




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

 Erik Brynjolfsson,^a Xiang Hui,^b Meng Liu^b
^a Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; ^b Marketing, Olin School of Business, Washington University in St. Louis, St. Louis, Missouri 63130

Contact: erikb@mit.edu,  <http://orcid.org/0000-0002-8031-6990> (EB); hui@wustl.edu,  <http://orcid.org/0000-0001-7595-3461> (XH); mengli@wustl.edu,  <http://orcid.org/0000-0002-5512-7952> (ML)

Received: April 18, 2019

Revised: April 18, 2019

Accepted: April 18, 2019

Published Online in Articles in Advance: September 3, 2019

<https://doi.org/10.1287/mnsc.2019.3388>
Copyright: © 2019 INFORMS

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 a new 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.

History: Accepted by Joshua Gans, business strategy.

Supplemental Material: The online appendix is available at <https://doi.org/10.1287/mnsc.2019.3388>.

Keywords: artificial intelligence • international trade • machine translation • machine learning • digital platforms

1. Introduction

Recent progress of AI and, in particular, machine learning (ML), has dramatically increased predictive power in many areas, including speech recognition, image recognition, and credit scoring (Mullainathan and Spiess 2017, Agrawal et al. 2018). Machines that automatically translated languages were a staple of science fiction for generations, but they are now reality. Unlike the last generation of information technology, ML is designed to learn patterns automatically from examples (Brynjolfsson and Mitchell 2017). This has opened a broad new frontier of applications with significant economic implications.

Has AI adoption begun to affect economic activities? Extensive resources are being allocated to AI research and implementation, and expectations are very high.¹ However, although AI capabilities are impressive, direct causal evidence of a relationship between the use of AI and economic activities is lacking. In particular, contributions from AI are not found in aggregate productivity measures. Brynjolfsson et al. (2019) argue that the gap between expectations and statistics is likely a result of lags in complementary innovations. In light of this, the first domains to be affected by AI are likely to be settings in which AI applications can be seamlessly embedded into existing systems without requiring new complementary innovations. In particular, digital platforms are pioneers of AI adoption, providing ideal opportunities for early assessment of AI's economic effects.

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 cross-border trade among more than 200 countries in 2014. The focal AI technology, eMT, is an in-house ML system that statistically learns language translation. It replaces and significantly improves translation quality relative to eBay's previous translation technology. We exploit the introduction of eMT for various language pairs, particularly for Spanish-speaking Latin American countries (henceforth "Latin America") as a natural experiment and study its consequence on U.S. exports to these countries on eBay. Language has long been recognized as a barrier to international trade, so it is plausible that any technology that significantly lowers this barrier will increase trade.

A standard approach of evaluating the effectiveness of machine translation (MT) would exploit across-country variation in the availability or quality of MT in this setting and broader settings. This approach has two key challenges. First, country pairs with better trade prospects may select into introducing or improving MT. Hence, what appears to be MT's effect among countries with higher translation quality could be self-selection. Second, contemporaneous events around the improvement of MT (e.g., changes in marketing activities or macroeconomic conditions) may bias the estimation.

We adopt a *within-country*, continuous difference-in-difference (DiD) approach that exploits the variation in the number of words across listing titles. Essentially,

we compare the postpolicy change in U.S. exports to Latin America for listings with longer versus shorter titles. We assume 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 length (in words).

Our within-country identification addresses the selection issue at the country level by holding the set of countries fixed. The approach also addresses the concern of contemporaneous events as long as they are orthogonal to title lengths, which is a reasonable assumption in our setting. For example, the effects of any change in advertising spending in Latin America should similarly affect exports of listings with different title lengths.

We estimate that the overall exports to Latin America are 10.9% after the introduction of eMT of title translation. This reflects an average increase of 1.06% for each additional word in the listing title. Furthermore, we identify heterogeneous treatment effects consistent with eMT's mechanism in reducing translation costs for market participants: the export increase is more pronounced for differentiated products (which rely more on accurate descriptions), cheaper products (for which the fixed cost of translation is proportionately greater), and less experienced buyers (who may be less familiar with the products and with alternative means for translating titles). The heterogeneous treatments effects each support the interpretation of a causal effect of eMT on international trade.

Despite its strengths, 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 (2), we perform various placebo tests using different months before eMT's introduction. To deal with (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.

Another concern is that sellers may strategically change their listings' title lengths to take advantage of eMT. To mitigate this concern, we study items that were listed *before* the policy change and were not modified afterward and find a consistent policy effect for this set of listings.

Finally, 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. In each case, our findings are consistent with a causal effect of eMT.

