Code-Mixed Probes Show How Pre-Trained Models Generalise On Code-Switched Text

Leon, Frances A. Laureano De; Madabushi, Harish Tayyar; Lee, Mark

DOI:
10.48550/arXiv.2403.04872

Code-Mixed Probes Show How Pre-Trained Models Generalise On Code-Switched Text

Frances A. Laureano De Leon*, **Harish Tayyar Madabushi**†, **Mark Lee***
*University of Birmingham, †University of Bath
*School of Computer Science, Birmingham, UK
†Department of Computer Science, Bath, UK
fxl846@cs.bham.ac.uk, htm43@bath.ac.uk, m.g.lee@bham.ac.uk

Abstract

Code-switching is a prevalent linguistic phenomenon in which multilingual individuals seamlessly alternate between languages. Despite its widespread use online and recent research trends in this area, research in code-switching presents unique challenges, primarily stemming from the scarcity of labelled data and available resources. In this study, we investigate how pre-trained Language Models handle code-switched text in three dimensions: a) the ability of PLMs to detect code-switched text, b) variations in the structural information that PLMs utilise to capture code-switched text, and c) the consistency of semantic information representation in code-switched text. To conduct a systematic and controlled evaluation of the language models in question, we create a novel dataset of well-formed naturalistic code-switched text along with parallel translations into the source languages. Our findings reveal that pre-trained language models are effective in generalising to code-switched text, shedding light on the abilities of these models to generalise representations to CS corpora. We release all our code and data, including the novel corpus, at https://github.com/francesita/code-mixed-probes.

Keywords: code-switching, probing language models, multilingualism

1. Introduction

Code-switching (CS) is the phenomenon in which multilinguals effortlessly alternate between languages in the same conversation or piece of writing (Joshi, 1982; Dogruoz et al., 2021). CS arises in multilingual communities over the world, such as the United States, Latin America, and India, and gives way to the emergence of mixed 'languages' such as Hinglish (Hindi-English mix) and Spanglish (Spanish-English mix). The recent adoption of Pre-trained Language Models (PLMs) has been driven in part by their ability to gain a significant amount of linguistic information (Clark et al., 2019; Tenney et al., 2019a) and world knowledge (Petroni et al., 2019) based purely on the pre-training. A significant question is how much information PLMs can gather about the meaning of words from being trained on text alone (Bender and Koller, 2020). CS data is especially useful in helping to answer this question due to the multilingual nature of CS text. The presence of multiple languages in the text will prevent the model from relying on spurious statistical correlations when generating meaning because (1) the presence of multilingual text will encourage the model to learn meaning across languages, and (2) the model needs to learn the context of each language switch, which may prevent it from using simple patterns more easily found in monolingual text due to language-specific patterns. Hence, the semantic representations of CS data provide us a way of exploring the true extent to which models are able to capture and generalise meaning.

Despite the potential significance of exploring CS data in evaluating PLMs, research in this area is challenging, not least due to the lack of labelled data and resources (Santy et al., 2021; Aguilar et al., 2020). As such, in this work, we focus on how PLMs interact with and encode code-switched text. To ensure a balanced evaluation of models, we look at both real and synthetic CS text and focus exclusively on Spanglish (Spanish-English). We choose Spanglish for 3 reasons: (1) Spanish and English share a script, (2) English words share many Spanish cognates. This overlap is due to a portion of English vocabulary having Latin roots, and Spanish having originated from Colloquial Latin, (Nagy et al., 1993), and (3) although there are differences in Spanish and English grammar, such as word-order, gender and number, there also are overlaps in the structure of both languages (Rivera, 2019). These similarities ensure that our evaluation of model capabilities are not confounded by other aspects of language use, such as, for example, difference in script as in the case of Hinglish. We utilise synthetic data in our experiments to investigate whether the presence of mixed language text alone leads to satisfactory probe performance, or whether the use of naturalistic CS examples significantly impact experimental results.

To evaluate the extent to which PLMs can identify and encode the correct representations of CS text,

we focus our efforts on three different dimensions: a) the ability of PLMs to *detect* CS text, b) possible variations in the grammatical structural that PLMs are able to capture from CS text, and c) consistency of the meaning representations of CS text when compared with monolingual text. Our experimental results along these lines indicate that PLMs have the potential to capture all three of these dimensions with reasonable exactness. While further experiments are required in this regard, our findings seem to indicate that PLMs are surprisingly good at generalising across CS text, which could shed light on the potential of PLMs to capture some generalisations pertaining to language use.

To this end, we perform multiple experiments, largely using probes, to evaluate each of these different dimensions. We probe popular pre-trained models: mBERT (Devlin et al., 2019), and XLM-RoBERTa base (XLM-R-base) (Conneau et al., 2020), which consist of 12 layers and 768 dimensions, and XLM-RoBERTa large (XLM-R-large) (Conneau et al., 2020) with 16 layers and 1024 dimensions. We begin by exploring relevant literature associated with CS text, probing, graph edit distance and existing datasets in section 2 before then describing the construction of a well-balanced corpus of CS text that we create due to existing limitations of availability of such data in section 3. We then detail our experiments in each of these directions in Sections 4, 5, and 6, where we present our methods and results. We follow by a discussion of these results in Section 7 and provide a summary of our findings and suggestions for future work in Section 8.

