the listing reduces the policy effect by 0.04% to 0.06%, which is not very large. This suggests that having good title translation is of first-order importance, because buyers decide whether to click a listing based on the listing title and the leading picture on the search result page. Having multiple photos do not change their decision in this first step.[6] Although intuitive, we should note that the estimated moderating effect is only suggestive, since the number of photos might be negatively correlated with title length of a listing, as sellers may feel less necessary to create long titles if they think that the photos will accurately describe the product.

Lastly, we have studied heterogeneous policy effect by different product and buyer types separately. One might worry that there is a systematic correlation between the two, i.e., more experienced buyers may be more likely to buy certain type of products.

In Panel C of Table 3, we replicate the heterogeneous policy effect by buyer experience separately for homogeneous and differentiated products. We find qualitatively similar results across both product types. In particular, export increase for experienced buyers, relative to that of in-

[6] One might be tempted to compare listings with no photo with listings with some photos. However, only 1% of listings have no photo, and they rarely show up in search result page due to eBay's search ranking algorithm.

experienced buyers, is smaller for homogeneous products than for differentiated products. This could come from that fact that experienced buyers were already purchasing many homogeneous products from U.S. sellers before the policy change. This exercise suggests that both buyer types and product types are important margins when considering the effect of better translation.

Lastly, we have repeated the same set of exercises for eMT's rollouts in the European Union and Russia. The effects on exports are comparable for other language pairs: English–French, English–Italian, and English–Russian, which are reported in the appendix.

7 Conclusion

We exploit natural experiments on eBay to study the effect of an AI-based machine translation tool on international trade. We show that a quality upgrade in listing title translation increases exports on eBay by 10.9%. The increase in exports is larger for differentiated products, cheaper products, and less experienced buyers. These heterogeneous effects are consistent with a reduction in costs of translating listing titles.

Our results have two main implications:

First, language barriers greatly hinder trade. This is true even for digital platforms where trade frictions are already smaller than offline. In our study, the quality upgrade in machine translation for listing titles is moderate: an increase in BLEU score from 41.01 to 45.25, or alternatively, an increase of human acceptance rate from 82.4% to 90.2%. However, this moderate quality improvement generated an export increase of 10.9% in exports. To put our result in context, Hui [2018] has estimated that a removal of export administrative and logistic costs increased export revenue on eBay by 12.3% in 2013, which is similar to the effect of eMT for title translation. Additionally, Lendle et al. [2016] have estimated that a 10% reduction in distance would increase trade revenue by 3.51% on eBay. This means that the introduction of eMT is equivalent to an export increase from reducing distances between countries by 26.1%.[7] These comparisons suggest that the trade-hindering effect of language barriers is of first-order importance. Improved machine translation has made the eBay world significantly more connected.

Second, AI is already affecting productivity and trade, and it has significant potential to increase them further. In November 2016, Google announced its neural machine translation (NMT) system based on deep learning that has significantly improved translation quality compared to the previous

[7]The estimated overall policy effect on *export revenue* based on Equation (1) is 9.16% (results reported in the appendix). The equivalent reduction in distance is computed as 9.16%/3.51%*10% = 26.1%.

generation of Google Translate among many language pairs. The estimates in this paper suggest that the effect of NMT on cross-border trade could be large. Besides machine translation, AI applications are also emerging in other fields such as speech recognition and computer vision, with applications ranging from medical diagnoses and customer support to hiring decisions and self-driving vehicles. As each of the new systems come online, they will provide new opportunities to assess the economic impact of AI via natural experiments such as the one examined in this paper.

References

Daron Acemoglu and Pascual Restrepo. Artificial intelligence, automation and work. Technical report, National Bureau of Economic Research, 2018.

Philippe Aghion, Benjamin F Jones, and Charles I Jones. Artificial intelligence and economic growth. Technical report, National Bureau of Economic Research, 2017.

