Assessing the Syntactic Capabilities of Transformer-based Multilingual Language Models

Laura Pérez-Mayos[1], Alba Táboas García[1], Simon Mille[1], Leo Wanner[2,1]

[1] TALN Research Group, Pompeu Fabra University, Barcelona, Spain
[2] Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain

{laura.perezm|alba.taboas|simon.mille|leo.wanner}@upf.edu

Abstract

Multilingual Transformer-based language models, usually pretrained on more than 100 languages, have been shown to achieve outstanding results in a wide range of cross-lingual transfer tasks. However, it remains unknown whether the optimization for different languages conditions the capacity of the models to generalize over syntactic structures, and how languages with syntactic phenomena of different complexity are affected. In this work, we explore the syntactic generalization capabilities of the monolingual and multilingual versions of BERT and RoBERTa. More specifically, we evaluate the syntactic generalization potential of the models on English and Spanish tests, comparing the syntactic abilities of monolingual and multilingual models on the same language (English), and of multilingual models on two different languages (English and Spanish). For English, we use the available SyntaxGym test suite; for Spanish, we introduce SyntaxGymES, a novel ensemble of targeted syntactic tests in Spanish, designed to evaluate the syntactic generalization capabilities of language models through the SyntaxGym online platform.

1 Introduction

Transformer-based neural models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), DistilBERT (Sanh et al., 2019), XLNet (Yang et al., 2019), etc. are excellent learners. They have been shown to capture a range of different types of linguistic information, from morphological (Edmiston, 2020) over syntactic (Hewitt and Manning, 2019) to lexico-semantic (Joshi et al., 2020). A particularly significant number of works study the degree to which these models capture and generalize over (i.e., learn to instantiate correctly in different contexts) syntactic phenomena, including, e.g., subject-verb agreement, long distance dependencies, garden path constructions, etc. (Linzen et al., 2016; Marvin and Linzen, 2018; Futrell et al., 2019; Wilcox et al., 2019a). However, most of these works focus on monolingual models, and, if the coverage of syntactic phenomena is considered systematically and in detail, it is mainly for English, as, e.g., (Hu et al., 2020a). This paper aims to shift the attention from monolingual to multilingual models and to emphasize the importance to also consider the syntactic phenomena of languages other than English when assessing the generalization potential of a model. More specifically, it systematically assesses how well multilingual models are capable to generalize over certain syntactic phenomena, compared to monolingual models, and how well they can do it not only for English, but also for Spanish.

Multilingual models such as mBERT (multilingual BERT, (Devlin et al., 2019)), XLM (Lample and Conneau, 2019) and XLM-R (Conneau et al., 2020) proved to achieve outstanding performance on cross-lingual language understanding tasks, including on low-resource languages for which only little training data is available. However, these models face the risk of running into what Conneau et al. (2020) refer to as "curse of multilinguality": adding languages to the model increases the performance on low-resource languages up to a point, after which the overall performance on monolingual and cross-lingual benchmarks degrades. The question is thus whether, and if yes to what degree, this degradation affects the syntactic generalization potential of multilingual models across languages.

The reason to extend the evaluation to other languages (in our case, Spanish) is that many existing syntactic phenomena such as determiner and adjective agreement within the noun phrase, subject pro-drop, or flexible word order – to name only a few – are not prominent or do not exist in English, while in Spanish all of them do.

Our evaluation methodology is similar to that by

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3799–3812

Hu et al. (2020a), who test 20 model type combinations and data sizes on 34 English syntactic test suites, and find substantial differences in the syntactic generalization performance across different models. We draw upon their tests to test the syntactic generalization potential of monolingual and multilingual transformer-based models for English, and upon the Spanish SyntaxGym introduced in this paper for Spanish. Tu run the tests, we use the SyntaxGym toolkit (Gauthier et al., 2020).