1.1. Contributions

Given the importance of research in CS, this work makes the following contributions: We create the first curated dataset of well-formed, naturalistic instances of Spanglish CS data with translations for both source languages to allow for a precise evaluation of grammatical structure and sentence meaning. We perform extensive experiments to determine the extent to which PLMs can detect CS text and capture both the structure and meaning associated with CS text. Additionally, we extend our manually curated dataset with synthetic data to allow for ablation studies which include various controls such as the mix of languages in CS text. We provide a template for future experimental verification of linguistic theories pertaining to CS based on the usage-based principle of language acquisition.

2. Related Work

In this section we discuss related research associated with CS data, probes and methods of comparing language structure.

2.1. Code-switching and data generation

As previously mentioned, CS has become more available thanks to the rise of social media and multilingual users (Winata et al., 2023). To facilitate research, datasets and evaluation benchmarks, such as LinCE, have been created in an effort to have a centralised evaluation platform for code-switching, (Aguilar et al., 2020). LinCE combines ten corpora covering four different code-switched language pairs and four tasks. Khanuja et al. (2020) also provide a generalised CS benchmark that is inspired by GLUE known as GLUE-CoS. Despite these efforts, research in the domain remains challenging due to reasons mentioned in Section 1. Our work in introducing a novel dataset is aimed at addressing this shortcoming. This has led to growing research in synthetic data generation for CS text, which motivates us to expand our manually curated dataset with synthetic data, see Section 3. Some of the techniques employed to generate synthetic CS data in previous works include: (1) Identification and replacement of noun-phrases in monolingual sentences with the translation of that phrase in the other language pair to be studied (Salaam et al., 2022). (2) Generation of CS text containing randomly selected languages to create a CS example containing switching in multiple languages (Krishnan et al., 2021). This data was used to create a model referred to as "modified mBERT", which is trained on synthetic and real code-switched data and then tested on NLI in Hinglish. (3) The use of models trained on CS text generation (Winata et al., 2019; Rizvi et al., 2021). We use the first two of these methods to augment our dataset with synthetic data.

Many of the synthetic data generation methods are inspired or driven by CS grammar theories developed in the field of linguistics (Bullock and Toribio, 2009; Sebba et al., 2012). There are two CS theories that take precedence within NLP, the Equivalence Constraint theory (EC), in which language switches occur when the surface structures of languages align (POPLACK, 1980) and Matrix Language Frame (MLF) model, in which one language is dominant and determines the syntax of a CS phrase (Joshi, 1982; McClure, 1995). Although there are other grammar theories explaining CS, these are the most used in NLP for the creation of synthetic data. EC theory states that alternations between languages occur when the surface structures of the languages align, therefore the grammar rules of both languages are obeyed. Broadly, the MLF theory holds that in CS sentences, there is a matrix language and an embedded language. The matrix language is that which provides the grammatical structure that accommodates words or phrases from another language (Dogruoz et al.,

2021). Our exploration of the manner in which the syntactic information pertaining to CS text is encoded in PLMs is driven by this theoretical work in linguistics. Given that there are competing theories explaining the use of CS languages, our experiments are designed to evaluate if the grammatical structure of CS data extracted by PLMs is independent of either source languages, see Section 5.

2.2. Probes

In this section, we introduce literature related to probes, which we use extensively in this work. Probes, also known as auxiliary or diagnostic classifiers (Adi et al., 2017), have been developed to investigate linguistic properties encoded in text representations (Tenney et al., 2019b). They have been used for extrinsic exploration, in which a machine learning model is used to determine whether a linguistic structure in present in representations through performance on a task such as named entity recognition (Hennigen et al., 2020) and intrinsic exploration, which looks to evaluate representations on benchmarks regarding the relationship between words or sequences (Lab et al., 2020). Probes have been used for a number of years and have been largely used to analyse morphological, semantic and syntactic language properties (Dalvi et al., 2019). Probes are necessarily simply classifiers used to predict a property of some input text (Adi et al., 2017) based on the representations generated by a model, and often consist of a linear layer, or multilayer perceptron on top of frozen representations. Generally, the word or sentence representations studied are frozen, in order to prevent further training of the representations and are used as the embedding inputs for the probe classifier. As the representations are frozen, if a probe classifier learns to predict the property it was trained on, it is an indication that there is a linear mapping between the internal representations of the model and the required output and so an indication of that property being embedded within the model.

Works relevant to us in the field of probing is the syntactic structural probe by Hewitt and Manning (2019), in which they find that syntax trees are embedded in a deep models' representations. This work is expanded on by Chi et al. (2020) , who use the structural probe and find that syntactic features overlap between languages, which agrees with universal dependencies' taxonomy in mBERT. Chi et al. (2020) also find that the structural probe most effectively recovers tree structure from the 7th or 8th mBERT layer, and that a maximum rank beyond 64 or 128 gives no further gains. Tenney et al. (2019b) introduce a framework they call "edge probing", which provides a uniform architecture across tasks. They use the edge-probing technique to do layer-wise explorations of the BERT model, in which they find that basic syntactic information appears earlier in the network, and high-level semantic information appears at the higher layers (Tenney et al., 2019a).

Prior work probing the syntactic structure of CS text has been limited: Pires et al. (2019), as part of their study, use a POS dataset to probe mBERT on code-switched text. A more detailed probe study was done by Santy et al. (2021), in which synthetically generated and real code-mixed data are used to probe mBERT. They compare the probe results for different tasks, such as POS, NER, LID to the fine-tuned version of the model trained for that task. They find that using synthetically generated data in certain tasks yield lower results than using naturally occurring code-switched data.