In addition to establishing causality, we also find that the effect of eMT is quite large. Drawing on

gravity models of international trade, 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. AI and Economic Welfare. The current generation of AI represents a revolution of prediction and classification capabilities (e.g., Brynjolfsson and McAfee 2017). Recent breakthroughs in ML, especially supervised learning systems using deep neural networks, have allowed 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. 2019).

Machine translation has also experienced significant improvement because of 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 2018,² according to the widely used BLEU score, which measures how close the MT translation output is to one or more reference translations by linguistic experts (for details, see Papineni et al. 2002). Much of the recent progress in MT has been a shift from symbolic approaches toward 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 relatively seamless. For instance, for product listings on eBay, users consume the output of the translation system but, otherwise, need not change their buying or selling process. Although 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. Although, so far, much of the work on the economic effects of AI has been theoretical (Sachs and Kotlikoff 2012, Aghion et al. 2017, Korinek and Stiglitz 2017, Acemoglu and Restrepo 2018, Agrawal et al. 2019) and notably (Goldfarb and Treffer 2018) in the case of global trade, the introduction of improved MT on eBay is an early opportunity to assess the economic effects of AI using plausible natural experiments.

1.1.2. Language Barriers in Trade. Empirical studies using gravity models, which are formally derived in Anderson and Van Wincoop (2003), have established a robust negative correlation between bilateral trade and language barriers. Typically, researchers regress

bilateral trade on a “common language” dummy and find that this coefficient is strongly positive (Egger and Lassmann 2012).³ 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 (e.g., the United Kingdom and Australia) can also be similar in preferences for food, clothing, entertainment, and so forth. Without exogenous variation in one or the other, it is impossible to tease out the language effect on trade.

Our paper exploits a natural experiment on eBay that provides exactly such an exogenous change, namely a large reduction in the language barrier, and assesses its effect on international trade. The online marketplace provides us with a powerful laboratory to study the consequences on bilateral trade after this decrease in language barriers for a given language pair. Our finding that a quality upgrade of machine translation could increase exports by about 10.9% is consistent with Lohmann (2011) and Molnar (2013), who argue that language barriers may be far more trade hindering than previously suggested.

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), Overby and Forman (2014), and Cavallo (2017). These reduced frictions also mitigate geographic inequality in economic activities in the case of ride-sharing platforms (Lam and Liu 2017, 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, Cowgill and Dorobantu 2018, Hui 2019). 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. 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 handcrafted 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 for title translation in July 2014 between the United States and Latin America. We discuss both introductions in this section to comprehensively illustrate buyers’ interaction with eMT although we mainly aim to identify the effect of eMT for item title translation. To shop on eBay, buyers in Latin America visit www.ebay.com and see items from sellers who sell to buyers’ countries. eBay recognizes buyers’ IP addresses from Latin America and shows buyers the website in Spanish. Note that the translation of nonuser-generated content on the website, such as product categories, existed before and was not affected eMT’s introduction. Instead, eMT affected translation quality of only *search queries* and *listing titles* in the period we study.⁴ 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 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 the other hand, although human evaluations are highly interpretable, they are very costly and can be less consistent.

A comparison of Bing and eMT translation for item titles from English into Spanish revealed 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 human translation. Therefore, in cases in which the grammar and style of translation are not of first-order importance, such as in listing titles, one might prefer using HAR for measuring translation quality.

When eBay first introduced eMT for query translation in May 2014, it also made other localization changes; it catered local deals to buyers’ home country and allowed buyers to see prices in local currencies.⁵

Also, eBay might have increased its advertising spending in Latin America. As is 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 the introduction of eMT for query translation and other localization effects, so we also shrink the estimation window to exclude this period.

In 2014, eBay 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 months apart in Latin America but 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.

3. Data and Empirical Strategy

We use administrative data from eBay, including detailed product, listing, and buyer characteristics. Importantly, we observe the number of words in listing titles, which provide information on the number of words translated from English into Spanish.⁶ 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 is larger for listings with higher ex ante translation costs. Because we do not directly observe translation costs, we use the number of words in listing titles (excluding numbers

and symbols) as a proxy and assume that ex ante translation costs are nondecreasing in the number of words in the title. 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 .⁷