Ajay Agrawal, J Gans, and Avi Goldfarb. Exploring the impact of artificial intelligence: Prediction versus judgment. *University of Toronto*, 2016.

James E Anderson and Eric Van Wincoop. Gravity with gravitas: A solution to the border puzzle. *The American Economic Review*, 93(1):170–192, 2003.

Bernardo S Blum and Avi Goldfarb. Does the internet defy the law of gravity? *Journal of International Economics*, 70(2):384–405, 2006.

Jeffrey R Brown and Austan Goolsbee. Does the internet make markets more competitive? evidence from the life insurance industry. *Journal of Political Economy*, 110(3):481–507, 2002.

Erik Brynjolfsson and Andrew McAfee. Whats driving the machine learning explosion? *Harvard Business Review*, pages 3–11, 2017.

Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534, 2017.

Erik Brynjolfsson and Michael D Smith. Frictionless commerce? a comparison of internet and conventional retailers. *Management Science*, 46(4):563–585, 2000.

Erik Brynjolfsson, Daniel Rock, and Chad Syverson. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. Technical report, National Bureau of Economic Research, 2017.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Christian Catalini and Xiang Hui. Can capital defy the law of gravity? investor networks and startup investment. 2017.

Alberto Cavallo. Are online and offline prices similar? evidence from large multi-channel retailers. *American Economic Review*, 107(1):283–303, 2017.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

Peter H Egger and Andrea Lassmann. The language effect in international trade: A meta-analysis. *Economics Letters*, 116(2):221–224, 2012.

Liran Einav, Theresa Kuchler, Jonathan Levin, and Neel Sundaresan. Assessing sale strategies in online markets using matched listings. *American Economic Journal: Microeconomics*, 7(2): 215–247, 2015.

Liran Einav, Chiara Farronato, and Jonathan Levin. Peer-to-peer markets. *Annual Review of Economics*, 8:615–635, 2016.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Chiara Farronato and Andrey Fradkin. The welfare effects of peer entry in the accommodation market: The case of airbnb. Technical report, National Bureau of Economic Research, 2018.

Avi Goldfarb and Daniel Trefler. Ai and international trade. Technical report, National Bureau of Economic Research, 2018.

Avi Goldfarb and Catherine Tucker. Digital economics. Technical report, National Bureau of Economic Research, 2017.

Xiang Hui. Facilitating inclusive global trade: Evidence from a field experiment. 2018.

Xiang Hui, Maryam Saeedi, Zeqian Shen, and Neel Sundaresan. Reputation and regulations: Evidence from ebay. *Management Science*, 2016.

Xiang Hui, Maryam Saeedi, Giancarlo Spagnolo, and Steve Tadelis. Certification, reputation and entry: An empirical analysis. *NBER Working Paper*, 2018.

Anton Korinek and Joseph E Stiglitz. Artificial intelligence and its implications for income distribution and unemployment. Technical report, National Bureau of Economic Research, 2017.

Chungsang Tom Lam and Meng Liu. Demand and consumer surplus in the on-demand economy: the case of ride sharing. 2017.

Alon Lavie. Evaluating the output of machine translation systems. *AMTA Tutorial*, page 86, 2010.

Andreas Lendle, Marcelo Olarreaga, Simon Schropp, and Pierre-Louis Vézina. There goes gravity: ebay and the death of distance. *The Economic Journal*, 126(591):406–441, 2016.

Meng Liu, Erik Brynjolfsson, and Jason Dowlatabadi. Technology, incentives, and service quality: the case of taxis and uber. *NBER Working Paper*, 2018.

Johannes Lohmann. Do language barriers affect trade? *Economics Letters*, 110(2):159–162, 2011.

Jacques Melitz. Language and foreign trade. *European Economic Review*, 52(4):667–699, 2008.

Jacques Melitz and Farid Toubal. Native language, spoken language, translation and trade. *Journal of International Economics*, 93(2):351–363, 2014.

