Agreement Attraction Effects in Multilingual BERT

Abby Bertics
abertics@mit.edu

Shinjini Ghosh
shinghos@mit.edu

Abstract

Multilingual BERT (mBERT) is a transformer model that is able to generate sentence representations and masked language modeling predictions for 104 languages. We investigate subject-verb agreement and related attraction effects in both monolingual and codeswitched data across three languages (English, French, and Russian). We first establish that mBERT has a cross-linguistic notion of agreement and that agreement attraction effects occur. We then investigate how and speculate as to why the agreement attraction effects in codeswitched sentences compare and contrast with the monolingual ones and how the pattern of these effects varies with the language of the verb.

1 Introduction

Codeswitching, or the use of two or more languages in discourse between bilingual or multilingual speakers, is a crucial area gaining greater popularity in the fields of linguistics and NLP. In this paper, we aim to investigate a specific form of agreement attraction in codeswitched sentences on a state-of-the-art language model, and delve deeper into its predictions, how those compare with human errors, and why such effects might occur. Deeper insights into this field will not only help us refine our models for application purposes such as automatic speech recognition and synthesis, but also to move closer towards understanding the human cognition behind agreement attraction errors and how we might build more human-like language models.

2 Background

It is now in the vogue to run psycholinguistic experiments on Natural Language Processing models and see how they fare. This is popular to do with RNNs, transformers, and other models (Linzen et al., 2016; Futrell et al., 2018; Arehalli and Linzen, 2020; Gulordava et al., 2018; Ettinger, 2020; Gauthier et al., 2020).

2.1 Agreement Attraction

Attraction is a common type of error in language production, where a feature from one word in a sentence is incorrect extended to another (close) word. The most commonly seen form of attraction is agreement attraction (a word takes on a feature based on some aspect of agreement with another word, be it number, tense, gender, class, etc.), but there also exist other forms of attraction such as case attraction (wherein a feature is assigned to a word based on its grammatical role or theta-role in a sentence) and negative attraction (which extends negation particles (Labov, 1972)).

2.1.1 Agreement Attraction in Linguistics

The erroneous agreement of a verb with a noun that is not its subject (i.e., the specifier in its Binding Domain) is termed as agreement attraction. The most famous example of this is the wrongly printed phrase in the New Yorker, 'Efforts to make English the official language *is* gaining strength throughout the US', where the verb is wrongly in its singular form, being 'attracted' by 'language', when it should have agreed with the word at the head of the noun phrase, in Spec, TP position, 'efforts'. Agreement attraction has also been observed in object-verb agreement when wh-movement and fronting occurs in English (Dillon et al., 2017), in SOV constructions in Dutch and in other languages, but the majority of this paper rests on agreement attraction concerning

subject-verb agreement in various languages.

2.1.2 Agreement Attraction in NLMs

Recent studies have observed that RNNs capture some human agreement effects (Arehalli and Linzen, 2020; Futrell et al., 2018). Neural networks were trained only on word-prediction over large corpora and were sound to capture significant subject-verb agreement patterns, albeit with occasional errors. (Arehalli and Linzen, 2020) found that LSTMs captured critical human behavior in 3 out of 6 experiments performed, behaving similarly to humans in that a) attractors in a prepositional phrase general stronger attraction effects than in relative clauses, b) attractors closer to the verb generate stronger attraction effects linearly and c) attractors outside of the clause where the agreement is computed cause attraction effects. Contrasting, LSTMs had opposite (or no significant) effects as compared to humans in a) attractors closer to the verb generating stronger attraction effects linearly, b) collective subjects with distributive readings having higher rates of plural agreement than those with collective readings and c) attractors in oblique arguments creating a larger attraction effect than those in core arguments.

2.2 Codeswitching

Codeswitching (CS) is the mixing of two or more languages in discourse, often used within the same context or topic, as is used by bilingual and multilingual people (Poplack, 2001), (Deuchar, 2020). While inter-sentential codeswitching occurs quite frequently in bilingual speech, it is the intra-sentential codeswitching that has garnered greater attention in the community recently, and is of main interest in this paper as well. A real-life example in Italian-English codeswitching from (Rosignoli, 2011)'s transcripts is as follows.

