# Lázaro: An automatic observatory of anglicism usage in the Spanish press

Elena Álvarez Mellado
NLP&IR group at UNED (Madrid)

# A little bit of context

- What I'm presenting here started as my MS thesis at Brandeis University in Massachusetts, is now part of my PhD thesis at UNED University in Madrid.

- This is work in progress

# Aim of this project

Creating a computational model that can detect and track new(ish) anglicisms in Spanish newspapers.

# What is an anglicism?

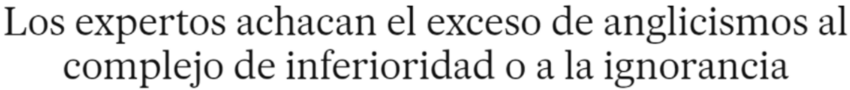
# What is an anglicism?

An anglicism is a word that comes from English and is used in another language

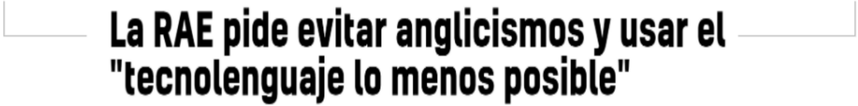
*podcast, app, online, crowdfunding, spin-off, big data, fake news…*

- English is a prolific source of new words in many languages.
  - Particularly in the press
- Our aim is to monitor recent anglicisms that are being incorporated in the Spanish press
  - Our approach is purely descriptivist
  - Understand borrowing as a linguistic phenomenon

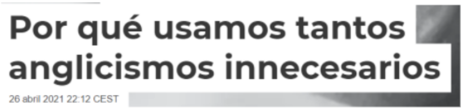
# Los anglicismos: ¿una amenaza para la lengua española?



Los expertos achacan el exceso de anglicismos al complejo de inferioridad o a la ignorancia

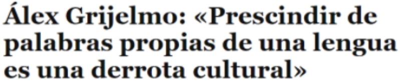


La RAE pide evitar anglicismos y usar el "tecnolenguaje lo menos posible"



Por qué usamos tantos anglicismos innecesarios

26 abril 2021 22:12 CEST



Álex Grijelmo: «Prescindir de palabras propias de una lengua es una derrota cultural»

# Why is borrowing interesting in Linguistics

- Borrowing is a manifestation of how languages change

# Why is borrowing interesting in Linguistics

- Borrowing is a manifestation of how languages change

- New realities → new words (relevant for lexicography)
  *online, software, streaming...*

# Why is borrowing interesting in Linguistics

- Borrowing is a manifestation of how languages change

- New realities → new words (relevant for lexicography)

  *online, software, streaming...*

- Old realities → new words (relevant for sociolinguistics)

  *barato → low-cost* *olio (from lat. 'oleum') → azeyte ('aceite')*

# Why is borrowing interesting in Linguistics

- Borrowing is a manifestation of how languages change

- New realities → new words (relevant for lexicography)

  *online, software, streaming...*

- Old realities → new words (relevant for sociolinguistics)

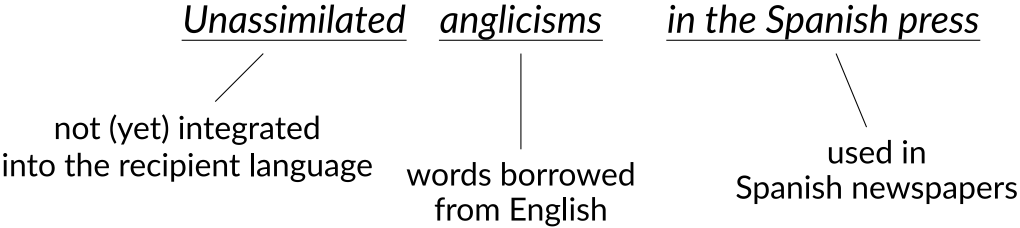
  *barato → low-cost*    *olio (from lat. 'oleum') → azeyte ('aceite')*

- Linguistic adaptation:

  *football → fútbol*
  *spaghetti → espaguetis*

# The task

We want to build a system that can automatically detect unassimilated anglicisms in the Spanish press



*Unassimilated*

not (yet) integrated
into the recipient language

*anglicisms*

words borrowed
from English

*in the Spanish press*

used in
Spanish newspapers

Ex: *Las prendas bestsellers se estampan con motivos florales, 'animal print' o a retales tipo patchwork*

Best-seller clothes are printed with flowers, animal print or patchwork style

# How could we build it?

# How could we build it?

- Dictionary lookup

  Is this word in a SPA/EN dictionary/corpus?

# How could we build it?

- Dictionary lookup

  Is this word in a SPA/EN dictionary/corpus?

- Pattern matching

  *wh-, -sh-, -ing*

# How could we build it?

- Dictionary lookup

  Is this word in a SPA/EN dictionary/corpus?

- Pattern matching

  *wh-*, *-sh-*, *-ing*

- Word probability

  What is the probability of this word being EN/SPA?

