Tuiteamos o pongamos un tuit? Investigating the Social Constraints of Loanword Integration in Spanish Social Media

Ian Stewart*
University of Michigan
ianbstew@umich.edu

Diyi Yang
Georgia Institute of Technology
diyi.yang@cc.gatech.edu

Jacob Eisenstein
Google Research
jeisenstein@google.com

Abstract

Speakers of non-English languages often adopt loanwords from English to express new or unusual concepts. While these loanwords may be borrowed unchanged, speakers may also integrate the words to fit the constraints of their native language, e.g. creating Spanish *tuitear* from English *tweet*. Linguists have often considered the process of loanword integration to be more dependent on language-internal constraints, but sociolinguistic constraints such as speaker background remain only qualitatively understood. We investigate the role of social context and speaker background in Spanish speakers' use of integrated loanwords on social media. We find first that newspaper authors use the integrated forms of loanwords and native words more often than social media authors, showing that integration is associated with formal domains. In social media, we find that speaker background and expectations of formality explain loanword and native word integration, such that authors who use more Spanish and who write to a wider audience tend to use integrated verb forms more often. This study shows that loanword integration reflects not only language-internal constraints but also social expectations that vary by conversation and speaker.

1 Introduction

Languages exchange loanwords constantly as multilingual people adopt words from other languages to express themselves in their native language (Haspelmath, 2009). The English word *tweet* has been adopted into many other languages following the success of Twitter, e.g. producing the Spanish verb *tuitear*. One form of adoption is known as *integration* by which a speaker adapts the loanword to the underlying grammar of the language, e.g. adding the Spanish verb ending *-ear* to the loanword *tweet* to help the word adhere to Spanish grammar (Poplack and Dion, 2012). Speakers may choose to use loanwords with the prescriptively correct form, in this case adding verbal morphology, or with less standard forms, in this case using a paraphrase such as *send a tweet*. We show several examples of this alternation in Table 1. To further the theoretical understanding of the process of loanword integration, this work assesses this process from a speaker's perspective.

Loanword	Verbs	Count
Connect	*conectear*, *hacer un conexión*	7785
Like	*likear*, *dar un like*	5666
Stalk	*stalkear*, *ser un stalker*	5455
Flash	*flashear*, *hacer flash*	4521
Ship	*shippear*, *hacer ship*	4079

Table 1: Top 5 most frequent loanwords on social media and corresponding verb forms.

Researchers have often studied the process of loanword adoption and integration from a language-internal perspective, such as phonological constraints on loanword use (Kang, 2011). However, loanwords also carry *social meaning* (Levendis and Calude, 2019) that relates to formality and standard language norms, and speakers may have their own intuitions about the "correct" way to use a loanword. Therefore, a speaker's background, such as their multilingual knowledge (Poplack, 1988), and the social context of a conversation (Lev-Ari and Peperkamp, 2014) may also play a role in the integration of loanwords. Such social and behavioral factors may also help explain the long-term *acceptance* of loanwords into a language (Chesley, 2010; Zenner et al., 2012). To that end, we leverage multilingual data from social media to assess the speaker-level factors that underlie loanword integration.

Our study provides the following contributions:

- We first collect verb forms for a variety of

*Work completed at Georgia Institute of Technology.

English loanwords related to technology and social life online, as well as similar *control* pairs for native Spanish verbs (§ 3.1, § 3.2).

- To test for the effect of formality, we compare the rate of integrated verb use for loanwords and native verbs between social media posts and newspaper articles (§ 4.1). We find that loanwords and native verbs are integrated at a higher rate in newspaper articles, suggesting that integration is associated with more formal language registers.
- Drawing on this finding, we test the role of different contextual and speaker-background factors as they explain the choice to use integrated verbs for loanwords (§ 3.4, § 4.2). With regression analysis on social media data, we show that speaker background plays a large role: Latin American speakers and high-Spanish speakers tend to choose integrated verbs for loanwords and native words. We also find that the context of a post explains integration, because posts with a larger presumed audience have higher rates of integration. Lastly, we find several points of divergence between loanwords and native verbs, suggesting some differences in social perception of the word groups.

2 Related work

Loanword integration has mainly been studied from the perspective of *pronunciation*, i.e. whether a loanword adheres to the phonology of the source or target language (Kang, 2011). Speakers may have to choose between different valid pronunciations, e.g. pronouncing the word *Iraq* with an American English "short-A" (/ɪɹæk/) or an Arabic "long-A" (/ɪɹɑk/) (Hall-Lew et al., 2010). Traditional studies of loanword integration relied on sociolinguistic interviews and elicitation, which often lack spontaneous loanword use (Poplack, 1988). With the growing availability of large-scale written corpora, researchers have tracked the adoption of loanwords over time, particularly English loanwords into other languages (Chesley, 2010; Garley and Hockenmaier, 2012; Zenner et al., 2012). Such large-scale corpora also allow researchers to track *morphological* integration (Coats, 2018; Kilgarriff, 2010), which is a word's ability to combine with bound morphemes from the target language (e.g. *tuitear* ["to tweet"] = *tuit* ["tweet"] + *-ear* [VERB.INF]). We continue this line of work and study the role of contextual and speaker-background factors in loanword integration. This helps test theories related to multilingual decisions (Poplack et al., 2020) and how loanwords are collectively adopted into a language (Levendis and Calude, 2019).