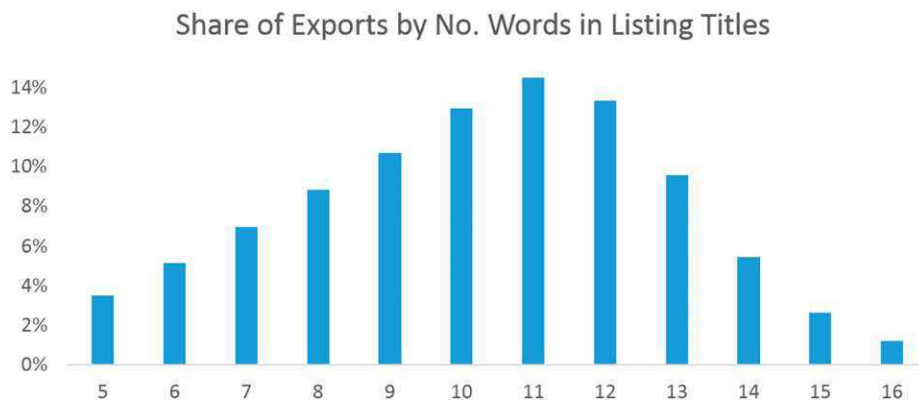
We aim to create treatment and control groups using variations in title lengths across different listings. Figure 1 plots the share of exports across listings with 5–16 words in the title (95% of U.S. exports to affected countries). Note that exports refer to the export quantity throughout the paper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

$$\log(Y_{cIt}) = \beta \text{Num_Words}_{cIt} \times \text{Post}_t + \gamma \text{XR}_{ct} + \eta_c + \text{Num_Words}_l + \xi_t + \epsilon_{cIt}. \quad (1)$$

The regression is performed at the country–title length–time period level. We use Y_{cIt} to denote the exports to country c of title length l in period t , Num_Words_{cIt} 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_l are title length fixed effects, and ξ_t are time fixed effects. The coefficient β represents the 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 correlation of exports.

In a nutshell, we estimate the policy effect by comparing the postpolicy change in U.S. exports to Latin America 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

Figure 1. (Color online) Share of Exports by Title Length (in Words)



Note. Exports are measured in quantity; 95% of exports have title lengths between 5 and 16 words.

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: title length 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 eMT’s introduction. Finally, to deal with time-variant and serially uncorrelated omitted variables, we include many time-varying, title length–specific market characteristics in the regressions as robustness checks.

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

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

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 the parallel trend assumption for our continuous DiD specification. Figure 2(a) plots the average monthly U.S. exports to Latin America by the number of words in 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 $t = -1$. The figure shows 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 eMT’s introduction reduces translation costs in understanding item titles.

We estimate the policy effect with Equation (1) using the same data as in Figure 2(a). Specifically, the sample includes 14 months \times 18 Latin American countries \times 12 categories of title lengths. In Table 1, “*Post*” dummy turns to one at $t = 2$ when eMT title translation is introduced. Column (1) in panel A shows that the export increase is 1.06% larger for listings with one

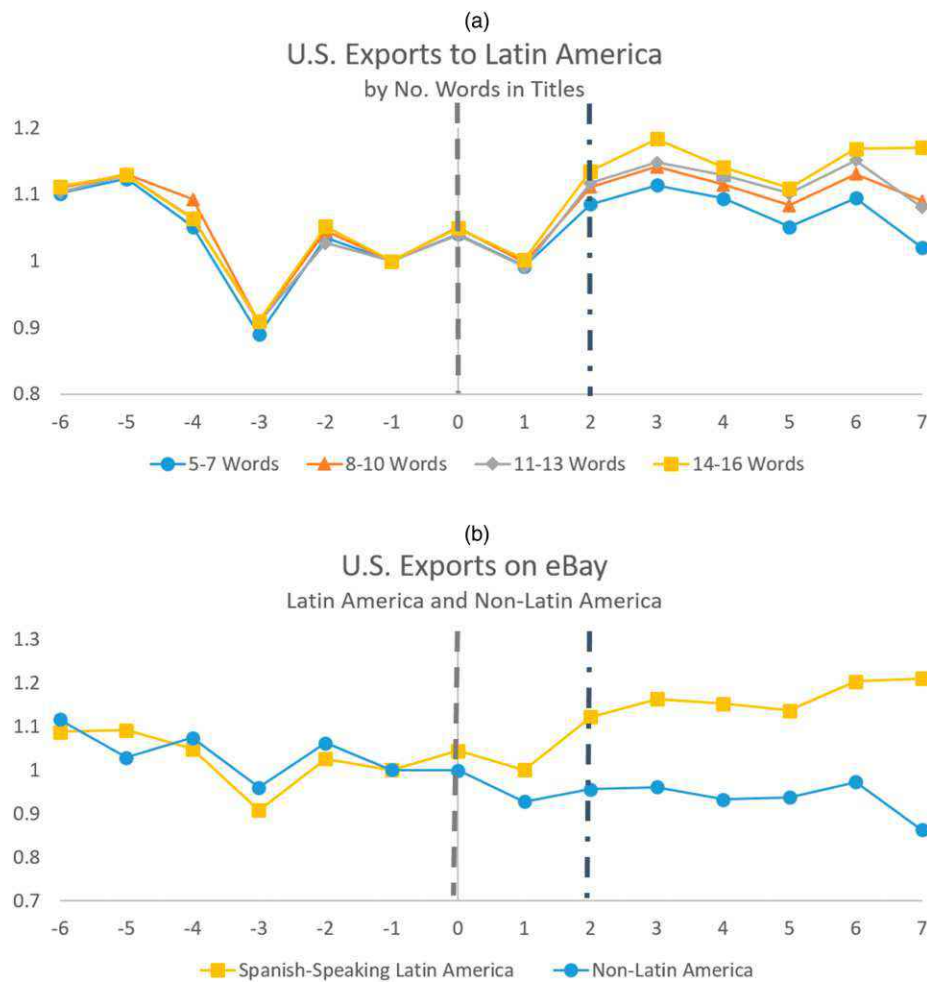
more word in the title. This implies a 10.9% overall export increase given that the average title contains 10.26 words.

