Lázaro: An Extractor of Emergent Anglicisms in Spanish Newswire

Elena Álvarez Mellado

What this thesis is about

Emergent *anglicisms* *in Spanish Newswire*

new or newish

words borrowed from English

used in Spanish newspapers

Lexical borrowings from English that have recently been imported into Spanish and that are being used in Spanish newspapers

podcast, *crowdfunding*, *spin-off*, *big data*, *fake news*...

Why?

- There has been a growing interest in the influence of English in other languages (Görlach, 2002).

- The influence of English on Spanish produces great interest both among scholars and non-specialized public.
  - Spanish prescriptivist institutions (such as *Real Academia Española* or *Fundéu*) admonish against the usage of English borrowings.

- Despite this growing interest, there is a lack of data-driven approaches to anglicism tracking.

Why?

The motivation behind this thesis is to produce a computational model that can detect and track new(ish) anglicisms being used in Spanish newspapers.

Contributions of this thesis

- A corpus of Spanish newswire annotated with anglicisms.
- Two models that perform automatic extraction of anglicisms:
  - A conditional random field model with handcrafted features
  - A BiLSTM-CRF model with word and character embeddings
- An automatic pipeline that performs daily extraction of anglicisms from the main national newspapers of Spain.

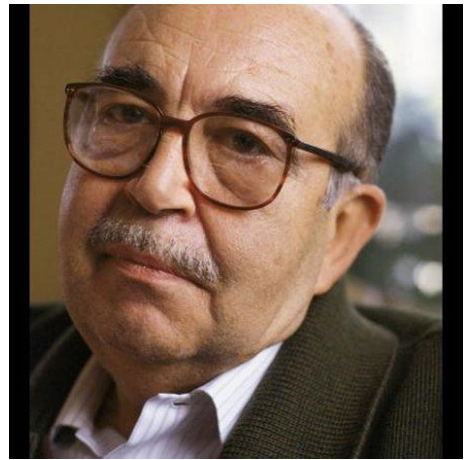
Outline

1. Anglicism: definition and scope
2. Previous work on anglicism detection
3. Corpus creation and annotation
4. Models for automatic extraction of anglicisms
   - 4.1 Non-neural model: CRF baseline
   - 4.2 Neural model: BiLSTM-CRF model
5. An automatic pipeline for anglicism extraction
6. Conclusions

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.



Anglicism: definition and scope

Anglicism: definition and scope

Anglicisms: lexical borrowings from English.

- 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)

- Lexical borrowing is the incorporation of lexical units form one language into another language.
  - Borrowing and code-switching are related and have frequently been described as a continuum (Clyne et al., 2003)

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*)	extraction of spans of interest (*à la NER*)

Previous work on anglicism detection

Previous work on anglicism detection

Work	Pattern matching	Lexicon/corpus lookup	Char n-grams probability	Machine Learning model	Language
Alex (2008)		✓			German
Andersen (2012)	✓	✓	✓		Norwegian
Chesley (2010)	✓				French
Furiassi and Hofland (2007)		✓	✓		Italian
Garley and Hockenmaier (2012)		✓	✓	Maxent	German
Koo (2015)			✓	EM	Korean
Leidig et al. (2014)		✓	✓	DT, SVM	German
Losnegaard and Lyse (2012)			✓	k-NN	Norwegian
Mansikkaniemi and Kurimo (2012)			✓		Finnish
Serigos (2017)	✓	✓	✓		**Spanish**

Serigos (2017)

- A module calculates the probability of a word being English or Spanish based on character n-grams.
- If the difference between both probabilities is smaller than a given threshold, a lexicon lookup module adjudicates the label.
- A capitalization module was set to ignore titlecased words.
- F1 score of 74.50 on unseen data (token level evaluation).

Limitations in the work by Serigos (2017)

1. Each word is analyzed in isolation.
2. All tokens in titlecase were ignored.
3. Only two language tags considered: Spanish and English (*gourmet* would be considered Spanish)
4. All adjacents anglicisms were considered a single multiword borrowing.
5. The system was evaluated on token level exclusively (poor evaluation of multiword borrowings)

This thesis seeks to address the shortcomings derived from these assumptions.

