A Naive Hybrid Approach to Borrowing Detection

Fernando Sánchez-León

Unaffiliated, Spain

Abstract

In this paper, we report on our approach to automatic borrowing detection and retrieval as implemented for ADoBo 2025 shared task. Our more outstanding hypothesis is that news text is edited according to strict editing guidelines that may be leveraged to extract borrowing candidates from them. The list of candidates so compiled can be (semi-)automatically turned into a gazetteer that allows us to experiment with a rule-based approach (simple lexicon lookup with blocked-for-labeling regions) that may complement a deep learning (DL) model for the task. Our main goal was to explore to what extent a naive hybrid system using contextual clues could improve the performance of an existing neural model. Our findings using the mid-size datasets provided for the task suggest that even a rule-based solution alone may outperform the DL model in some cases, and that it improves performance of the original model in any setting. The system ranked second in the shared task, with an F1 score of 95.93. Finally, our approach naturally extends to the retrieval of loanwords from languages other than English.

Keywords

Hybrid borrowing detection system, Gazetteers, Word-level language identification

1. Introduction

As part of IberLEF 2025 shared evaluation campaign of Natural Language Processing systems in Spanish and other Iberian languages [1], a second edition of ADoBo, a shared task on automatic detection of borrowings has been launched [2]. This paper reports on our results as participants of this task.

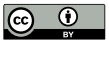
Lexical borrowing, the transfer of single words or multi-words from one language to another, has received much attention by NLP community, given its potential benefits to NLP applications such as machine translation and speech recognition and synthesis, as well as classical applied linguistics areas like lexicography and language change monitoring. Moreover, in a world linguistically dominated by English, anglicisms, that is, unadapted English loanwords, are scattered throughout journalistic texts, many a time even before the denotation that the loanword refers to has penetrated into the Spanish-speaking culture. Therefore, there are reasons beyond the purely linguistic to remain attentive to the use of lexical borrowing (LB) as a particular and very extended phenomenon surrounded by the wider landscape of neology.

Although vaguely related from a linguistic point of view to Named Entity Recognition (NER), LB detection is generally approached in the same way as this by the NLP community. Hence, in an era dominated by machine learning and prediction methods, it comes as no surprise that current approaches to LB detection frame it as a sequence labeling task using probabilistic graphical models like Conditional Random Fields (CRFs) and/or architectures based on deep learning (DL) like Long Short-Term Memory (LSTM) or the self-attention mechanism (as implemented in the `Transformer` architecture), both capable of learning the context and relationships in sequential data.

In these approaches, everything is left to the whim of prediction, and so the model, on the basis of the samples used to characterize a given task and domain, must predict the sequence of labels that most accurately replicates the annotation by an expert, using advanced pattern recognition mechanisms and producing state-of-art results for anglicism detection. Thus, [3] report an F1 score of 87.16 on the task for anglicism detection in Spanish journalistic texts using a corpus constructed for this purpose, the COALAS corpus. These authors highlight the model's capability to generalize to previously unseen borrowings. Nonetheless, this model has a number of drawbacks, some of them not mentioned by these authors in their error analysis section. For instance, this model inconsistently labels some tokens,

IberLEF 2025, September 2025, Zaragoza, Spain

f.sanchez.lcmcvp@gmail.com (F. Sánchez-León)



published 2025-12-10

probably due to low training data (*kit*, an adapted borrowing, occurs once in the training set labeled as O and twice in the test set, one of them incorrectly labeled as ENG), and/or to biased learning of contextual features (*Big Data* occurs only once in the development set, in a quoted context, and it is labeled as O in unquoted context in both cases where it appears in the test set).

In light of these inconsistencies, and following our view of the LB detection task, we are convinced that a specific lexical resource of known LBs combined with a DL model could boost score of the best performing model reported in [3]. However, rather than integrating gazetteer features in the DL model, we think that a much simpler, ensemble design with an oracle orchestrating the decision making will favor interpretable results for already known loanwords, away from the opaque heuristics of DL models. As for the lexical resource building, we plan to extract a large gazetteer from a huge journalistic corpus leveraging typographic conventions used by journalists and newspaper editors to drive the extraction routines. We project to implement three LB detection systems: (1) a baseline system labeling the text for LB using exclusively the gazetteer, (2) a system using also NER combined with the baseline system, and (3) a system that leverages contextual clues when using the gazetteer (with or without NER). All three systems can be benefited from the DL model to deal with OOV words and homographic forms.

2. Related work

There exists much work on the integration of lexicons and/or gazetteers in a DL architecture, at least for Named Entity Recognition (NER), although its benefits to task performance are arguable. In this regard, [4] provide a gentle overview of methods for incorporating gazetteers and entity segmentation in NER and conclude that gazetteer-enhanced models are more stable across datasets and perform better in the different scenarios evaluated in their work, improving entity segmentation and not just entity typing. On the other hand, [5], recognizing that extracting features from a rich model of the gazetteer they build and then concatenating such features with the input embeddings of a neural network significantly outperform other, more conventional approaches, state that large gazetteers may "overwhelm the training data" and degrade system performance, due to 'feature under-training'.

Most experiments on the combination of gazetteers with DL approaches are devised at the DL architecture level. [6] enhance a BiLSTM-CRF architecture for NER for kiwifruit diseases and pests with two new layers, specifically one AttSoftLexicon Layer that integrates in the original model the character and word information in the lexicon. In their study, [4] explore variants of the BiLSTM-CRF architecture for NER in order to integrate features and segmentation derived from GloVe vectors trained on Common Crawl and contextual ELMo representations trained on the 1B Word Benchmark. This information, as well as character-based word representations, is concatenated as input to the bidirectional LSTMs.

