Tuiteamos o pongamos un tuit?

Comparing social factors in loanword integration on social media

Ian Stewart
Georgia Institute of Technology
22 Sept 2020
Loanwords in linguistics

Speakers of different languages often borrow words and phrases as a result of cultural contact (Haspelmath 2009), which spurs language change (Poplack and Dion 2012).

A speaker’s use of loanwords can reveal their phonological expectations (Kang 2011), L2 ability (Poplack 1988), and cultural expectations (Zenner et al. 2012).
Loanwords in linguistics

Speakers of different languages often borrow words and phrases as a result of cultural contact (Haspelmath 2009), which spurs language change (Poplack and Dion 2012).

A speaker’s use of loanwords can reveal their phonological expectations (Kang 2011), L2 ability (Poplack 1988), and cultural expectations (Zenner et al. 2012).
Loanwords vs. code-switching

Loanwords are distinct from code-switching, which is the alternation between multiple varieties in the same utterance (Haspelmath 2009).

A loanword acts as an independent unit that follows the matrix language’s grammar.

 yo veo un tweet ≠ yo veo a tweet
Speakers do not always adopt loanwords “as-is” but often *integrate* them to align to their native language grammar (Kang 2011).

- phonology (*football* => *fútbol*)
- morphology (*tweet* => *tweetear*)
- syntax (*cool* => *los sombreros cools*)

While loanwords’ phonology is integrated gradually (Poplack et al. 1988), their morphology is integrated “instantly” (Poplack et al. 2012).
Loanword integration

Speakers do not always adopt loanwords “as-is” but often integrate them to align to their native language grammar (Kang 2011).

- phonology (football => fútbol)
- morphology (tweet => tweetear)
- syntax (cool => los sombreros cools)

While loanwords’ phonology is integrated gradually (Poplack et al. 1988), their morphology is integrated “instantly” (Poplack et al. 2012).
Loanword integration: dynamic?

Poplack et al. (1988)

Older loanwords may be more **well-known** globally and therefore more likely to be integrated.
Loanword integration: dynamic?

Older loanwords may be more **well-known** globally and therefore more likely to be integrated.
Speaker-level social factors may also play a role in loanword integration! E.g. **social acceptance** and **language ability**.
Loanword integration is often viewed as a natural process for native speakers (Poplack and Dion 2012), may represent more formal register.
Loanword integration is often viewed as a natural process for native speakers (Poplack and Dion 2012), may represent more formal register.

1. Does loanword integration correlate with formal or informal writing domain?
Speaker factors?

Speakers who feel more aligned to their native language’s culture (Lev-Ari and Peperkamp 2014) may also adhere to more formal language standards.

If integration is considered more formal, then more pro-native culture speakers may use integrated verbs more often.
Speaker factors: social attitudes

Positive or negative evaluation that a speaker expresses regarding a social group to which they belong or don't belong (Olson and Zanna 1993).

Multilingual speakers may choose between languages based on attitudes connected to culture (Auer 2013).
Speaker factors: social attitudes

Positive or negative evaluation that a speaker expresses regarding a social group to which they belong or don't belong (Olson and Zanna 1993).

Multilingual speakers may choose between languages based on attitudes connected to culture (Auer 2013).
Speaker factors?

Speakers who feel more aligned to their native language’s culture (Lev-Ari and Peperkamp 2014) may also adhere to more formal language standards.

If integration is considered more formal, then more pro-native culture speakers may use integrated verbs more often.

2. What (a) demographic and (b) attitude-related speaker attributes explain integration in loanwords and native words?
A multilingual speaker’s choice of language can reflect their attitudes about other social groups (Auer 2013); not easily quantified in spoken studies.
Study goals

RQ1: Does loanword integration correlate with formal or informal writing domain?

RQ2: Which (a) demographic and (b) attitude-related speaker attributes explain integration in loanwords and native words?
In spoken language (e.g. interviews), loanwords are often rare (Poplack and Dion 2012), difficult to study from a quantitative social perspective.

Social media provides (1) a large volume of data to sample loanword use and (2) rich background for speakers, e.g. sociolinguistic background.
Data: identifying words

We focus on alternation between integrated loanword verbs and light verb constructions (tweetear vs. poner un tweet).

We need a control condition to determine whether social factors affect just loanwords or verb integration in general.

Data goal: identify loanword verb pairs and native verb verb pairs.
Data: identifying words

We focus on alternation between integrated loanword verbs and light verb constructions (tweetear vs. poner un tweet).