# Limitations of these approaches

- *Prime time* is a borrowing:
    - *prime* is form of the verb *primar*
    - *time* is form of the verb *timar*

# Limitations of these approaches

- *Prime time* is a borrowing:
  - *prime* is form of the verb *primar*
  - *time* is form of the verb *timar*

- *Social media* is a borrowing:
  - But both *social* and *media* are also Spanish words

# Limitations of these approaches

- *Prime time* is a borrowing:
  - *prime* is form of the verb *primar*
  - *time* is form of the verb *timar*

- *Social media* is a borrowing:
  - But both *social* and *media* are also Spanish words

- Not every English word is necessarily a borrowing.
  - *Sgt.Peppers Lonely Hearts Club Band*
  - *Eternal sunshine of the spotless mind*
  - Quotations, metalinguistic usage, etc

# Context is key

Me gusta mucho Johnny Cash `[not an anglicism]`

I really like Johnny Cash

Estoy sin **cash** `[anglicism]`

I have no money

There is not a single rule we can use to determine whether a given word in isolation is an anglicism, because it depends on the context

When rules can’t take you there,
machine learning might take you close enough

When rules can’t take you there,
machine learning might take you close enough

Set of rules → data-driven approach

# The data driven approach (a.k.a machine learning)

1. We gather a corpus that is rich in anglicisms
2. We manually annotate the corpus
3. We give the annotated corpus to a ML model
4. Hopefully, the model will find statistical correlations in the annotated data and will “learn” to recognize anglicisms

# Machine learning model for anglicism detection

We frame the task as a sequence labeling problem:

- The model takes a sentence as input
- Each word receives a tag (anglicism/not anglicism)
- Each tag is dependent on the nearby tags and assigned to maximize global probability of the sequence ($\rightarrow$ context)
- Same approach as in other NLP tasks: Named Entity Recognition and Part of Speech tagging

# BIO encoding

```
Este  O
mes  O
os  O
sugerimos  O
probar  O
el  O
batch  B-ENG
cooking  I-ENG
```

[a]

[a] *This month we recommend you try* ***batch cooking***

```
La  O
era  O
de  O
las  O
apps  B-ENG
para  O
ligar  O
```

[a]

[a] *The era of* ***apps*** *for dating*

```
La  O
era  O
de  O
las  O
dating  B-ENG
apps  I-ENG
```

[a]

[a] *The era of* ***dating apps***

# 1st model: CRF

- A Conditional Random Fields model (“classic” ML)

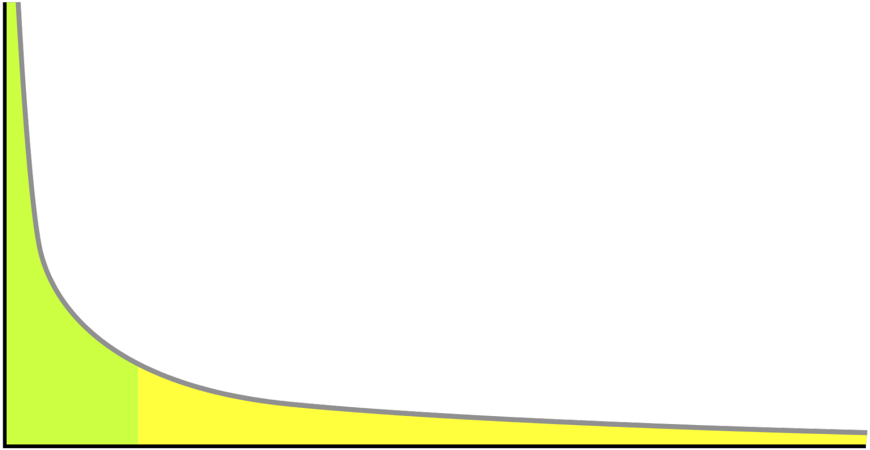
- Each word is represented by a set of handcrafted features
  - Token, shape, punctuation, titlecase, char trigram, quotation, etc

- Trained and tested on an corpus of Spanish press annotated with anglicisms (325,000 tokens)

- Results on the test set F1=87 (100 would mean perfect)

- Developed in 2020, in production 2020-2022

# Long tail distribution



# 2nd model: BiLSTM-CRF

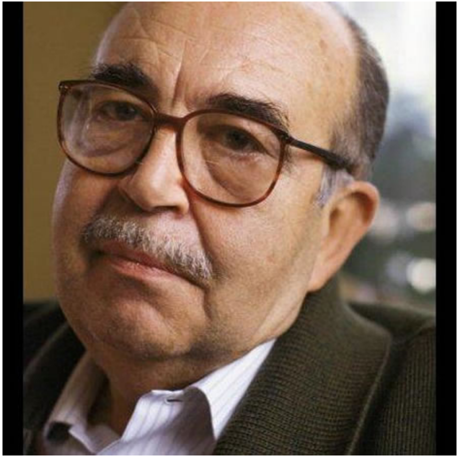
- Deep learning model: BiLSTM-CRF
- Fed with Transformer-based word embeddings pretrained on codeswitched data, along with subword embeddings
- Train and tested on COALAS (COrpus of AngLicisms in the SpAnish Press) (370,000 tokens, more diverse)
- Results on the test set F1 = 85 (CRF model F1 = 55)
- Developed in 2021, in production 2022-