Other hybrid approaches in modern NER include dynamic gazetteer integration, where mechanisms for token-level gating are defined to allow for models to selectively combine gazetteer and contextual information. In this regard, [7] propose GEMNET, a novel approach for gazetteer knowledge integration that uses a Mixture-of-Experts gating network that allows the model to learn conditionally this combined information. [8] inject gazetteers via adaptive layers to improve NER performance across languages and domains, achieving a 17.6 improvement to F1 in low-resource languages through knowledge transfer.

[5] use a kind of mixed approach for NER by combining the outputs of a BiLSTM-based NER model and a separate gazetteer model, and then feeding both outputs into a CRF layer in order to get a final prediction. These authors also point that it is a relatively common practice in NER systems to use the presence of a token in a gazetteer as an additional feature for a classifier. However, the ensemble oracles so developed are always integrated in the DL, so in no way is the gazetteer the key information for the decision.

3. Assessment

The task of automatic LB detection can be seen as a special case of word-level language identification. If we take the following sentence from the ADoBo development set

> El PP se ha aliado en esta ocasión con los socialistas al justificar que Juan Carlos I cuente "con un pequeño 'staff' de Patrimonio Nacional" en Abu Dabi.

at word level, we can identify three languages: Spanish, for most of the words in the sentence; English, for the word *staff*; and a non-relevant (or universal) language for tokens that are part of a named entity (NE). The downstream task of LB detection is, very broadly formulated, that of isolating single tokens or multi-tokens from a vocabulary other than Spanish but ignoring NEs.

For a vast number of loanwords, its belonging to one language or another is independent of the context in which it appears, as long as it is inserted in another primary language. This contrasts with NER, since in the latter a given single token or multi-token could be an NE, even when written in lowercase (at least in some texts). If this assertion is true, then prediction is a desirable labeling method for guessed cases when there is no previous language identification (LI) knowledge for a given token, but not for already known ones.

To have a clear view of the knowledge of LBs that the model described in [3] has acquired (*flair-cs*), we use it to label the COALAS full dataset,[1] an unorthodox exercise, given that two thirds of the dataset have served as training and development sets. We provide a brief summary of our findings, that intend to be complementary to the thorough analysis of errors performed by these authors.

English words graphophonologically noncompliant with Spanish rules found in the train+dev sets include *youtuber* and *hall*: theses forms are non-systematically tagged as either ENG or O in the test set.

When rerun on the sets used for parameter estimation, a similar behavior is observed: for words like *funk* (seen once in a quoted context in the training set and also once, unquoted, in the development set), only the former is labeled as ENG; the same is true for *webcams* (once in a quoted context in the training set, labeled as O in the test set); the singular *webcam* is seen only once in the development set in an unquoted context, unspecified enough typographic environment so as for the word being labeled as O both of the times it appears in the test set. These findings reveal the special weight quotes have on the model predictions. The ENG bias has so much strength that the word *viral*, which is a case of homography between Spanish and English occurring 10 times in the train+dev sets (always labeled as O) is tagged as ENG in a quoted context in the full dataset run. The same is true for the NE *Seat*, a car brand.[2] Similar cases include *cool* and *offshore*.

As regards English loanwords that comply with graphophonological rules in Spanish, model behavior is alike: *bot* (5 times in the training set, 2 in the test set) is labeled as ENG only once; *jersey*, an adapted borrowing included in dictionaries, labeled as O twice in the training+dev sets is tagged as ENG in one of its two occurrences; other adapted anglicisms include *fans* (seen 4 times in training+dev sets, but relabeling these same contexts, gets ENG in one of them), and *gags* (seen only once in training set, as O, but labeled as ENG if the same sentence is retagged).

On the other hand, this model does show remarkable capabilities with OOV cases, although, once again, predictions are not systematic: *curly* is considered an English loanword, but only in one out of three cases in the test set; *halter* shows a 1/1 for the 2 times it is found; *pinballs* is labeled as ENG while its singular form is considered O. Finally, *denim* is tagged as part of an LB when collocated in the multiword *total look denim* (a secondary predication where *total* is not labeled as ENG), and as O in one of its two remaining occurrences.

In contrast, a clear example of the learning capabilities of the model is offered by words that exist both in Spanish and English. Given no low-data issues happen, as with the word *post* (only labeled as LB when it refers to content shared on a social media network, but not when it is a detachable prefix), the system performs on par with its overall score, resulting in 4 false positives out of 30 occurrences. The word *film*, an assimilated English word however labeled as ENG in the complex form *papel film* (2 times in the training set) and as OTHER when part of the form *film noire* (only once), when the model is run over the complete dataset it mislabels the French complex form as ENG and outputs one false positive, for a 2/10 error rate.

[1] https://github.com/lirondos/coalas

[2] The quotes in this example correspond to a highlighted span in the original webpage (https://www.infolibre.es/videolibre/playlist-de/ramoncin-he-hecho-he-querido_1_1192339.html).

