Impact of Spanish Dialect in Deep Learning Next Sentence Predictors

Pablo Duboue
Textualization Software Ltd.
Vancouver
British Columbia
CANADA

***Abstract*—With the second largest number of native speakers in the world, Spanish exhibits a large dialectical difference, both in grammar and word forms. We analyze the impact of these differences in a Natural Language Processing task using the proceedings of two parliamentary bodies from Costa Rica and Argentina. The task chosen is to determine whether to pieces of text are coherent, that is, whether they can follow each other in normal discourse. We train a deep learning network to perform next sentence prediction using the BERT model. Our experiments indicate that task adaptation presents a more substantive impact than regional differences but dialect still can account for up to 8% difference in F-measure.**

***Index Terms*—Natural Language Processing, Dialect Differences, Coherence Assessment, Deep Learning, BERT, Civic Applications of AI**

I. Introduction

The Cambridge Advanced Learner's Dictionary [1] defines a dialect as

> a form of a language that is spoken in a particular part of a country or by a particular group of people and that contains some words, grammar, or pronunciations that are different from the forms used in other parts or by other groups.

The key difference between dialects and languages is that speakers of different dialects of the same language find their utterances to be **mutually intelligible**, that is, to be understandable back and forth.[1] As Spanish is the world's second-most spoken native language [3], it presents itself with a wide variety of dialects. While these dialects all pass the mutual intelligibility test with human hearers and speakers, it begs the question of whether the dialectical differences impact computer systems negatively. In this work, we have sought to answer this question using the language as expressed in the parliamentary proceedings (floor discussion transcriptions) from Costa Rica and Argentina.

To study the impact, we use the task of predicting textual coherence, that is, taking two contiguous text segments, distinguish them from two segments of text that are not contiguous. We evaluate the impact using BERT models for Natural Language Processing [4].

As we work with written language, the differences in pronunciation are not our concern, but the difference in grammar and, particularly, in word forms might have a rather large impact in NLP tasks, specially if pre-trained embeddings diverge significantly from one dialect to the other. The question we are trying to elucidate in this work is:

> Can pre-trained models for Spanish perform well across dialectical differences?

This question arises as a result in our involvement with the Voz y Voto project, that seeks to analyze the impact of mandated equal representation in the discourse at parliamentary bodies. All the data and scripts employed in this paper are available for download at the project's website:

http://vozyvoto.ie4opendata.org

By narrowing to parliamentary discussions, we focus on a text genre that can exhibit more dialectical elements, in a controlled environment. Still, there is the question of whether differences between regions in the same country might be greater than differences between countries [5]. Moreover, each person has their own dialect (known as an *idiolect*) and that will also impact the data. Parliamentary discussions have the advantage of multiple speakers and large amount of available data. We hope this will help iron out the regional/personal differences and allow to study the differences in national dialects.

By comparing a general model trained on Wikipedia versus country-specific model, we found that there is a much bigger difference between the genre than dialect but that regional differences can still account for 8 percentage points in F-measure. Moreover, the directionality matters, a model trained on Argentinian Spanish seems to perform better overall than one trained on Costa Rican Spanish. We plan to further explore this finding by adding more dialects.

This paper is structured as follows. In the next section, we will discuss related work on dialect impact in NLP processing and impact of sub-languages in NLP tasks. The experiments performed in this paper rely on a crawl of the proceedings of the parliaments of Argentina and Costa Rica discussed in Section III. The task we analyze in this work, textual coherence using the BERT model is presented in Section IV. Section V showcases our experiments and analysis.

II. Previous Work

We will look at related work both in dealing with dialects in NLP and the impact of sub-languages (of which dialects are

[1]The mutual aspect is key, as Portuguese speakers are said to be able to understand Spanish but the opposite is seldom true [2].

particular type) for specific tasks. Also, as our project involves civic applications of NLP technology, we will briefly touch upon this in Section II-C.

A. Dialects and NLP

Dealing with dialects is also further complicated because identifying the dialect is much harder than language identification [6], [7]. The most spoken languages also have the most dialects. We will look into some languages with large communities of speakers in turn: English, Arabic, Chinese, Spanish and Portuguese.

1) English: Interestingly, English dialects present quite a variation in pronunciation but lexical and grammatical variations are less confined to regional but rather social groups. The impact of dialectical pronunciation differences in automatic transcription for captioning has been studied by Rachel Tatman in the context of Ethics in NLP [8]. She found robust differences in accuracy for transcription of speakers from Scotland and the Deep South of US.

Working on analyzing the impact of language as used by different social groups, Blodgett and O'Connor found that tweets written by African American are less likely to be identified as English by existing tools and online APIs [9].

2) Arabic: Arabic is an amalgamation of local dialects that are mutually intelligible within geographic neighbours but the farther away the communities of speakers are, the more difficult the communication arises [10]. There is still the concept of Standard Arabic as dictated by the media, and most NLP tools cater to that dialect [11]. However, processing local dialects (like Moroccan, Egyptian or Saudi Arabian) poses multiple challenges [12], given the fact that research work has been earmarked to any of specific 20 dialects studied in the literature. Given such diversity, dialect identification is key to usable spoken interfaces [13].

3) Chinese: Chinese is unique in that a writing system based in ideograms allows non-mutually intelligible dialects to remain dialects through a common written form. The differences among dialects in written form has been investigated by Zhang [14] using corpus linguistics and statistical methods, allowing for a visual representation of dialectical differences. Reproducing such work in Spanish is among our long term interests.

While the writing system is not phonetic, that is not entirely true for transliterating foreign words, which exhibit dialectical and regional differences. Their impact in entity clustering was studied by Peng and colleagues [15] who found that using exact matching resulted on a 16% drop in F-measure.

As the language spoken natively by the larger number of speakers in the world, it has many dialects, identifying them correctly [16] poses many Chinese-specific challenges which render traditional character n-gram methods unsuccessful. The state of the art uses point-mutual information and an assortment of features to achieve 82% accuracy distinguishing among the main six dialects. That speaks of the complexity of the task.