# Building Lázaro: an observatory of anglicism usage in the Spanish press

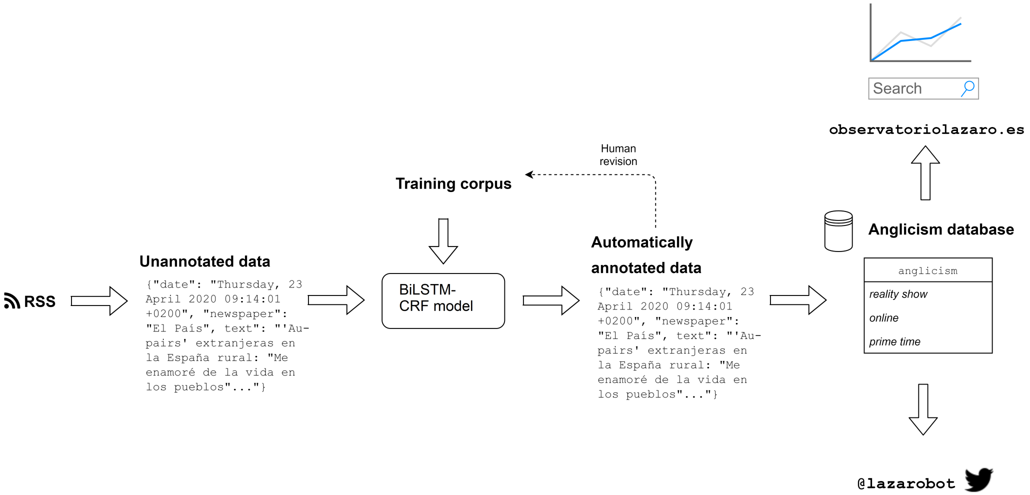
# Why *Lázaro*?

- A tribute to Spanish linguist Fernando Lázaro Carreter.

- His newspaper columns admonishing against the usage of anglicisms in the Spanish press became very popular in 1980s-1990s.



# Extraction pipeline: from RSS to `@lazarobot`



RSS

**Unannotated data**

```
{"date": "Thursday, 23
April 2020 09:14:01
+0200", "newspaper":
"El País", text": "'Au-
pairs' extranjeras en
la España rural: "Me
enamoré de la vida en
los pueblos"..."}
```

**Training corpus**

BiLSTM-CRF model

Human revision

**Automatically annotated data**

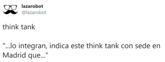
```
{"date": "Thursday, 23
April 2020 09:14:01
+0200", "newspaper":
"El País", text": "'Au-
pairs' extranjeras en
la España rural: "Me
enamoré de la vida en
los pueblos"..."}
```

**Anglicism database**

| anglicism |
|---|
| *reality show* |
| *online* |
| *prime time* |

`observatoriolazaro.es`

`@lazarobot`

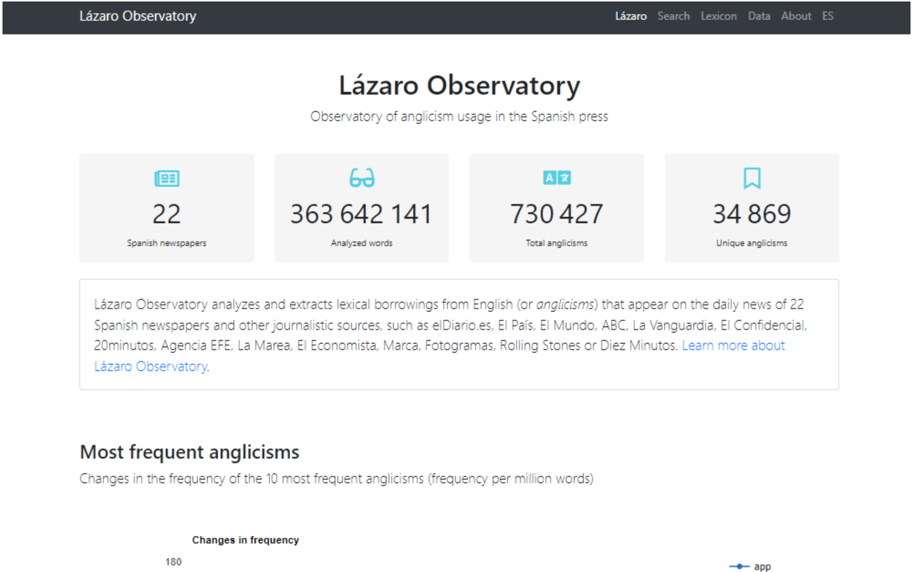


lazarobot
@lazarobot

think tank

"...lo integran, indica este think tank con sede en Madrid que..."

# observatoriolazaro.es



## Lázaro Observatory

Observatory of anglicism usage in the Spanish press

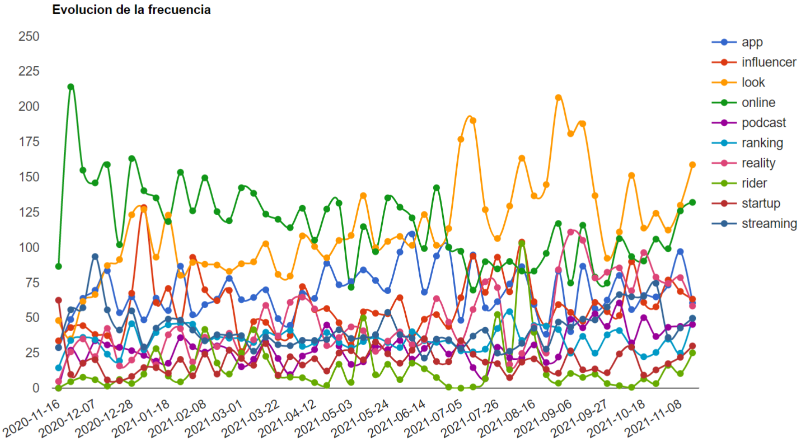
| | | | |
|---|---|---|---|
| 22<br>Spanish newspapers | 363 642 141<br>Analyzed words | 730 427<br>Total anglicisms | 34 869<br>Unique anglicisms |

Lázaro Observatory analyzes and extracts lexical borrowings from English (or *anglicisms*) that appear on the daily news of 22 Spanish newspapers and other journalistic sources, such as elDiario.es, El País, El Mundo, ABC, La Vanguardia, El Confidencial, 20minutos, Agencia EFE, La Marea, El Economista, Marca, Fotogramas, Rolling Stones or Diez Minutos. Learn more about Lázaro Observatory.