(1) they'll have uh kind of metodo di studio
they'll have uh kind of methods of study
and uh linguaggio
and uh language
they'll have uh kind of study skills and uh language skills

2.2.1 Codeswitching in Humans

Multilingual humans codeswitch frequently, and certain codeswitched sentences sound correct or grammatical, while others do not. All combinations of replacing words from one language to another would not result in a grammatical, correct-sounding codeswitched sentence. So what makes swapping out certain words, phrases and clauses in a sentence to a different language feel 'grammatical', and when is it not? There also remains more subtle question of how to deal with morphoogical clashes, grammatical category mismatches, idiomatic expressions, and so on. Another interesting question in the field of codeswitching is what makes a word codeswitched rather than just borrowed?

Hundreds of linguistic models of codeswitching have been developed to attempt to succinctly and accurately capture these phenomena. (Sitaram et al., 2019) presents a survey of current linguistic models (as well as NLP), a lot of whom fall into one of two main categories—those based on the Free Morpheme (that it is grammatical to codeswitch constituents or full sentences in the presence of a free morpheme) and Equivalence Constraint (codeswitching occurs at places where there is no syntactic rule violation of either language) and the Constraint-free or Minimalist Models.

2.2.2 Codeswitching in NLMs

Because so much of multilingual human speech is codeswitched, and a sizable population of the world uses mixed speech on a daily basis, this topic is of considerable interest to the NLP community. (Sasidharan and Poornachandran, 2018) and (Sitaram et al., 2019) provide a survey of some of the existing codeswitched models and their applications (importantly, Automatic Speech Recognition and Speech Decoding & Synthesis) in NLP. Currently relevant codeswitched NLMs include those with functional head constraints (Li and Fung, 2014), RNNs trained with interleaved monolingual data in both languages (Choudhury et al., 2017) (which is also the basis for this paper), factored language models with re-ranking techniques (Gebhardt, 2011), RNNLMs with n-gram models (Adel et al., 2013), and bilingual attention language models (Lee and Li, 2020).

2.3 Multilingual Language Models

Multilingual models seem to be the new rage. There are multiple ways to go about this, and

most of them involve training on interleaved monolingual data. This has been done with RNNs (Samih et al., 2016; Sercu et al., 2017) and transformers (Devlin et al., 2018; Conneau et al., 2019).

2.3.1 Multilingual BERT

This transformer model that we will primarily concern ourselves with for this paper, Multilingual BERT (mBERT), is one of the many released by Devlin et al. (2018). It is a 12 layer transformer model that was trained on the Wikipedia corpora of 104 different languages. There are no markers denoting which language is being input, and there is a single multilingual 110k-token shared WordPiece vocabulary. For all intents and purposes, this model is completely language agnostic; it is not programmed to know or care what language is being fed to it (although, for all we know, it learns to pick up on this information implicitly).

For this experiment we will use the pre-trained PyTorch implementation[1], using the `bert-base-multilingual-cased` mode. Code-switched data has been used to fine-tune mBERT (Pan et al., 2020; Qin et al., 2020), but I have not found cases where the model tested alone on it.

3 General Methods

In psycholinguistics research, behavioral data is often used in order to investigate how humans process language. The amount of time it takes to read or process a word reflects how much the human is expecting that word. This can be quantified behaviorally via eye-tracking, timed reading, etc. Unfortunately, NLP models neither have eyes to read with nor fingers to press buttons, so researchers have had to be inventive. In order to try to ascertain knowledge of language, surprisal values, which is linearly correlated with human reading time, can be used (Levy, 2013; Hale, 2001)[2]. This behavioral proxy is well defined for probabilistic language models. A sentence is a string of words $w_1, ..., w_n$. The surprisal of a word w_i following words $w_1, ..., w_{i-1}$ is

$$\log \frac{1}{P(w_i|w_1, ..., w_{i-1})} \quad (1)$$

This works less well for BERT, because BERT is bidirectional in nature and not a language model. We hence have to limit our tests to minimize the sketchiness. This also brings up questions of whether or not it is even reasonable to assess mBERT's knowledge of language if it is not modeling language.

3.1 Task

Here is an example of what we will feed the model. If we have the following two sentences, with a minimal difference of verb inflection:

(2) The senator near the taxi drivers has flowers.

(3) *The senator near the taxi drivers have flowers.

We remove the inflected verb and feed the following sentence into BERT:

(4) The senator near the taxi driver [MASK] flowers.