Atif Mian and Amir Sufi. The effects of fiscal stimulus: Evidence from the 2009 cash for clunkers program. *The Quarterly journal of economics*, 127(3):1107–1142, 2012.

Alejandro Molnar. Language barriers to foreign trade: evidence from translation costs. *Nashville: Vanderbilt University*, 2013.

Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

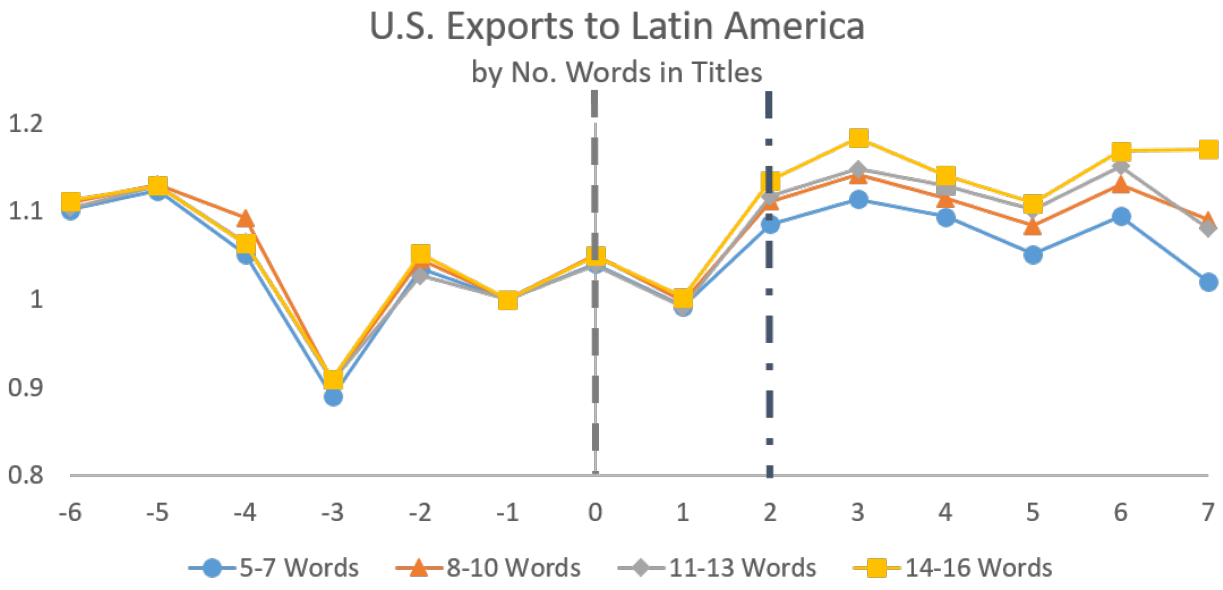
Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Jeffrey D Sachs and Laurence J Kotlikoff. Smart machines and long-term misery. Technical report, National Bureau of Economic Research, 2012.