It is in light of the above predictions that we argue that, although the system does show remarkable generalization capabilities when trained with a dataset rich in borrowing density, it is, as such, inadequate for LB retrieval if used as standalone component and propose our hybrid solution. Moreover, to increase even more the challenge of the prediction, for the correct identification of LBs we must consider the set of nuances described in the task guidelines,[3] that forces any program/model to change the prediction according to third-party information or the belonging of the loanword to certain subsets. In this regard, a lexicon-based component could provide a systematic, interpretable labeling for a significant subset of the borrowings commonly used in journalistic texts, thus complementing a predictive system and improving the overall performance.

4. Method and materials

4.1. General method

We plan to extract a large gazetteer from a huge journalistic corpus using typographic conventions used by journalists and newspaper editors to guide the extraction routines. A set of language wordlists for various languages (and not only English) will be used to perform a word-level language identification on the extracted text fragments, removing from the final extracted list all sequences labeled as being written is Spanish with the aid of a morphological analyzer for Spanish. The borrowing gazetteer[4] will be used as a backbone information for the ensemble oracle. This information and the predictions from other NLP modules will be kept in memory in a tabular tokenized version of the input text. The other modules used for decision making are an NER classifier, the language(s) to which the token belongs and the prediction from `flair-cs`.

4.2. Newspaper text corpus

A set of nearly 13M news pages from European Spanish sources has been compiled. These include newspapers, both nation wide (El País, ABC, La Vanguardia, El Periódico de España, OK Diario, El Confidencial, 20 Minutos, among them) and "regional" newspapers (El Correo, La Vox de Galicia, La Verdad, Diario de Almería, Diario Vasco); magazines (Hola, Cosmopolitan, Elle, Forbes, Vogue, National Geographic); radio stations (Cadena SER, COPE, OndaCero); and other miscellaneous sources (EFE, Zenda).

Files are preprocessed with a custom-made version of `trafilatura`[5] in order to archive an XML version with no boilerplate text but keeping original categories and/or keywords and, most importantly for this task, a normalized form of highlighted spans. The text size, once clean and without XML markup, reaches 6,600M tokens.

Although recognizing that anglicisms are constantly included in Spanish journals, we consider this a not so dynamic genre as social media, for instance. It is under this spirit that our huge corpus tries to mitigate the out-of-vocabulary (OOV) challenges that lexicon-based systems pose, which, nevertheless, is more prominent in specialized domains.

4.3. Donor languages wordlists and Spanish morphological analyzer

Wordlists have been collected for a set of donor languages observed when reading Spanish newspapers. We have, however, lists for other languages not very common in Spanish newswire since wordlist compilation is a task we started to perform some time ago. In fact, the list curation started five years ago using *unmunched* versions of `hunspell`[6] dictionaries for some of the languages we were working on. New words have been added to the initial lists over time mainly from two sources: the internet, and our

[3] Available at https://adobo-task.github.io/docs/guidelines.pdf

[4] We will use *gazetteer* and *lexicon* interchangeably in this text, since the list includes some other linguistic features not used in the context of ADoBo shared task like the relation between correct form and incorrect variants.

[5] https://trafilatura.readthedocs.io

[6] https://github.com/hunspell/hunspell

work with texts in languages other than Spanish. A simple search on `github.com` domain can point the reader in the right direction to discover many valuable word compilation efforts. The drawback of many of these lists, nonetheless, is that they are usually polluted of (lowercased) proper names and words from a donor language (most of the times English). For instance, Italian lists generally contain many English words, due to the ease with which Italian speakers integrate unadapted anglicisms in their language. In this respect, text processing for these languages is a vital task to purge these lists, using a methodology similar to the one described in this paper. Finally, the lists used in our experiments developed for ADoBo 2025 shared task have been non-systematically reviewed so as to remove from each of them words showing (some of the) character sequences not observed natively in each given language. The subset of languages used in these experiments include Basque, Catalan, English, French, Galician, German, Italian, Japanese, Korean, Latin and Portuguese.

For Spanish, we use a fully fledged morphological analyzer covering inflection, derivation and compounding. Inflectional morphology uses a full form generator and its output is sent to a key-value store for form lookup during analysis. Productive derivation and compounding use a high level grammar-like description that is compiled to C for efficient processing. Validation of word formation rules is performed via full form DB lookup, where forms are labeled with lexical features blocking/allowing specific word formation processes.

Currently, the lexicon contains 154,871 lemma-category entries. These have been obtained from major dictionaries like *Diccionario de la lengua española* (RAE), *Diccionario de uso del español* (Gredos), *Diccionario del español actual* (Santillana), *Diccionario de la lengua española* (Espasa Calpe), *Diccionario Salamanca de la lengua española* (Santillana), but also from specialized dictionaries: *Diccionario de términos médicos* (RANME), *Diccionario Enciclopédico Ilustrado de Medicina* (Dorland), *Diccionario Español de Ingeniería* (RAING). A significant number of entries come from our own effort collecting words from the Spanish texts we process. The lexicon is the backbone of all our NLP work in Spanish, as it overcomes the limitations of wordlists while developing the strategy described in this paper.

As regards ADoBo 2025 specific guidelines for loanwords included in *DLE*, these have only partially met in our Spanish lexicon. Where available, this information, however, is used in all labeling configurations described in this paper, mainly as a blocking mechanism for certain tokens (*bebop*, considered as a *realia word*; *petabyte*, as a unit of measurement in computers).

The tools and resources described in this section are currently unpublished research.

4.4. Corpus processing

The route we take in our system implementation is that of traditional corpus-based methods, which can be characterized as hypothesis-driven. Our working hypothesis is three fold:

a) Journalistic text shows typographic clues that allow the isolation of borrowing candidates.
b) A span of text is a borrowing if it is a lexical form or construction in a donor language;
c) and it is inserted in a Spanish discourse.