We then look at and compare the scores assigned to the words "have" and "has" at the masked position. This varies from standard surprisal calculations because the entire sentence minus the inflected verb is conditioned on, rather than just the preamble. In the case when both verbs are one token, this is what was done in Goldberg (2019).

3.2 Experimental Methodology

We are trying to mimic a production, forced-choice task here. The winning word is the one that has the higher probability. This is straightforward for the LSTM case because it is built straight into the model.

$$S(x_i) = -\log_2 p(x_i|h_{i-1}) \quad (2)$$

BERT is a masked language model, so one would think it would be as simple as getting probabilities for the masked word. But, it is only reasonable to compare "probabilities" for sentences of the same length (Wang and Cho,

[1] https://github.com/huggingface/pytorch-pretrained-BERT

[2] We do not want to pretend that these language models are humans. Instead, we hope to use this measurement to see where models are falling short in terms of true knowledge of language.

2019). This becomes an issue because BERT uses subword tokens. For example, there is no way to compare the probability of "wait" with "waits", because "waits" gets separated into two tokens "wait" and "##s".

To deal with this, a couple of solutions have been suggested. One option is just to completely ignore all the words that get tokenized into more than one token (Goldberg, 2019). Another option is to look at confusion values (Wang and Cho, 2019; Lin et al., 2019), training another classification layer on top.

What we ended up doing was only considering cases where the number of tokens for each target word was the same. This allows a direct comparison, because the lengths of the sentences are the same. We use the chain rule. Let's say, we have a sentence $x_1, ..., x_n$ and a word x_i that corresponds to the tokens $t_1, ..., t_m$. After feeding the sentence with x_i masked (M_i) through BERT, we get a layer, L that we can apply softmax to. to get the "probabilities" for each token in the vocabulary. We can then calculate the word's pseudo log probability as follows

$$PLP(x_i|M_i) = \sum_{i=1}^{m} \log(\sigma(L[t_i])) \qquad (3)$$

Again, this probability comparison is only valid when we are looking at sentences of equal length, so we are limited to when the words we are trying to predict are of the same length post-tokenization[3].This was fine because we were able to create the dataset, but this is definitely a limitation of these types of investigations into transformers. Normally, we assume that we are dealing with probabilistic language models to even start thinking about these things.

4 Multilingual Agreement Attraction

Because we are dealing with multiple languages and sentences that contain multiple languages, we want to stay with the simplest possible experiment. First, we confirm that mBERT has a notion of subject-verb agreement in monolingual and codeswitched situations. Only then can we look into agreement attraction.

for language combination (27) **do**
  for number combination (4) **do**
    for subject (10) **do**
      for prepositional phrase (36) **do**
        for verb phrase (8) **do** make
        end for
      end for
    end for
  end for
end for

Figure 1: Data generation process for 311,440 training examples

4.1 Data generation

In order to elicit agreement attraction effects, we are going to rely on an attractor in a prepositional phrase on the subject so that it is linearly between the subject and the verb. Both of the subject and the attractor will be animate and plausibly be the subject of the verb. The codeswitching only occurs at phrase boundaries; there will be no language switching within the subject, prepositional, or verb phrases.

The generative process for the data is fairly straightforward, and accomplished with a many-times-nested for-loop. To start with, we have 10 possible different subjects, 6 prepositions and 6 attractors ($6 \times 6 = 36$ prepositional phrases), and 8 verb phrases[4]. For each item, we have the corresponding English, French, and Russian translations in both singular and plural.

To keep things understandable, let think of all of the possible sentences in English with only singular nouns and verbs; using all combinations will result in $10 \times 6 \times 6 \times 8 = 2880$ possible base sentences. Then, we vary the number of the subject and the attractor. For each base sentence, we have $2 \times 2 = 4$ number combinations. Then, for each number combination, we want to generate all of the language combinations: $3 \times 3 \times 3 = 27$. This results in a grand total of $2880 \times 4 \times 27 = 311440$ possible sentences. This process is outlined in Figure 1.

[3]A lot aren't. For example, wait/waits, listen/listens, eat/eats, etc.

[4]The verbs had to be carefully token so that both the singular and the plural option resulted in the same number of subword tokens.

The great thing about using a computer model is that we can show the same model all combinations, whereas with a human, one would only be able to show each of them one of the combinations.