The loanword integration process relates partly to structure: if the source and target language are similar (e.g. Italian and Spanish) then a speaker may have little difficulty in integrating the loanword (Boersma et al., 2009; Peperkamp, 2004). However, a speaker's decision to integrate a loanword also depends on the speaker's prior experiences and the social context of the conversation (Wohlgemuth, 2009). For one, the choice of using an integrated loanword depends on the speaker's own background with the source language (Poplack, 1988) and their willingness to uphold linguistic standards for the loanword. In addition, the process of loanword integration may be related to the *domain* of speech, as some writing domains such as newspapers have strong norms (Biber and Conrad, 2019) and therefore may prefer the formal version of the loanword. Lastly, the social expectations of a given *conversation* may convince a speaker to use the integrated form (Lev-Ari and Peperkamp, 2014), e.g. if their listeners are expecting a less formal response and therefore a non-integrated loanword. While some work has tested both linguistic and social constraints on the integration of loanwords (Garley, 2014; Sanchez, 2005), linguists generally lack access to speech across a variety of speakers and social contexts. This work addresses the social meaning of loanwords by drawing on the rich speaker-level data available from social media.

3 Data

3.1 Identifying Loanwords

The use of a loanword is considered distinct from code-switching (switching between languages), because a loanword is produced in isolation within the "matrix" language (Poplack, 1988; Cacoullos and Aaron, 2003). This study concerns the alternation between integrated verbs, i.e. those in which the loanword has been morphologically integrated into the language (*tuitear* "to tweet") and light verbs, i.e. phrases in which the loanword is used as a noun (*poner un tweet* "to send a tweet"). We seek light verb phrases that are semantically similar to the integrated verbs, to avoid possible

confounds on the choice between forms.

The list of loanword integrated verbs was identified from two resources: Wiktionary and social media. We first collected all verbs on Spanish-language Wiktionary that are English-origin loanwords and end in one of the standard verb suffixes (*-(e)ar*).[1] Using a sample of Reddit and Twitter data,[2] we collected all words in Spanish-language posts tagged using `langid` (Lui and Baldwin, 2012) that match the structure `ENGLISH_WORD` + *-(e)ar*,[3] under the assumption that most loanword verbs use the *-(e)ar* conjugation (Rodney and Jubilado, 2012). From the combined set of verbs, we removed all cases of ambiguity, e.g. *plantar*, which can be formed by English *plant* + *-ar*, is also a native Spanish word.

For each loanword, we identified a corresponding light verb phrase with a meaning similar to the integrated form. Spanish has a closed class of light verbs used to form phrases with nouns (Buckingham, 2013), such as *tomar* ("take") in *tomar un viaje* ("take a vacation"). We used dictionary definitions from Wiktionary and WordReference to identify valid light verb forms, and we queried the internet for the remaining loanwords to determine their validity (e.g. comparing search results for *hacer un tweet* versus *poner un tweet*). We validated the loanword pairs with Spanish linguistics experts familiar with the process of loanword integration. The experts removed several loanwords that may have been considered native words by Spanish speakers.[4]

This process yielded 120 integrated and light verb pairs that we used to define the dependent variable of the study, i.e. integrated verb use vs. light verb use. We show examples of the most frequent loanword and light verb pairs in Table 1. Many of the words identified relate to technology and online behavior (e.g. *likear* "to like (on social media)"), which represents a sample bias. Because we study loanword use specifically on Twitter, it is likely that the loanwords here relate more to the interests of the platform community rather than the general population.

Native word	Verbs	Count
Dream	*soñar*, *tener un sueño*	39,392
Buy	*comprar*, *hacer la compra*	36,337
End	*terminar*, *poner término*	34,234
Use	*usar*, *hacer uso*	30,834
Test	*probar*, *poner a prueba*	29,930

Table 2: Top 5 most frequent native word pairs and corresponding verb forms on social media.

3.2 Identifying Native Verbs

Studying loanwords in isolation can yield interesting results, but we must also determine whether the patterns of usage reflect constraints on Spanish verbs in general (Wichmann and Wohlgemuth, 2008). To address this concern, we collect an additional set of verbs that are native to Spanish.

We first identified light verb constructions from several grammar blogs and dictionaries,[5] and generated the corresponding integrated verb by adding a standard verb suffix to the noun phrase and verifying with a dictionary.[6] This process yielded 49 pairs of native integrated and light verbs that serve as a baseline to compare with loanword use. We extracted all uses of these native verbs from the set of loanword-using authors mentioned above. As shown in Table 2, the native verbs occur more frequently than the loanword verbs, which compensates for the fact that we have fewer word types for native verbs.

The complete list of loanwords and native verbs is provided in Appendix A for replicability and for linguists to build upon in future work.

3.3 Collecting Loanword Author Data

For our social media data, we collect posts from a 1% Twitter archive sample of Spanish-language posts, ranging from 1 July 2017 to 30 June 2019. We match all original (non-RT) posts that contain at least one loanword verb form, either in the

[1] Accessed 1 Jan 2020: https://es.wiktionary.org/wiki/Categoria:ES:Palabras_de_origen_ingles.

[2] Data sample of Spanish-language posts ranges from 1 July 2017 to 30 June 2019. For Reddit this includes all comments (∼560,000), for Twitter this includes a 1% sample from the Twitter stream (∼110,000,000).

[3] English words collected from a standard spellcheck dictionary and filtered to exclude words shorter than $n = 4$ characters. Accessed 1 Nov 2019: http://wordlist.aspell.net/dicts/.

[4] E.g., Spanish speakers may not consider *flipar* ("to flip") to be a loanword due to its older status.

[5] E.g. "support verbs" mentioned here, accessed 1 Jan 2020: https://comunicarbien.wordpress.com/2011/08/06/verbos-de-apoyo/.

[6] E.g. for the light verb construction *tomar un viaje* ("to take a trip") with the noun *viaje*, we generated the integrated verb *viajar* ("to travel").

integrated form or light verb form.[7] This yields roughly 87,000 posts from 80,000 unique authors over the period of study, from which roughly 23,000 posts from 20,000 authors were used in the regression, after filtering for available variables described in § 3.4.

