ADoBo: Automatic Detection of Borrowings

Detecting unassimilated lexical borrowings in the Spanish press

Elena Álvarez Mellado[1] Luis Espinosa Anke[2]
Julio Gonzalo Arroyo[1] Constantine Lignos[3] Jordi Porta Zamorano[4]

[1]NLP & IR group, UNED

[2]School of Computer Science and Informatics, Cardiff University

[3]Michtom School of Computer Science, Brandeis University

[4]Centro de Estudios de la RAE

Iberian Languages Evaluation Forum (IberLEF 2021)

Table of Contents

1. What is lexical borrowing (and why it matters as an NLP task)

2. ADoBo 2021 shared task

Table of Contents

1. What is lexical borrowing (and why it matters as an NLP task)

2. ADoBo 2021 shared task

What is lexical borrowing?

Lexical borrowing is the incorporation of words form one language into another language.

For ex., using in Spanish words that come from English:
podcast, app, online, crowdfunding, spin-off, big data, fake news...

- Lexical borrowing is a type of linguistic borrowing.
  - Linguistic borrowing is the process of reproducing in one language the patterns of other languages Haugen (1950)
- Borrowing and code-switching are related and have frequently been described as a continuum Clyne et al. (2003)
  - Code-switching = mixing two languages in one sentence.
    Ex: *You start a sentence in English y la acabas en español*
    Poplack (1980); Poplack et al. (1988)

Lexical borrowing vs Code switching

	Code Switching	**Lexical Borrowing**
Speaker	bilinguals	monolinguals
Grammar compliance	both languages	recipient language
Level of integration	not integrated	can be integrated
NLP approach	one tag per token (*à la POS-tagging*)[1]	extraction of spans of interest (*à la NER*)

[1] see Computational Approaches to Linguistic Code-Switching workshops (CALCS) Solorio et al. (2014); Diab et al. (2016); Aguilar et al. (2018); Solorio et al. (2020, 2021)

Why is borrowing interesting in Linguistics

- Borrowing is a manifestation of how languages change and interact (diachronic linguistics and contact linguistics) Weinreich (1963)

Why is borrowing interesting in Linguistics

- Borrowing is a manifestation of how languages change and interact (diachronic linguistics and contact linguistics) Weinreich (1963)
- New realities → new words (relevant for lexicography)

  online, software, streaming...

Why is borrowing interesting in Linguistics

- Borrowing is a manifestation of how languages change and interact (diachronic linguistics and contact linguistics) Weinreich (1963)
- New realities → new words (relevant for lexicography)

  online, software, streaming...
- Old realities → new words (relevant for sociolinguistics)

  color carne → nude
  barato → low-cost
  olio (from lat. 'oleum') → azeyte (current 'aceite')

Why is borrowing interesting in Linguistics

- Borrowing is a manifestation of how languages change and interact (diachronic linguistics and contact linguistics) Weinreich (1963)
- New realities → new words (relevant for lexicography)

  online, software, streaming...
- Old realities → new words (relevant for sociolinguistics)

  color carne → nude

  barato → low-cost

  olio (from lat. 'oleum') → azeyte (current 'aceite')
- Linguistic adaptation:

  football → fútbol

  spaghetti → espaguetis

Why is borrowing relevant in NLP

- Borrowings are a common source of out-of-vocabulary words Gerding Salas et al. (2018).

Why is borrowing relevant in NLP

- Borrowings are a common source of out-of-vocabulary words Gerding Salas et al. (2018).
- Automatically detecting lexical borrowings from text has proven to be relevant for NLP downstream tasks:
  - Parsing Alex (2008)
  - Text-to-speech synthesis Leidig et al. (2014)
  - Machine translation Tsvetkov and Dyer (2016)