4.2 Results

From the predictions that mBERT generates[5], we can make the following inferences about the nature of how that model processes language. First, agreement across languages occurs. With a matching attractor (SS and PP), the model is able to make the subject and verb agree, better than chance, even when the subject and verb are not in the same language. In the presence of a mismatching attractor (SP and PS), accuracy is much worse, dipping below chance in some cases. Second, the pattern of monolingual agreement for mBERT is different for different languages. For example, in French, it is best at agreeing singular subjects and verbs, and in English, it is best at agreeing plural subjects and verbs. We will expound on this later. Third, mBERT is most easily "tricked", or makes the most agreement errors, when the attractor is in the same language as the subject.

In Table 1, we can see the general trends of agreement attraction. We have grouped the data by language of the verb. The cases where there is a number match (SS and PP) act as a baseline to establish that each language knows how to agree subject and verb. Then we can compare SS with SP (where the grammatical choice is a singular verb) and PP with PS (where the grammatical choice is a plural verb). If agreement attraction is occurring, we expect the accuracies for SP to be lower than those for SS, and for PS to be lower than for PP, which is what we see.

In the following analyses, we use linear mixed-effects regressions and Wald or likelihood ratio tests to determine the significance of the relevant fixed variables, with a random factor intercept (but not slope due to non-convergence) for each base item. Because we have binary predictions from mBERT, we will use binomial logistic regression.

[5]Github repo for data generation, mBERT predictions, and analysis is located here: https://github.com/abertics/mbert-aa-cs

Lang. of Verb	SS	SP	PS	PP
Mono. English	0.55	0.22	0.84	0.98
Mono. French	0.86	0.66	0.46	0.74
Mono. Russian	0.78	0.72	*0.71*	*0.74*
English Verb	0.53	0.36	0.72	0.88
French Verb	0.84	0.75	0.33	0.50
Russian Verb	0.81	0.66	0.59	0.75

Table 1: Agreement accuracies. (SP means singular subject and plural attractor.) The accuracies in italics if they are not statistically significant ($p >.01$ in a t-test).

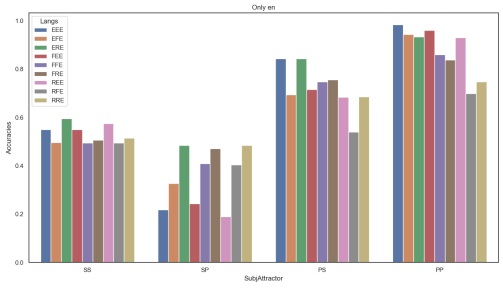
4.2.1 Monolingual

The first order of business is confirm the existence of agreement attraction effects with this model in monolingual contexts. Subject-Attractor number match is a statistically significant fixed effect ($|z|$=2.84, p $<$ 0.005), meaning that there are more agreement errors when the subject and the attractor do not match in number, and the attractor is able to "attract" the verb. This is summarized in Table 2.

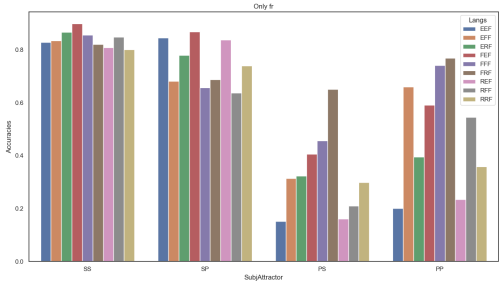
- In English, the model is overall worse at agreeing with singular subjects than plural ones ($|z|$=43.6; $p < .001$). There is a statistically significant interaction between attractor number and subject-attractor match (β=-.18; $|z|$=4.20; $p < .001$). The most attraction occurs when the attractor is plural and the subject is singular.
- In French, the model is overall worse at agreeing with plural subjects than singular ones ($|z|$=18.7; $p < .001$). There is no statistically significant ($p >.05$) interaction between subject number and match.
- In Russian, the model is also overall worse at agreeing with plural subjects ($|z|$=4.0; $p < .001$). There is a mild ($p <.05$) interaction between subject number and match; a plural attractor has more sway.