Next, we collect all available prior posts from these loanword authors using both the original archive sample (2017-2019) and from the authors' full timelines (2014-2019).[8] We recovered roughly 10 million posts from the authors (about 100 extra posts per author) from which we extracted native verb use and speaker background variables for analysis (see Table 3).

3.4 Extracting Speaker-Level Variables

For the speaker-level analysis, we seek to assess the relative importance of several author-level and post-level factors in explaining loanword integration. Following prior work in loanword use, we investigate factors related to **formality** (Biber and Conrad, 2019) and aspects of **speaker background** (Poplack and Dion, 2012) that reflect support for language standards. We therefore use the following metrics to predict verb integration.

- **Formality**:
  - **Post features**: First, we approximate a post's intended *audience* by marking the presence of a hashtag (larger audience) and the presence of an @-mention (smaller audience). We also use the length of a post — excluding the verb phrase — to identify posts that are longer and therefore potentially more formal, following prior work in perceptions of formality in online communication (Chhaya et al., 2018; Pavlick and Tetreault, 2016).
- **Speaker background**:
  - **Posting behavior**: Authors who post frequently may have more extensive knowledge of linguistic norms online and therefore adhere to the standard integrated verb form. For this metric, we extract the author's mean number of prior posts per day. In addition, authors who share more content online may also be more connected to online norms and may therefore adopt the more standard verb form. We compute an author's rate of sharing as (1) the percentage of prior posts that contain a URL and (2) the percentage of prior posts that are retweets.
  - **Location**: The Spanish dialects spoken in Latin America have diverged significantly from Castilian Spanish (Lipski, 1994), which may result in different patterns of loanword adoption. We identify authors' location[9] at the region level: Latin America, US, Europe, or other.[10]
  - **Language use**: Bilingual speakers may be more likely to use the light verb forms of the loanwords , because bilingual speakers often use paraphrases to address unfamiliar concepts (Jenkins, 2003) and may perceive light verb constructions differently (Doğruöz and Nakov, 2014). We tag the authors' prior posts using `langid`,[11] and compute the rate of Spanish use for all authors who have written at least 5 posts. We then bin language use under the assumption that language use may not be linear. Authors who use exclusively Spanish (100%) are assumed to be "strict" monolingual speakers as compared to more "relaxed" bilingual (0-50%) or mid-range bilingual (50-100%) speakers.

    In addition to language choice, speakers who use more integrated native verbs may also use more integrated forms for loanwords. We compute the authors' rate of prior integrated verb use as the number of integrated native verb tokens (§ 3.2) normalized by the total number of native verb tokens.

All variables in the social media data are summarized in Table 3. Note that we choose not to analyze individuals' gender and age due to the

[7] We searched for the most frequently inflected forms of each verb, which include all forms of indicative present, simple past and imperfect. We also remove all verb forms that are ambiguous: e.g. the verb *acceso* ("I access") has the same spelling as the noun *acceso* ("access").

[8] Collected in Mar 2020.

[9] Following prior work (Kariryaa et al., 2018), we use an author's self-reported location in their profile as a location marker. We define an author as a resident of a particular country based on the presence of unambiguous country, state or city keywords in their profile location.

[10] We acknowledge the considerable diversity of Spanish dialects spoken in Latin America (Buckingham, 2013), but we use the level of region in our analysis to avoid data sparsity.

[11] We filter to posts with a confidence score above 90% to reduce likelihood of code-switching.

Variable type	Name	Description	Mean / distribution	
Formality			**Loanword posts**	**Native word posts**
Post content	Hashtag	Whether post contains a hashtag.	8.1%	6.6%
	Mention	Whether post contains an @-mention.	35.2%	7.4%
	Post length	Length of post in characters, excluding the verb phrase.	88	131
Background			**All authors**	
Posting behavior	Activity	Mean posts per day.	8.5	
	Content re-sharing	Percent of prior posts that are retweets.	35.2%	
	Link sharing	Percent of prior posts that contain a URL.	0.5%	
Location	Location	Author's geographic region based on self-reported location.	54.6% UNK, 34.7% Latin America, 7.0% Europe, 2.7% US, 0.9% Other	
Language	Language type	Percent of prior posts written in Spanish.	83.8% high Spanish, 15.5% medium Spanish, 0.7% low Spanish	
	Verb use	Percent of prior native verb posts that contain an integrated verb.	95.4%	

Table 3: Summary of all social media variables used in study.

relative difficulty of extracting such information from social media data, particularly in non-English contexts (Wang et al., 2019).

4 Results

4.1 Domain Differences in Loanword Integration

The first hypothesis to test concerns the role of domain. As newspapers are generally considered more formal than social media (Biber and Conrad, 2019; Pavlick and Tetreault, 2016), we expect that loanwords and native verbs to be produced with the presumably more formal integrated forms.

H1: Writers in a more formal domain will tend to use the integrated form of loanwords at a higher rate than writers in a less formal domain.