In-house journal editorial recommendations rely on educated linguistic norm as defined by institutions like the Royal Spanish Academy for Spanish speaking countries. As regards the use of unadapted borrowings, there are clear typographic guidelines to be followed that can be reduced to the need to highlight, either with a font style or with certain quotation marks, these lexical items (*DPD*[7], [9, § 2.1.1], [10, *passim.*]). We use three such contexts for our corpus exploration and extraction, namely text in italics (`<hi rend="#i"></hi>` in our normalized XML format), or within left-and-right (‘’) or left-and-right-pointing double angle («») quotation marks. We only extract lowercase fragments (≤ 4 tokens) with no internal punctuation. With these constraints, we believe that this oversimplified approach to LB extraction may rarely interpret as borrowing code-switched fragments or quotations, although this point has not been verified.

[7] https://www.rae.es/dpd/comillas, d).

The XML text is scanned to isolate the relevant contexts. For each of them, the extraction program tries to identify the language it is written, using the lexical resources at our disposal. Key to this task is the use of both the language wordlists and the Spanish morphological tagger. The computing mechanism is simple: each language counter is incremented for each word in the expression found in the corresponding language wordlist (overlapping). The program is instructed to reject Spanish expressions, promote Latin in cases of tie when this language is present,[8] and finally output a (possibly ambiguous regarding its assignment to language) borrowing candidate with its frequency and concordance, if requested.

The reason for using a lexical approach for automatic LI is also threefold: (a) we are interested in real words possibly loaned to Spanish and not on predictions for non-existent and/or incorrect words, so in an ideal implementation of our system all foreign words would have been previously added to the various lexicons; (b) there are significant accuracy drops for LI systems with very short texts ([11] report on accuracy drops for average language data of only 5.1 words and a great number of languages to be evaluated; [12] recognize also this issue at the word level[9]); and (c) we believe that a lexical approach is easier to maintain when development is in its early stages.

During the scanning phase and in order to reduce as much as possible incomplete collection of borrowing candidates due to wrong highlighting of the span, in the extraction phase we allow for one character to the left and/or right of the `hi` content. As shown in the following examples, this simple additional preprocessing step will result in correct gathering of some borrowings:

> acusando a ERC de excluir casos de l<hi rend="#i">awfare</hi> de la Ley de amnistía que afectan a su formación (*La Vanguardia*, 11/06/2023, (https://www.lavanguardia.com/politica/20231106/9355779/marta-rovira-avisa-junts-amnistia-personas-debe-facilitarse-impugnacion.html).
>
> Junto a los ejemplares eróticos, el cineasta reunió unas 1.500 revistas pornográficas, especialmente las dedicadas al <hi rend="#i">bondag</hi>e, (*El País*, 01/15/2018, https://elpais.com/cultura/2018/01/15/actualidad/1516013266_129849.html).

If so desired, the program can extract spans labeled as being written in one language but where some of the words are unknown to the system. Examples of this type of span are *big pharmas*, *biogluten free*, *enclothed cognition* (where *pharmas*, *biogluten* and *unclothed* are not included in the English wordlist). However, this feature has not been used since it may introduce errors in the lexicon (*transfonning growth faetors*), so the processing of these candidates is left as future work.

Once identified as a borrowing candidate, we think that the language label a span receives with our extraction method can be projected to every context it is used in, as long as it is inserted in a Spanish discourse.

4.4.1. Issues with the candidate list

The candidate list so compiled has 42,802 English entries, but it is incomplete and/or incorrect in at least two respects, namely, it contains no borrowings whose constituting words are homographs to Spanish words, and it suffers from the same "over-merging" or "over-aggregation" problem observed in transformer-based models [13, 14]. These issues are detailed below.

The extraction procedure provides no treatment for homographs, since Spanish is promoted in case of a form homography or a tie for multi-word candidates. Hence, if a word or word sequence could belong to either language and it is used in an English context, it will be incorrectly classified by this approach and it is expected to be labeled by the predictive model (*flair-cs*), since it shows a significant improvement in recall on homographic forms ([3]). To reduce the impact of this drawback, we have carefully reviewed the 100 most frequent candidates where LI is ambiguous between Spanish and

[8] It is a common problem of wordlists available on the internet to have plenty of Latin words.

[9] Although these authors report good results in the literature they reviewed using CRFs for word-level LI.

Table 1
Sizes of our various datasets

Material	Size
News pages processed	12,840,556
Word count	6,695,555,903
Relevant contexts extracted	8,533,352
Unique relevant contexts	1,461,207
Different text strings in those contexts	1,281,384
English language strings (bulk output)	42,802
English language strings (after semi-automatic review)	37,212
Other languages strings (unambiguous, bulk)	18,897
Rest of non-Spanish strings (ambiguous, bulk)	6,618

English. Candidates like *prime time*, *total looks*, *real estate* or *late show* get its way to the borrowing lexicon after this human inspection.