Corpus creation and annotation

Corpus description

- A corpus of newpaper headlines in European Spanish
  - 21,570 headlines; 325,665 tokens
- Extracted from the Spanish newspaper *eldiario.es*
  - it is one of the main newspapers of Spain
  - it publishes all of its content under a Creative Common license :)
- Why only headlines?
  - faster and easier than annotating full articles
  - anglicisms are abundant in headlines (Furiassi and Hofland, 2007)
  - borrowings that make it to the headline are likely to be salient

Annotation process

ENG: unadapted emerging anglicisms (Gómez Capuz, 1997)

- ✓ 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*
- ✗ incorporated anglicisms that comply with Spanish spelling *bar, club*
- ✗ syntactic anglicisms, literal translations
- ✗ proper names

0THER: borrowings from other languages *gourmet, tempeh*

Corpus split

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.

Models for automatic extraction of anglicisms

Models for automatic extraction of anglicisms

1. A CRF model with handcrafted features
2. A neural BiLSTM-CRF model

CRF model

- Handcrafted features (similar to NER features)
  - Token, shape, titlecase, char trigram, quotation, word embedding
- Grid search for hyperparameters and embeddings
  - c1 = 0.05, c2 = 0.01, scaling = 0.5
  - word2vec embeddings from the Spanish Billion Words Corpus (Cardellino, 2019)
- BIO encoding (adapted from Ramshaw and Marcus (1999))
- Feature extractor: a two-token window in each direction

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 + Digit	**97.85**	**83.11**	**89.88**	**+0.28**
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
Baseline + Perplexity threshold	**97.86**	**83.56**	**90.15**	**+0.55**

CRF model results

Set	Precision	Recall	F1 score
Development set	97.84	82.65	89.60
Development set (inc. 0THER)	96.86	79.40	87.26
Test set	95.05	81.60	87.82
Test set (inc. 0THER)	95.19	79.11	86.41
Supplemental test set	83.16	62.70	71.49
Supplemental test set (inc. 0THER)	87.62	57.14	69.17

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*

Neural model

- Implemented using NCRF++
  - PyTorch-based library for sequence-labeling (Yang et al., 2018)
- Three layers: character sequence layer, word sequence layer and inference layer
  - character CNN + word LSTM + CRF model
  - Successful architecture on NER (Lample et al., 2016)
- Hyperparameters
  - Hidden dim = 200; iterations = 100, char dim = 30; lr = 0.0075

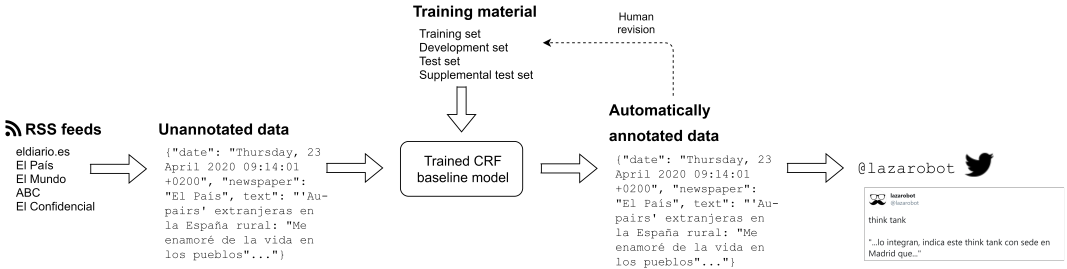
Best results obtained with the neural model

Set	Precision	Recall	F1 score
Dev set	96.49	75.34	84.62
Test set	93.89	80.48	86.67

(CRF baseline results: F1 on dev = 89.60, F1 on test = 87.82)

An automatic pipeline for anglicism extraction

Extraction pipeline: from RSS to @lazarobot



Anxiety baking

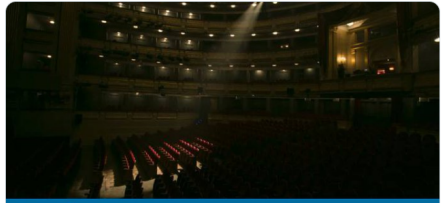
Old school

lazarobot
@lazarobot

old school

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

Translate Tweet



10:43 AM · May 14, 2020 · lazarobot

Neologisms: *Brilli-brilli*

Future work

- The phenomenon of anglicism is much wider
  - This thesis is only concerned with unadapted lexical borrowings
  - Models for other phenomena (borrowing adaptation, syntactic borrowing, semantic calques)

- Improve the BiLSTM-CRF model
  - Pretrained character embeddings
  - The pipeline output could be used as training material after human careful revision

- Tracking and frequency analysis over time of the anglicisms detected by the extraction pipeline.