To test this hypothesis, we collect data from a corpus of Spanish language newspapers from 21 different Spanish-speaking countries and regions.[12] We collect the 50 most frequent loanword pairs and native verb pairs from the social media data and compute their raw frequencies in the newspaper data. For each pair of integrated verb and light verb, we compute the rate of integrated verb use as the normalized frequency of the integrated verb. Formally, for a word base w, the set of all integrated verb forms $\mathcal{W}_{i,w}$, and the set of all light verb forms for the word $\mathcal{W}_{l,w}$, the rate of integrated verb use I_w is defined as:

$$I_w = \frac{\sum_{w_i \in \mathcal{W}_{i,w}} \text{count}(w_i)}{\sum_{w' \in W_{i,w} \cup W_{l,w}} \text{count}(w')}$$

We show the rates of integration across domains and locations in Figure 1. The first key finding is that the rate of integration is not significantly different for newspapers across locations, despite known dialect differences across regions. In addition, we see that for loanwords both social media and newspapers favor the integrated form over the light verb form, in correspondence with the expected "hierarchy" of loanword adaptation that places light verbs below integration (Wohlgemuth, 2009). With respect to **H1**, we see that newspaper writers consistently use the integrated form of loanwords and native verbs more frequently than the social media authors. Loanwords are integrated at a mean per-word rate of 91% in the newspapers as compared to 82% in social media, while native verbs have a rate of 93% in the newspapers and 82% in social media.[13] We show in Figure 1 that this difference holds across all regions.[14]

[12] News On the Web Spanish, approximately 7 billion tokens over 25 million documents, accessed May 2020: https://www.corpusdelespanol.org/now/.

[13] Both cases had a significant difference with $p < 0.01$ by Wilcoxon's signed-rank test.

[14] We find $p < 0.05$ across all location pairs except loanwords in US America and native verbs in Latin America, by Wilcoxon's test with Bonferroni correction.

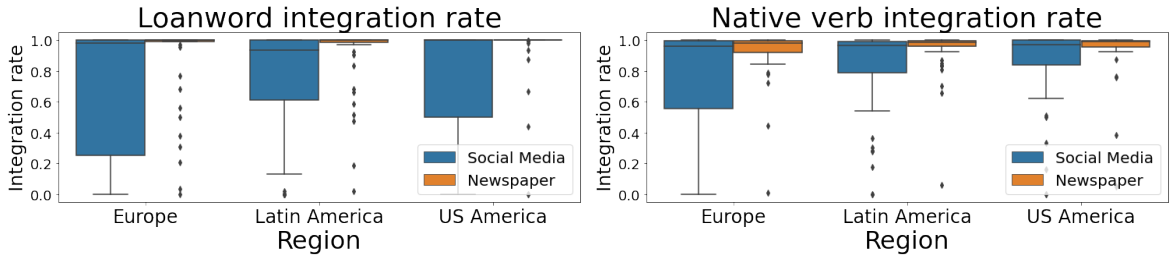

Figure 1: Integrated verb use across social media text (blue/left) and newspaper text (orange/right). Each unit is the ratio of integrated verb use for a single word type.

The consistent difference between social media and newspaper writing suggests that the domain of newspaper writing has more formal standards with respect to the use of both loanwords and native words (Geeraerts, 2003). Such consistency may reflect differences in how newspaper writers are expected to cover emerging phenomena such as new loanwords. A newspaper writer might be encouraged to use the formal version of a newer loanword to maximize the likelihood of their readers' understanding the word (Iwasaki, 1994; Llopis and Sánchez-Lafuente, 2009). To investigate this in more detail, we show the loanwords with the highest absolute difference in integration rate across social media and newspapers in Table 4. The loanwords that are integrated more often in newspapers seem to be relatively newer and possibly related more to online social media activity (e.g. *block*, *hype*), while the loanwords that are integrated more often on social media seem to be somewhat older and relevant to a wider range of activities (e.g. *host*, *rock*). This finding about domain reinforces the *social* meaning of loanword use, which informs the following speaker-level analysis.

Word	$I_{w,\text{social media}}$	$I_{w,\text{newspaper}}$	ΔI_w
zap	0.179	1.000	-0.821
block	0.153	0.857	-0.704
hype	0.393	0.995	-0.602
link	0.335	0.872	-0.536
like	0.115	0.649	-0.534
...	...	...	...
pitch	0.998	0.988	0.011
host	0.990	0.972	0.018
google	0.561	0.531	0.030
rock	0.787	0.648	0.139
DM	1.000	0.120	0.880

Table 4: Loanwords with biggest differences in integration between newspaper and social media.

4.2 Speaker-level factors in loanword integration

We now turn to speaker-level data to assess the relative impact of different social factors in the use of integrated loanwords. If integrated verbs are considered more formal than light verbs (§ 4.1), then we expect factors relevant to formality and speech standards to predict integrated verb use for both loanwords and native verbs:

H2: Speakers in social contexts that prefer formal language standards, and with backgrounds that support more standard language use, will tend to use integrated loanwords.

We use logistic regression to predict the use of an integrated verb (1/0) for a given loanword or native word, using different subsets of post-level and speaker-level features specified in § 3.4. We add fixed effects for all sufficiently frequent authors and word types.[15] To avoid overfitting the fixed effect variables, we choose an L2 weight for ridge regression, in order to maximize likelihood on held-out data.[16] For the default values of categorical variables in the regression, we specify "Unknown" for author location and "low Spanish" for prior language use. All scalar variables (post length, post activity, content sharing, link sharing, native integrated verb use) were log-transformed and Z-normalized before regression.

We show the social media regression results in Table 5. The following significant results emerge

[15] All authors and words with a count less than N=5 were assigned to a `RARE` category to avoid sparsity.

[16] Weight selected from grid search to maximize held-out likelihood on a 10% test split of the data, for each separate regression.

from the analysis.

4.2.1 Speaker-level Factors: Formality

First, we find the following trends with respect to formality.