Finally, as some dialects, like Mandarin and Cantonese, do not pass the mutual intelligibility test, they are more apt to be called separate languages with resources including parallel corpora [17], [18].

4) Spanish: Closer to the present work, Spanish has also been the focus of dialect analysis, particularly rural dialects [19].

Working on Multi-Word Expressions (MWE), Bogantes and colleagues [20] analyzed the impact of four Latinamerican dialects (Colombia, Costa Rica, Mexico and Peru). For example, "to be out of money" in Costa Rica is expressed by the MWE "estar limpio" (lit., to be clean) while in Colombia, that meaning is captured by "estar pelado" (lit., to be naked). Similarly, in the present work we address two dialects (Argentina and Costa Rica), but we are interested in expanding to other dialects. They manually categorized 40 such MWEs.

Finally, working on dialect and author identification, Sanchez-Perez and colleagues [21] worked on eight varieties of the Spanish language (Argentinian, Mexican, Colombian, Chilean, Venezuelan, Panamanian, Guatemalan, and Peninsular Spanish) as presented in news articles. They found that a combination of character n-grams and lexical features attained the best performance in dialect identification.

5) Portuguese: There has been considerable work on dialect-specific word embeddings for Portuguese [22], [23], together with Portuguese-specific dialect differences [24].

B. Impact of sub-language for NLP tasks

The impact of sub-language for NLP tasks has received particular attention in domain specific areas of NLP. The impact for NLP tasks of the domain from where the text is extract is a well studied topic, particularly for tasks such as Part-of-Speech tagging [25] and domains such as the medical domain [26], [27], [28] and software engineering [29]. This could be particularly important for sub-problems such as negation detection [30].

The domain impact is particularly crucial when dealing with languages with fewer annotated resources [31]. This problem is also crucial in Deep Learning [32].

Domain differences can be used to generate "ungrammatical" sentences for a parser trained on a different domain, therefore analyzing the performance of a parser under the presence of grammatically different sentences [33].

The lack of adaptability for models trained on limited corpora is often mentioned as a problem for commercial deployment of NLP systems [34].

C. Civic applications of NLP technology

NLP technology helps reshape the political debate by allowing crowd-sourced deliberation [35] and feedback collection [36]. Digital simulations of the democratic process are also useful to engage citizens in political discourse [37].

On the other hand, parliamentary proceedings are a staple of NLP research [38], [39], [40].

We are interested in analyzing political debates, similar to the work of Onyimadu and colleagues [41] that used sentiment

analysis on parliamentary debates over the Canadian hansard. Our concern about the dialectical differences for the Spanish language triggered the current work. Analysis of political discourse is a staple of both political science [42], [43], [44] and linguistics [45], [46], [47].

1) Voz y Voto: The ongoing Voz y Voto project is a study of group participation in the proceedings of government assemblies, both in terms of speaking and being addressed by other speakers. The current demo was put together at a hackathon called Hack(at)ONG in Cordoba, Argentina in September, 2016. It analyzes participation of female representatives in the congress floor, a topic of interest for the FUNDEPS NGO[2] which studies the impact of mandatory gender quotas. Its long term vision is to allow for experimentation using different subgroups (user-defined) of representative assemblies from other countries and government levels.

The system is a combination of shell scripts, pandoc, iconv, perl and OpenNLP. The gender identification component is based on the Spanish Wikipedia list of unambiguous male and female names. The entity linking (person disambiguation) component uses a heuristic point system, attributing a phrase identified as a person mentioned by OpenNLP to one of the representatives present. Topic identification per session uses TF*IDF. A live web-graph using plot.ly can be seen at http://vozyvoto.ie4opendata.org/demo.html, also in Figure 1.

III. Data: Parliamentary Proceedings

The data for the experiments presented in this paper are two historical crawls of the Argentinian (Honorable Camara de Diputados[3]) and Costa Rican (Asamblea Legislativa Republica de Costa Rica[4]) congress.

The total dataset sizes are listed in Table I. While Costa Rica had more data available, it was much difficult to crawl, as it was behind a Rich Internet Application that necessitated a headless browser[5] to successfully crawl it.

The text from the crawls was then fed through spaCy [48] sentence boundary detector, using their provided es_core_news_sm model for Spanish (10Mb in size). This corpus is available at the IE4OpenData project.[6] The parliamentary proceedings include a variety of data, not only speech transcripts. Other data include topics to be discussed on each session, attendance lists and, the case of the Costa Rican data at least, the full text of all proposed legislation. As we are only interested on transcribed speech data, a heuristic using formatting was devised, specific to each data source. This heuristic was more successful extracting sentences in the AR source than in the CR source. Therefore, even if the AR source covers less years, it has more usable data. The extracted speech is done in segments, with all sentences in a segment being spoken by the same person as one reported, contiguous unit. To assemble the training and test data for the next sentence prediction task, we sample contiguous sentences in-segment (isNext=true) and sentences in separate segments (isNext=false).

Besides the dialect-specific data, we also used 1M sample sentences from the Spanish Wikipedia to train a base system. We used the cirrussearch Wikipedia dump, that contains the text templates already expanded.

An example of the corpus is shown in Figure 2.

TABLE I
CRAWLED DATA

Metric	Argentina	Costa Rica
Raw data (incl. formatting)	161Mb	382Mb
Years	2001-2016	1993-2016
Textual data		
Words	2M	64M
Spoken transcripts		
Segments	17,986	9,965
Sentences	25,133	19,099
Avg. Sent. per Segment	1.40	1.92
Words	1.12M	1.24M
Avg. Words per Sent.	44.95	65.30

For the Voz y Voto project, we also crawled attendee sheets and use gender identification information to obtain an analysis of the number of attendees and speakers per session, per gender as shown in Figure 1.