4.2.2 Verb Language

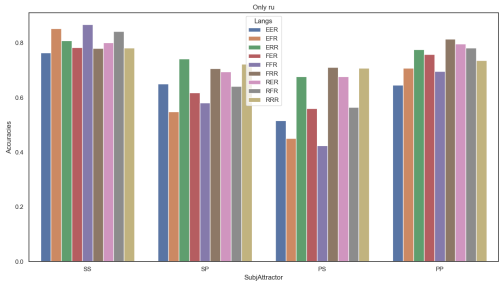
Now, let's move onto the combinations where there might be two or three languages in one utterance. For example, there might be an English subject, a French attractor, and a Russian verb that has to try to agree with the



(a) English



(b) French



(c) Russian

Figure 2: Monolingual Agreement Attraction Trends By Language. The left-most cluster is SS, then SP, then PS, then PP. Different colors represent different language combinations.

English subject. mBERT has not seen training data of this type (beyond what exists in Wikipedia), so agreement of this sort would constitute a sort of zero-shot learning task. When there is no possible confounding mismatch between subject and attractor, and neither the subject nor the attractor match the verb in language, the agreement accuracies are 0.64, 0.56, and 0.76 for English, French, and Russian verbs, respectively. This implies that mBERT has attained a language-agnostic understanding[6] of agreement.

[6]I use this word in the loosest possible sense. I am not implying that mBERT is capable of actually understanding anything at all.

We conducted t-tests to determine if agreement attraction occurs for each combination of singular and plural subject and verb language. The differences in accuracies, as seen in Table 1 were all statically significant ($p < .0001$).

Unsurprisingly, the language of the verb has significant effects on agreement. The aptitude for plural verbs that was shown in monolingual English data extends to all combinations where English is the verb. This also holds true for singular verbs and French verbs.

Let's look at the top three and bottom three language combinations for each verb, ranked by strength of agreement attraction[7] and see if there are any patterns. The winners (and losers) are written from most effect to least.

- **English Verb:** (most attraction) REE, EEE, FEE, ... RRE, FRE, RRE (least)
- **French Verb:** (most attraction) EFF, RFF, FFF, ..., FEF, EEF, REF (least)
- **Russian Verb:** (most attraction) EFR, FFR, RFR, ..., FRR, ERR, RRR (least)

English and French exhibit similar patterns, while Russian does not. In general, when the language combination is ABB (SV mismatch but AV match), there is the most agreement attraction effect, followed by the monolingual AAA combination. The least attraction occurs with a BBA pattern, followed by ABC, and ABA.

4.2.3 Attractor Language

Next, we wanted to see whether or not the language of the attractor matters.

Verb Lang.	$L\&N$	$\bar{L}\&N$	$\bar{L}\&\bar{N}$	$L\&\bar{N}$
English	**0.76**	0.68	0.57	**0.48**
French	**0.75**	0.63	0.56	**0.49**
Russian	**0.78**	0.77	**0.58**	0.71

Table 2: Agreement accuracies vary based on the language of the attractor and number mismatch of attractor and subject. Bolded are the best and worst for each language. L means the attractor and verb are in the same language. N means the subject and attractor match in number.

[7]I measure decrease in accuracy going from subject-attractor number match to mismatch. For example, I would take the arg{max|min} of (accuracy(SS) - accuracy(SP)).

Looking at the numbers in Table 2, the accuracies are best in the left-most column and get progressively worse as one looks to the right. The first factor that makes mBERT worse at agreement here is whether or not the attractor matches in number with the subject; this is the canonical cause of agreement attraction errors in production. The second is whether the attractor is in the same language as the verb. When it is, it appears to be a more potent attractor, and the combination of these two factors results in the worst accuracies, slightly below chance.

We conducted a likelihood-ratio test with respect to the attractor matching the verb in language, grouped by the language of the verb. Let us define a variable AVMatch that has a value of 1 if the verb and attractor are in the same language and -1 otherwise. For English verbs, this variable alone does not have a strong effect (χ^2=2.7, p >0.1). However, the interaction between this variable and the subject and attractor (mis)matching in number was statistically significant (χ^2=764.8, p <.0001). So, the language of the attractor only affects the accuracy if the attractor does not match the subject in number. For both French and Russian verbs, the language of the attractor alone and the interaction of it with number match are both statistically significant effects (χ^2 >90, p <.0001), except for Russian, both the number and language matching seemed to hurt agreement[8]

4.2.4 Subject Language

Continuing with the trend here, we wanted to see if the language of the subject (in relation to the language of the verb) has an effect on agreement accuracies. If we ignore the language of the attractor, if the language of the subject matches the verb, that has a facilitatory effect on matching (β=.42, $|z|$=45, p <.001); the interaction between subject-verb language match and subject-attractor number match is weak.