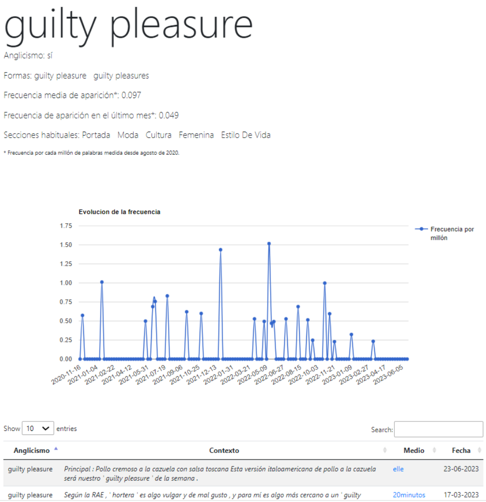
## Most frequent anglicisms

Changes in the frequency of the 10 most frequent anglicisms (frequency per million words)

# observatoriolazaro.es



# observatoriolazaro.es



## guilty pleasure

Anglicismo: sí

Formas: guilty pleasure guilty pleasures

Frecuencia media de aparición*: 0.097

Frecuencia de aparición en el último mes*: 0.049

Secciones habituales: Portada Moda Cultura Femenina Estilo De Vida

* Frecuencia por cada millón de palabras medida desde agosto de 2020.

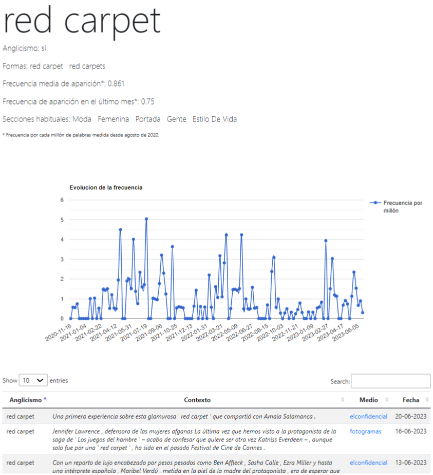
Evolucion de la frecuencia

Show 10 entries

Search:

| Anglicismo | Contexto | Medio | Fecha |
|---|---|---|---|
| guilty pleasure | *Principal : Pollo cremoso a la cazuela con salsa toscana Esta versión italoamericana de pollo a la cazuela será nuestro ' guilty pleasure ' de la semana .* | elle | 23-06-2023 |
| guilty pleasure | *Según la RAE , ' hortera ' es algo vulgar y de mal gusto , y para mí es algo más cercano a un ' guilty* | 20minutos | 17-03-2023 |

# observatoriolazaro.es



## red carpet

Anglicismo: sí

Formas: red carpet red carpets

Frecuencia media de aparición*: 0.861

Frecuencia de aparición en el último mes*: 0.75

Secciones habituales: Moda Femenina Portada Gente Estilo De Vida

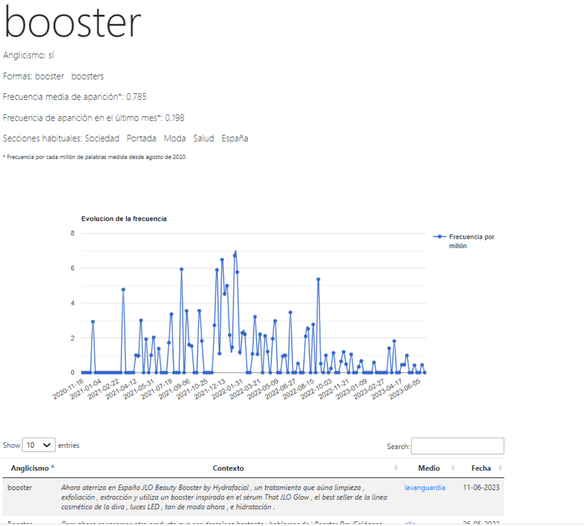
* Frecuencia por cada millón de palabras medida desde agosto de 2020.

Show 10 entries

Search:

| Anglicismo | Contexto | Medio | Fecha |
|---|---|---|---|
| red carpet | *Una primera experiencia sobre esta glamurosa ' red carpet ' que compartió con Amaia Salamanca .* | elconfidencial | 20-06-2023 |
| red carpet | *Jennifer Lawrence , defensora de las mujeres afganas La última vez que hemos visto a la protagonista de la saga de ' Los juegos del hambre ' – acaba de confesar que quiere ser otra vez Katniss Everdeen – , aunque solo fue por una ' red carpet ' , ha sido en el pasado Festival de Cine de Cannes .* | fotogramas | 16-06-2023 |
| red carpet | *Con un reparto de lujo encabezado por pesos pesados como Ben Affleck , Sasha Calle , Ezra Miller y hasta una intérprete española , Maribel Verdú , metida en la piel de la madre del protagonista , era de esperar que* | elconfidencial | 13-06-2023 |

# observatoriolazaro.es



## booster

Anglicismo: sí

Formas: booster boosters

Frecuencia media de aparición*: 0.785

Frecuencia de aparición en el último mes*: 0.198

Secciones habituales: Sociedad Portada Moda Salud España

* Frecuencia por cada millón de palabras medida desde agosto de 2020.