References

Alex, B. (2008). *Automatic detection of English inclusions in mixed-lingual data with an application to parsing*. PhD thesis, University of Edinburgh.

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.

Cardellino, C. (2019). Spanish Billion Words Corpus and Embeddings. `https://crscardellino.github.io/SBWCE/`.

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.

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.

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

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

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

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

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

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.

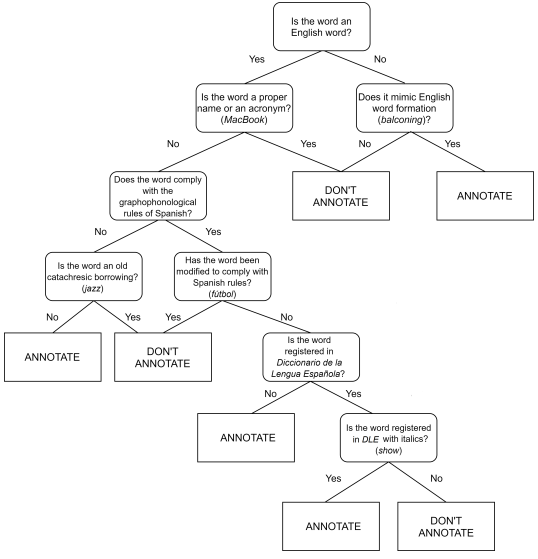
Mansikkaniemi, A. and Kurimo, M. (2012). Unsupervised vocabulary adaptation for morph-based language models. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 37–40. Association for Computational Linguistics.

Ramshaw, L. A. and Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

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.

Yang, J., Liang, S., and Zhang, Y. (2018). Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*.

Annotation decision steps



Unseen anglicisms

	True Positives	False Negatives
Dev set	*arcade*, *balconing*, *brain hacking*, *breaking of*, *call centers*, *communities*, *dating show*, *daytime*, *deep learning*, *docushows*, *edredoning*, *fast food*, *fitness*, *fracking*, *game jams*, *gin-tonic*, *girl-group*, *hardcore-punk*, *hip*, *hip hop*, *influencers*, *invent*, *machine learning*, *made in China*, *merchandising*, *microvlogging*, *morning show*, *networking*, *punky*, *redneck*, *roll-on*, *routers*, *skaters*, *stick*, *tickets*, *webcam*	*backstage*, *chat*, *deluxe*, *drone*, *ethio-jazz*, *femtech*, *geek*, *golden visa*, *he*, *her*, *him*, *influencer*, *influencers*, *instagramer*, *made in China*, *mindfulness*, *noodles*, *off*, *ok*, *ok*, *boomer*, *packs*, *share*, *she*, *showman*, *social media*, *social trading*, *spoiler*, *spoilers*, *spray*, *trailer*, *vamping*

Out of 104 unique true positives on the development set, 36 were not present in the training set

Headlines with anglicisms per section

Section	Percentage of anglicisms
Opinion	2.54%
Economy	3.70%
Lifestyle	6.48%
TV	8.83%
Music	9.25%
Technology	15.37%

People engaging :__)

Song titles


lazarobot
@lazarobot
Crying
"...:. . ORIGINAL:. . Cryings not for me. . ..."
Esta es la playlist que necesitas para salir a la calle y disfrutar del aire fresco
Si buscas temazos para practicar ejercicio al aire libre o, simplemente, quieres dar un paseo y mantener el ánimo arriba, aquí tienes lo que necesitas.
20minutos.es
12:18 PM · May 14, 2020 · lazarobot

Crowdfunding

lazarobot
@lazarobot

crowdfunding

"...2019 crearon una campaña de crowdfunding y con lo recaudado regalaron..."

11:59 AM · May 14, 2020 · lazarobot

Perfect match

lazarobot
@lazarobot

perfect match

"...aún no has encontrado tu perfect match debes echarle un vistazo al..."

Translate Tweet

11:40 AM · May 14, 2020 · lazarobot

Typos: *smpartphones*