Let us now consider these two language match factors in tandem (both subject-verb language match and attractor-verb language match), because if they both have predictive significance independently, they might interact together. Indeed, this is the case ($|z|$ =17.3, p <.001).

To summarize, we have three conditions that reduce mBERT's accuracy in subject-verb agreement: 1) when the subject and attractor mismatch in number, 2) when the attractor matches the verb in language, and 3) when the subject does not match the verb in language. In a fitted mixed effects model with these three variables, the strongest effect is (1), subject-attractor number mismatch, with a β-value of 0.38 logits ($|z|$ =85, p <.001); this is the factor most affecting monolingual agreement attraction as well.

5 Discussion

This experiment illuminated many different interesting patterns. Throughout this section, we will wildly speculate as to possible explanations for why these occur and possible inferences that can be made from them. Describing this section as anything other than wild speculation would not be scientifically responsible, because all we have are correlations and patterns; there is no causation. You, the reader, have been duly warned and should proceed with caution.

First, considering purely monolingual data, it is interesting that mBERT has different agreement patterns for different languages. Why is there such a strong singular preference in French and a strong plural one in English? Is this the result of patterns in the training data? Or could this be a result of the morphology of the languages? In English, regular singular verbs are the plural ones plus the $-s$ suffix morpheme. In French, regular plural verbs are the singular ones plus an $-nt$. Would there therefor be a preference for the "default" conjugation?

Along those lines, the ability of English verbs to agree with singular subjects and French to agree with plural subjects even in the presence of matching (helping) attractors is concerning at best. Without a distracting attractor, the predictions operate at a chance level or marginally better. With a distracting attractor (number mismatch), the model performs significantly worse than chance. In English, for example, when the attractor is plural, the verb agrees with the attractor instead of

[8] I can not think of a reasonable explanation for why this might be the case.

the subject 78% of the time. That casts doubt on the claim that mBERT knows what subject-verb agreement is in the first place. Same as with singular attractors in French, the preference for the verb show the number of the attractor seems to outweigh the fact that the subject is not of that number. This is not characteristic of human language competence or performance. There is something weird going on here, and if mBERT is to be used, it should be probed further.

Enough with English and French; there is one interesting thing here with Russian: the attractors do not bear nominative case (which signals that a noun is a subject and should agree with the verb), and should therefore be less attractive to the verb. The data does exhibit weaker agreement attraction effects than English or French, but the fact that they occur at all suggests that there might be distinct underlying number and case features on nouns in this model. It can not be pure n-gram-esque memorization, because the plural ending in the dative case will not pattern only with plural verbs; they should be independent in the data. The Russian data is kind of more over the place than the English or French.

If the attractor and verb matching in language produces more agreement errors, it might be because the verb is relying more heavily on memorization. We see this pattern with English and French verbs, but not with Russian ones. With Russian, it is curious that a French attractor should exhibit result in the most agreement errors. One reason for this could tie back into case marking; a Russian attractor is weaker, because in the cases we tested, it does not look like a subject. The French attractor, however does. If the model has a language agnostic notion of number, then the Russian verb could be swayed. But the fact that French and English verbs are more prone to same language attractors provides opposing evidence.

Pires et al. (2019) hypothesized that a sentence representation in mBERT is composed of a language-specific component, which identifies the language of the sentence, and a language neutral component, which captures the meaning of the sentence in a language-independent way. Further work should look into the nature of this language-specific component when the sentence is in more than one language. One hypothesis, along the lines of the Matrix Language Frame Model (Myers-Scotton, 2001), could be that the language of the inflected verb determines the main language of the utterance. How would that work, though, in the case of this experiment, when the sentence fed to mBERT includes everything but the finite inflection?

To end on a more promising note, the fact that mBERT is able to agree subject and verb beyond chance in codeswitched data is quite impressive. This could technically be considered zero-shot learning, because never in training has it encountered codeswitched text. This suggests that mBERT is doing more than explicit pattern memorization; at the very least, it is inferring some sort of notion of number.