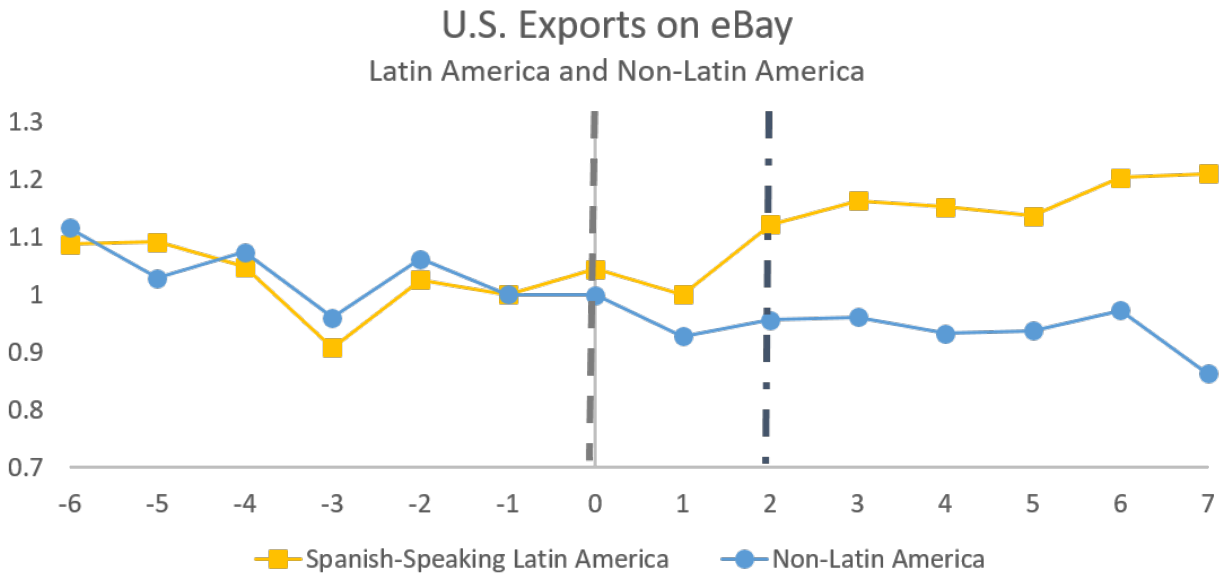
David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Figure 1: Share of Exports by Title Lengths

Notes: Exports are measured in quantity. 95% of exports have title lengths between 5 and 16.



(a)



(b)

Figure 2: Parallel Trends Assumption

Notes: Exports are measured in quantity and are normalized to the level in April 2014 (t=-1). The dashed and dot-dashed lines indicate the introduction of query and item title translations, respectively.

Table 1: Overall Policy Effect

	(1)	(2)	(3)	(4)
Panel A. Overall Effect				
	All Data		+/- 6 Weeks	
	Main Spec	Add'l Controls	Main Spec	Add'l Controls
No. Words*Post	0.0106	0.014	0.0079	0.0123
	(0.0014)	(0.0025)	(0.0021)	(0.0043)
Obs	3024	3024	2592	2592
Panel B. By Homogeneity of Products				
	All Data		+/- 6 Weeks	
	Main Spec	Add'l Controls	Main Spec	Add'l Controls
No. Words*Post	0.0185	0.0224	0.014	0.0155
	(0.0022)	(0.0035)	(0.0032)	(0.0036)
No. Words*Post*Homogeous	-0.012	-0.0165	-0.01	-0.0098
	(0.0031)	(0.0048)	(0.0044)	(0.0049)
No. Obs	6048	6048	5184	5184
Panel C. By Product Value				
	All Data		+/- 6 Weeks	
	Main Spec	Add'l Controls	Main Spec	Add'l Controls
No. Words*Post	0.0144	0.0169	0.0091	0.0134
	(0.0011)	(0.0013)	(0.0013)	(0.0024)
No. Words*Post*Value $\in$ [10,50)	-0.0002	-0.0018	-0.0009	-0.0013
	(0.0016)	(0.0021)	(0.0025)	(0.0034)
No. Words*Post*Value $\in$ [50,200)	-0.0025	-0.0039	-0.0026	-0.0044
	(0.0016)	(0.0021)	(0.0026)	(-0.0032)
No. Words*Post*Value $\geq$ 200	-0.0052	-0.0063	-0.0046	-0.0053
	(0.0019)	(0.0022)	(0.0022)	(0.0031)
No. Obs	12096	12096	10368	10368
Panel D. By Buyer Experience				
	All Data		+/- 6 Weeks	
	Main Spec	Add'l Controls	Main Spec	Add'l Controls
No. Words*Post	0.0124	0.0144	0.0083	0.0133
	(0.0011)	(0.0021)	(0.0016)	(0.002)
No. Words*Post*Experienced	-0.0055	-0.0063	-0.0055	-0.0056
	(0.002)	(0.0029)	(0.0022)	(0.0027)
Obs	6048	6048	5184	5184