Post context matters Speakers tend to use the integrated form more often for native verbs when using hashtags (β=0.099) and less often for both loanwords and native verbs when using @-mentions (β=-0.087 loanwords, β=-0.050 native verbs). Prior work demonstrated a similar effect with nonstandard English words on Twitter (Pavalanathan and Eisenstein, 2015) and found that hashtags and @-mentions correlated with larger and smaller audience expectations. Since formal language is often expected with a larger audience (Bell, 1984), Spanish speakers may naturally choose the integrated verb forms to adapt to a larger potential audience. For post length, we find that longer posts tend to have integrated verbs more often for loanwords (β=0.051) and less often for native verbs (β=-0.046). This effect may be related to post content (e.g. including direct objects for loanword verbs) but it may also reflect inherent differences in the perceptions of loanwords and native verbs.

4.2.2 Speaker-level factors: Background

For loanword and native verb integration, we find the following trends with respect to speaker background.

Information sharing affects integration differently We find that the frequent URL-sharing speakers are more likely to use the integrated form for loanwords (β=0.024), and less likely to use the integrated form for native verbs (β=-0.015). If we assume that people who share more URLs are more interested in sharing new information (Holton et al., 2014), then these people may also be more likely to use formal verb forms for newer words (loanwords) and informal forms for older words (native verbs), due to the speakers' increased awareness of how new information should be treated. For RT sharing, we find that authors who frequently retweet others are more likely to use the integrated form of native verbs (β=0.025), which suggests that authors with more social ties (higher network embeddedness; cf. Milroy and Milroy 1985) tend toward more standard language choices for frequently used words, i.e. native verbs.

Latin American authors prefer integration For both word groups, Latin American authors use integrated verbs at a higher rate (β=0.228 for loanwords, β=0.133 for native verbs). Prior studies in World Englishes have found that dialects in post-colonial countries such as India sometimes adopt more linguistically conservative features (Sharma, 2017), which may be reflected in the higher rate of verb integration in Latin America (cf. conservative pronunciation in Latin American Spanish; Guy 2014). In contrast, authors from Europe tend to use less verb integration (β=-0.367 for loanwords, β=-0.223 for native verbs), which suggests that using standard forms is less important for mainland Spain authors due to the dialect's relative prestige (Hernández-Campoy and Villena-Ponsoda, 2009).

More integration for monolinguals For loanwords, high-Spanish authors use integrated verbs at a higher rate than low-Spanish authors (β=0.589), and medium-Spanish authors use integrated verbs at a slightly higher rate (β=0.424). For native verbs, both high-Spanish and medium-Spanish authors use integrated verbs at a higher rate than low-Spanish authors (β=0.606 high-Spanish, β=0.687 medium-Spanish). Integrated verbs may be considered canonical and therefore more accessible for monolingual speakers, while light verbs could be more readily accessible to bilingual speakers who may default to simpler light verb constructions (González-Vilbazo and López, 2011). For example, the loanword phrase *dar un like* may sound more natural to a bilingual speaker who is uncertain of the acceptability of *likear*.

We note that for some of the variables such as post length and URL sharing, the effect direction for loanword integration is the opposite of the direction for native word integration. The use of loanwords may bear a different social meaning for speakers as compared to native words (e.g. speakers consider loanwords to be newer in their vocabulary, Levendis and Calude 2019), which results in different effects on integration for the same social variable. However, we leave more careful investigation of the differences between the word types for future work.

5 Discussion

We investigate the tendency for Spanish-speaking authors to use integrated verb forms for English

		Native words		Loanwords	
Variable type	Variable	β	S.E.	β	S.E.
	Intercept	2.572*	0.030	1.376*	0.234
Formality					
Post features	Has hashtag	0.099*	0.010	0.079	0.026
	Has mention	-0.050*	0.009	-0.087*	0.015
	Post length	**-0.046***	0.002	**0.051***	0.008
Background					
Author behavior	Post activity	0.006	0.003	-0.034	0.011
	URL sharing	**-0.015***	0.003	**0.024***	0.010
	RT sharing	0.025*	0.003	-0.010	0.009
Location	Latin America	0.133*	0.005	0.228*	0.016
	Europe	-0.223*	0.010	-0.367*	0.033
	US	0.008	0.015	-0.143	0.048
	Other	0.171*	0.025	-0.193	0.082
Language	High Spanish	0.606*	0.031	0.589*	0.110
	Medium Spanish	0.687*	0.030	0.424*	0.107
	Integrated verb use			-0.006	0.007
Sample size		235969		25436	
Likelihood ratio (vs. null)		2427*		3995*	

Table 5: Regression results for predicting integrated verb use for loanwords. * indicates $p < 0.01$, otherwise $p > 0.01$; Bonferroni correction applied for significance testing for individual coefficients. **Bold** indicates variables for which effects are significant across both conditions and point in opposite directions.

loanwords, with a corpus of social media data augmented with speaker-level information.

The study provides a data set of loanwords and native words that linguists can use to investigate specific contexts of usage (e.g. in quotations, Iwasaki 1994). The study also offers a pipeline for collecting various forms of loanwords using structured data (dictionaries) and data "in the wild." More broadly, our work demonstrates the utility of social media as a window into speaker-level and contextual factors that underlie multilingual phenomena such as loanwords.

Our analyses show that integrated verb use for loanwords is clearly connected to underlying expectations of formality and standardness in language use, which also apply to native verbs. The findings of this study provide additional context to prior work that showed some social correlates of loanword integration such as neighborhood composition (Poplack, 1988). The decision to use integrated verb forms appears to rely not just on the speakers' background (e.g. linguistic knowledge) but even utterance-level context (e.g. audience), suggesting that the process is not "inevitable" (Poplack and Dion, 2012). Furthermore, the differences in domain-level and speaker-level effects across word groups (and within word groups, e.g. Table 4) suggest different social perceptions, i.e. "marked" loanwords versus older, well-accepted native verbs. Such implicit social evaluations can help predict the long-term entrenchment of loanwords in a speech community (Chesley, 2010; Zenner et al., 2012), and shed light on processes of cross-cultural contact and attitudes (Lev-Ari and Peperkamp, 2014).