We need a control condition to determine whether social factors affect just loanwords or verb integration in general.

Data goal: identify loanword verb pairs and native verb verb pairs.
Data: identifying words

\{\text{ENGLISH\_WORD}\} + (e)ar
Data: identifying words

{ENGLISH_WORD} + (e)ar

<table>
<thead>
<tr>
<th>Loanword</th>
<th>Verbs</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Like</td>
<td>likear, dar un like</td>
<td>13,154</td>
</tr>
<tr>
<td>Connect</td>
<td>conectear, hacer un conexión</td>
<td>7857</td>
</tr>
<tr>
<td>Flip</td>
<td>flipear, hacer flip</td>
<td>6904</td>
</tr>
<tr>
<td>Stalk</td>
<td>stalkear, ser un stalker</td>
<td>5508</td>
</tr>
<tr>
<td>Tweet</td>
<td>tweetear, poner un tweet</td>
<td>5294</td>
</tr>
</tbody>
</table>

N=124
loanword verb pairs

<table>
<thead>
<tr>
<th>Native word</th>
<th>Verbs</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dream</td>
<td>soñar, tener un sueño</td>
<td>39,392</td>
</tr>
<tr>
<td>Buy</td>
<td>comprar, hacer la compra</td>
<td>36,337</td>
</tr>
<tr>
<td>End</td>
<td>terminar, poner término</td>
<td>34,234</td>
</tr>
<tr>
<td>Use</td>
<td>usar, hacer uso</td>
<td>30,834</td>
</tr>
<tr>
<td>Test</td>
<td>probar, poner a prueba</td>
<td>29,930</td>
</tr>
</tbody>
</table>

N=49
native verb pairs
Data: identifying words

NLP helps!

\{ENGLISH\_WORD\} + (e)ar

<table>
<thead>
<tr>
<th>Loanword</th>
<th>Verbs</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Like</td>
<td>likear, dar un like</td>
<td>13,154</td>
</tr>
<tr>
<td>Connect</td>
<td>conectear, hacer un conexión</td>
<td>7857</td>
</tr>
<tr>
<td>Flip</td>
<td>flipear, hacer flip</td>
<td>6904</td>
</tr>
<tr>
<td>Stalk</td>
<td>stalkear, ser un stalker</td>
<td>5508</td>
</tr>
<tr>
<td>Tweet</td>
<td>tweetear, poner un tweet</td>
<td>5294</td>
</tr>
</tbody>
</table>

N=124
loanword verb pairs

<table>
<thead>
<tr>
<th>Native word</th>
<th>Verbs</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dream</td>
<td>soñar, tener un sueño</td>
<td>39,392</td>
</tr>
<tr>
<td>Buy</td>
<td>comprar, hacer la compra</td>
<td>36,337</td>
</tr>
<tr>
<td>End</td>
<td>terminar, poner término</td>
<td>34,234</td>
</tr>
<tr>
<td>Use</td>
<td>usar, hacer uso</td>
<td>30,834</td>
</tr>
<tr>
<td>Test</td>
<td>probar, poner a prueba</td>
<td>29,930</td>
</tr>
</tbody>
</table>

N=49
native verb pairs
Data: identifying domains, speakers

- **Domain**
  - Corpus del Español: NOW
  - Newspapers from 21 countries 2012-2019

- **Speakers**
  - 1% Twitter sample
  - July 2017 - June 2019
  - langid
  - lang=ES
  - 87K tweets w/ 1+ loanword
  - 80K authors
Data: identifying domains, speakers

Domain

Corpus del Español: NOW

newspapers from 21 countries 2012-2019

Speakers

1% Twitter sample July 2017 - June 2019

langid
lang=ES

87K tweets w/ 1+ loanword

80K authors

NLP helps!
RQ1: integration across domains?

Test: compare rate of integration for loanwords, native verbs between newspapers and social media.

\[
\text{integration}(w) = \frac{\# \text{ integrated } w \text{ tokens}}{\# \text{ integrated } w \text{ tokens} + \# \text{ non-integrated } w \text{ tokens}}
\]
RQ1: domain differences

Newspapers use integrated verbs at higher rate than social media, i.e. integration relates to formality.
Newspapers use integrated verbs at higher rate than social media, i.e. integration relates to formality.
Newspapers tend to integrate technology-related loanwords much more often than social media. Newspaper writers may assume that some loanwords need to be treated more formally to justify choosing loanword over native equivalent. E.g. likear vs. gustar.