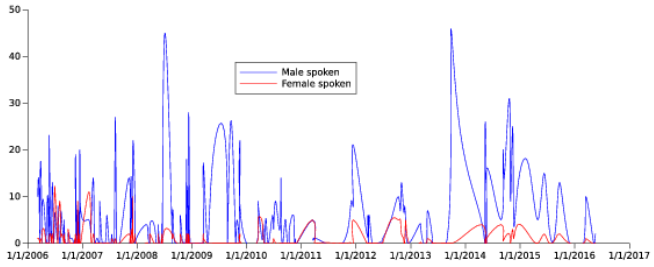

Fig. 1. Representatives participation by gender in the Argentinian chamber of deputies

IV. Next Sentence Prediction

The task we are using to study dialect difference is one of the family of so called "self-supervised tasks," that is, tasks where labels can be obtained directly from text without the need of annotators. The particular task is next sentence prediction: given two sentences, determine whether they appear contiguous to each other in text, or they are far apart or from different documents. This task is a surrogate for textual coherence and has been successfully employed to train sentence orderings for multi-document summarization [49].

Moreover, the newly available BERT [4] pre-trained models use this task as their main pre-training objective. Therefore, we can profit from BERT performance without heavy fine-tuning. As our system relies heavily on BERT, we will discuss it succinctly. It is based on CNNs with positional encodings and a deep attention mechanism [50], as part of the Transformers formalism [51]. A BERT model is trained to predict with a

[2]http://www.fundeps.org/en/home/
[3]https://www.diputados.gov.ar/
[4]http://www.asamblea.go.cr/SitePages/Inicio.aspx
[5]https://phantomjs.org
[6]http://ie4opendata.org

masked language model (i.e., to predict missing words explicitly hidden during training) for a training instance consisting of a class pseudo-token, [CLS], a sequence of tokens, a separator pseudo-token, [SEP], and a second sequence of tokens. The tokens are represented using Google's WordPiece tokenizer [52], that splits words based on frequencies (marked with '##' in the examples). During BERT training, the embeddings are learned based on the auxiliary task, predicting the isNext value using the embedding of the [CLS] token as input through an extra dense layer of size 768 (this layer is not distributed with the embeddings, sadly) plus predicting the masked words (these two predictions is what drives the model to learn the embeddings). Therefore, the embeddings for the [CLS] token are thus trained to predict the isNext class.

The BERT model is pre-trained for days on clusters of high-end GPUs with 64Gb of RAM, over billions of words. Such models are made available for download by Google LLC. In this work, we are using the multilingual cased model released in Nov. 2018 (659Mb). This single model handles English, Chinese, Spanish and a variety of other languages.

The two sequences of tokens in BERT input represent two sentences or text segments that can be contiguous (on which case the class value isNext is set to true) or not (isNext is false). With fine-tuning, the two sequences of tokens can be made to be a question and an answer or a piece of text and its paraphrase. Such fine-tuning currently requires access to GPUs with 64Gb of RAM. Luckily, as we are building upon the original isNext task, we did not need to resort to such fine-tuning.

Instead, we use the BERT model as a feature extractor. In the original BERT paper, the authors compare the BERT model to ELMO feature extractors [53] and found them to be relatively similar, with BERT producing better results. The recommended approach is to produce a BERT embedding (of 3,072 dimensions) per token and then train a bidirectional LSTM over them [54]. This is very onerous in terms of storage space (it will require about a terabyte of space for each of the systems used in the current experiments). Luckily, the embedding for the class pseudo-token [CLS] captures the isNext task and we only need to store 3,072 floating point values per pair of sentences we want to run.

These embeddings contain all the information to do isNext prediction but a second system needs to be trained on top of them. We thus arrive to the system used in this work:

1) Take pairs of sentences, some contiguous (isNext = true), some not (isNext = false), using a 50% sampling for each class.
2) Compute BERT embeddings for each pair (3,072 dimensions).
3) Train a network to predict isNext. We used a dense layer sized 64 using GeLU activation [55] (similar to BERT auxiliary task[7]) and a single neuron as output, using sigmoid activation. The network is trained for 30 epochs using the Adam optimizer and a binary cross-entropy loss function.

As our training data is very small compared to the size of BERT embeddings, we apply a heavy dropout with a keep probability of 25%. Otherwise the results are too unstable. The final system takes two sentences and determines how likely are they to follow each other, a surrogate for a coherence metric.

Conceptually, this is akin to take a Wikipedia sentence s_1 in page P_1 and take the next sentence s_2 in page P_1. Then train to say that the pair $\langle s_1, s_2 \rangle$ is coherent. And taking a sentence s'_1 in page P_2 and a sentence s'_2 in page P_3 and set to train that $\langle s'_1, s'_2 \rangle$ is not coherent. In the next section we will discuss how we trained the three different systems (Wikipedia, AR and CR) using this approach.

V. Experiments

Using the data described in Section III, we split 20% of the sessions for a held-out evaluation as test test (6.2k sentences for AR, 3.6k sentences for CR). From the held-out we sampled 10k sentence pairs for each country.

Over the remaining training data (18.8k sentences AR, 15.5k sentences CR), we sampled 100k sentence pairs for each country.

Finally, we took the sample of 1M sentences from Spanish Wikipedia and created 100k sentence pairs.

All the pairs had 50% isNext=true and 50% isNext=false pairs. For each of these five sets (Wikipedia train 100k, AR train 100k, CR train 100k, AR test 10k, CR test 10k), we extracted the 3,072 BERT embedding using a 6Gb GPU. From the train datasets, we use the isNext labels and embeddings to train an isNext detector, obtaining three trained systems. The training was done for 30 epochs. As Wikipedia is part of the BERT training data, it achieved a validation accuracy of 97%. The AR model achieved a validation accuracy of 93% and the CR of 85%. The average training pair in Wikipedia has 118 tokens, while AR has 93 tokens and CR has 133. We believe the difference in length makes the CR data more challenging. This might be an artifact of the transcription process or a difference in the level of verbosity between the two speaker communities.

We then proceed to run each of the three models over the two evaluation sets, with the results in Table II. Earlier experiments were very unstable so we run five different training runs, reporting the mean and the maximum boundary that encompasses all five runs as taken around the mean.