Some of the spans belong to two contiguous borrowings, since they are examples of secondary predication and not of compounding. This is true, for instance, for the extracted candidate *look denim*. In order to deal with cases like this, we take advantage of the fact that Spanish is a left-headed language whereas English is right-headed and implement a simple method that read the candidate lexicon and, for each word to the right of the candidate, checks whether it is used alone (in a relevant context) and the number of entries (and its frequency) where it in a non-head position (assuming a correct English word order). This information is collected and presented to the user for revision. Using the example above, it turns out that *denim* is found alone 1,930 times and it is also used to the left in a number of candidates like *denim dress*, *denim jacket*, *denim shorts*, and less clear-cut cases like *denim over denim*. The information gathered is helpful in the human decision on keeping this candidate in the lexicon or remove it. A fragment of the top-most frequent sorted list was also manually reviewed to remove some weird candidates or add textual variants to the existing ones. The time spent in this human revision was around three days' work (20 hours), which competes favorably with that required to annotate the training data for a machine learning model and requires significantly less computational effort.

This solution avoids the systematic labeling of word sequences like *blazer oversize* or *light healthy* as a single ENG span, as it happens when tagging the COALAS corpus. Nonetheless, we are aware, as already stated, that this solution does not generalize to new LB sequences.

Note in passing that the resulting list may still contain a number of errors (*crowd founding*, *doo woo*), but this is considered a minor problem since sequence models would agree on labeling this errors as English borrowings. Nevertheless, we gather some of these errors performing a second pass over the first candidate list allowing, for each remaining candidate ($\geq$ 5 chars), a Levenshtein distance of 1 to any of the borrowings obtained in the first pass over the corpus. These candidates are linked in the lexicon to the corresponding approximate matched strings for future revision.

Table 1 shows some figures of the different resources used or generated in our current solution to the ADoBo task.

5. Experiments and Results

5.1. On the development set

We use the mid-size development set (1,836 sentences) provided by the ADoBo shared task organizers for experimentation on various configurations of our elementary system. In a first, simplistic attempt, we use exclusively our borrowing lexicon/gazetteer to unconditionally label every matching sequence in the text (`vanilla`) in order to prove our hypothesis that a given string identified as a borrowing candidate can be labeled always as such.

Table 2
`vanilla` and `flair-cs` results on the dev set

Tagger/model	Precision	Recall	F1
`vanilla`	71.12	**98.31**	82.53
`flair-cs`	**85.86**	88.47	**87.15**

We are fairly impressed by the recall of this first configuration, even though precision is, as expected, quite poor. Our high score on recall has probably to do with the overlap of the (guessed) sources used to derive the development set and the journals we used to generate our borrowing resources. However, this also suggests that alive borrowings constitute an aprioristic retrievable set.

Results for this baseline labeler and `flair-cs` model on the development set are shown in Table 2.

5.1.1. NER systems

In a 50M word subcorpus from our news texts, we label more than 2.2M NEs spanning to 3.8M tokens (7.6 per 100 tokens are part of an entity) of which 0.22M are English words overlapping with one token from our English wordlist. Person names and many other named classes are not part of this wordlist, so most of these tokens may be incorrectly matched by our `vanilla` system, which does not use any sort of contextual modeling. To avoid this undesired behavior, we perform NER on the text.

Four different models for NER have been tested. They are used in a black-box manner, being our only interest for this shared task to block the labeling of any (part of a) NE as a borrowing. For this purpose, we use `Stanza`'s pretrained NER model for Spanish.[10] `Stanza` uses a standard BiLSTM-CRF architecture at prediction and it is trained with forward and backward character-level LSTM language models. The other three models are based on the `Transformer` architecture. These are: (1) XLM-Roberta-large-NER-Spanish,[11] an XLM-Roberta-large model fine-tuned for NER with the same dataset used for the `Stanza` model (XLM); (2) Spanish BERT (BETO) + NER,[12] a Spanish BERT model fine-tuned also with the same dataset (BERT); and (3) RoBERTa-base-BNE-Capitel-NER,[13] a Spanish RoBERTa-base trained on BNE fine-tuned for CAPITEL Named Entity Recognition (NER) dataset[14] (BNE).

As Table 3 shows, `Stanza` is the most precise model for our use case, while BERT ranks best on recall in our constrained configuration, where some of the labels are ignored. Both reach its respective high scores at the expense of hurting the other. XLM, however, obtains the best F1 scores, as it is the most balanced. No improved performance is observed when our system is run with all pairs combinations of the NER models in our task.[15] Incidentally, although not tested systematically with the `flair-cs` model, when various combinations of NER systems are combined with it, performance tends to be slightly improved (same recall with BERT model as the original `flair-cs` model, but various combinations outperform this model). Combined with the BNE NER model, performance improvements of +0.46 on F1 can be obtained at the cost of a small loss in recall.

As suggested by one of the reviewers, we provide some additional information on the results obtained using `flair-cs` combined with some of the NER systems. When XLM model is also used, the number of false positives drop compared to the `flair-cs` model alone (*Hook*, *BDSM*, *organic light-emitting diode*, *street view*, among them) at the cost of producing new true negatives (*photocall*, *jet stream*, *Big Data*, *Batch Cooking*); the scores, however, slightly favor a combined system. On the other hand, in combination with BERT there is no gain with respect to the DL model alone. This is due to our default

[10] https://stanfordnlp.github.io/stanza/ner_models.html. We use the model trained with CoNLL02 corpus (https://huggingface.co/datasets/tomaarsen/conll2002), in spite of performing -0.5 worse than the one trained with AnCora corpus (https://clic.ub.edu/corpus/system/files/2022-01/ancora-es-2.0.0_2.zip).

[11] https://huggingface.co/MMG/xlm-roberta-large-ner-spanish

[12] https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner

[13] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-capitel-ner

[14] This dataset is not publicly available.