<table>
<thead>
<tr>
<th>Word</th>
<th>$I_{w,social\ media}$</th>
<th>$I_{w,newspaper}$</th>
<th>$\Delta I_w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>block</td>
<td>0.105</td>
<td>0.857</td>
<td>-0.752</td>
</tr>
<tr>
<td>hype</td>
<td>0.267</td>
<td>0.995</td>
<td>-0.728</td>
</tr>
<tr>
<td>link</td>
<td>0.227</td>
<td>0.872</td>
<td>-0.645</td>
</tr>
<tr>
<td>like</td>
<td>0.051</td>
<td>0.649</td>
<td>-0.598</td>
</tr>
<tr>
<td>perform</td>
<td>0.065</td>
<td>0.561</td>
<td>-0.496</td>
</tr>
<tr>
<td>access</td>
<td>0.523</td>
<td>1.000</td>
<td>-0.477</td>
</tr>
<tr>
<td>tweet</td>
<td>0.129</td>
<td>0.598</td>
<td>-0.470</td>
</tr>
<tr>
<td>boycott</td>
<td>0.593</td>
<td>0.989</td>
<td>-0.396</td>
</tr>
<tr>
<td>fangirl</td>
<td>0.632</td>
<td>1.000</td>
<td>-0.368</td>
</tr>
<tr>
<td>post</td>
<td>0.676</td>
<td>0.999</td>
<td>-0.323</td>
</tr>
</tbody>
</table>
RQ2: methods for attitudes

We know that integration relates to formality. Does this imply that integration relates to more formal speaker attributes (e.g. attitude)?

Attitude: support of the culture associated with one’s native language over another language’s culture.
RQ2: methods for attitudes

We know that integration relates to formality. Does this imply that integration relates to more formal speaker attributes (e.g. attitude)?

Attitude: support of the **culture** associated with one’s native language over another language’s culture.
Music ~ attitude

We approximate attitude with music sharing:

- No linguistic, financial barrier
- Popular across demographics
- Strong connection with sociolinguistic attitudes (Low et al. 2009; Dubois and Horvath 1999; Hernandez 2010)
Music ~ attitude

We approximate attitude with **music sharing**:

- No linguistic, financial barrier
- Popular across demographics
- **Strong connection with sociolinguistic attitudes** (Low et al. 2009; Dubois and Horvath 1999; Hernandez 2010)
Music ~ attitude

We approximate attitude with **music sharing**:

- No linguistic, financial barrier
- Popular across demographics
- Strong connection with sociolinguistic attitudes (Low et al. 2009; Dubois and Horvath 1999; Hernandez 2010)

\[
\text{SLA media sharing} = \frac{\# \text{ Spanish/Latin links}}{\# \text{ Spanish/Latin links + \# US/UK links}}
\]
Music ~ attitude

We approximate attitude with music sharing:

- No linguistic, financial barrier
- Popular across demographics
- Strong connection with sociolinguistic attitudes (Low et al. 2009; Dubois and Horvath 1999; Hernandez 2010)

SLA media sharing = (# Spanish/Latin links) / (# Spanish/Latin links + # US/UK links)

Time-consuming!
Attitude: construct validity

How do we know that people who share Spanish/Latin American media have consistent pro-Spanish culture attitude?

High SLA media authors use more standard Spanish words in their profile descriptions (compared to low-media authors).
Attitude: construct validity

How do we know that people who share Spanish/Latin American media have consistent pro-Spanish culture attitude?

High SLA media authors use more standard Spanish words in their profile descriptions (compared to low-media authors).
A person may share a media link because they are a fan of the musician (Katz 2017), not necessarily because they identify with culture.

We approximate each musician’s expected fan ages from Facebook advertising estimates.
A person may share a media link because they are a fan of the musician (Katz 2017), not necessarily because they identify with culture.

We approximate each musician’s expected fan ages from Facebook advertising estimates.
A person may share a media link because they are a fan of the musician (Katz 2017), not necessarily because they identify with culture.

We can address this partly by controlling for age: younger people will generally share music from younger musicians.

We compute age distributions from Facebook advertising audience estimates.
We collect global age counts for all musicians in data, using four age bins: 15-25, 25-35, 35-45, 45-65+.

Age distributions reliably identify musicians with known younger fan bases (e.g. Lil Yachty) and older fan bases (Don McLean).
We balance the age distributions by (greedy) pairing each US American link with a Latin American link with the most similar age distribution (lowest distance).
RQ2: evaluation

Logistic regression
(verb = integrated) ~ speaker attributes + fixed effects
RQ2: evaluation

Logistic regression
(verb = integrated) ~ speaker attributes + fixed effects
**RQ2: demographic results**

**High Spanish** authors use more integrated loanwords ($\beta=1.99$, $p < 0.001$), following hypothesis about conservative linguistic behavior (Poplack et al. 1988).