2.3. Syntax and Graph Edit Distance

An important aspect of our work is in evaluating the similarity of syntactic structure extracted by probes. In this section, we review relevant work pertaining to the comparison of such structures. Graph Edit Distance (GED) is a metric commonly used for structural pattern recognition and analysis of graphs (Gao et al., 2010). GED is used on dependency parses, where the parses are represented by unordered directed trees in order to filter out sentence pairs that cannot be compared syntactically (Kroon et al., 2019). Kroon et al. (2019) utilise this method for the massive automatic syntactic comparison of languages. Unordered graphs make it so that the GED algorithm is more robust between different languages, which is a reason they find GED to be a good technique for syntax comparison between different languages. They favour the use of parallel corpora for automatic comparisons because it facilitates finding the contexts in which differences in syntax occur (Kroon et al., 2019).

2.4. Existing CS Datasets

In this section, we discuss existing datasets for Spanglish text. Although CS datasets are generally scarce, Spanglish is a popular language pair, in which some CS data can be found (Winata et al., 2023). Many of the publicly available datasets are from shared tasks, such as CALCS workshops (Winata et al., 2023). Some of the most used data in research include language identification (LID) data from a shared task in 2016 by Molina et al. (2016), SentiMix 2020 sentiment analysis dataset by Patwa et al. (2020), and datasets for part-of-speech classification (POS) (AlGhamdi et al., 2016) and named entity recognition (NER) (Aguilar et al., 2018) created for shared tasks. All these datasets consist of tweets, apart from the POS dataset, which is derived from the Miami Bangor Corpus, and consists of bilingual and CS conversations from four

speakers. This dataset is annotated with Universal POS tags by Soto and Hirschberg (2017). All the aforementioned datasets contain LID labels. There is also a machine translation dataset for Spanglish available that was created for CALCS 2021, but this dataset does not contain parallel translations (Chen et al., 2022). All of these datasets are available on the LinCE website [1].

As far as we know, there is no available naturalistic Spanglish dataset that includes translations for BOTH source languages. Such a dataset is essential for conducting a systematic and controlled evaluation of the PLMs under investigation. Hence, we create such a dataset, which stands as one of our contributions to the research.

3. Dataset Creation

Due to the absence of naturalistic datasets containing Spanglish data with associated parallel Spanish and English, we construct a novel dataset to address this shortcoming, see table 1.

3.1. CS Data Collection

We collect CS data from X, previously known as Twitter, using the techniques described in De Leon et al. (2020). This method uses a keyword file that contain the most commonly used words in one of the language pairs (i.e. Spanish). We use the top 100 most frequent words used in Spanish according to the Dictionary of the Royal Spanish Academy [2], and filter out words that contain 4 letter or less to prevent overlap with other languages and remove articles and pronouns. To ensure a CS output, the search query should specify the other language pair to be studied (i.e. English). We select a random subset of posts from the collected tweets to be part of the CS dataset that we use to test our probes. A person fluent in Spanish and English helped check this subset of tweets for real occurrences of CS in Spanglish and to discard any unusable or incoherent posts. These CS posts were then translated into Spanish and English by a speaker of both languages, with the aid of Google Translate API [3]. We ultimately obtain a total of 316 posts after quality checks and translations. A subset of this collected data is used as part of our syntax and semantic experiments. Specifically, we choose examples containing intra-sentential CS, the type of CS in which language alternations happens within a sentence. Intra-sentential instances of CS are essential to observe with confidence the interaction between two grammars (Joshi, 1982), and are, therefore, key for the syntax experiments, see Section 5. We gather 254 intra-sentential examples to use for syntax experiments, and refer to these as ***r-CS*** to denote that they are real instances of CS. These 254 examples were chosen on the basis of whether they were instances of intra-sentential CS. We remove hashtags, links from these examples. The examples were also re-written by bilinguals in Spanish and English, in order to create well-formed sentences, which is challenging to find in social media. View table 2 for examples of the original posts, the edited CS text and translations into source languages.

3.2. CS data generation

We utilise the parallel translations of the CS data we collect (***r-CS***) to generate synthetic CS data using two different techniques found in literature, random replacement of a token in either of the language pairs (Krishnan et al., 2021), and the noun-phrase replacement technique (Salaam et al., 2022). The source data to generate the synthetic examples come from the English and Spanish translations of the ***r-CS*** dataset. For the random generation method, we tokenize the examples, and randomly choose whether that token should be translated or not. If translated, that token is replaced by the translation. For the noun-phrase synthetic dataset, we follow a 3-step process as described in Salaam et al. (2022). (1) Noun-phrase identification, we use the spaCy library to do this [4], (2) translate the noun-phrase into the desired language, (3) replace the correct span with the translated noun-phrase. We generate noun-phrase synthetic examples with both Spanish monolingual and English monolingual translations of our ***r-CS***, to ensure we have examples with majority Spanish (***NP-CS-es***) and majority English tokens (***NP-CS-en***).

4. Detection

We use probes to conduct a layer-wise exploration of the PLMs in order to find if models are able to differentiate between monolingual and CS input. These experiments fall under (1) sentence classification, in which we train probe classifiers to differentiate between monolingual and CS sentences, and (2) an LID task, in which a probe is trained to detect the natural language of a token, given a CS sentence. These experiments are designed to understand whether PLMs have access to source language information in processing CS data, and if so, we wish to determine if the information pertaining to language varies between the layers of different language models.