[15] When NER system outputs are combined in the ensemble oracle, we ignore the NE type, being irrelevant for our purposes, and convert all tags to inside (I) or outside (O) of an NE.

configuration with this model, which differs from out-of-the-box behavior: from past experience using this model, we know it displays a bias towards labeling MISC spans and lowercase sequences. Hence, we ignore this label and force the first token of an NE span to be uppercase. The effect of this constraint is that BERT does not contribute to improving flair-cs system. Yet, better F1 score (F1: 87.54) can be obtained with this small dev set, if NE predictions are accepted on all four NE types but the lowercase restriction is kept.

Even though, according to our experiments, performance improvements seem marginal when using an NER with flair-cs, this type of setting may open paths for further experimentation with this performant model. Note, in this respect, that our experiments have not been focused on model combinations around flair-cs. Thus, different settings might be tested in order to find out the best performing combination with the available training data.

We believe that the most outstanding fact that Table 3 reveals is that our very simple approach, paired with the XLM NER system, outperforms original flair-cs by +5.98 in F1 (+5.52 when compared with our best performing LB+NER combination), a score that should be taken with caution given the relatively small size of the development set. We use XLM NER for the rest of our experiments, in order to minimize processing load and time for the same input file.

Finally, as the high recall score of our system suggests, no performance improvement is obtained when the flair-cs predictions are also sent to the oracle to be considered as the final prediction model for unlabeled tokens.

A closer look to some of the differences when labeling the development set using flair-cs combined with XLM NER, on one hand, and vanilla with the same NER model, on the other, reveals quite interesting facts. A set of common English words like, among others, *instagramers*, *yankee*, *webcam*, *tracks*, *riff*, *offshore*, *masters*, *rallies*, *hall* (the latter two, in *DLE* in italics; we assume a regular plural for the singular *rally*) are ignored by flair-cs, while they are correctly labeled by *vanilla*[16]. This is striking as many of these words show graphotactic combinations not found in the Spanish vocabulary. Besides, *vanilla* has no major problems adhering to the nuances laid down in the ADoBo guidelines and it does not label as LBs words or expressions like *brexit* (event), *frikies* (adapted English borrowing), *rock and roll* (*realia word*), *cui prodest* (Latin expression), *zigzag* (also as separate words, *zig zag*, a French loanword documented in Spanish since the XIX century). All these expressions are false positives for flair-cs. These two cases alone explain the performance gap between both system combinations.

As described in Section 3, non-systematic labeling is another drawback of DL models, since they tend to overfit to specific examples and are sometimes unable to generalize to our trivial hypothesis for any given LB —a borrowing must be always labeled as such if found in the primary language. In this respect, flair-cs correctly labels the word *K-Pop* in the following context: "El K-Pop no tiene un género musical específico ...", but not in these two, very similar, contexts: "Sara lleva siete años estudiando coreano y el K-Pop es su pasión ...", "... el 'fandom' en el K-Pop." Given that XLM model does not consider this word an NE and it is included in the gazetteer, vanilla labels it as LB in all three contexts.

A final, paradigmatic case is that of quotes and the role they play in LB detection for flair-cs model. Only *spin* is labeled in the context "... ejerció de 'spin doctor,' y le susurró al oído ...', with an usual place for the closing quote. It seems that these quotes come from a preprocessing of the text converting the original HTML markup for italics in the newspaper webpages to single quotes for text annotation, overlooking that the rules for highlighting and quoting placement are not the same. The span is also mislabeled by the model if quotes are removed. Only when the closing quote is placed before the comma will flair-cs label the complete span as ENG, sustaining the hypothesis that a strictly alphabetic quoted environment is the most salient feature the model has learned for LB borrowing. Once more, vanilla labels this LB correctly.

[16]Some of these words were already commented upon in Section 3.

Table 3
vanilla and flair-cs + NER systems results on the dev set

Model combination	Precision	Recall	F1
vanilla + Stanza	**92.61**	89.15	90.85
vanilla + XLM	92.05	94.24	**93.13**
vanilla + BERT	86.05	**98.31**	91.77
vanilla + BNE	91.72	93.90	92.80
vanilla + Stanza + BERT	92.61	89.15	90.85
vanilla + Stanza + XLM	92.53	88.14	90.28
vanilla + Stanza + BNE	92.58	88.81	90.66
vanilla + XLM + BERT	92.05	92.24	93.13
vanilla + XLM + BNE	92.23	92.54	92.39
vanilla + BERT + BNE	91.72	93.90	92.80
vanilla + XLM + BERT + BNE	92.23	92.54	92.39
flair-cs + Stanza	**89.45**	83.39	86.32
flair-cs + XLM	88.24	86.44	87.33
flair-cs + BERT	85.86	**88.47**	87.15
flair-cs + BNE	87.76	87.46	**87.61**
flair-cs + XLM + BNE	88.81	86.10	87.44
vanilla + XLM + flair-cs	92.05	94.24	**93.13**

5.1.2. Contextual rules

Our all-context labeling approach is still a source for unwanted errors. Given that we consider LB detection as a type of word-level LI, it makes sense to take context into account for best performance. [15] argue in favor of the use of "clues in a word's context" in order to improve independent word level classification. Following their suggestion, we add some contextual knowledge to our lexical approach to LB. Only three such context types are implemented as rules and a new version of our system is developed (ctx-lex). The types implemented are quoted contexts (force a known LB in quoted context to span over the entire quoted text; otherwise, extend a partially known LB in quoted context to span over the entire quoted text if and only if added tokens are labeled as being in English); first sentence position (simply ignore NE prediction in this context); and LBs with surrounding Spanish language context (force an LB to be inserted in a context labeled as Spanish). The latter restriction is checked within the former two contexts, as well as the blocking of any span being part of an NE, if this option is active. A final non-contextual application is the default case. The quoted context described above is the only environment in which we perform dynamic identification of LB (OOV cases). The extraction of LBs without typographic clues is left as a future exercise.