This study has several limitations that merit further research. First, the findings are narrowly focused on one form of integration, i.e. the alternation between different verb forms. Future work should consider other forms of loanword integration on social media, including in orthography (Eng. *football* → Sp. *fútbol*) and syntax (*el key* vs. *la key* "the key") (Montes-Alcalá and Shin, 2011; Vendelin and Peperkamp, 2006). It may be the case that some forms of loanword integration are more socially salient than others (Myers-Scotton, 1998) and therefore more strongly constrained by factors

such as audience expectations. In addition, this analysis found some location-level effects but did not zoom in to the level of the community, which is important since different speech communities may have different perceptions of the social value of loanwords (Aaron, 2015; Garley, 2014). As people of different linguistic backgrounds continue to interact on social media (Kim et al., 2014), it will be important to consider how different sub-communities on the platform adopt loanwords from one another, as such processes can lead to long-term language change. Lastly, different languages may have different expectations about the social meaning of integrated loanword use, e.g. integrated verbs in Japanese may seem less formal than their light verb equivalent (Tsujimura and Davis, 2011). More cross-linguistic work is needed to understand how well the social ramifications of loanword integration can be generalized (Haspelmath, 2009) and whether they reflect culture-specific norms rather than inherent trends about language and socialization.

Acknowledgments

This project was funded under NSF CAREER grant #1452443 to JE and a Data Curation Award from Georgia Institute of Technology's Institute for Data Engineering and Science (IDEaS) to DY. The authors thank Dr. Cecilia Montes-Alcalá and Dr. Lewis Chad Howe for their feedback on the validity of the loanword and native word pairs, as well as their feedback on early paper drafts. The authors also thank members of the Computational Linguistics lab and the SALT Lab at Georgia Institute of Technology for their feedback.

References

Jessi Elana Aaron. 2015. Lone English-origin nouns in Spanish: The precedence of community norms. *International Journal of Bilingualism*, 19(4):459–480.

Allan Bell. 1984. Language style as audience design. *Language in society*, 13(2):145–204.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Paul Boersma, Silke Hamann, et al. 2009. Loanword adaptation as first-language phonological perception. *Loanword phonology*, pages 11–58.

Louisa Buckingham. 2013. Light verb constructions in Latin American newspapers: Creative variants and coinages. *Spanish in Context*, 10(1):114–135.

Rena Torres Cacoullos and Jessi Elana Aaron. 2003. Bare English-origin nouns in Spanish: Rates, constraints, and discourse functions. *Language Variation and Change*, 15(3):289–328.

Paula Chesley. 2010. Lexical borrowings in French: Anglicisms as a separate phenomenon. *Journal of French Language Studies*, 20(3):231–251.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 76–86.

Steven Coats. 2018. Variation of New German Verbal Anglicisms in a Social Media Corpus. In *Proceedings of the 6th conference on CMC and social media corpora for the humanities*.

A. Seza Doğruöz and Preslav Nakov. 2014. Predicting dialect variation in immigrant contexts using light verb constructions. In *EMNLP*, pages 1391–1395.

Matt Garley. 2014. Seen and not heard: The relationship of orthography, morphology, and phonology in loanword adaptation in the german hip hop community. *Discourse, Context & Media*, 3:27–36.

Matt Garley and Julia Hockenmaier. 2012. Beefmoves: dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In *ACL*, pages 135–139.

Dirk Geeraerts. 2003. Cultural models of linguistic standardization. In René Dirven, Roslyn Frank, and Martin Pütz, editors, *Cognitive models in language and thought. Ideology, metaphors and meanings*, volume 2568.

Kay González-Vilbazo and Luis López. 2011. Some properties of light verbs in code-switching. *Lingua*, 121(5):832–850.

Gregory Guy. 2014. Variation and change in Latin American Spanish and Portuguese. In *Portuguese-Spanish interfaces: Diachrony, synchrony, and contact*, pages 443–464.

Lauren Hall-Lew, Elizabeth Coppock, and Rebecca L Starr. 2010. Indexing political persuasion: Variation in the Iraq vowels. *American Speech*, 85(1):91–102.

Martin Haspelmath. 2009. Lexical borrowing: Concepts and issues. In *Loanwords in the world's language: A Comparative Handbook*, pages 944–967.

Juan Manuel Hernández-Campoy and Juan Andrés Villena-Ponsoda. 2009. Standardness and nonstandardness in Spain: dialect attrition and revitalization of regional dialects of Spanish. *International Journal of the Sociology of Language*, 2009(196-197):181–214.

Avery E Holton, Kang Baek, Mark Coddington, and Carolyn Yaschur. 2014. Seeking and sharing: Motivations for linking on Twitter. *Communication Research Reports*, 31(1):33–40.

Yasufumi Iwasaki. 1994. Englishization of Japanese and acculturation of English to Japanese culture. *World Englishes*, 13(2):261–272.

Devin L. Jenkins. 2003. Bilingual Verb Constructions in Southwestern Spanish. *Bilingual Review*, pages 195–204.

Yoonjung Kang. 2011. Loanword phonology. *The Blackwell companion to phonology*, pages 1–25.

Ankit Kariryaa, Isaac Johnson, Johannes Schöning, and Brent Hecht. 2018. Defining and predicting the localness of volunteered geographic information using ground truth data. In *CHI*, pages 1–12.

Adam Kilgarriff. 2010. Google the verb. *Language Resources and Evaluation*, 44(3):281–290.