4.1. Methods

For the experiments dealing with sentence and token classification (LID), we use the following CS datasets: SentiMix 2020 Patwa, Parth and

[1] https://ritual.uh.edu/lince/datasets
[2] https://corpus.rae.es/frec/5000_formas.TXT
[3] https://pypi.org/project/googletrans/
[4] https://spacy.io/

dataset	# tokens	en	es	other	ne	unk
Real CS data	4302	2174 (50.53%)	1397 (32.47%)	657 (15.27%)	73 (1.69%)	1 (0.02%)
Random CS data	4649	2039 (43.85%)	1867 (40.16%)	662 (14.24%)	80 (1.72%)	1 (0.02%)
En noun phrase synthetic CS data	4233	1338 (31.61%)	2145 (50.67%)	673 (15.90%)	76(1.79%)	1(0.02%)
Es noun phrase synthetic CS data	4640	2411 (51.96%)	1456 (31.38%)	657 (14.16%)	116 (2.5%)	0

Table 1: Created CS datasets: en stands for English, es for Spanish, other for largely punctuation, ne for named-entities and unk for unknown.

Original post	Edited post	Spanish translation	English translation
siempre me dicen que no sea tan inseguro, i'M tRyInG mY bESssT.	Siempre me dicen que no sea tan inseguro, I'm trying my best.	Siempre me dicen que no sea tan inseguro, Estoy tratando.	They always tell me not to be so insecure, I'm trying my best.
NO HABIA VISTO QUE HE WAS ALMOST SHIRTLESS https://t.co/d5tKtpzPMw	No había visto que he was almost shirtless.	No había visto que estaba casi sin camisa.	I hadn't seen that he was almost shirtless.
@rcknatsu first u gotta inhalar el aire hacia los pulmones	First you gotta inhalar el aire hacia los pulmones.	Primero tienes que inhalar el aire hacia los pulmones.	First you gotta inhale the air into your lungs.

Table 2: Examples of original posts collected from X, and minimal editions and translations.

Aguilar, Gustavo and Kar, Sudipta and Pandey, Suraj and PYKL, Srinivas and Gambäck, Björn and Chakraborty, Tanmoy and Solorio, Thamar and Das, Amitava (2020), and CALCS 2016 LID dataset Chen et al. (2022). Additionally, we use monolingual datasets in Spanish and English created for the ProfNER 2021 shared task Miranda-Escalada et al. (2021). The ProfNER dataset consists of tweets in Spanish and English. We use the ProfNER data together with the SentiMix data to create a balanced dataset containing text in Spanish (es) (4,000 examples with label 0), English (en) (4,000 examples with label 0), and CS (8,000 examples with label 1). This way we are sure to have balanced classes for training the probe classifier. We use an 80-10-10 split to train, validate and test the classifier. This combined dataset is used on the *sentence classification* task, in which we train a probe classifier to distinguish between monolingual and code-switched sentences.

For the LID *token classification task*, we use two datasets, CALCS 2016 LID dataset, and SentiMix 2020, which contain language ID tags for each token. The possible labels for the LID task are *lang1* (en), *lang2* (es), *other*, *ne* (named entities), *fw* (a language different from *lang1* and *lang2*), *mixed* (partially in both languages), *unk* (unrecognizable words), *ambiguous* (either one language or another) (Aguilar et al., 2020). For the LID task, we train probes separately on the datasets to see how probe performance changed, if at all. In the CALCS 2016 dataset, 7,986 examples contain both source languages in the same sentence. There are 21,030 train examples in this dataset, meaning that 38% of the data the probe was trained and tested with contained true instances of intra-sentential CS. The SentiMix dataset, on the other hand, contains 11,783 intra-sentential CS examples. In total, 96% of the examples used to train, validate and test probes on the SentiMix dataset contain instances of CS. For both the sentence and token classification tasks, we report the average F1 score across 5 seeds for each layer and model. For each probe, we use a batch size of 32, and learning rate of 1e-3.

4.2. Results

The results of the detection experiments are displayed in figures 1 and 2. The probe results indicate that PLMs are, in general, able to distinguish between CS text and monolingual text. Interestingly, for the sentence classification task, CLS pooling for XLM-R-base causes the probe to struggle with the sentence classification task, although by the latter layers (10, 11) the F1 score begins to match that of the other probe classifiers corresponding to those layers. This could be because the CLS token may not fully capture information relating to the differences in multiple languages, while mean pooling considers the entire input sequence, thereby capturing the differences in languages better. On the other hand, we can see that XLM-R-large CLS pooling is effective for the task, indicating, perhaps, that models with more parameters are able to encode this information in the CLS token. Overall, it seems that for the base models, the mean pooling strategy is more effective than using the CLS token, likely because mean pooling allows us to consider the full input sequence.

For the LID task, our results indicate that PLMs seem to have language information at the token level embedded within them from early layers in the models, see figure 2. This indicates that PLMs may have encoded knowledge on features such as vocabulary or morphology for different languages. Given that the probe classifiers achieve high F-1 scores for both datasets, SentiMix2020 and CALCS 2016, it may be the case that this information is used throughout all layers. In our experiments, mBERT seems to struggle when compared to the other models on the SentiMix dataset. This could

Category	Experiment	Explanation	Aim
Detection	Sentence Classification	Train probe classifiers for each PLM and layer to detect whether a sentence is monolingual or code-switched.	Find if models can distinguish between monolingual and CS sequences
Detection	Language Identification (LID)	Train probe classifiers for each PLM per layer to learn the language ID of tokens of CS text	Find if models can distinguish between all the languages in a CS input at the token level.
Syntax	Dependency parse from structural probe	Train a structural probe to extract the dependency parse of sentences in English and Spanish. The probe is used on CS data and the translations.	Study the structures of CS input and compare them with the structure of the monolingual translations.
Semantics	Semantic Text Similarity (STS)	Fine-tune PLMs on STS task in Spanish and English, which assigns a score on the similarity of two texts. Use CS data and Spanish and English data to get scores on different language pairs and sentence pairs.	Determine whether PLMs are consistent in encoding meaning of CS text compared to monolingual representations.