This table volunteers some important results but it also highlights questions that will need further data and experiments to be elucidated. First, the difference between using a general model (Wikipedia) vs. a genre specific model is crucial: using a generic model on Argentina results in a drop from 83 F-measure to 68 in Argentina and from 80 to 69 in Costa Rica, a relative difference of 17.3% and 12.9%, respectively. The drop from using a custom model is much higher, in the case of Argentina of 21 points of absolute F-measure performance. Also, while the numbers have some variability

[7]In the BERT auxiliary task, a dense layer of size 768 is used, but for our training data such network resulted in unstable results.

between runs, it is possible to conclude that Costa Rica is closer to standard Spanish as embodied by Wikipedia, by about 0.8% F-measure difference (relative, computed over the higher and lower boundaries 68.89 for Argentina and 69.49 for Costa Rica). This result is non-obvious, as the Costa Rica dataset has longer sentences and produced much lower validation accuracy when training.

The validation results while training track the performance on unseen data quite well (85% to 83 F-measure in Costa Rica and 93% to 90 F-measure in the case of Argentina). Interestingly, when applied to the Argentinian data, the Costa Rican model produced results indistinguishable from the model applied to Costa Rican data. It thus exhibits no dialectical differences. It is not a very good model, but it seems to be equally poor on the two datasets. The Argentinian model, however, it is 8% better (relative, 7.55 absolute) on Argentinian data than on Costa Rican data. That means that for Argentinian data, it pays off to use a custom model, but for general use, a model trained on a more neutral dialect is to be preferred. But these are preliminary results that need more dialects for contrast. Also, this methodology is very sensitive to changes in the training data and corpus cleaning.[8] Adding more dialects should help further the understanding on the impact.

Are these differences due to differences in language or other differences? To help start answering that question, we looked into a sample of the 306 cases where the AR data had isNext=true and the AR model predicted it correctly **but** the CR model did not.[9] A clear pattern from that data is the fact that speakers in the Argentinian congress like to address their words to the president of the chamber, including calls to action in the middle of a speech thus addressed directly to the president's chamber. We believe that might be confusing to models where the speakers do not use this style. These are particularities of a group of speakers that we believe qualify for dialectical differences. More traditional differences include the use of the lexical items such as the verb "desandar" (to go back, to undo), which is not present on the CR data, the word "irrisorio" (risible) that is present twice as much in AR data than in CR data, the word "zapatilla" (sneaker) that is called very differently in other areas of Latinamerica, the word "patronal" (company management, the opposing party in labour union negotiations), "conurbano" (metropolitan area), not present in CR data and "corralito" (small corral) which refers to a banking freeze specific to Argentina. Some examples deemed dialectical by an Argentinian native speaker are shown in Figure 3.

To make these intuitions more systematic, we turned to a simple domain specificity for term extraction [56] using the 100k training sentences for each country as foreground model and the 100k sentences from Wikipedia as background model. For all nouns, we computed the ratio of their foreground frequencies over the background frequencies. We kept the nouns with a domain specificity above 3.0 (about 3,000 per country). We also computed bigrams of noun/adjective pairs where one of the nouns appear in the list of country-specific nouns but the results were not as informative as the nouns. Many of the nouns that differ in distribution from the background were not country specific, just parliamentary discourse specific. We thus proceed to further filter them as country specific or parliamentary specific by focusing on the top 1,000 terms for each country and only considering as country specific the ones that appear only in one country list or, if they appear in the other country list, they have a 10-fold difference in score. This totalled 316 joint terms, 761 Argentinian terms and 807 Costa Rican terms.

The top terms are shown in Table III. From the table, we can see that the terms that diverge from the background distribution and appear at both are clearly of a conversational, legislative type. The table for Argentina shows some political terms specific to Argentina but also dialect-specific word use: "modificatorio" as "thing that modifies something else" and "apartamiento" as "the action of separating," other dialects might prefer "separación" (separation) but it might be the term has a precise legal meaning if Argentina. The Costa Rica list is an excerpt as there seems some all uppercase titles have leaked into the data and confound spaCy lemmatizer, including many proper nouns as regular nouns. For a speaker of a different dialect, the spelling of "tal vez" (maybe) as one word, and "posposición" (the act of postponing) and "irrespeto" (the act of engaging in a lack of respect) seem unusual and dialectical.

The table also includes a diminutive form for the word "poco" (a little something), "poquito" (a very little something). Costa Ricans are known to be very fond of diminutives, there are 22 terms ending in -ito, -ita marked as Costa Rican-specific terms (compared to 7 for Argentina). It is reassuring to see this phenomenon captured in our data.

Finally, we took the 814 base forms associated with the 686 Argentinian terms and see how often did they appear on the whole test set: 93.67% of the test sentence pairs contained one of the 814 forms. We then looked into the 306 sentence pairs that the Argentinian model found but were missed by the Costa Rican model and analyzed whether they had a higher percentage of Argentinian-specific terms. We found they have a 95.75% presence, giving further evidence that dialect was at least a partial culprit to the observed differences. Similarly, analyzing the 3,667 sentence pairs from the Argentina test set that both Argentina and Costa Rican models got correct, the incidence of Argentinian specific went down to 89.99%.

If the difference observed is not due to dialectical differences, the key question is: why did the Costa Rican model on Costa Rican data performed below the Argentinian model on Argentinian data? From Table I, we know that it has much longer sentences. The impact of this phenomenon can be ablated by splitting the segments into other type of chunks, for example, elementary discourse units (EDUs) [57]. Such approach was outside the scope of the current work. Similarly, it should be possible to establish by analyzing other regional texts whether they exhibit longer sentence length on average than texts from Argentina (comparing, for example, newspaper

[8]As the heuristics to isolate spoken speech improved, we saw multiple changes to the table.

[9]This analysis was done on an earlier run of the system.