Suin Kim, Ingmar Weber, Li Wei, and Alice Oh. 2014. Sociolinguistic analysis of Twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 243–248.

Shiri Lev-Ari and Sharon Peperkamp. 2014. An experimental study of the role of social factors in language change: The case of loanword adaptations. *Laboratory Phonology*, 5(3):379–401.

Katharine Levendis and Andreea Calude. 2019. Perception and Flagging of Loanwords–A diachronic case-study of Māori loanwords in New Zealand English. *Ampersand*, 6:100056.

John Lipski. 1994. *Latin American Spanish*. Longman, New York.

María Ángeles Orts Llopis and Ángela Almela Sánchez-Lafuente. 2009. Translating the Spanish economic discourse of the crisis: Dealing with the inevitability of English loanwords. *International Journal of English Studies*, 9(3):133–158.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *ACL*, pages 25–30.

James Milroy and Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of linguistics*, 21(2):339–384.

Cecilia Montes-Alcalá and Naomi Lapidus Shin. 2011. Las keys versus el key: Feminine gender assignment in mixed-language texts. *Spanish in context*, 8(1):119–143.

Carol Myers-Scotton. 1998. A theoretical introduction to the markedness model. In *Codes and consequences: Choosing linguistic varieties*.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech*, 90(2):187–213.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Sharon Peperkamp. 2004. A psycholinguistic theory of loanword adaptations. In *Annual Meeting of the Berkeley Linguistics Society*, volume 30, pages 341–352.

Shana Poplack. 1988. Contrasting patterns of code-switching in two communities. *Codeswitching: Anthropological and sociolinguistic perspectives*, 48:215–244.

Shana Poplack and Nathalie Dion. 2012. Myths and facts about loanword development. *Language Variation and Change*, 24(3):279–315.

Shana Poplack, Suzanne Robillard, Nathalie Dion, and John C. Paolillo. 2020. Revisiting phonetic integration in bilingual borrowing. *Language*, 96(1):126–159.

C Rodney and C Jubilado. 2012. Morphological Study of Verb of Anglicisms in Spanish Computer Language. *Polyglossia*, 23:43–47.

Tara Sanchez. 2005. The (socio-)linguistics of morphological borrowing: A quantitative look at qualitative constraints and universals. *University of Pennsylvania Working Papers in Linguistics*, 11(2):12.

Devyani Sharma. 2017. English in India. In *Varieties of English*, pages 311–329.

Natsuko Tsujimura and Stuart Davis. 2011. A construction approach to innovative verbs in Japanese. *Cognitive Linguistics*, 22(4):799–825.

Inga Vendelin and Sharon Peperkamp. 2006. The influence of orthography on loanword adaptations. *Lingua*, 116(7):996–1007.

Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The Web Conference*, pages 2056–2067.

Søren Wichmann and Jan Wohlgemuth. 2008. Loan verbs in a typological perspective. *Empirical approaches to language typology*, 35:89.

Jan Wohlgemuth. 2009. *A typology of verbal borrowings*, volume 211. Walter de Gruyter.

Eline Zenner, Dirk Speelman, and Dirk Geeraerts. 2012. Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics*, 23(4):749–792.

A Appendix

A.1 All integrated and light verb pairs

To assist study replication, we list all pairs of integrated and light verbs for loanwords and native verbs used in this study. We list them in alphabetical order (by integrated verb) in the format:
loanword/translation: integrated verb ; light verb phrase(s)