Table 3: Summary of tested dimensions and associated conducted experiments.

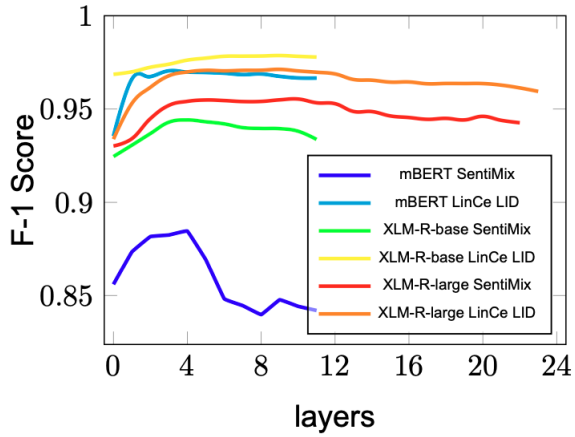

Figure 1: Mean F-1 Scores across layers for the sentence classification task for each of the PLMs studied. In this task, probe classifiers learn to distinguish between CS and monolingual text.

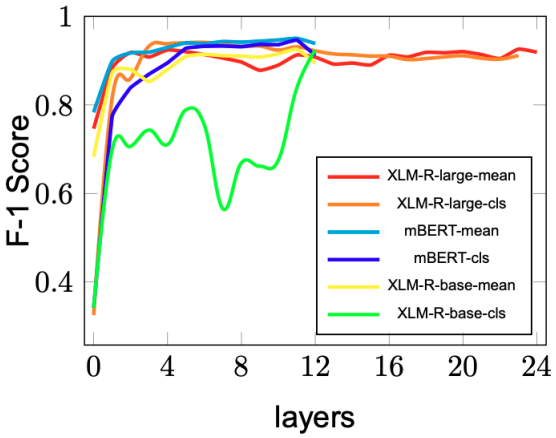

Figure 2: LID model mean F-1 Scores across layers for the probe classifiers. In this task, probe classifiers learn the LID of the tokens in CS sentences.

be due to a combination of two things: (1) XLM-RoBERTa generally outperforms mBERT on cross-lingual classification (Conneau et al., 2020), and (2) the SentiMix dataset may be more challenging than CALCS 2016 because SentiMix contains more CS examples. Regardless, the average F1 score for mBERT on the SentiMix dataset remains at 0.84 and above, indicating that the model still has some information at the token level to do well at the LID task. Generally, the probe classifiers trained and tested on the SentiMix dataset exhibit a drop in performance in contrast with the probes trained on the CALCS 2016 dataset. This likely due to the amount of CS examples in each of the datasets 4.1, which may mean that the SentiMix dataset may be more representative of the LID task for CS text. Overall, these experiments show that PLMs are very effective at detecting CS text at both the sentence and token levels, even with a more challenging dataset.

5. Syntax

To evaluate the effective generalisability of the inferred structure of CS data, we evaluate the extent to which CS data is similar to the majority language in that text. We do this using our dataset ***r-CS*** which has translations into the source languages. We set this up using the structural probe which was developed by Hewitt and Manning (2019) to allow us to extract the structural information captured in CS text and evaluate if the structure is closer to one language compared to another. We repeat this experiment with synthetic data to ensure that we have a more controlled way of measuring this. The results of these experiments are presented in Section 5.2.

5.1. Methods

We use the structural probe developed by Hewitt and Manning (2019), and use the code base by Chi et al. (2020) to train a probe to recreate the dependency tree structure using Universal Dependencies (UD) datasets (Nivre et al., 2020). We train the probe using mBERT on both UD Spanish ancora Taulé et al. (2008) and UD English EWT Silveira et al. (2014), and validate the probe on the test partition of these datasets to ensure high performance on monolingual data, based on two evaluation metrics: Spearman correlation between predicted and true word pair distances, and on undirected, unlabelled attachment score

(UUAS), the percentage of undirected edges placed correctly (Chi et al., 2020). We train the structural probe on layer 7 of mBERT and use a maximum rank of 128 to recover the path length between each pair of words in a sentence (Hewitt and Manning, 2019). We do this because of reasons found in Section 2. Once we have ensured that the probe recovers appropriate dependency parses for monolingual data, we extract dependency parses from the probe using the CS examples from our ***r-CS*** dataset. We then generate dependency parses for the translations in English and Spanish from our dataset.

Due to the lack of gold labels for CS-dependency parses, we have decided to use the graph edit distance (GED) between the dependency parse of a code-mixed sentence and the dependency parse of the monolingual translations as given by the syntax probe. Using GED tells us how many changes a dependency parse need to undergo to resemble another dependency parse. Therefore, we analyse the distances between the GEDs of a code-mixed sentence and the translated monolingual sentences, with the aim of finding if the CS structure aligns more with one of the source languages when compared to the other. To do this, we use the NetworkX python library (Hagberg et al., 2008). Finding the GED between two graphs can often be slow for graphs containing more than 10 nodes (Hagberg et al., 2008). This is the reason we select a subset of the CS dependency parses derived from the structural probe, specifically, examples that contain 10 nodes or fewer, totalling 118 examples. We then extract the dependency parses for the translations of the 118 CS examples and compare the distances . We repeat these experiments with the synthetically generated CS data: ***randCS*** and ***NP-CS-es*** and ***NP-CS-en*** 3.