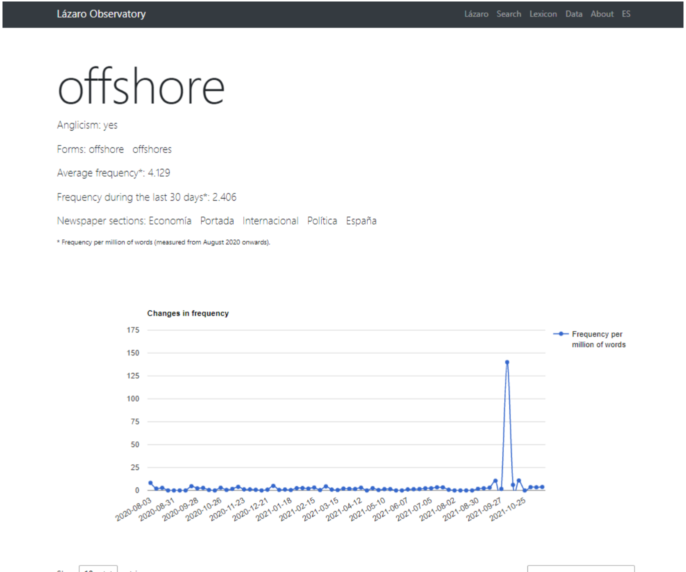
Evolucion de la frecuencia

Show 10 entries

Search:

| Anglicismo | Contexto | Medio | Fecha |
|---|---|---|---|
| booster | *Ahora aterriza en España JLO Beauty Booster by Hydrafacial , un tratamiento que aúna limpieza , exfoliación , extracción y utiliza un booster inspirado en el sérum That JLO Glow , el best seller de la línea cosmética de la diva , luces LED , tan de moda ahora , e hidratación .* | lavanguardia | 11-06-2023 |
| Booster | *Para ahora conocemos otro producto que nos desvelan bastante : hablamos de ' Booster Pro-Colágeno* | elle | 26-05-2023 |

# observatoriolazaro.es



# offshore

Anglicism: yes

Forms: offshore offshores

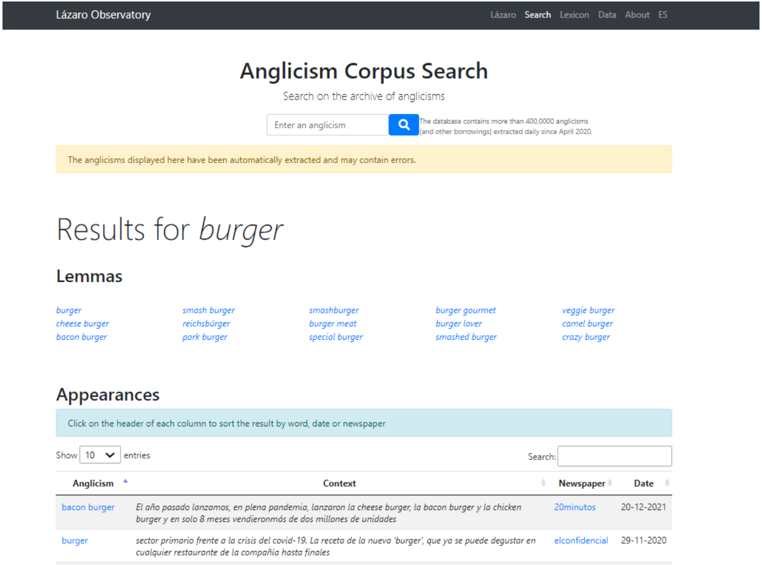
Average frequency*: 4.129

Frequency during the last 30 days*: 2.406

Newspaper sections: Economía Portada Internacional Política España

* Frequency per million of words (measured from August 2020 onwards).

# Database query



## Anglicism Corpus Search

Search on the archive of anglicisms

The database contains more than 400,0000 anglicisms (and other borrowings) extracted daily since April 2020.

The anglicisms displayed here have been automatically extracted and may contain errors.

## Results for *burger*

### Lemmas

*burger*
*cheese burger*
*bacon burger*
*smash burger*
*reichsbürger*
*pork burger*
*smashburger*
*burger meat*
*special burger*
*burger gourmet*
*burger lover*
*smashed burger*
*veggie burger*
*camel burger*
*crazy burger*

### Appearances

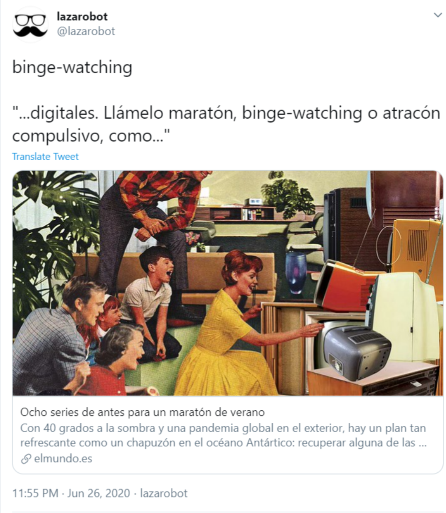
Click on the header of each column to sort the result by word, date or newspaper