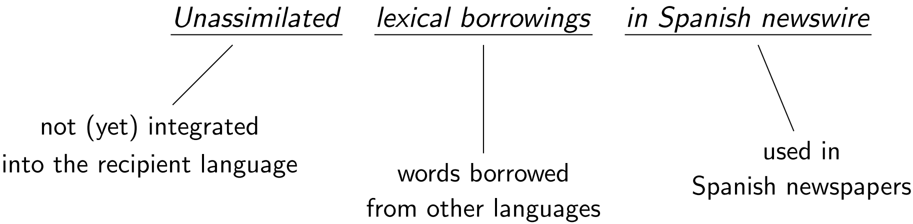
Why is borrowing relevant in NLP

- Borrowings are a common source of out-of-vocabulary words Gerding Salas et al. (2018).
- Automatically detecting lexical borrowings from text has proven to be relevant for NLP downstream tasks:
    - Parsing Alex (2008)
    - Text-to-speech synthesis Leidig et al. (2014)
    - Machine translation Tsvetkov and Dyer (2016)
- In the last decade has been a growing interest in the influence of English in other languages Görlach (2002).
    - Previous work on automatic detection of borrowings in different European languages: German, French, Italian, Norwegian, Finnish Andersen (2012); Chesley (2010); Furiassi and Hofland (2007); Garley and Hockenmaier (2012); Losnegaard and Lyse (2012); Mansikkaniemi and Kurimo (2012)
    - In Spanish, the automatic detection of anglicisms has been seldom explored Serigos (2017); Álvarez Mellado (2020)

Table of Contents

1. What is lexical borrowing (and why it matters as an NLP task)
2. ADoBo 2021 shared task

The task



Unassimilated *lexical borrowings* *in Spanish newswire*

not (yet) integrated
into the recipient language

words borrowed
from other languages

used in
Spanish newspapers

Words from other languages (mainly English) that have recently been imported into Spanish and that are being used in Spanish newspapers

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

Borrowing detection is harder than it seems

Why dictionary lookup is not enough:

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

Borrowing detection is harder than it seems

Why dictionary lookup is not enough:

- *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
  - *Media social* is not a borrowing.

Borrowing detection is harder than it seems

Why dictionary lookup is not enough:

- *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
  - *Media social* is not a borrowing.
- Not every English word is necessarily a borrowing.
  - *Sgt.Peppers Lonely Hearts Club Band*
  - *Eternal sunshine of the spotless mind*
  - *Stranger Things*

Borrowing detection is harder than it seems

Why dictionary lookup is not enough:

- *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
  - *Media social* is not a borrowing.
- Not every English word is necessarily a borrowing.
  - *Sgt.Peppers Lonely Hearts Club Band*
  - *Eternal sunshine of the spotless mind*
  - *Stranger Things*
- Prior work based on dictionary lookup produced very modest results (F1=47, F1=26) Serigos (2017); Álvarez Mellado (2020)

The corpus

We distributed a corpus:

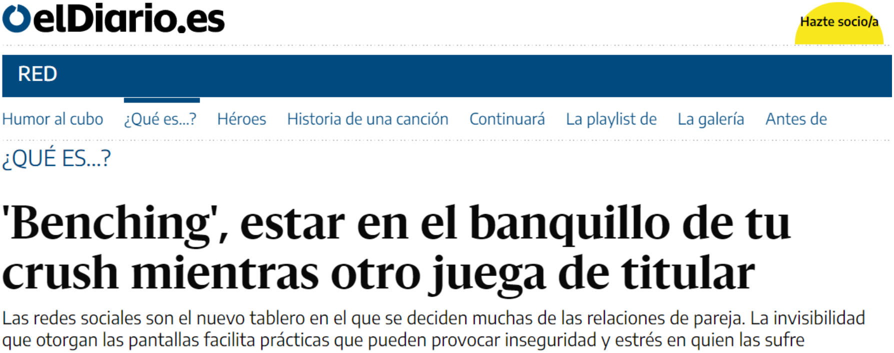
- Composed of Spanish newspapers
- Annotated with lexical borrowings with 2 tags:
  - `ENG`: for English borrowings
  - `OTHER`: for borrowings from other languages
- In CoNLL format
- With BIO encoding
  Because borrowings can be single token (*app*) or multitoken (*machine learning*)

The corpus: example

```
En   O
este   O
mes   O
especialmente   O
puede   O
ser   O
de   O
utilidad   O
apuntarnos   O
al   O
batch   B-ENG
cooking   I-ENG
```

```
Benching   B-ENG
,   O
estar   O
en   O
el   O
banquillo   O
de   O
tu   O
crush   B-ENG
mientras   O
otro   O
juega   O
de   O
titular   O
```

The corpus: example



¿QUÉ ES...?