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. The inclusion of these characteristics does not reduce the estimated policy effect. The estimated coefficients of these characteristics are also sensible: number of distinct products, average seller size, and share of top rated sellers are positively correlated with exports, and 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) to exclude the period containing the introduction of eMT for query translation (May 2014) among other advertising and localization efforts. 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 the policy effect on a per-word basis. Alternatively, we estimate it by comparing U.S. exports across treated and nontreated countries. In Figure 2(b), we plot monthly U.S. exports to 18 Latin American and 86 non–Latin American countries. Exports are normalized by the value in the month before 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 nontreated countries. The gap between the two series became even larger after the introduction of improved title translation ($t = 2$). The observations suggest that the introduction of improved query and title translation increased exports although the latter seems to have had a bigger effect.

Interestingly, although improved query translation appears to increase U.S. exports when we compare across countries ($t = 0$ in Figure 2(b)), it does not lead to differential export changes across title lengths within affected countries ($t = 0$ in Figure 2(a)). This result makes sense because improved *query* translation should not have a differential effect on *titles* with different lengths. 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.

Figure 2. (Color online) 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.

4.2. Heterogeneous Policy Effects

If the observed export changes estimated from Equation (1) are caused by the introduction of eMT for title translation, we should expect to see more pronounced effects in categories with higher translation costs (see a theoretical framework in the online 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). Because 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 the export increase for each additional word is 1.85% for differentiated products but only 0.65% for homogeneous products. This difference is consistent with a causal effect of eMT and is robust to the inclusion of title length characteristics and using shorter estimation windows.

Next, we explore how the policy effect differs by product value. Because 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 the same title length, one worth \$5

Table 1. Overall Policy Effect

	(1)	(2)	(3)	(4)
Panel A. Overall effect				
	All data		±6 weeks	
	Main spec	Additional controls	Main spec	Additional controls
<i>Number of words × post</i>	0.0106*** (0.0014)	0.014*** (0.0025)	0.0079*** (0.0021)	0.0123*** (0.0043)
Adjusted R ²	0.97	0.97	0.97	0.97
Observations	3,024	3,024	2,592	2,592
Panel B. By homogeneity of products				
	All data		±6 weeks	
	Main spec	Additional controls	Main spec	Additional controls
<i>Number of words × post</i>	0.0185*** (0.0022)	0.0224*** (0.0035)	0.014*** (0.0032)	0.0155*** (0.0036)
<i>Number of words × post × homogenous</i>	-0.012*** (0.0031)	-0.0165*** (0.0048)	-0.01** (0.0044)	-0.0098** (0.0049)
Adjusted R ²	0.97	0.98	0.97	0.98
Observations	6,048	6,048	5,184	5,184
Panel C. By product value				
	All data		±6 weeks	
	Main spec	Additional controls	Main spec	Additional controls
<i>Number of words × post</i>	0.0144*** (0.0011)	0.0169*** (0.0013)	0.0091** (0.0013)	0.0134*** (0.0024)
<i>Number of words × post × value ∈ [10,50)</i>	-0.0002 (0.0016)	-0.0018 (0.0021)	-0.0009 (0.0025)	-0.0013 (0.0034)
<i>Number of words × post × value ∈ [50,200)</i>	-0.0025 (0.0016)	-0.0039** (0.0021)	-0.0026 (0.0026)	-0.0044 (-0.0032)
<i>Number of words × post × value ≥ 200</i>	-0.0052** (0.0019)	-0.0063*** (0.0022)	-0.0046** (0.0022)	-0.0053* (0.0031)
Adjusted R ²	0.98	0.98	0.99	0.99
Observations	12,096	12,096	10,368	10,368
Panel D. By buyer experience				
	All data		±6 weeks	
	Main spec	Additional controls	Main spec	Additional controls
<i>Number of words × post</i>	0.0124*** (0.0011)	0.0144*** (0.0021)	0.0083*** (0.0016)	0.0133*** (0.002)
<i>Number of words × post × experienced</i>	-0.0055*** (0.002)	-0.0063** (0.0029)	-0.0055** (0.0022)	-0.0056** (0.0027)
Adjusted R ²	0.96	0.98	0.97	0.98
Observations	6,048	6,048	5,184	5,184

Notes. We control for variables according to Equation (1). In panel B, we additionally control for the dummy for homogeneous products, its interaction with *Number of words*, and its interaction with *Post*. In panel C, we additionally control for the dummies for the four value ranges, their interaction with *Number of words*, and their interaction with *Post*. In panel D, we additionally control for the standalone dummy variable *Experienced*, its interaction with *Number of words*, and its interaction with *Post*. Standard errors clustered at the country level. Spec, specification.