Show 10 entries

Search:

| Anglicism | Context | Newspaper | Date |
|---|---|---|---|
| bacon burger | *El año pasado lanzamos, en plena pandemia, lanzaron la cheese burger, la bacon burger y la chicken burger y en solo 8 meses vendieronmás de dos millones de unidades* | 20minutos | 20-12-2021 |
| burger | *sector primario frente a la crisis del covid-19. La receta de la nueva 'burger', que ya se puede degustar en cualquier restaurante de la compañía hasta finales* | elconfidencial | 29-11-2020 |

# *Binge-watching*



lazarobot
@lazarobot

binge-watching

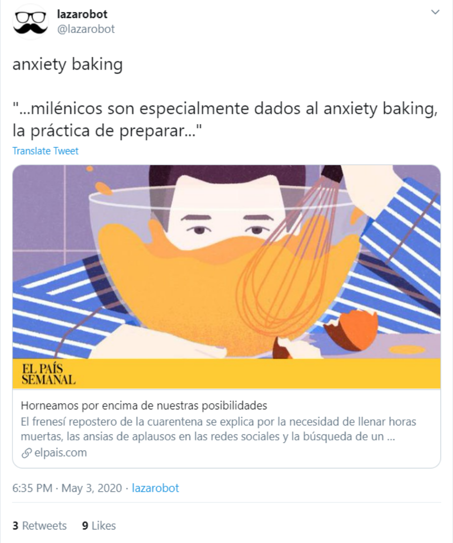
"...digitales. Llámelo maratón, binge-watching o atracón compulsivo, como..."

Translate Tweet

Ocho series de antes para un maratón de verano
Con 40 grados a la sombra y una pandemia global en el exterior, hay un plan tan refrescante como un chapuzón en el océano Antártico: recuperar alguna de las ...
elmundo.es

11:55 PM · Jun 26, 2020 · lazarobot

# *Anxiety baking*



lazarobot
@lazarobot

anxiety baking

"...milénicos son especialmente dados al anxiety baking, la práctica de preparar..."

Translate Tweet

Horneamos por encima de nuestras posibilidades

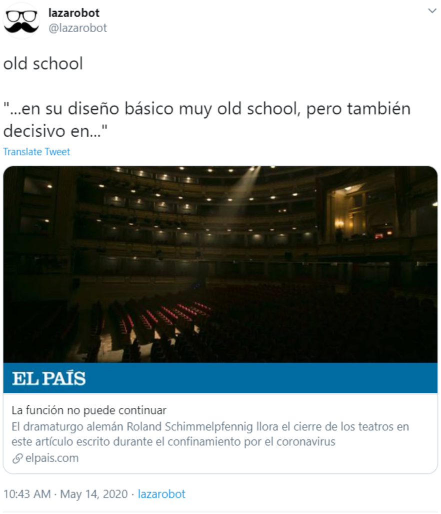
El frenesí repostero de la cuarentena se explica por la necesidad de llenar horas muertas, las ansias de aplausos en las redes sociales y la búsqueda de un ...

elpais.com

6:35 PM · May 3, 2020 · lazarobot

3 Retweets 9 Likes

# Old school



lazarobot
@lazarobot

old school

"...en su diseño básico muy old school, pero también decisivo en..."

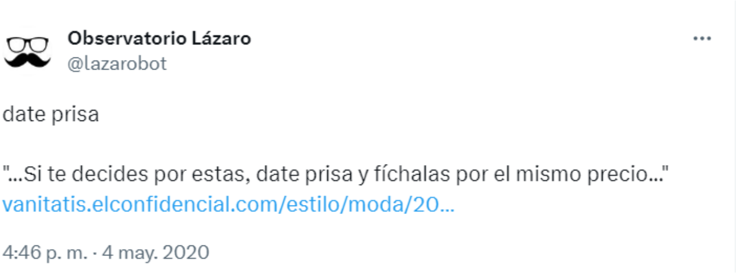
La función no puede continuar

El dramaturgo alemán Roland Schimmelpfennig llora el cierre de los teatros en este artículo escrito durante el confinamiento por el coronavirus

elpais.com

10:43 AM · May 14, 2020 · lazarobot

# Date prisa



Observatorio Lázaro
@lazarobot

date prisa

"...Si te decides por estas, date prisa y fíchalas por el mismo precio..."
vanitatis.elconfidencial.com/estilo/moda/20...

4:46 p. m. · 4 may. 2020

# Some data from May 2023

- 30,000 articles crawled
- 40,000 borrowings extracted
- 6,000 of them were unique
- 2,000 of them were seen for the first time

# Lázaro’s satellite projects

- CSV files with +700,000 anglicisms since 2020 on the website
- Models, guidelines and annotated corpora available on GitHub and HuggingFace
- `Pylazaro`, a Python library that performs anglicism detection in Spanish `https://pylazaro.readthedocs.io/`
- ADoBo, Automatic Detection of Borrowings shared task (2021)
- Papers at CALCS2020, SCiL2021, SEPLN2021, LREC2022, ACL2022.