TABLE II
EVALUATION RESULTS

Training Data	Evaluation Set	Prec	Rec	F1
Wikipedia	Argentina	52.83 ± 0.28	98.18 ± 0.23	68.70 ± 0.19
Wikipedia	Costa Rica	54.14 ± 0.44	98.12 ± 0.31	69.78 ± 0.29
Argentina	Argentina	88.46 ± 1.22	93.01 ± 2.48	90.62 ± 0.56
Argentina	Costa Rica	72.16 ± 2.58	90.00 ± 3.26	80.07 ± 0.38
Costa Rica	Argentina	83.74 ± 1.77	82.44 ± 2.76	83.07 ± 0.65
Costa Rica	Costa Rica	86.39 ± 3.06	81.14 ± 3.38	83.65 ± 0.33

articles). It is unclear whether such differences would fall into what is normally considered dialectical differences. Another possibility is that the Argentinian data has some particularity that allows the model to easily outperform a general case. Again, adding more dialects will help further understand the boundaries of this task and build more informed expectations.

VI. CONCLUSIONS

In this work, we have looked into the impact of dialects in the isNext prediction using parliamentary proceedings as training data. We found that task-specific differences might be more important than dialect, accounting for up to 21 points of absolute F-measure difference. Nevertheless, dialects can also have an impact, accounting for up to 7 points of absolute F-measure difference. It also opens up the question of which type of dialect is to be preferred for training "generic" models, a topic that will necessitate experimentation with more Spanish dialects.

A. Further work

In future work, we would like to expand this research to other Spanish dialects.[10] We are interested in furthering the visualization of dialectical differences, perhaps using [14] work as a starting point. Other methodological improvements include changing the base model from Wikipedia (encyclopedic text) to movie subtitles (spoken transcripts) [58].

We are also interested in incorporating new techniques for domain adaptation appearing in the field [59] and techniques to measure other types of biases on word representations [60].

Our goal is to use the isNext detector to gauge consensus in a two party system by looking at how compatible the debates from one party are when compared to the opposite party.

ACKNOWLEDGEMENTS

We would like to acknowledge to the original Voz y Voto team, Annie Ying and Mauricio Korach. To Dr. Javier Sanchez and Dennys Gajdamaschko, for encouragement and discussion. To the people at FUNDEPS, particularly Virginia Pedraza and to the people at TEC Costa Rica, particularly Prof. Eddy Ramirez. Finally, to the anonymous reviewers for CLEI, for useful discussion and ideas.

[10]We hope to have Panamanian Spanish in time for CLEI.

REFERENCES

[1] C. McIntosh, Ed., *Cambridge Advanced Learner's Dictionary*. Cambridge University Press, 2013.

[2] A. Akmajian, A. K. Farmer, L. Bickmore, R. A. Demers, and R. M. Harnish, *Linguistics: An introduction to language and communication*. MIT press, 2017.

[3] I. Cervantes, "Informe 2018. el español: una lengua viva." Instituto Cervantes, Madrid, Tech. Rep., 2018.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] M. L. Murphy, "British english? american english? are there such things?: It's hard to talk about "national" englishes," *English Today*, vol. 32, no. 2, pp. 4–7, 2016.

[6] A. M. Ciobanu and L. P. Dinu, "A computational perspective on the romanian dialects." in *LREC*, 2016.

[7] W. Radford and M. Gallé, "Discriminating between similar languages in twitter using label propagation," *arXiv preprint arXiv:1607.05408*, 2016.

[8] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 53–59. [Online]. Available: https://www.aclweb.org/anthology/W17-1606

[9] S. L. Blodgett and B. O'Connor, "Racial disparity in natural language processing: A case study of social media african-american english," *CoRR*, vol. abs/1707.00061, 2017. [Online]. Available: http://arxiv.org/abs/1707.00061

[10] M. Diab and N. Habash, "Natural language processing of arabic and its dialects," in *Tutorial at EMNLP 2014*. ACL.

[11] K. Shaalan, S. Siddiqui, M. Alkhatib, and A. A. Monem, "Challenges in arabic natural language processing," *Computational Linguistics, Speech And Image Processing For Arabic Language*, vol. 4, p. 59, 2018.

[12] A. Shoufan and S. Alameri, "Natural language processing for dialectical arabic: A survey," in *Proceedings of the Second Workshop on Arabic Natural Language Processing*, 2015, pp. 36–48.

[13] Y. Lei and J. H. Hansen, "Factor analysis-based information integration for arabic dialect identification," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4337–4340.

[14] Z.-s. Zhang, *Dimensions of variation in written Chinese*. Routledge, 2017.

[15] N. Peng, M. Yu, and M. Dredze, "An empirical study of chinese name matching and applications," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, vol. 2, 2015, pp. 377–383.

[16] F. Xu, W. Mingwen, and M. Li, "Sentence-level dialects identification in the greater china region," *International Journal on Natural Language Computing*, vol. 5, pp. 9–20, 12 2016.

[17] T.-s. Wong, K. Gerdes, H. Leung, and J. Lee, "Quantitative comparative syntax on the cantonese-mandarin parallel dependency treebank," in *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 2017, pp. 266–275.

[18] F. Xu, M. Wang, and M. Li, "Building parallel monolingual gan chinese dialects corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[19] C. de Benito Moreno, J. Pueyo, and I. Fernández-Ordóñez, "Creating and designing a corpus of rural spanish," in *Procceedings of the 13th Conference on Natural Language Processing (KONVENS)*, 2016, pp. 78–83.