'Benching', estar en el banquillo de tu crush mientras otro juega de titular

Las redes sociales son el nuevo tablero en el que se deciden muchas de las relaciones de pareja. La invisibilidad que otorgan las pantallas facilita prácticas que pueden provocar inseguridad y estrés en quien las sufre

Figure: Published at elDiario.es on December 2020[2]

[2] https://www.eldiario.es/red/que-es/benching-banquillo-crush-juega-titular_1_6484819.html

The corpus: counts

Set	**Tokens**	ENG	OTHER	**Unique**
Train	231,126	1,493	28	380
Dev.	82,578	306	49	316
Test	58,997	1,239	46	987
Total	372,701	3,038	123	1,683

Table: Corpus split and counts.

Evaluation

- Results of the task were computed using SeqScore[3], a Python package for evaluating sequence labeling tasks, configured to emulate the conlleval evaluation script (Palen-Michel et al., 2021).
- F1-measure was used as the official evaluation score for the final ranking.
- Evaluation was done exclusively at the span level. This means that only exact matches were considered, and no credit was given to partial matches.
- Additional evaluation was done removing orthographic cues: removing all quotation marks and converting all text to lower case.

[3] https://github.com/bltlab/seqscore

Shared task results

Team	System	Type	Prec.	Rec.	F1	Ref.	Pred.	Corr.
Marrouviere	(1)	ALL	88.81	81.56	85.03	1,285	1,180	1,048
		ENG	90.70	82.65	86.49	1,239	1,129	1,024
		OTHER	47.06	52.17	49.48	46	51	24
Versae	(2)	ALL	88.77	81.17	84.80	1,285	1,175	1,043
		ENG	90.31	82.73	86.35	1,239	1,135	1,025
		OTHER	45.00	39.13	41.86	46	40	18
Marrouviere	(3)	ALL	89.40	66.30	76.14	1,285	953	852
		ENG	90.98	67.55	77.54	1239	920	837
		OTHER	45.45	32.61	37.97	46	33	15
Marrouviere	(4)	ALL	92.28	61.40	73.74	1,285	855	789
		ENG	93.43	63.12	75.34	1,239	837	782
		OTHER	38.89	15.22	21.88	46	18	7
Versae	(5)	ALL	62.76	46.30	53.29	1,285	948	595
		ENG	62.97	47.62	54.23	1,239	937	590
		OTHER	45.45	10.87	17.54	46	11	5
Mgrafu	(6)	ALL	65.15	37.82	47.86	1,285	746	486
		ENG	65.31	38.90	48.76	1,239	738	482
		OTHER	50.0	8.69	14.81	46	8	4
BERT4EVER	(7)	ALL	75.27	27.47	40.25	1,285	469	353
		ENG	75.43	28.25	41.10	1,239	464	350
		OTHER	60.00	6.52	11.76	46	5	3
BERT4EVER	(8)	ALL	76.29	25.29	37.99	1,285	426	325
		ENG	76.48	25.99	38.80	1,239	421	322
		OTHER	60.00	6.52	11.76	46	5	3
BERT4EVER	(9)	ALL	76.44	24.75	37.39	1,285	416	318
		ENG	76.64	25.42	38.18	1,239	411	315
		OTHER	60.00	6.52	11.76	46	5	3

Álvarez Mellado et al. (2021)

BERT4EVER team: CRF model with data augmentation

Jiang et al. (2021)

- Combined several CRF models trained on different portions of the task's training data
- The models were used to label a freely-available open corpus in Spanish
- Models were then re-trained on the output
- F1 score of 40.25

Versae submission: using STILTs

De la Rosa (2021)

- Experimented with using supplementary training on intermediate label-data tasks
- Fine-tuned several multilingual language models (mBERT, RoBERTa)
- F1 score of 84.80

Some final thoughts on ADoBo shared task

- This was the first edition of ADoBo

  As far as we know, ADoBo is the first shared task on borrowing detection whatsoever.

Some final thoughts on ADoBo shared task

- This was the first edition of ADoBo

  As far as we know, ADoBo is the first shared task on borrowing detection whatsoever.

- We had a moderate turnout

  50 registered participants, 9 submissions, from 4 different teams, 2 paper submissions

Some final thoughts on ADoBo: future editions?

A post-competition questionnaire showed that participants would like to see future editions of ADoBo. Here are some of the topics that were suggested:

- Lexical borrowing detection in more languages

Some final thoughts on ADoBo: future editions?

A post-competition questionnaire showed that participants would like to see future editions of ADoBo. Here are some of the topics that were suggested:

- Lexical borrowing detection in more languages
- Semantic borrowing detection

Some final thoughts on ADoBo: future editions?