This is work in progress, hopefully the list will grow

# Thank you! Questions?

Thank you very very much to eLex conference, the Board of Trustees and the sponsors of the Adam Kilgarriff Prize.

This work is advised by Julio Gonzalo from UNED and Constantine Lignos from Brandeis University.

# References I

[1] Beatrice Alex. Automatic detection of English inclusions in mixed-lingual data with an application to parsing. PhD thesis, University of Edinburgh, 2008.

[2] Gisle Andersen. Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In Cristiano Furiassi, Virginia Pulcini, and Félix Rodríguez González, editors, The anglicization of European lexis, pages 111–130. 2012.

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

[4] Cristiano Furiassi and Knut Hofland. The retrieval of false anglicisms in newspaper texts. In Corpus Linguistics 25 Years On, pages 347–363. Brill Rodopi, 2007.

[5] Matt Garley and Julia Hockenmaier. Beefmoves: Dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 135–139, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[6] Juan Gómez Capuz. Towards a typological classification of linguistic borrowing (illustrated with anglicisms in romance languages). Revista alicantina de estudios ingleses, 10:81–94, 1997.

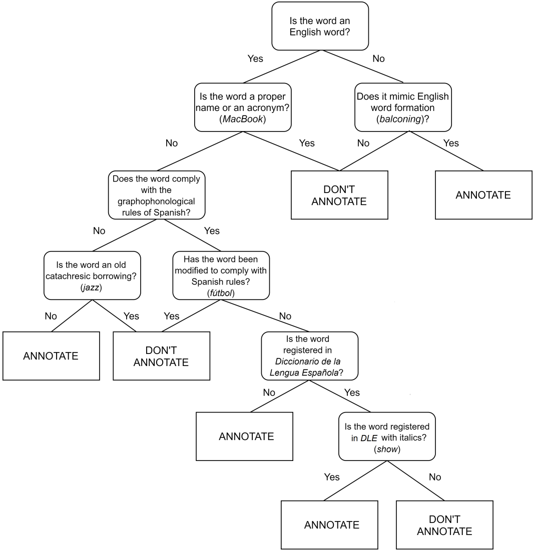
[7] Hahn Koo. An unsupervised method for identifying loanwords in Korean. Language Resources and Evaluation, 49(2):355–373, 2015.

[8] Sebastian Leidig, Tim Schlippe, and Tanja Schultz. Automatic detection of anglicisms for the pronunciation dictionary generation: a case study on our German IT corpus. In Spoken Language Technologies for Under-Resourced Languages, 2014.

[9] Gyri Smordal Losnegaard and Gunn Inger Lyse. A data-driven approach to anglicism identification in Norwegian. In Gisle Andersen, editor, Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian, pages 131–154. John Benjamins Publishing, 2012.

[10] André Mansikkaniemi and Mikko Kurimo. Unsupervised vocabulary adaptation for morph-based language models. In Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, pages 37–40. Association for Computational Linguistics, 2012.

[11] Jacqueline Rae Larsen Serigos. Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish. PhD thesis, The University of Texas at Austin, 2017.



Is the word an English word?

Yes

No

Is the word a proper name or an acronym? (*MacBook*)

Does it mimic English word formation (*balconing*)?

No

Yes

No

Yes

Does the word comply with the graphophonological rules of Spanish?

DON'T ANNOTATE

ANNOTATE

No

Yes

Is the word an old catachresic borrowing? (*jazz*)

Has the word been modified to comply with Spanish rules? (*fútbol*)

No

Yes

Yes

No

ANNOTATE

DON'T ANNOTATE

Is the word registered in *Diccionario de la Lengua Española*?

No

Yes

ANNOTATE

Is the word registered in *DLE* with italics? (*show*)

Yes

No

ANNOTATE

DON'T ANNOTATE

# Previous work on anglicism detection

| Work | Pattern matching | Lexicon/corpus lookup | Char n-grams probability | Machine Learning model | Language |
|---|---|---|---|---|---|
| [1] | | ✓ | | | German |
| [2] | ✓ | ✓ | ✓ | | Norwegian |
| [3] | ✓ | | | | French |
| [4] | | ✓ | ✓ | | Italian |
| [5] | | ✓ | ✓ | Maxent | German |
| [7] | | | ✓ | EM | Korean |
| [8] | | ✓ | ✓ | DT, SVM | German |
| [9] | | | ✓ | k-NN | Norwegian |
| [10] | | | ✓ | | Finnish |
| **[11]** | ✓ | ✓ | ✓ | | **Spanish** |

# Corpus split (CRF)

| Set | Headlines | Tokens | Headlines with anglicisms | Anglicisms | Other borrowings |
|---|---:|---:|---:|---:|---:|
| Train | 10,513 | 154,632 | 709 | 747 | 40 |
| Dev | 3,020 | 44,758 | 200 | 219 | 14 |
| Test | 3,020 | 44,724 | 202 | 212 | 13 |
| Suppl. test | 5,017 | 81,551 | 122 | 126 | 35 |
| **Total** | 21,570 | 325,665 | 1,233 | 1,304 | 102 |

Number of headlines, tokens and anglicisms per corpus subset.