TABLE III
COUNTRY AND DOMAIN SPECIFIC TERMS, WITH DOMAIN SPECIFICITY SCORES

spaCy Lemma	Gloss	Example	Specificity Score
BOTH			
referirme	*to emphasize a topic*	'referirme'	922.50
quórum	*quorum*	'quórum'	522.50
moción	*formal proposal*	'Moción', 'mociones', 'moción'	125.50
ponernos	*set ourselves up to*	'ponernos'	210.50
votarlo	*to vote something*	'votarlo'	186.00
permitirme	*allow myself to*	'permitirme'	183.00
apelar	*to invoke*	'apelo', 'apela', 'apelé'	165.50
inciso	*section*	'inciso', 'incisos', 'incisas'	160.23
oficialismo	*official party*	'oficialismo'	158.44
decirnos	*to tell us*	'decirnos'	151.50
Argentina			
solicitar	*to request*	'solicito', 'solicitadas', 'solicitados', 'solicité'	1294.00
renegociación	*renegotiation*	'renegociaciones', 'renegociación'	843.00
interbloque	*multy-party agreement*	'interbloque'	779.00
modificatorio	*amending*	'modificatoria', 'modificatorias', 'modificatorios'	622.00
alícuota	*debt rate*	'alícuotas', 'alícuota'	584.00
coparticipación	*tax provincial share*	'coparticipación'	532.00
afirmativo	*affirmative*	'afirmativos', 'afirmativa'	449.00
planteos	*ideas presented*	'planteos'	394.00
apartamiento	*separation*	'apartamientos', 'apartamiento'	331.25
informarse	*to inform oneself*	'informarse'	285.34
Costa Rica			
colón	*colon (CR currency)*	'colones', 'colón'	1574.50
posposición	*act of posponing*	'posposiciones', 'posposición'	1288.00
talvez	*maybe*	'talvez'	1055.00
costarricense	*relative to CR*	'costarricense', 'costarricenses'	751.77
señoría	*your highness*	'señoría', 'señorías'	639.17
poquito	*a little bit*	'poquita', 'poquitas', 'poquitos', 'poquito'	597.00
cafetín	*small coffee*	'cafetín'	595.00
irrespeto	*lack of respect*	'irrespeto'	554.00
solidarismo	*solidarity*	'solidarismo'	503.00
portillo	*small door*	'portillo', 'portillos'	338.00

[20] D. Bogantes, E. Rodríguez, A. Arauco, A. Rodríguez, and A. Savary, "Towards lexical encoding of multi-word expressions in spanish dialects," in *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[21] M. A. Sanchez-Perez, I. Markov, H. Gómez-Adorno, and G. Sidorov, "Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2017, pp. 145–151.

[22] E. R. Fonseca, J. L. G. Rosa, and S. M. Aluísio, "Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese," *Journal of the Brazilian Computer Society*, vol. 21, no. 1, p. 2, 2015.

[23] N. Hartmann, E. Fonseca, C. Shulby, M. Treviso, J. Rodrigues, and S. Aluisio, "Portuguese word embeddings: evaluating on word analogies and natural language tasks," *arXiv preprint arXiv:1708.06025*, 2017.

[24] D. W. Castro, E. Souza, D. Vitório, D. Santos, and A. L. Oliveira, "Smoothed n-gram based models for tweet language identification: A case study of the brazilian and european portuguese national varieties," *Applied Soft Computing*, vol. 61, pp. 1160–1172, 2017.

[25] J. P. Ferraro, H. Daumé III, S. L. DuVall, W. W. Chapman, H. Harkema, and P. J. Haug, "Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 931–939, 2013.

[26] R. Baud, A.-M. Rassinoux, and J.-R. Scherrer, "Natural language processing and semantical representation of medical texts," *Methods of information in medicine*, vol. 31, no. 02, pp. 117–125, 1992.

[27] A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P. Duboue, W. Weng, W. J. Wilbur *et al.*, "Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data," *Journal of biomedical informatics*, vol. 37, no. 1, pp. 43–53, 2004.

[28] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell, "Extracting information from the text of electronic medical records to improve case detection: a systematic review," *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 1007–1015, 2016.

[29] A. Ferrari, B. Donati, and S. Gnesi, "Detecting domain-specific ambiguities: an nlp approach based on wikipedia crawling and word embeddings," in *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2017, pp. 393–399.

[30] T. Miller, S. Bethard, H. Amiri, and G. Savova, "Unsupervised domain adaptation for clinical negation detection," in *BioNLP 2017*, 2017, pp. 165–170.

[31] B. Plank, "What to do about non-standard (or non-canonical) language in nlp," in *In Proc. KOVENS 2016*, 2016.

[32] L. Qu, G. Ferraro, L. Zhou, W. Hou, N. Schneider, and T. Baldwin, "Big data small data, in domain out-of domain, known word unknown word: The impact of word representations on sequence labelling tasks," in *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 2015, pp. 83–93.

[33] H. B. Hashemi and R. Hwa, "An evaluation of parser robustness for ungrammatical sentences," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1765–1774.

[34] D. Dahlmeier, "On the challenges of translating nlp research into commercial products," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 92–96.

[35] T. Aitamurto, K. Chen, A. Cherif, J. S. Galli, and L. Santana, "Civic crowdanalytics: Making sense of crowdsourced civic input with big data tools," in *Proceedings of the 20th International Academic Mindtrek Conference*. ACM, 2016, pp. 86–94.

[36] J. Whittle, W. Simm, M.-A. Ferrario, K. Frankova, L. Garton, A. Woodcock, J. Binner, A. Ariyatum *et al.*, "Voiceyourview: collecting real-time feedback on the design of public spaces," in *Proceedings of the 12th*

ACM international conference on Ubiquitous computing. ACM, 2010, pp. 41–50.

[37] K. D. Poole, M. J. Berson, and P. Levine, "On becoming a legislative aide: Enhancing civic engagement through a digital simulation," *Action in Teacher Education*, vol. 32, no. 4, pp. 70–82, 2010.

[38] M. Carpuat, "Mixed language and code-switching in the canadian hansard," in *Proceedings of the first workshop on computational approaches to code switching*, 2014, pp. 107–115.

[39] S. Wattam, P. Rayson, M. Alexander, and J. Anderson, "Experiences with parallelisation of an existing nlp pipeline: Tagging hansard." in *LREC*, 2014, pp. 4093–4096.

[40] K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, and Y. Den, "Balanced corpus of contemporary written japanese," *Language resources and evaluation*, vol. 48, no. 2, pp. 345–371, 2014.

[41] O. Onyimadu, K. Nakata, T. Wilson, D. Macken, and K. Liu, "Towards sentiment analysis on parliamentary debates in hansard," in *Joint international semantic technology conference*. Springer, 2013, pp. 48–50.

[42] C. Ilie, "Insulting as (un) parliamentary practice in the british and swedish parliaments," *Cross-cultural perspectives on parliamentary discourse*, vol. 10, p. 45, 2004.