*** Significant at the 1% level; ** significant at the 5% level; * significant at the 10% level.

and the other \$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 six months preceding the policy change. Column (1) in panel C of Table 1 shows that the export increase for cheap products is 1.44% for each additional word in listing titles but decreases to 0.92% for expensive products. This finding is consistent across columns (2)–(4) as we include more controls and shrink the estimation window.

Besides heterogeneous translation costs across product types, buyers themselves 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 were more likely to incur the hassle cost of using translation tools. Therefore, improved translation quality should mainly affect inexperienced buyers who may have a higher benefit from eMT.

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. 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 it is only 0.69% for experienced ones. The qualitative finding is persistent when we add title length-specific controls and narrow the estimation windows.

5. Placebo Tests

The key identification assumptions in Equation (1) are 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.⁸ However, the model might 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 changes in exports related to title lengths even before the policy change. If there exist time-varying characteristics that correlate simultaneously with title length and exports, then this confounding relationship should

spuriously drive differential export changes even before the policy change, when no actual treatment took place. This test assumes that the confounding correlation has 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 ($t = 0$) when eMT for query translation was introduced. We have seen in Figure 2, (a) and (b), that exports increased at both $t = 0$ and $t = 2$, but the effects differed across title length only after the actual treatment. Testing this formally, we included "No. of Words \times Placebo" in Equation (1), where "Placebo" equals one on or after $t = 0$. The coefficient of this interaction captures the potential policy effect for improved query translation.

The results analogous to Figure 2(a) are reported in columns (1) and (2) in panel A of Table 2. 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 2(b), estimated using Equation (2), are reported in columns (1) and (2) in panel B. 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.

Our second placebo treatment is six months before the actual treatment ($t = -4$). In columns (3) and (4) in panel A, "Placebo" equals one on or after $t = -4$, and we use data from 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 find no effect on exports.

Finally, our third placebo treatment is 12 months before the actual treatment ($t = -10$). In this case, "Placebo" equals one on or after $t = -10$. Using data from six months before and after this placebo month, we do not detect any export increase as indicated in columns (5) and (6). This is the case according to both Equations (1) and (2), shown in panels A and B, respectively.

In summary, our preferred continuous DiD specification estimates a 10.9% overall export increase after the introduction of eMT for title translation. Interestingly, the across-country DiD specification yields an estimate of 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 do not find any spurious relationship between title lengths and exports in the 6 and 12 months before the introduction of eMT for title translation. These null results based on

Table 2. Placebo Tests

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Continuous DiD						
	All data		6 months before		12 months before	
	Main spec	Additional controls	Main spec	Additional controls	Main spec	Additional controls
Panel A. Overall effect						
<i>Number of words × placebo</i>	0.0010 (0.0025)	0.0038 (0.0028)	−0.0038 (0.0029)	−0.0030 (0.0044)	−0.0011 (0.0033)	−0.0012 (0.0039)
<i>Number of words × post</i>	0.0106*** (0.0026)	0.0125*** (0.0028)				
Adjusted R^2	0.98	0.98	0.97	0.98	0.98	0.98
Observations	3,024	3,024	2,592	2,592	2,592	2,592
Panel B. DiD: U.S. exports to Spanish-speaking Latin America and to other countries						
	All data		6 months before		12 months before	
	Main spec	Additional controls	Main spec	Additional controls	Main spec	Additional controls
Panel A. Overall effect						
<i>LatAm × placebo</i>	0.0609** (0.0237)	0.0513** (0.0242)	−0.0332 (0.0330)	−0.0308 (0.0331)	−0.0138 (0.0136)	−0.0137 (0.0137)
<i>LatAm × post</i>	0.1193*** (0.0218)	0.0987*** (0.0224)				
Adjusted R^2	0.99	0.99	0.99	0.99	0.99	0.99
Observations	1,456	1,456	1,248	1,248	1,248	1,248

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. Spec, specification.

*** Significant at the 1% level; ** significant at the 5% level; * significant at the 10% level.

Equation (1) are reassuring evidence of the validity of our main identification strategy.