# The corpus: counts

| Set | Tokens | ENG | OTHER | Unique |
|---|---:|---:|---:|---:|
| Training | 231,126 | 1,493 | 28 | 380 |
| Development | 82,578 | 306 | 49 | 316 |
| Test | 58,997 | 1,239 | 46 | 987 |
| Total | 372,701 | 3,038 | 123 | 1,683 |

Corpus splits with counts

# Annotation process

- In CoNLL format, with BIO encoding
  Because borrowings can be single token (*app*) or multitoken (*machine learning*)

ENG: unadapted emerging anglicisms [6]

✓ unadapted lexical anglicisms *show, smartphone, prime time*

✓ pseudoanglicisms *puenting, balconing*

✗ anglicisms that have been orthographically adapted *fútbol, mitin*

✗ anglicisms that have been morphologically adapted *hackear*

✗ proper names

OTHER: borrowings from other languages *gourmet, tempeh*

# Ablation study results

| Features | Precision | Recall | F1 score | F1 change |
|---|---:|---:|---:|---:|
| All features | 97.84 | **82.65** | **89.60** | |
| – Bias | 96.76 | 81.74 | 88.61 | –0.99 |
| – Token | 95.16 | 80.82 | 87.41 | –2.19 |
| – Uppercase | 97.30 | 82.19 | 89.11 | –0.49 |
| – Titlecase | 96.79 | **82.65** | 89.16 | –0.44 |
| – Char trigram | 96.05 | 77.63 | 85.86 | –**3.74** |
| – Quotation | 97.31 | **82.65** | 89.38 | –0.22 |
| – Suffix | 97.30 | 82.19 | 89.11 | –0.49 |
| – POS tag | **98.35** | 81.74 | 89.28 | –0.32 |
| – Word shape | 96.79 | **82.65** | 89.16 | –0.44 |
| – Word embedding | 95.68 | 80.82 | 87.62 | –1.98 |

# Additional features tried

| Features | Precision | Recall | F1 score | F1 change |
|---|---:|---:|---:|---:|
| Baseline | 97.84 | 82.65 | 89.60 | |
| Baseline + Bigram | 95.16 | 80.82 | 87.41 | −2.19 |
| Baseline + 4-gram | 97.28 | 81.74 | 88.83 | −0.77 |
| Baseline + Lemma | 97.81 | 81.74 | 89.05 | −0.55 |
| Baseline + Punctuation | 96.26 | 82.19 | 88.67 | −0.93 |
| Baseline + Sentence position | 96.76 | 81.74 | 88.61 | −0.99 |
| Baseline + Graphotactic shape | 94.27 | 82.65 | 88.08 | −1.52 |
| Baseline + Lexicon (ES) | 94.76 | 82.65 | 88.29 | −1.31 |
| Baseline + Lexicon (EN) | 96.76 | 81.74 | 88.61 | −0.99 |
| Baseline + Probability (ES) | 97.84 | 82.65 | 89.60 | 0.00 |
| Baseline + Probability (EN) | 97.84 | 82.65 | 89.60 | 0.00 |
| Baseline + Probability EN > ES | 96.22 | 81.28 | 88.12 | −1.48 |

# CRF model results

| Set | Precision | Recall | F1 score |
|---|---:|---:|---:|
| Development set | 97.84 | 82.65 | 89.60 |
| Development set (inc. OTHER) | 96.86 | 79.40 | 87.26 |
| Test set | 95.05 | 81.60 | 87.82 |
| Test set (inc. OTHER) | 95.19 | 79.11 | 86.41 |
| Supplemental test set | 83.16 | 62.70 | 71.49 |
| Supplemental test set (inc. OTHER) | 87.62 | 57.14 | 69.17 |

# Models results on COALAS

| Model | Word emb. | BPE emb. | Char emb. | Development | | | Test | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Prec. | Recall | F1 | Prec. | Recall | F1 |
| CRF | w2v (spa) | - | - | 74.13 | 59.72 | 66.15 | 77.89 | 43.04 | 55.44 |
| BETO | - | - | - | 73.36 | 73.46 | 73.35 | 86.76 | 75.50 | 80.71 |
| mBERT | - | - | - | 79.96 | 73.86 | 76.76 | 88.89 | 76.16 | 82.02 |
| BiLSTM-CRF | BETO+BERT | en, es | - | **85.84** | 77.07 | **81.21** | 90.00 | 76.89 | 82.92 |
| BiLSTM-CRF | BETO+BERT | en, es | ✓ | 84.29 | **78.06** | 81.05 | 89.71 | 78.34 | 83.63 |
| BiLSTM-CRF | Codeswitch | - | - | 80.21 | 74.42 | 77.18 | 90.05 | 76.76 | 82.83 |
| BiLSTM-CRF | Codeswitch | - | ✓ | 81.02 | 74.56 | 77.62 | 89.92 | 77.34 | 83.13 |
| BiLSTM-CRF | Codeswitch | en, es | - | 83.62 | 75.91 | 79.57 | 90.43 | 78.55 | 84.06 |
| BiLSTM-CRF | Codeswitch | en, es | ✓ | 82.88 | 75.70 | 79.10 | **90.60** | **78.72** | **84.22** |

Scores for the development and test sets across all models.

# CRF model error analysis

1. Neologisms in Spanish

   *puntocom, pin parental*

2. Proper names or entities:

   *lorazepam*

3. Orthographically adapted borrowings:

   *láser*

4. Titles from songs, films or series

   *it darker* in*'You want it darker', la despedida de Leonard Cohen*

5. Partial matches from multi-token anglicisms:

   *marketing* instead of *email marketing*