Besides, leveraging word-level LI functionality, we also implement a sentence-level LI, since the input file may contain full sentences written in a language other than Spanish. For sentence-level LI, we ignore punctuation and short tokens ($\leq$ 3 chars), and increment a language counter for each token included in each of the wordlists or analyzable by the Spanish morphological tagger. If Spanish language counter is not the highest, the sentence is ignored by the LB module. However, the sentence is passed to the LB module in case of a tie among various languages including Spanish.

Table 4 shows our performance metrics for the trivial rule-based contextual implementation with and without NE span blocking. The simplest (and fastest) implementation, without NE prediction, outperforms the original flair-cs model in precision (86.57 vs 85.86) with no loss in recall. However, the best performing combination, using NE span blocking, reaches an F1 of 95.38, +8.23 improvement as compared with the original flair-cs model F1 score. Nonetheless, as with experiments with our first system, these results must be taken with caution due to the reduced size of the development set.

Also included in this table are the results obtained when combining the context-aware model with flair-cs. As it happened when testing on our non-contextual system padded with NE prediction,

Table 4
Results with our `ctx-lex` system (and `flair-cs`) on the dev set

Model combination	Precision	Recall	F1
`ctx-lex`	86.57	**98.31**	92.06
`ctx-lex` + XLM	**92.93**	97.97	**95.38**
`ctx-lex` + `flair-cs`	85.04	**98.31**	91.19
`ctx-lex` + XLM + `flair-cs`	91.75	97.97	94.75

Table 5
Results for ADoBo test set

Model combination	Precision	Recall	F1
`ctx-lex`	94.20	**95.56**	94.88
`ctx-lex` + XLM	**97.73**	82.88	89.70
`ctx-lex` - LBs with f=1	96.49	95.37	**95.93**
`flair-cs`	85.15	23.77	37.17

this combination, rather than resulting in an improved performance, slightly degrades the score.

5.2. On the test set

The shared task webpage[17] states that "[t]he dataset for ADoBo 2025 will consist of a collection of sentences from the journalistic domain written in European Spanish. Each sentence in the dataset may contain one anglicism, several anglicisms or none." A similar statement is specifically referred to the test set.[18] However, the actual test dataset consists of a small collection of sentences distorted in relation to typography (specially word casing) and peppered with multiple quotation marks. Not very representative of the well-formed news text to be expected as they are, it turns out that these lab sentences do not confuse much our purely lexical+contextual approach, although NE prediction is completely mixed up. In fact, our results are reversed, being the configuration without NE blocking the best performant with an F1 score of 94.88, boosted by our excellent score on recall. However, without NE labeling precision decreases consequently.

Since our (very incomplete) revision of the gazetteer was performed over a frequency sorted version, we can expect a wide variety of specimens at the bottom of this file. Hence, we performed another execution where *hapax legomena* (that may be considered *nonce borrowings* for our task) have been pruned from it. The edited 37,212 entry list was reduced by approximately half (19,709) following Zipfian distribution for frequency-1 words. With this new lexicon, we rerun our system, scoring our best results for the task. Table 5 shows our results for the various configurations. Scores obtained by the original `flair-cs` model evidence the challenge this set poses for any borrowing detection system.

5.3. On COALAS test set

We finally evaluate our system with an OOV-rich setting like the one posed by the COALAS test set. This feature derives, actually, from the fact that it has been created from sources other than mainstream newspapers, with the main goal of providing better assessment of generalization. The evaluation is performed exclusively for the ENG label. As the Table 6 shows, this OOV-richness is responsible for an unsurprising recall drop of the LB+NER combination, which, however, still keeps a remarkable precision. When the oracle has access to `flair-cs` predictions, F1 is improved by +2.13 with respect to our naive combination, boosting recall due to the good generalization capabilities of `flair-cs`.

[17] https://adobo-task.github.io/
[18] "The test set will consist of a collection of sentences written in European Spanish from the journalistic domain."

Table 6
Results for COALAS test set

Model combination	Precision	Recall	F1
`ctx-lex` + XLM (LB+NER)	**95.57**	86.00	90.79
`ctx-lex` + XLM + `flair-cs` (LB+NER+flair)	95.37	**90.49**	**92.92**
`flair-cs`	90.11	83.81	86.97

Table 7
Other languages candidates

Language	Candidates
Catalan (CAT)	4,601
Italian (ITA)	4,581
French (FRA)	3,742
Latin (LAT)	3,272
Portuguese (POR)	946
Galician (GLG)	833
Japanese (JPN)	455
German (DEU)	301
Korean (KOR)	104
Basque (EUS)	62

Finally, the key fact revealed by this table is that both combinations outperform the DL model used as baseline in our experimentation.

5.4. Detection of borrowings from other languages

A final fundamental advantage of our approach is the possibility to extract borrowings from languages other than English during corpus processing. Current DL models for borrowing detection show low performance on this type of borrowings, that are annotated with the generic label OTHER by [3]. These authors report for their best embedding combination an F1 score of 15.13 for this label, a very low performance as compared to the F1 of 85.76 for ENG with the same model.