6. Other Robustness Checks

One concern is that some sellers could endogenously change their listing attributes, such as title lengths, to take advantage of the improved translation system. This could cause bias because some of the estimated export increase could come from changes in seller behavior rather than a reduction in translation cost.

To mitigate 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 panel A of Table 3 show a policy effect of similar magnitude, suggesting that any bias from sellers' strategic behavior is small.

Next, we include additional fixed effects in Equation (1) to test the robustness of our estimate. Column (1) in panel B reports the estimated policy effect using Equation (1), which is the same as column (1) in panel A of Table 1. In column (2), we additionally control for country by month fixed effects. The identification of

the policy effect here comes from variation in title lengths within country-month cells. The estimated coefficient barely changed and remained highly statistically significant. In column (3), we additionally control for a separate “Num_Words” regressor for each country and a separate “Num_Words” regressor for each month. This allows for the correlation between title lengths and sales to differ across countries and across time. We see that, although the estimated policy effect becomes less statistically significant (t -value = 1.86), the estimated coefficient barely changed. Finally, we fully saturated the model by including all possible fixed effects: country by month fixed effects, country by number of words fixed effects, and month by number of words fixed effects. The estimated coefficient becomes insignificant (t -value = 1.29), which is expected given that we saturate the model with the full set of two-way fixed effects (the total number of fixed effects in the saturated model is 636, which comes from 14 months, 18 countries, and 12 title length categories). Despite the lack of statistical significance, the estimated policy effect changes very little in size. The robustness of our estimate across specifications is a reassuring.

Subsequently, we study how the number of photos in a listing moderates the effect of improved title

Table 3. Other Robustness Analyses

	(1)	(2)	(3)	(4)
Panel A. Listings without modification (± 4 weeks)				
	Main spec	Additional controls		
<i>Number of words</i> \times <i>post</i>	0.0149*** (0.0037)	0.0218*** (0.0068)		
Adjusted R^2	0.98	0.98		
Observations	1,728	1,728		
Panel B. Additional fixed effects				
	Main spec	Main spec	Main spec	Main spec
<i>Number of words</i> \times <i>post</i>	0.0106*** (0.0014)	0.0106*** (0.0039)	0.0106* (0.0057)	0.0094 (0.0073)
<i>Country</i> \times <i>Months FE</i>		✓	✓	✓
<i>Country-specific number of words</i>			✓	
<i>Month-specific number of words</i>			✓	
<i>Country</i> \times <i>number of words FE</i>				✓
<i>Month</i> \times <i>number of words FE</i>				✓
Adjusted R^2	0.97	0.99	0.99	0.99
Observations	3,024	3,024	3,024	3,024
Panel C. By number of photos				
	0–8 photos		1–4 photos	
	Main spec	Additional controls	Main spec	Additional controls
<i>Number of words</i> \times <i>post</i>	0.0126*** (0.0017)	0.0153*** (0.0016)	0.0128*** (0.0031)	0.0149*** (0.003)
<i>Number of words</i> \times <i>post</i> \times <i>Number of photos</i>	–0.0006*** (0.0002)	–0.0004*** (0.0002)	–0.0004** (0.0001)	–0.0003* (0.0001)
Adjusted R^2	0.97	0.97	0.97	0.97
Observations	27,216	27,216	12,096	12,096
Panel D. Buyer experience \times product type				
	Homogeneous product		Differentiated product	
	Main spec	Additional controls	Main spec	Additional controls
<i>Number of words</i> \times <i>post</i>	0.0102*** (0.0021)	0.0113*** (0.0021)	0.0179*** (0.0024)	0.0193*** (0.0029)
<i>Number of words</i> \times <i>post</i> \times <i>Experienced</i>	–0.0061** (0.003)	–0.0065** (0.003)	–0.0095*** (0.0033)	–0.0064** (0.0031)
Adjusted R^2	0.97	0.97	0.95	0.95
Observations	6,048	6,048	6,048	6,048

Notes. We control for variables according top. In Panel B, we additionally control for the dummy for number of pictures, its interaction with *Number of words*, and its interaction with *Post*. In panel C, we additionally control for the standalone dummy variable *Experienced*, its interaction with *Number of words*, and its interaction with *Post*. Standard errors clustered at the country level. Spec, specification.

*** Significant at the 1% level; ** significant at the 5% level; * significant at the 10% level.

translation. Because both photos and titles provide information on item characteristics, the two instruments might be substitutes. We, therefore, further interact “No. Words \times Post” with “No. Photos” in Equation (1) and control for all the stand-alone and two-way interaction dummies.

The results are reported in panel C of Table 3. We estimate the moderating effect using listings with both zero to eight photos (95% of listings) and one to four photos (80% of listings). We find that having one more photo in the listing reduces the policy effect by 0.04%–0.06%, which is not large. This suggests that

having a good title translation is of first-order importance because buyers decide whether to click a listing based on its title and the leading picture on the search result page. The number of photos does not change their decision in this first step.⁹ Although intuitive, note that the estimated moderating effect is only suggestive because the number of photos might be endogenous to listings’ title lengths.