A post-competition questionnaire showed that participants would like to see future editions of ADoBo. Here are some of the topics that were suggested:

- Lexical borrowing detection in more languages
- Semantic borrowing detection
- Diachronic assimilation of borrowings

Some final thoughts on ADoBo: future editions?

A post-competition questionnaire showed that participants would like to see future editions of ADoBo. Here are some of the topics that were suggested:

- Lexical borrowing detection in more languages
- Semantic borrowing detection
- Diachronic assimilation of borrowings
- Code-switching

Some final thoughts on ADoBo: future editions?

A post-competition questionnaire showed that participants would like to see future editions of ADoBo. Here are some of the topics that were suggested:

- Lexical borrowing detection in more languages
- Semantic borrowing detection
- Diachronic assimilation of borrowings
- Code-switching
- Other ideas? Feel free to reach out!

References

Aguilar, G., AlGhamdi, F., Soto, V., Diab, M., Hirschberg, J., and Solorio, T. (2018). Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task. In Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Alex, B. (2008). Comparing corpus-based to web-based lookup techniques for automatic English inclusion detection. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Álvarez Mellado, E. (2020). Lázaro: An extractor of emergent anglicisms in Spanish newswire.

Álvarez Mellado, E., Espinosa-Anke, L., Gonzalo Arroyo, J., Lignos, C., and Porta Zamorano, J. (2021). Overview of ADoBo 2021: Automatic detection of unassimilated borrowings in the Spanish press.

Andersen, G. (2012). Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In Furiassi, C., Pulcini, V., and Rodríguez González, F., editors, The anglicization of European lexis, pages 111–130.

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

Clyne, M., Clyne, M. G., and Michael, C. (2003). Dynamics of language contact: English and immigrant languages. Cambridge University Press.

De la Rosa, J. (2021). Adobo 2021: The futility of stilts for the classification of lexical borrowings in spanish.

Diab, M., Fung, P., Ghoneim, M., Hirschberg, J., and Solorio, T., editors (2016). Proceedings of the Second Workshop on Computational Approaches to Code Switching, Austin, Texas. Association for Computational Linguistics.

Furiassi, C. and Hofland, K. (2007). The retrieval of false anglicisms in newspaper texts. In Corpus Linguistics 25 Years On, pages 347–363. Brill Rodopi.

Garley, M. and Hockenmaier, J. (2012). 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. Association for Computational Linguistics.

Gerding Salas, C., Cañete González, P., and Adam, C. (2018). Neología sintagmática anglicada en español: Calcos y préstamos. Revista signos, 51(97):175–192.

Görlach, M. (2002). English in Europe. OUP Oxford.

References (cont.)

Haugen, E. (1950). The analysis of linguistic borrowing. Language, 26(2):210–231.

Jiang, S., Cui, T., Fu, Y., Lin, N., and Xiang, J. (2021). Bert4ever at adobo 2021: Detection of borrowings in the spanish language using pseudo-label technology.

Leidig, S., Schlippe, T., and Schultz, T. (2014). 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.

Losnegaard, G. S. and Lyse, G. I. (2012). A data-driven approach to anglicism identification in Norwegian. In Andersen, G., editor, Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian, pages 131–154. John Benjamins Publishing.

Mansikkaniemi, A. and Kurimo, M. (2012). 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.

Palen-Michel, C., Holley, N., and Lignos, C. (2021). SeqScore. `https://github.com/bltlab/seqscore`.

Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1.

Poplack, S., Sankoff, D., and Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. Linguistics, 26(1):47–104.

Serigos, J. R. L. (2017). Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish. PhD thesis, The University of Texas at Austin.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In Proceedings of the First Workshop on Computational Approaches to Code Switching, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Solorio, T., Chen, S., Black, A. W., Diab, M., Sitaram, S., Soto, V., Yilmaz, E., and Srinivasan, A., editors (2021). Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, Online. Association for Computational Linguistics.

Solorio, T., Choudhury, M., Bali, K., Sitaram, S., Das, A., and Diab, M., editors (2020). Proceedings of the The 4th Workshop on Computational Approaches to Code Switching, Marseille, France. European Language Resources Association.

Tsvetkov, Y. and Dyer, C. (2016). Cross-lingual bridges with models of lexical borrowing. Journal of Artificial Intelligence Research, 55:63–93.

Weinreich, U. (1963). Languages in contact (1953). The Hague: Mouton.