[43] B. L. Monroe and P. A. Schrodt, "Introduction to the special issue: The statistical analysis of political text," *Political Analysis*, vol. 16, no. 4, pp. 351–355, 2008.

[44] P. Rasiah *et al.*, "Can the opposition effectively ensure government accountability in question time?: an empirical study," *Australasian Parliamentary Review*, vol. 25, no. 1, p. 166, 2010.

[45] J. Bara, A. Weale, and A. Bicquelet, "Deliberative democracy and the analysis of parliamentary debate," in *Workshop on Advanced Empirical Study of Deliberation*, 2007, pp. 1–47.

[46] P. Rasiah, "A framework for the systematic analysis of evasion in parliamentary discourse," *Journal of Pragmatics*, vol. 42, no. 3, pp. 664–680, 2010.

[47] C. F. Rodríguez, "Cortesía e imagen en las preguntas orales del parlamento español," *Cultura, Lenguaje y Representación/Culture, Language and Representation*, vol. 9, no. 9, pp. 53–79, 2011.

[48] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, 2017.

[49] L. Logeswaran, H. Lee, and D. Radev, "Sentence ordering and coherence modeling using recurrent neural networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[51] L. Logeswaran and H. Lee, "An efficient framework for learning sentence representations," in *Proc. of ICLR2018*, 2018.

[52] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[53] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[54] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[55] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[56] Y. Park, S. Patwardhan, K. Visweswariah, and S. C. Gates, "An empirical analysis of word error rate and keyword error rate," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[57] H. Schauer, "From elementary discourse units to complex ones," in *Proceedings of the 1st SIGdial workshop on Discourse and dialogue-Volume 10*. Association for Computational Linguistics, 2000, pp. 46–55.

[58] M. MÃŒller and M. Volk, "Statistical machine translation of subtitles: From opensubtitles to ted." in *GSCL*, ser. Lecture Notes in Computer Science, I. Gurevych, C. Biemann, and T. Zesch, Eds., vol. 8105. Springer, 2013, pp. 132–138.

[59] J. Lee, R. Sarikaya, and Y.-B. Kim, "Locale-agnostic universal domain classification model in spoken language understanding," *arXiv preprint arXiv:1905.00924*, 2019.

[60] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring Bias in Contextualized Word Representations," in *Proceedings of the First ACL Workshop on Gender Bias for Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, 2019.

Argentina

[CLS] Yo qui ##ero rei ##vind ##ica ##r a la mayoría de los ju ##ece ##s , que son pro ##bos , que trabaja ##n , que se dedica ##n , aunque como en todo ámbito donde participa el ser humano también existen actitud ##es pat ##ológicas , raya ##nas en las irregular ##idades , y para esto es buen ##o que exist ##an instrumentos para que estas conducta ##s sean [SEP] La necesidad de que un poder que parece ai ##sla ##do en relación con la participación popular se am ##pl ##í ##e , se debat ##a y se discu ##ta ya constituye un valor en sí mismo , más allá de los dis ##ens ##os y disc ##re ##pan ##cias que pod ##amos tener . [SEP] (isNext = true)

I would like to defend most judges that are exemplar, working hard and devoting themselves, even though there exist also, as in any human endeavour, almost illegal pathological attitudes, thus is good to have tools to have such behaviour... [truncated] ||| *The need for a government body that appears to be isolated from popular involvement to become wider and to debate and discuss it, that on itself is valuable, beyond our potencial disagreements.*

[CLS] Señor presidente : sin án ##imo de quer ##er rei ##tera ##r la reunión de la Comisión de Labor Parlament ##aria , deb ##o decir que en realidad fu ##i mos muchos los presidente ##s de bloque que nos mani ##festa ##mos en esa dirección . [SEP] Lam ##enta ##blem ent ##e esta com ##puls ##ión que ten ##emos los argentino ##s de llegar tarde siempre a todo no sólo se rep ##ite esta vez ; además , ahora corre ##mos el riesgo de perder una nueva oportunidad de tomar un tema tan violenta ##mente dolor ##oso . [SEP] (isNext = false)

Mr. President: without recounting the last parliamentary group meeting, I want to say we have been many caucases presidents that expressed ourselves that way. ||| *Sadly, there is this compulsion we, the Argentinians, have to arrive always late. That repeats itself here, moreover, we risk to miss a new opportunity to deal with a violently painful topic.*

Costa Rica

[CLS] Y la Sala Constitucional nos indica lo siguiente : hace quince días , en el presente as ##unto , el re ##cu ##rren ##te ad ##uce el in ##cum ##pli ##miento a lo ordenado por la Sala mediante la sente ##ncia número 2013 - 003 ##69 ##1 de las once horas treinta minutos del 15 de marzo del 2013 , toda vez que [SEP] Y en este pronuncia ##miento de hace quince días de la Sala Constitucional el por tanto es el siguiente : se le rei ##tera al ministro de Hacienda , He ##lio Fall ##as Vene ##gas , que pro ##ced ##a al cum ##pli ##miento de lo dis ##puesto en la sente ##ncia número 2013 - 003 ##36 ##91 , de las once [SEP] (isNext = true)

And the constitutional chamber remind us of the following: fifteen days ago, on the current topic, the appealer calls for the lack of action as mandated by setence number 2013-0036911 of the eleven hours, thirty minutes of March 15th, 2013, and such that [truncated] ||| *And in such sentence from fifteen days ago, the constitutional chamber, it follows that: it repeats itself to the minister of economic affairs, Helio Fallas Venegas, that he ought to abide by the sentence number 2013-0033691 of the eleven [truncated]*

[CLS] En realidad el ex ##ped ##iente 16 . 662 , no ha sido enviado a publicación todavía , por lo que le soli ##cita ##mos que ins ##tru ##ya us ##ted a Servicio ##s Parlament ##arios , a efectos de que sea enviado a la menor breve ##dad , a efectos de su publicación porque no ha sido en ##viada la publicación . [SEP] Es para ap ##rove ##char que el día de hoy en la Fra ##cción del Partido Liber ##ación Nacional tu ##vimo ##s la oportunidad de tener a varios grupos indígenas junto con el alcalde de Buenos Aires , José Rojas , estuvieron ex ##pon ##iendo la situación que viven ellos en Bu ##eno Aires [UNK] [SEP] (isNext = false)