Furthermore, we study the policy effect by different product and buyer types separately. One might worry that there is a systematic correlation between the two (experienced buyers purchase certain product types more).

In panel D 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, the export increase for experienced buyers, relative to that of inexperienced buyers, is smaller for homogeneous products than for differentiated products. This might be because 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.

Finally, we repeat the 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, as reported in the online appendix.

7. Conclusion

We exploit a set of natural experiments on eBay to study the effect of an AI-based machine translation tool on international trade. We show that the introduction of eMT—a significant quality upgrade in translation—increases exports on eBay by 10.9%. The export increase is larger for items with longer titles, differentiated products, cheaper products, and less experienced buyers. Each of these heterogeneous effects is consistent with a causal effect from the reduction in the cost of translating listing titles.

7.1. Our Results Have Two Main Implications

First, language barriers greatly hinder trade. This is true even for digital platforms in which trade frictions are already smaller than off line. The quality upgrade in machine translation for listing titles is moderate: the BLEU score increases from 41.01 to 45.25 or, alternatively, the HAR increases from 82.4% to 90.2%. This quality improvement generated an export increase of 10.9%. Putting our result in context, Hui (2019) 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 implies that the introduction of eMT is equivalent to an export increase from reducing distances between countries by 26.1%.¹⁰ 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. Besides machine translation, AI applications are also emerging in other fields, such as speech

recognition and computer vision, with applications including medical diagnoses, customer support, hiring decisions, and self-driving vehicles. As each of the new systems come online, they will provide new opportunities to assess the economic effect of AI via natural experiments, such as the one examined in this paper.

Endnotes

¹ Overall investments in AI startups increased by 150% globally from 2016 to 2017. Source: <https://www.abiresearch.com/market-research/product/1030415-artificial-intelligence-investment-monitor/>.

² Source: <http://matrix.statmt.org/matrix>.

³ In Melitz (2008) and Melitz and Toubal (2014), the authors argued for a continuous variable instead of a common language dummy.

⁴ eBay prioritized search queries and item titles because searching for a product and viewing search results in the buyers' language allows consumers to make informed decisions on which listings to open.

⁵ Source: <https://www.ebayinc.com/stories/news/ebay-delivers-localized-shopping-experiences-latin-america/>.

⁶ A good translation should preserve brand names.

⁷ On eBay, phrases in listing titles are usually nonrepetitive and are about different characteristics of an item. For example, "Diamond-Cut Stackable Thin Wedding Ring New .925 Sterling Silver Band Sizes 4-12" and "Alpine Swiss Keira Women's Trench Coat Double Breasted Wool Jacket Belted."

⁸ We have also identified the effect for new versus used items. The estimates are 0.0018 and 0.0013, both highly significant.

⁹ One might be tempted to compare listings with photos and without photos. However, only 1% of listings have no photo, and they rarely show up in the search result page because of eBay's search ranking algorithm.

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

References

- Acemoglu D, Restrepo P (2018) Artificial intelligence, automation and work. NBER Working Paper No. 24196, National Bureau of Economic Research, Cambridge, MA.
- Aghion, 2017 Aghion P, Jones BF, Jones CI (2017) Artificial intelligence and economic growth. NBER Working Paper No. 23928, National Bureau of Economic Research, Cambridge, MA.
- Agrawal A, Gans J, Goldfarb A (2018) Exploring the Impact of Artificial Intelligence: Prediction versus Judgment. NBER Working Paper No. 24626, National Bureau of Economic Research, Cambridge, MA.
- Agrawal A, Gans JS, Goldfarb A (2019) Artificial intelligence: The ambiguous labor market impact of automating prediction. NBER Working Paper No. 25619, National Bureau of Economic Research, Cambridge, MA.
- Anderson JE, Van Wincoop E (2003) Gravity with gravitas: A solution to the border puzzle. *Amer. Econom. Rev.* 93(1):170–192.
- Blum BS, Goldfarb A (2006) Does the Internet defy the law of gravity? *J. Internat. Econom.* 70(2):384–405.
- Brown JR, Goolsbee A (2002) Does the Internet make markets more competitive? Evidence from the life insurance industry. *J. Political Econom.* 110(3):481–507.
- Brynjolfsson E, McAfee A (2017) What's driving the machine learning explosion? *Harvard Business Review* (July 18), <https://hbr.org/2017/07/whats-driving-the-machine-learning-explosion>.