5.2. Results

Our results associated with syntax are presented in table 4. These results show that there is a strong correlation in the graph edit distances between real CS text and the monolingual translations of that text. In order to assess the potential correlation between the distances for the different language pairs, we use Spearman correlation. For example, to compare monolingual text to real instances of CS text, we use the***r-CS*** data and find the distance between those CS examples and the corresponding Spanish translations, then we do the same for the English translations. We then find the Spearman statistic between these two sets of distances. The results indicate that the model generates CS dependency parses that are similar in distance to the monolingual parses, that is, the dependency parses are not closer in distance to one language compared to another. This is the case despite our dataset containing more English tokens than Spanish tokens, see table 1. The results also show that when synthetic CS examples are used, the correlation of distances between the CS examples and parallel translations drops, perhaps indicating that some of the synthetic CS examples lack syntactic structure.

6. Semantics

One of our aims is to discover whether PLMs are able to effectively capture the meaning of code-mixed sequences. We carry out an intrinsic exploration to see how the representations of code-switched sentences compare with monolingual sentences. We want to find if PLMs are *consistent* in representing semantic information in CS text when compared to semantic representations of monolingual text. To do this, we fine-tune all the PLMs on the semantic text similarity (STS) task using monolingual benchmark STS data in Spanish and English.

6.1. Methods

PLMs do not generate semantically meaningful sentence embeddings unless specifically trained for this, therefore we must fine-tune the models on the STS task. We build on work by Tayyar Madabushi et al. (2022) to set up the semantic experiments. Tayyar Madabushi et al. (2022) developed a method to find whether a PLM is consistent in scoring two sentences or expressions with similar meaning. Given two input sentences, the models must return an STS score between 0 (least similar) and 1 (most similar). We adopt this method to find if a model, after it is fine-tuned on the STS task, is consistent in scoring monolingual sentences and CS sentences. The PLMs fine-tuned on the STS task should be consistent in scoring monolingual sentences and CS sentences. That is, the sentence similarities of (i_{es}, j_{es}) and (i_{en}, j_{en}) , should approximate the similarities between (i_{cs}, j_{cs}) and (i_{es}, j_{es}) and (i_{en}, j_{en}). We formalise this in Eq. 1.

$$sim(S_i^{l_1}, S_j^{l_2}) = sim(S_i^{cs}, S_j^{l}) \quad (1)$$

where sim represents the cosine similarity. S is a sentence in the dataset. The languages of the sentences are encoded by $(l_1, l_2, l) \in \{es, en\} \times \{es, en\} \times \{es, en, cs\}$.

The indexes represented by $(i, j) \in N^2$ correspond to the sentences in the dataset of length N.

We use the dataset ***r-CS*** to get similarities between the language pairs listed in table 5. The similarities output by the fine-tuned PLMs are compared to each other using Spearman Rank Correlation. All PLMs were fine-tuned using a batch size of 8. The base models were fine-tuned using

lang-pair 1	lang-pair 2	Spearman statistic
cs vs. en	cs vs. es	0.8308
NP-CS-en vs. en	NP-CS-en vs. es	0.6876
NP-CS-es vs. en	NP-CS-en vs. es	0.7564
randCS vs. en	randCS vs. es	0.6983

Table 4: Spearman rank for correlation between distances of code-mix and monolingual text. Results on real CS data is highlighted.

l-pair-1	l-pair-2	cosine spearman		
		mBERT	XLM-R-base	XLM-R-large
en-en	cs-cs	0.8503	0.8208	0.8256
es-es	cs-cs	0.7892	0.7655	0.7799
en-es	cs-en	0.8695	0.8656	0.8704
en-es	cs-es	0.7266	0.6947	0.7200

Table 5: Spearman rank statistic for the cosine similarity between language pair 1 (l-pair-1) and language pair 2 (l-pair-2).

a learning rate of 2e-5, and XLM-R-large was fine-tuned with a learning rate of 2e-6.

6.2. Results

Our results associated to semantics are presented in tables 5 and 6. In general, these results show that the models are able to capture the meaning of naturally occurring code-mixed sentences in a way that aligns with how they capture the meanings in monolingual sentences. These results show that the strongest correlations are between $sim(cs_i, cs_j)$ - $sim(en_i, en_j)$ and between $sim(cs_i, en_j)$ - $sim(en_i, es_j)$. In general, though, for all models, the Spearman rank statistic comparing all language pairs is high, meaning that the PLMs, fine-tuned on monolingual data, have the capacity to effectively capture and represent semantic relationships between CS text and monolingual text in a manner consistent with how they represent those relationships between the monolingual pairs. We also conduct these experiments using synthetic CS data, ***randCS*** and ***NP-CS-es*** and ***NP-CS-en***; table 6 contain the results for the experiments with the synthetic data. These results may indicate that the model is not able to capture meaning consistent to the monolingual translations of these examples. This may be for a number of reasons, perhaps because these generations are not guaranteed to be well-formed CS text, it may indicate that the model relies on the syntactic structure of a sentence to provide semantic similarity. Further experiments with different types of synthetically generated CS would be needed for proper analysis.