**Latin American** authors use more integration ($\beta=0.177$, $p < 0.001$; $\beta=0.124$, $p < 0.001$).

- **Possible reason**: LA dialects are more conservative, less idiosyncratic than expected (Buckingham 2011).
RQ2: speaker results

For media-sharing authors, more prior use of integrated native verbs leads to **less integration for loanword verbs** ($\beta=-0.168$, $p < 0.01$).

- Integrated loanword verbs ("likear") are often used as **default** in online platforms, may sound less formal than light verbs ("dar un like").
RQ2: attitude results

Media does not explain loanword integration \( (\beta=0.108, p > 0.05) \) but does explain native verb integration \( (\beta=0.126, p < 0.01) \).

High SLA media = 90+% Spanish/Latin American links

Medium SLA media = 10-50% Spanish/Latin American links
RQ2: attitude results

Media does not explain loanword integration \( (\beta=0.108, \ p > 0.05) \)
but does explain **native verb integration** \( (\beta=0.126, \ p < 0.01) \).

- Same effect holds even after **filtering** for only high-Spanish authors.

- Native verbs are more well-known among speakers and therefore may have **stronger connection** to language norms and latent social attitudes.
No idea when it started but I imagine that if you Google it, it will pop out at you.

Cómo me enferma que se tomen todo a pecho NO SABES QUE ES EL SARCASMO BUENO BUSCA EN GOOGLE Y DEJATE DE JODER gracias.
(I mean, I work with mass every day, if someone doesn’t see the price of a product, they won’t buy it because they would rather not make the effort to ask questions)

Go on WhatsApp, damn. I asked you a question.

Quiero decir trabajo con masa todos los días si alguien no ve el precio de un producto, pasa de llevarselo porque pasa de esforzarse en preguntar.

Entra a WhatsApp, caraja. Te hice una pregunta.
RQ1: Loanword integration, like native verb integration, is likely considered more formal.

RQ2: Author demographics (Spanish use, location) are surprisingly important for explaining integrated verbs.

RQ2: Authors who share more Spanish/Latin American music use more native integrated verbs: stronger connection between attitude and formality for (well-attested?) native verbs.
Results: cultural attitudes

Cultural attitudes may not explain integration in **loanwords** because:

1. Music sharing is **less clear signal** of social attitude than e.g. political content sharing: people may share music for enjoyment (context collapse; Marwick and boyd 2011).

2. Loanword integration is not strongly marked as a multilingual choice, as compared to e.g. code-switching which is generally **marked** as a social signal (Myers 1988).
Results: cultural attitudes

Cultural attitudes may not explain integration in loanwords because:

1. Music sharing is less clear signal of social attitude than e.g. political content sharing: people may share music for enjoyment (context collapse; Marwick and boyd 2011).

2. Loanword integration is not strongly marked as a multilingual choice, as compared to e.g. code-switching which is generally marked as a social signal (Myers 1988).

Less marked

Integrated vs. light verb

Loanword spelling

Code-switching

Metalinguistic commentary
(“English is awesome”)
Conclusion

This study provides **quantitative evidence** to explain how loanwords become integrated.

- Social factors and especially formality play a role; morphological integration is not just “instant” (Poplack and Dion 2012).

The study proposes a **new method** for understanding possible motivations for multilingual choices, i.e. music sharing.
Future work: types of integration

This study focuses on just one form of loanword integration, i.e. morphological integration of verbs.

Alternation between integrated and light verbs may fall on spectrum of language choices.

Future work should use variety of data available online to compare different forms of integration.
Future work: types of integration

This study focuses on just one form of loanword integration, i.e. morphological integration of verbs.

Alternation between integrated and light verbs may fall on spectrum of language choices.

Future work should use variety of data available online to compare different forms of integration.
Future work: cultural attitudes

This study focuses on music as a highly available, flexible expression of cultural attitudes (Dubois et al. 2018).

Potential problems: context collapse (e.g. sharing music for fun), low coverage, possible population bias.

Multilingualism intersects with many aspects of cultural identity (Auer 2013): sports, film/TV, politics, food…

Need more generic/scalable metrics for cultural attitude to understand sociolinguistic variation.
Questions?

istewart6@gatech.edu