- Brynjolfsson E, Mitchell T (2017) What can machine learning do? Workforce implications. *Science* 358(6370):1530–1534.
- Brynjolfsson E, Smith MD (2000) Frictionless commerce? A comparison of Internet and conventional retailers. *Management Sci.* 46(4):563–585.
- Brynjolfsson E, Rock D, Syverson C (2019) Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago), 23–57.
- Callison-Burch C, Osborne M, Koehn P (2006) Re-evaluation the role of Bleu in machine translation research. *Proc. 11th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 249–256.
- Catalini C, Hui X (2017) Can capital defy the law of gravity? Investor networks and startup investment. Working paper, Massachusetts Institute of Technology, Cambridge.
- Cavallo A (2017) Are online and offline prices similar? Evidence from large multi-channel retailers. *Amer. Econom. Rev.* 107(1): 283–303.
- Cowgill B, Dorobantu C (2018) The US-Canada border effect: Evidence from online commerce. Working paper, Columbia Business School, New York.
- Denkowski M, Lavie A (2014) Meteor universal: Language specific translation evaluation for any target language. *Proc. 9th Workshop Statist. Machine Translation*, 376–380.
- Egger PH, Lassmann A (2012) The language effect in international trade: A meta-analysis. *Econom. Lett.* 116(2):221–224.
- Einav L, Farronato C, Levin J (2016) Peer-to-peer markets. *Annual Rev. Econom.* 8:615–635.
- Einav L, Kuchler T, Levin J, Sundaresan N (2015) Assessing sale strategies in online markets using matched listings. *Amer. Econom. J. Microeconom.* 7(2):215–247.
- Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118.
- Farronato C, Fradkin A (2018) The welfare effects of peer entry in the accommodation market: The case of Airbnb. NBER Working Paper No. 24361, National Bureau of Economic Research, Cambridge, MA.
- Goldfarb A, Treffer D (2018) AI and international trade. NBER Working Paper No. 24254, National Bureau of Economic Research, Cambridge, MA.
- Goldfarb A, Tucker C (2017) Digital economics. NBER Working Paper No. 23684, National Bureau of Economic Research, Cambridge, MA.
- Hui X (2019) Facilitating inclusive global trade: Evidence from a field experiment. *Management Sci.* Forthcoming.
- Hui X, Saeedi M, Shen Z, Sundaresan N (2016) Reputation and regulations: Evidence from ebay. *Management Sci.* 62(12): 3604–3616.
- Hui X, Saeedi M, Spagnolo G, Tadelis S (2018). Certification, reputation and entry: An empirical analysis. NBER Working Paper No. 24916, National Bureau of Economic Research, Cambridge, MA.
- Korinek A, Stiglitz JE (2017) Artificial intelligence and its implications for income distribution and unemployment. NBER Working Paper No. 24174, National Bureau of Economic Research, Cambridge, MA.
- Lam CT, Liu M (2017) Toward Inclusive Mobility: Ridesharing Mitigates Geographical Disparity in Transportation. Working paper, Clemson University, Clemson, SC.
- Lavie A (2010) Evaluating the output of machine translation systems. AMTA Tutorial 86.
- Lendle A, Olarreaga M, Schropp S, Vézina P-L (2016) There goes gravity: ebay and the death of distance. *Econom. J. (London)* 126(591):406–441.
- Liu M, Brynjolfsson E, Dowlatabadi J (2018) Do digital platforms reduce moral hazard? The case of Uber and taxis. NBER Working Paper No. 25015, National Bureau of Economic Research, Cambridge, MA.
- Lohmann J (2011) Do language barriers affect trade? *Econom. Lett.* 110(2):159–162.
- Melitz J (2008) Language and foreign trade. *Eur. Econom. Rev.* 52(4): 667–699.
- Melitz J, Toubal F (2014) Native language, spoken language, translation and trade. *J. Internat. Econom.* 93(2):351–363.
- Mian A, Sufi A (2012) The effects of fiscal stimulus: Evidence from the 2009 cash for clunkers program. *Quart. J. Econom.* 127(3):1107–1142.
- Molnar A (2013) *Language barriers to foreign trade: Evidence from translation costs*. Working paper, Vanderbilt University, Nashville, TN.
- Mullainathan S, Spiess J (2017) Machine learning: An applied econometric approach. *J. Econom. Perspect.* 31(2):87–106.
- Overby E, Forman C (2014) The effect of electronic commerce on geographic purchasing patterns and price dispersion. *Management Sci.* 61(2):431–453.
- Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: A method for automatic evaluation of machine translation. *Proc. 40th Annual Meeting Assoc. Comput. Linguistics* (Association for Computational Linguistics), 311–318.
- Sachs JD, Kotlikoff LJ (2012) Smart machines and long-term misery. NBER Working Paper No. 18629, National Bureau of Economic Research, Cambridge, MA.
- Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, et al. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587): 484–489.