Actually, the file 16,662 has not been sent for publication yet, therefore we ask the parliamentary services to send them for publication as soon as possible, so it can be published as it has not been sent for publication ||| *This is to inform that today in the fraction of the Liberacion Nacional party, we had the opportunity of receiving several indigenous people groups together with the mayor of Buenos Aires city, Jose Rojas. They were presenting their current living conditions in Buenos Aires*

Wikipedia

[CLS] El juego fue convertido al 3D ##O el mismo año , con una exclusiva banda sonora re ##mez ##cla ##da , la cual fue a ##ña ##dida a las otras com ##pila ##ciones de la fra ##n ##qui ##cia Street Fighter incluyendo la versión de con ##sola del Hy ##per Street Fighter II . [SEP] Fue más tarde convert ##ida al PC , Sega Dream ##cast como parte del Capcom ' s Match ##ing Service la cual permitió combates online , PlayStation y Sega Saturn como parte de la Street Fighter Collection , y PlayStation 2 y Xbox como parte de la Capcom Classics Collection Vol . [SEP] (isNext = true)

Super Turbo was originally ported to the 3DO Interactive Multiplayer, followed by the PlayStation and Sega Saturn (under the title of Super Street Fighter II Turbo: The Ultimate Championship) as part of the Street Fighter Collection, and for the Dreamcast in Japan under the title of Super Street Fighter II X for Matching Service. A remake of the game was released for the PlayStation 3 and Xbox 360 titled Super Street Fighter II Turbo HD Remix.

[CLS] En el Anti ##guo Egipto se in ##venta ##ron pro ##ced ##imientos para ela ##bor ar pan con leva ##dura , así como vino y ce ##rve ##za . [92] Como en tantos otros terrenos , fue en la Antigua Grecia donde se sent ##aron las bases de la gas ##tron ##om ##ía como ciencia , y fueron los creador ##e s de los re ##cet ##arios y la literatura gas ##tron [SEP] En principio , los griegos despre ##ciar ##on el p esca ##do como comida de pobres , pero alrededor de los siglos III y II a . C . se rev ##alo ##riz ##ó , pasando posteriormente a la cocina romana , que valor ##ó mucho el pesca ##do y el mari ##sco . [SEP] (isNext = true)

In Ancient Egypt, they invented methods for breadmaking using yeast, together with vine and beer. as in many other areas, Ancient Greece was the base for gastronomical science and the creators of recept books and literature ||| *Initially, Greeks despised fish as poor people's food, but around the second and third century BC, the fish took a new value. This phenomenon ontinues with Roman cuisine, where both fish and sea food where highly valued.*

Fig. 2. Example training data

- (0.06568766) En varios medios de comunicación totalmente objetivos, que no distorsionan la información y se expresan sin ningún sentimiento o corrimiento ideológico o propagandístico, como Página/12 o 6, 7, 8, se ha señalado que por culpa de las patronales o de los gremios o de la burocracia el empleado rural es el peor pago. ||| Es cierto, el empleado rural hoy no percibe grandes salarios, y podría estar gozando de un 37,5 por ciento de aumento de convenio colectivo de trabajo, entre la patronal y los empleados, pero no es así porque el Ministerio de Trabajo no homologa más que un 25 por ciento.
  In many objetive media sources that do not distort information nor express an ideology or propaganda, such as Pagina/12 or 6,7,8 it has been pointed out that due to the owners or the labour unions or bureocracy, the rural worker is the worse paid one. ||| That is true, the rural workers today do not receive large salaries and they could be enjoying a 37.5% improvement on their collective bargaining agreement between owners and workers, but that is not happening because the ministry of labour does not want to go beyong 25%.

- (0.00884765) Particularmente, un tema de actualidad -que merece un análisis más exhaustivo, toda vez que la Cámara deberá expresarse sobre su contenido- está vinculado con el decreto 905, referente a la salida de los llamados "corralito" y "corralón" y al plan que se ha elaborado para ello. ||| Por intermedio de las comisiones de Presupuesto y Hacienda y de Finanzas se ha invitado al secretario de Finanzas o al ministro de Economía para que suministren la información que todos deseamos conocer.
  In particular, a topic of current affairs –that deserves an in deep analysis, each time this chamber discusses the topic– is related to decree number 905, related to the exit of the so called "corralito" and "corralon" and the plan thus drafted. ||| Through the comissions for budget and finances, we have invited the finance secretary or the economic minister to provide the information we are all eager to find learn about.

- (0.2188364) Se transfieren riquezas al exterior despojándonos de la cultura del trabajo, cercenando oportunidades a las nuevas generaciones y también agrandando las asimetrías entre las regiones porque en el puerto quedan las ganancias, fruto de un intercambio desigual con quienes nos venden sus productos manufacturados con trabajo y tecnología incluidos. ||| Señor presidente: no es casual que tengamos casi 2 millones de desocupados, 40 por ciento de trabajadores en negro, miles de subsidios y seguros de desempleo, 30 por ciento de trabajadores con ingresos por debajo de sus necesidades e indigentes arracimados en los conurbanos a merced del clientelismo.
  Untold riches are being transferred outside the country, stripping us away from the culture of work, destroying the opportunities for future generations and increasing the regional asymmetries as the profits remain in the sea port, as a result of an unequal exchange with the ones that sell us manufactured goods with work and technology included ||| Mr. President [of the chamber]: it it not a surprise that we have almost 2 million unemployed people, 40% undeclared workers, thousand of subsidies and unemployment benefits, 30% of workers with salaries below their needs and destitute people hoarded in the peripheria falling prey to populism.

Fig. 3. Dialectical AR examples correctly classified by the AR model and missed by the CR model (number between parenthesis is the probability predicted by the CR model)