Loanwords

- *access*: accesar ; hacer/tener acces
- *aim*: aimear ; hacer/tener aim
- *alert*: alertear ; hacer alert
- *audit*: auditar ; hacer (un) audit
- *ban*: banear ; hacer un ban
- *bang*: bangear ; hacer bang
- *bash*: bashear ; hacer/dar bash
- *block*: blockear ; hacer/dar (un) block
- *boycott*: boicotear ; hacer (un) boicot
- *box*: boxear ; hacer (el) box/boxing
- *bully*: bulear ; hacer/ser (el) bully
- *bust*: bustear ; hacer (el) bust
- *cast*: castear ; hacer cast/casting
- *change*: changear ; hacer change
- *chat*: chatear ; hacer chat
- *check*: chequear ; hacer un cheque
- *shoot*: chutar ; hacer/tomar el shot
- *combat*: combatear ; hacer (el) combat
- *connect*: conectar ; hacer (un) conexión
- *crack*: crackear ; hacer crack
- *customize*: customizar ; hacer custom/customized
- *default*: defaultear ; hacer default
- *delete*: deletear ; hacer/poner delete
- *DM*: dmear ; mandar/enviar/poner un dm
- *dope*: dopar ; hacer doping
- *downvote*: downvotear ; poner/dar (un) downvote
- *draft*: draftear ; hacer/tener draft
- *drain*: drenar ; hacer (el) dren
- *smash*: esmachar ; hacer smash
- *sniff*: esnifar ; hacer sniff
- *standard*: estándar ; hacer (un) standard
- *exit*: exitear ; hacer exit
- *export*: exportear ; hacer export
- *externalize*: externalizar ; hacer external
- *fangirl*: fangirlear ; hacer/ser fangirl
- *film*: filmar ; tomar (un) film
- *flash*: flashear ; hacer (un) flash
- *flex*: flexear ; hacer (un) flex
- *flirt*: flirtear ; hacer flirt
- *focus*: focalizar ; hacer focus
- *format*: formatear ; hacer/dar (el) formato
- *form*: formear ; hacer form
- *freak*: friquear ; estar freaked
- *freeze*: frizar ; hacer freeze
- *fund*: fundear ; dar/hacer fund/funding
- *gentrify*: gentrificar ; hacer/tener gentrificación
- *ghost*: gostear ; hacer gost/ghost
- *google*: googlear ; buscar en google
- *hack*: hackear ; hacer hack
- *hail*: hailear ; hacer hail
- *hang*: hanguear ; hacer hang
- *harm*: harmear ; hacer harm
- *hypnosis*: hipnotizar ; hacer hipnosis
- *host*: hostear ; hacer host
- *hype*: hypear ; hacer hype
- *intercept*: interceptear ; hacer/tirar interception
- *hang*: janguear ; hacer hang (out)
- *lag*: lagear ; hacer (un) lag
- *like*: likear ; dar/poner (un) like
- *limit*: limitear ; hacer (un) limit
- *lynch*: linchar ; hacer lynch
- *link*: linkear ; dar/poner (un) link
- *love*: lovear ; hacer love
- *look*: luquear ; dar/hacer (un) look
- *make*: makear ; hacer make
- *melt*: meltear ; hacer melt
- *mope*: mopear ; hacer mope
- *nag*: nagear ; hacer nag
- *knock*: noquear ; dar/hacer (un) knockout
- *pack*: packear ; hacer pack
- *pan*: panear ; hacer/dar (un) panorama
- *panic*: paniquear ; tener panic
- *park*: parquear ; hacer parking
- *perform*: performar ; hacer (un) performance
- *pitch*: pichear ; hacer (un) pitch
- *pin*: pinear ; hacer pin
- *PM*: pmear ; enviar/mandar (un) pm
- *punch*: ponchar ; hacer un punch
- *post*: postear ; dar/poner (un) post
- *posterize*: posterizar ; hacer poster
- *print*: printear ; hacer print
- *protest*: protestear ; hacer (un) protest
- *push*: puchar ; hacer un push
- *pump*: pumpear ; hacer pump(s)
- *quote*: quotear ; hacer quote
- *rank*: rankear ; hacer rank
- *rant*: rantear ; hacer (un) rant
- *rape*: rapear ; hacer (un) rape
- *record*: recordear ; hacer (un) recording
- *render*: renderizar ; hacer render(ed)
- *rent*: rentear ; hacer rental/renting
- *report*: reportear ; hacer (un) report
- *reset*: resetear ; hacer reset
- *respect*: respectear ; hacer respect
- *ring*: ringear ; hacer ring
- *rock*: rockear ; hacer rock
- *roll*: rollear ; hacer roll
- *sample*: samplear ; hacer (un) sample
- *selfie*: selfiar ; tomar (un) selfie
- *sext*: sextear ; dar/mandar un sext
- *ship*: shippear ; hacer ship
- *shitpost*: shitpostear ; hacer/poner un shitpost
- *shock*: shockear ; hacer shock
- *sign-in*: signear ; hacer sign-in
- *stalk*: stalkear ; actuar como un stalker
- *strike*: strikear ; hacer/dar un strike
- *surf*: surfear ; hacer surf
- *tackle*: taclear ; hacer tackle
- *text*: textear ; mandar/enviar un text
- *tick*: ticar ; hacer (un) tick
- *torment*: tormentear ; hacer torment
- *touch*: touchear ; hacer (un) touch
- *transport*: transportear ; hacer transport
- *travel*: travelear ; hacer travel
- *troll*: trolear ; actuar como un trol
- *tweet*: tweetear ; poner/enviar/hacer (un) tweet
- *twerk*: twerkear ; hacer twerk
- *upvote*: upvotear ; dar (un) upvote
- *vape*: vapear ; hacer/tomar vape/vaping
- *zap*: zapear ; hacer zap/zapping

Native verbs

- *admire*: admirar ; tener admiración
- *befriend*: amistar ; tener amistad
- *encourage*: animar ; subir el ánimo
- *note*: anotar ; tomar nota
- *land*: aterrizar ; hacer un aterrizaje
- *joke*: bromear ; hacer bromas
- *mock*: burlarse ; hacer burla
- *punish*: castigar ; poner un castigo
- *buy*: comprar ; hacer la compra
- *copy*: copiar ; hacer una copia
- *tickle*: cosquillar ; hacer cosquillas
- *blame*: culpar ; echar la culpa
- *damage*: dañar ; hacer daño
- *decide*: decidir ; tomar decisiones
- *apologize*: disculparse ; pedir disculpas
- *shower*: ducharse ; darse una ducha
- *question*: dudar ; poner en duda
- *exemplify*: ejemplificar ; poner un ejemplo
- *estimate*: estimar ; tener estima
- *explain*: explicar ; dar una explicación
- *finish*: finalizar ; poner fin
- *photograph*: fotografiar ; tomar fotos
- *escape*: fugarse ; darse a la fuga
- *mention*: mencionar ; hacer mención
- *look at*: mirar ; echar una mirada
- *penalize*: multar ; poner una multa
- *negotiate*: negociar ; hacer negocios
- *originate*: originar ; dar origen
- *participate*: participar ; tomar parte
- *walk*: pasear ; dar un paseo
- *step*: pisar ; poner el pie
- *value*: preciar ; poner precio
- *ask*: preguntar ; hacer (una) pregunta
- *anticipate*: prever ; tener previsto
- *test*: probar ; poner a prueba
- *recommend*: recomendar ; hacer recomendación
- *write*: redactar ; hacer una redacción
- *cure*: remediar ; poner remedio
- *breathe*: respirar ; dar un respiro
- *jump*: saltar ; dar un salto
- *nap*: sestear ; echar una siesta
- *dream*: soñar ; tener un sueño
- *end*: terminar ; poner término
- *use*: usar ; hacer uso
- *travel*: viajar ; hacer un viaje
- *see*: vistar ; echar un vistazo
- *fly*: volar ; tomar un vuelo