As already stated, our corpus-based approach to borrowing extraction is far from fully polished, and it still leaves plenty of room for improvement, specifically in the correct assignment of donor language when the output is ambiguous. For one-word candidates, it is sometimes very difficult, if not impossible, to correctly assign a language, either because languages are typologically and/or geographical very close (*ferreiros*, or *varadoiro* are examples of ambiguity between Galician and Portuguese), or because of simple homography (*adults*, for Catalan/English ambiguity, or *jucha*, for a French/Korean example). Finally, misspelled words are definitely a source of noise in this (and any other) list extracted from running text, yet a careful inspection of this dataset can reveal new borrowings, and also allow for a better curation of language wordlists.

Table 7 shows some statistics on the size of this sublexicon in its current, bulk form. We have used this lexicon to tag COALAS full dataset (training, development and test partitions) in order to have a first impression on its quality. To allow scoring, our fine-grained language labels have been converted to OTHER for this test. Also, Latin expressions have been removed from our lexicon, since they are not labeled in this corpus. With the same configuration used for ADoBo shared task, results outperform `flair-cs` model by +66.54 (P: 83.76, R: 79.67, F1: 81.67)

6. Conclusions

We report on a naive hybrid system developed on top of a DL model for borrowing detection that outperforms the original model. The approach is very simple, with little computational complexity and no need for training data, albeit it relies on lightly curated language resources and a huge text corpus and requires a few hours of work to purge downloaded/extracted gazetteer in order to work with competitive performance. However, it has a very limited (or no) predictive capabilities, although this limitation has to be further explored in the future making a smart use of its word-level LI feature. Since its main component is symbolic and the labeling decisions are performed by a high-level oracle, its behavior shines in interpretability when compared with machine learning model architectures.

The hybrid system developed demonstrates that rule-based methods to borrowing detection can be combined with DL-based models using an ensemble oracle for improved performance and better compliance with proposed annotation guidelines.

Acknowledgments

We thank two anonymous reviewers for their valuable suggestions and typographical corrections on an earlier draft of this paper.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[2] E. Álvarez-Mellado, J. Porta-Zamorano, C. Lignos, J. Gonzalo, Overview of ADoBo at IberLEF 2025: Automatic Detection of Anglicisms in Spanish, Procesamiento del Lenguaje Natural 75 (2025).

[3] E. Álvarez-Mellado, C. Lignos, Detecting unassimilated borrowings in Spanish: An annotated corpus and approaches to modeling, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3868–3888. URL: https://aclanthology.org/2022.acl-long.268.

[4] O. Agarwal, A. Nenkova, The utility and interplay of gazetteers and entity segmentation for named entity recognition in English, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3990–4002. URL: https://aclanthology.org/2021.findings-acl.349/. doi:10.18653/v1/2021.findings-acl.349.

[5] S. Magnolini, V. Piccioni, V. Balaraman, M. Guerini, B. Magnini, How to use gazetteers for entity recognition with neural models, in: L. Espinosa-Anke, T. Declerck, D. Gromann, J. Camacho-Collados, M. T. Pilehvar (Eds.), Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5), Association for Computational Linguistics, Macau, China, 2019, pp. 40–49. URL: https://aclanthology.org/W19-5807/.

[6] L. Zhang, X. Nie, M. Zhang, M. Gu, V. Geissen, C. J. Ritsema, D. Niu, H. Zhang, Lexicon and attention-based named entity recognition for kiwifruit diseases and pests: A Deep learning approach, Frontiers in Plant Science 13 (2022) 1053449. URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.1053449/full. doi:https://doi.org/10.3389/fpls.2022.1053449.

[7] T. Meng, A. Fang, O. Rokhlenko, S. Malmasi, GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 1499–1512. URL: https://aclanthology.org/2021.naacl-main.118/. doi:10.18653/v1/2021.naacl-main.118.

[8] B. Fetahu, A. Fang, O. Rokhlenko, S. Malmasi, Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2777–2790. URL: https://aclanthology.org/2022.naacl-main.200/. doi:10.18653/v1/2022.naacl-main.200.

[9] Real Academia Española, Asociación de Academias de la Lengua Española, Ortografía de la lengua española, Espasa, 2010.

[10] El País, El Libro de estilo de El País, Aguilar, 2021.

[11] F. Xia, W. Lewis, H. Poon, Language ID in the context of harvesting language data off the web, in: A. Lascarides, C. Gardent, J. Nivre (Eds.), Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Association for Computational Linguistics, Athens, Greece, 2009, pp. 870–878. URL: https://aclanthology.org/E09-1099/.

[12] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin, K. Lindén, Automatic language identification in texts: A survey, 2018. URL: https://arxiv.org/abs/1804.08186. arXiv:1804.08186.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[14] F. Li, Z. Lin, M. Zhang, D. Ji, A span-based model for joint overlapped and discontinuous named entity recognition, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 4814–4828. URL: https://aclanthology.org/2021.acl-long.372/. doi:10.18653/v1/2021.acl-long.372.

[15] B. King, S. Abney, Labeling the languages of words in mixed-language documents using weakly supervised methods, in: L. Vanderwende, H. Daumé III, K. Kirchhoff (Eds.), Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 1110–1119. URL: https://aclanthology.org/N13-1131/.