7. Discussion

The results across all categories of experiments seem to indicate that PLMs are likely to have the potential to generalise to being able to handle CS

l-pair-1	l-pair-2	cosine spearman		
		mBERT	XLM-R-base	XLM-R-large
randCS-randCS	en-en	0.0106	-0.0028	0.0105
randCS-randCS	es-es	0.0091	0.0177	0.0189
NPesCS-NPesCS	en-en	0.0027	0.0009	-0.0029
NPesCS-NPesCS	es-es	0.0188	0.0208	0.0021
NPenCS-NPenCS	en-en	0.0151	0.0009	0.0048
NPenCS-NPenCS	es-es	0.0188	0.0205	0.0065
en-es	randCS-en	0.0184	0.0114	0.0212
en-es	randCS-es	0.0043	0.0154	0.0156
en-es	NPesCS-en	0.0102	0.0030	0.0130
en-es	NPesCS-es	0.0011	0.0223	0.0108
en-es	NPenCS-en	0.0107	0.0111	0.0168
en-es	NPenCS-es	0.0056	0.0221	0.0152

Table 6: Semantic experiments results with synthetic CS data and the original monolingual translations of the r-CS-syn dataset.

text. We find that PLMs are effective at detecting CS text at a sentence level and token level in our detection experiment. We find that dependency parses generated by the model are not more similar in distance to one language or another in our syntax experiments that is, experimental results reveal a strong correlation in the distances of dependency parses between English (cs-en) and Spanish (cs-es). We find as well that the models are consistent in capturing meaning representations of real CS text, but are unable to do so for synthetically generated text using our generation methods. They seem to capture syntactic structure and semantic meaning across real CS text, without being trained on CS text, see tables 4 and 5.

Our findings show that PLMs are able to generalise across CS text containing Spanish-English language pair. They also show that in general, performance of the probes degrades when using synthetic CS text. In the syntax experiments, the correlation between the distances diminish, though this could be attributed to the difference in distribution of a synthetically generated CS example when compared to a well-formed CS example. Although we used methods found in literature to generate the synthetic examples, there is no guarantee of these methods producing a naturalistic CS sentence. In the semantic experiments, the Spearman correlation statistic drops to nearly zero across all models when using synthetically generated CS text, which may be due to the loss of grammatical correctness. This may show that PLMs rely on the syntactic structure of a sentence to provide the semantics. Further experiments with different types of synthetically generated CS would be needed for a proper analysis.

In general, the experimental results across detection, syntax and semantics, show that for real CS text containing languages that are closely related, such as Spanish and English, PLMs may contain enough linguistic information from the source languages to handle the mixed language text. This is promising, because if monolingual data can be

harnessed for some tasks, then the scarcity of data in certain CS language pairs can be mitigated by the PLMs ability to generalise. We would like to explore this idea in future research.

8. Conclusion and Future Work

In this paper, we present our finding on how pre-trained models handle code-switched text. Our contributions include a novel dataset of CS text and translations into the source languages, Spanish and English. Additionally, we extend probing work to code-switching in Spanglish in the areas of syntax and semantics. We carry out experiments in detection, syntax and semantics, to explore how PLMs capture CS text. We find that PLMs seem to be effective at detecting CS text. In the future, we hope to explore PLMs abilities to learn from monolingual data for use on CS text, experiment with further synthetic data generation methods, and to expand to other languages.

9. Acknowledgements

The computations described in this research were performed using the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/). Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham.

10. Optional Supplementary Materials

10.1. Limitations

Our work only explores how models embed code-switched data for Spanglish, although this was done so as not to confound model capabilities by other aspects of language use, in the future, we would like to extend our explorations to languages such as Hinglish. Doing so will allow us to see the extent to which PLMs generalise to different language pairs, especially pairs that are not closely related languages. Our work only explores auto-encoder models, such as mBERT and XLM-RoBERTa, which does not offer a comprehensive view of how different types of models encode CS-text. In the future, we would like to explore the capabilities and degree to which models such as GPT encode CS text.

10.2. Ethics Statement

We do not use any private data, all data used is publicly available, or will become available after the end of the anonymity period. The dataset that we create is collected from social media and may contain profanity or toxic content. We work with one language pair for code-switching, out of many, and hope in the future to expand this to other CS language pairs, especially low-resource pairs.

11. Bibliographical References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In *ArXiv*, volume abs/1608.04207.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *LREC*.

Fahad AlGhamdi, Giovanni Molina, Mona Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. 2016. Part of Speech Tagging for Code Switched Data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 98–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barbara E Bullock and Almeida Jacqueline Toribio. 2009. 1 Themes in the study of code-switching. In *The Cambridge Handbook of Linguistic Code-switching*, pages 1–10. Cambridge University Press.

Shuguang Chen, Gustavo Aguilar, Anirudh Srinivasan, Mona Diab, and Thamar Solorio. 2022. CALCS 2021 Shared Task: Machine Translation for Code-Switched Data.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding Universal Grammatical Relations in Multilingual BERT. In *ArXiv*, volume abs/2005.04511.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James R Glass. 2019. What Is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models. In *AAAI*.

Frances Adriana Laureano De Leon, Florimond Guéniat, and Harish Tayyar Madabushi. 2020. CS-Embed at SemEval-2020 Task 9: The effectiveness of code-switched word embeddings for sentiment analysis. In *arXiv preprint arXiv:2006.04597*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

A Seza Dogruoz, Sunayana Sitaram, Barbara E Bullock, and Almeida Jacqueline Toribio. 2021. A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies. In *ACL/IJCNLP*.

Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. 2010. A survey of graph edit distance. *Pattern Analysis and Applications*, 13(1):113–129.

Aric A Hagberg, Daniel A Schult, and Pieter J Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic Probing through Dimension Selection. *arXiv*, 2010.02812v1.

John Hewitt and Christopher D Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *NAACL*.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A Survey of Current Datasets for Code-Switching Research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141. IEEE.

Aravind K Joshi. 1982. Processing of Sentences With Intra-Sentential Code-Switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An Evaluation Benchmark for Code-Switched NLP. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585. Association for Computational Linguistics (ACL).

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual Code-Switching for Zero-Shot Cross-Lingual Intent Prediction and Slot Filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martin Kroon, Sjef Barbiers, Jan Odijk, and Stéphanie van der Pas. 2019. A filter for syntactically incomparable parallel sentences. *Linguistics in the Netherlands*, 36:147–161.

Ukp Lab, Tu Darmstadt, Clara Vania, and Ilia Kuznetsov. 2020. LINSPECTOR: Multilingual Probing Tasks for Word Representations. *Computational Linguistics*, 46(2).

Philip May. 2021. Machine translated multilingual STS benchmark dataset.

Erica McClure. 1995. DUELLING LANGUAGES: GRAMMATICAL STRUCTURE IN CODESWITCHING. Carol Myers-Scotton. Oxford: Clarendon Press, 1993. Pp. xiv + 263. *Studies in Second Language Acquisition*, 17(1):117–118.

Gideon Mendels, Julia Hirschberg, Victor Soto, and Aaron Jaech. 2018. Collecting Code-Switched Data from Social Media. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation ({LREC} 2018)*, Miyazaki.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima-López, Luis Gascó, Vicent Briva-Iglesias, Marvin M Agüero-Torales, and Martin Krallinger. 2021. The ProfNER shared task

on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *SMM4H*.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the Second Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the Second Workshop on Computational Approaches to Code-Switching*, pages 40–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

William E Nagy, Georgia E García, Aydin Y Durgunoğlu, and Barbara Hancin-Bhatt. 1993. Spanish-English bilingual students' use of cognates in English reading. *Journal of Reading Behavior*, 25(3):241–259.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Jung Pandey, Srinivas Pykl, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets. In *ArXiv*, volume abs/2008.04277.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Stroudsburg, PA, USA. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *ACL*.

SHANA POPLACK. 1980. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Joel Laffita Rivera. 2019. A Study Conception about Language Similarities. *Open Journal of Modern Linguistics*, 09(02):47–58.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A Toolkit for Generating Synthetic Code-mixed Text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 205–211, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cesa Salaam, Franck Dernoncourt, Trung Bui, Danda Rawat, and Seunghyun Yoon. 2022. Offensive Content Detection Via Synthetic Code-Switched Text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6617–6624.

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. BERTologiCoMix * How does Code-Mixing interact with Multilingual BERT? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121.

Mark Sebba, Shahrzad Mahootian, and Carla Jonsson. 2012. *Language Mixing and Code-Switching in Writing Approaches to Mixed-Language Written Discourse*, 1st edition.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad Alghamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the First Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72. Association for Computational Linguistics.

Victor Soto and Julia Hirschberg. 2017. Crowdsourcing Universal Part-of-Speech Tags for Code-Switching. In *Interspeech 2017*, pages 77–81, ISCA. ISCA.

Mariona Taulé, M Antònia Mart\'\i, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, Ellie Pavlick, and Google AI Language. 2019b. What do you learn from context? Probing for Sentence Structure in Contextualized Word Representations. In *International Conference on Learning Representations*.

David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. 2016. EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis. In *LREC*.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Stroudsburg, PA, USA. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-Switched Language Models Using Neural Based Synthetic Data from Parallel Sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280, Stroudsburg, PA, USA. Association for Computational Linguistics.

12. Language Resource References

Chen, Shuguang and Aguilar, Gustavo and Srinivasan, Anirudh and Diab, Mona and Solorio, Thamar. 2022. *CALCS 2021 Shared Task: Machine Translation for Code-Switched Data*. PID https://ritual.uh.edu/lince/datasets.

Miranda-Escalada, Antonio and Farré-Maduell, Eulàlia and Lima-López, Salvador and Gascó, Luis and Briva-Iglesias, Vicent and Agüero-Torales, Marvin and Krallinger, Martin. 2021. *The ProfNER shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora*. Association for Computational Linguistics. PID https://zenodo.org/records/4563995.

Patwa, Parth and Aguilar, Gustavo and Kar, Sudipta and Pandey, Suraj and PYKL, Srinivas and Gambäck, Björn and Chakraborty, Tanmoy and Solorio, Thamar and Das, Amitava. 2020. *SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets*. Association for Computational Linguistics. PID https://ritual.uh.edu/lince/datasets.

Natalia Silveira and Timothy Dozat and Marie-Catherine de Marneffe and Samuel Bowman and Miriam Connor and John Bauer and Christopher D. Manning. 2014. *A Gold Standard Dependency Corpus for English*. PID https://universaldependencies.org/treebanks/en_ewt.

Taulé, Mariona and Martí, M. Antònia and Recasens, Marta. 2008. *AnCora: Multilevel Annotated Corpora for Catalan and Spanish*. European Language Resources Association (ELRA). PID https://universaldependencies.org/treebanks/es_ancora.