Notes: We control for variables according to equation (1). In Panel B, we additionally control for the dummy for homogeneous products, its interaction with "No. Words", and its interaction with "Post". In Panel C, we additionally control for the dummies for the four value ranges, their interaction with "No. Words", and their interaction with "Post". In Panel D, we additionally control for the standalone dummy variable "Experienced", its interaction with "No. Words", and its interaction with "Post". Standard errors clustered at the country level.

Table 2: Placebo Tests

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Continuous Diff-in-Diff						
	All Data		-6 Months		-12 Months	
	Main Spec	Add'l Controls	Main Spec	Add'l Controls	Main Spec	Add'l Controls
Panel A. Overall Effect						
No. Words*Placebo	0.0010	0.0038	-0.0038	-0.0030	-0.0011	-0.0012
	(0.0025)	(0.0028)	(0.0029)	(0.0044)	(0.0033)	(0.0039)
No. Words*Post	0.0106	0.0125				
	(0.0026)	(0.0028)				
Obs	3024	3024	2592	2592	2592	2592
Panel B. Diff-in-Diff: U.S. Exports to Spanish-Speaking Latin America and to Other Countries						
	All Data		6 Months Before		12 Months Before	
	Main Spec	Add'l Controls	Main Spec	Add'l Controls	Main Spec	Add'l Controls
Panel A. Overall Effect						
LatAm*Placebo	0.0609	0.0513	-0.0332	-0.0308	-0.0138	-0.0137
	(0.0237)	(0.0242)	(0.0330)	(0.0331)	(0.0136)	(0.0137)
LatAm*Post	0.1193	0.0987				
	(0.0218)	(0.0224)				
Obs	1456	1456	1248	1248	1248	1248

Notes: We control for variables according to equation (1). In columns (1) and (2), "Placebo" refers to May 2014 when eMT for query translation was introduced. In columns (3) and (4), "Placebo" refers to January 2014, which is six months before the introduction of the eMT for title translation. In columns (5) and (6), "Placebo" refers to July 2013, which is one year before the introduction of the eMT for title translation. Standard errors clustered at the country level.

Table 3: Other Robustness Analyses

	(1)	(2)	(3)	(4)
Panel A. Listings without Modification (+/- 4 Weeks)				
	Main Spec	Add'l Controls		
No. Words*Post	0.0149	0.0218		
	(0.0037)	(0.0068)		
Obs	1728	1728		
Panel B. By No. Photos				
	0-8 Photos		1-4 Photos	
	Main Spec	Add'l Controls	Main Spec	Add'l Controls
No. Words*Post	0.0126	0.0153	0.0128	0.0149
	(0.0017)	(0.0016)	(0.0031)	(0.003)
No. Words*Post*No. Photos	-0.0006	-0.0004	-0.0004	-0.0003
	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Obs	27216	27216	12096	12096
Panel C. Buyer Experience X Product Type				
	Homogeneous Product		Differentiated Product	
	Main Spec	Add'l Controls	Main Spec	Add'l Controls
No. Words*Post	0.0102	0.0113	0.0179	0.0193
	(0.0021)	(0.0021)	(0.0024)	(0.0029)
No. Words*Post*Experienced	-0.0061	-0.0065	-0.0095	-0.0064
	(0.003)	(0.003)	(0.0033)	(0.0031)
Obs	6048	6048	6048	6048

Notes: We control for variables according to equation (1). In Panel B, we additionally control for the dummy for number of pictures, its interaction with "No. Words", and its interaction with "Post". In Panel C, we additionally control for the standalone dummy variable "Experienced", its interaction with "No. Words", and its interaction with "Post". Standard errors clustered at the country level.