But then, to bring us back down to earth, there are significant limitations to this experiment. First, although the number of combinations tested was significant, we used a fairly small number of verbs and nouns. Second, and perhaps most significantly, BERT isn't designed to be used as language model, in the traditional sense of the phrase[9]. If BERT and other MLMs, due to their bidirecitonality, do not generate proper probability distributions over sentences, it seems like this might lead to problems in analyses, given that most analyses assume a language model to have a proper probability distribution. But then again, there is no unequivocal evidence that humans behave like proper probability distributions, and we still have "knowledge of language."

6 Future Work

Future work to be done includes utilizing the metadata about the pre-trained BERT training dataset to see if there already exist effects of the statistical distribution of singular and plural verbs in the results—and if not, what scaffolding does to it. And also extending this project to a more multilingual version, incorporating datasets from more languages.

[9]From Devlin, himself, https://github.com/google-research/bert/issues/35.

7 Contributions

Shinjini wrote the Introduction and the entire Background section. She assisted in proofreading the paper. She attempted to convert existing Agreement Attraction datasets into English-French codeswitched language. She also (in progress) investigated the mBERT training sets to try to find more statistics about the number of VBZ and VBP in English and French Wikipedia dumps.

Abby wrote the Abstract, General Methods, Results, and Discussion sections. She developed the data generation algorithm and generated the data used in the experiment. She designed and ran the experiment on mBERT. And she was responsible for the analysis of the results.

References

H. Adel, Ngoc Thang Vu, F. Kraus, Tim Schlippe, Haizhou Li, and Tanja Schultz. 2013. Recurrent neural network language modeling for code switching conversational speech. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8411–8415.

Suhas Arehalli and Tal Linzen. 2020. Neural Language Models Capture Some, But Not All, Agreement Attraction Effects. preprint, PsyArXiv.

Monojit Choudhury, Kalika Bali, Sunayana Sitaram, and Ashutosh Baheti. 2017. Curriculum design for code-switching: Experiments with language identification and language modeling with deep neural networks. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 65–74, Kolkata, India. NLP Association of India.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Margaret Deuchar. 2020. Code-switching in linguistics: A position paper. *Languages*, 5(2):22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Brian Dillon, Adrian Staub, Joshua Levy, and Clifton Jr. Charles. 2017. Which noun phrases is the verb supposed to agree with?: Object agreement in american english.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv:1809.01329 [cs]*. ArXiv: 1809.01329.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Jan Gebhardt. 2011. Speech recognition on english-mandarin code-switching data using factored language models - with port-of-speech tags, language id and code-switch point probability as factors.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *arXiv:1901.05287 [cs]*. ArXiv: 1901.05287.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

William Labov. 1972. Negative attraction and negative concord in english grammar.

Grandee Lee and Haizhou Li. 2020. Modeling code-switch languages using bilingual parallel corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 860–870, Online. Association for Computational Linguistics.

Roger Levy. 2013. Memory and surprisal in human sentence comprehension.

Y. Li and P. Fung. 2014. Code switch language modeling with functional head constraint. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4913–4917.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside bert's linguistic knowledge. *CoRR*, abs/1906.01698.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to

learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Carol Myers-Scotton. 2001. The matrix language frame model: Developments and responses. *Codeswitching worldwide II*, pages 23–58.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Mo Yu, and Saloni Potdar. 2020. Multilingual bert post-pretraining alignment. *arXiv preprint arXiv:2010.12547*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is Multilingual BERT? *arXiv:1906.01502 [cs]*. ArXiv: 1906.01502.

Shana Poplack. 2001. *Code Switching: Linguistic*, pages 2062–2065.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402*.

Alberto Rosignoli. 2011. Flagging in english-italian code-switching.

Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. 2016. Multilingual code-switching identification via lstm recurrent neural networks. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 50–59.

Thara Sasidharan and Prabaharan Poornachandran. 2018. Code-mixing: A brief survey. pages 2382–2388.

T. Sercu, G. Saon, J. Cui, X. Cui, B. Ramabhadran, B. Kingsbury, and A. Sethy. 2017. Network architectures for multilingual speech representation learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5295–5299.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W. Black. 2019. A survey of code-switched speech and language processing. *CoRR*, abs/1904.00784.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a markov random field language model. *CoRR*, abs/1902.04094.