Our results show that, indeed, there is a substantial difference between the syntactic generalization potential of monolingual and multilingual models. But this difference depends on the language: While for English monolingual models (BERT and RoBERTa) offer a higher syntactic generalization than multilingual models (mBERT and XLM-R), this is not the case for Spanish, for which multilingual models (XLM-R) generalize better. Furthermore, multilingual models do not generalize equally well across languages, with mBERT generalizing, in general, better in English and XLM-R better in Spanish. Our experiments also show that it depends on the language how well a multilingual model captures a specific syntactic phenomenon such as, e.g., Agreement, Center-embedding or Garden Path.

The remainder of the paper is structured as follows. Section 2 introduces the work that is related to ours in terms of the evaluation methodology and, in particular, in terms of the assessment of multilingual language models. Section 3 describes the English test suites, and presents the novel Spanish SyntaxGym test suites. Section 4 details the models that we tested and outlines how we use them to evaluate the probability of a text sequence. Section 5 offers a detailed analysis of the syntactic generalization abilities of the monolingual and multilingual versions of BERT and RoBERTa, and Section 6 summarizes the implications that our work has for the use of multilingual language models.

2 Related Work

Our work on the evaluation of the capability of monolingual and multilingual transformer-based LMs to capture syntactic information is in line with a number of previous works, including, e.g., those that are based on psycholinguistic experiments, focusing on highly specific measures of language modeling performance and allowing to distinguish models with human-like representations of syntactic structure (Linzen et al., 2016; Lau et al., 2017; Gulordava et al., 2018; Marvin and Linzen, 2018; Futrell et al., 2019). Supervised probing models have been used to test for the presence of a wide range of linguistic phenomena (Conneau et al., 2018; Hewitt and Manning, 2019; Liu et al., 2019a; Tenney et al., 2019; Voita and Titov, 2020; Elazar et al., 2020). Warstadt et al. (2020) isolate specific phenomena in syntax, morphology, and semantics, finding that state-of-the-art models struggle with some subtle semantic and syntactic phenomena, such as negative polarity items and extraction islands.

Recently, a number of works also address the cross-language assessment of models. Hu et al. (2020b) introduces XTREME, a multi-task benchmark for evaluating the cross-lingual generalization capabilities of multilingual representations across 40 languages and 9 tasks. They show that while XLM-R reduces the difference between the performance on the English test set and all other languages compared to mBERT for tasks such as XQuAD and MLQA, it does not have the same impact on structured prediction tasks such as PoS and NER. Mueller et al. (2020) introduces a set of subject-verb agreement tests, showing that mBERT performs better than English BERT on Sentential Complements, Short VP Coordination, and Across a Prepositional Phrase, but worse on Within-an-Object Relative Clause, Across-an-Object Relative Clause and in Reflexive Anaphora Across a Relative Clause, and offers high syntactic accuracy on English, but noticeable deficiencies on other languages, most notably on those that do not use Latin script, as also noted by Hu et al. (2020b). On the same line, Rönnqvist et al. (2019) concludes that mBERT is not able to substitute a well-trained monolingual model in challenging tasks.

As already mentioned in Section 1, Hu et al. (2020a) assembled a set of English syntactic tests in order to assess the syntactic generalization potential of a number of different neural LMs (LSTM, ON-LSTM, RNNG and GPT-2). The tests are accessible through the SyntaxGym toolkit (Gauthier et al., 2020); cf. also Section 3.1. Our methodology is analogous, although our objective is different. Rather than comparing the performance of several monolingual models, we contrast the performance of monolingual and multilingual transformer-based models. Furthermore, while their only test suite source is the English SyntaxGym, we create and

use also a Spanish SyntaxGym; cf. Section 3.2.

3 Test Suites

For the English test suites, we used SyntaxGym,[1] an online platform that compiles a variety of linguistic tests used by Hu et al. (2020a) to assess the syntactic coverage of language models. It contains 34 suites, grouped into 6 different so-called *circuits*, a classification based on what is required from the models to process the targeted constructions. For the Spanish test suites, we created SyntaxGymEs, adapting 11 of the existing suites for English and building 15 new ones, including a whole new circuit. In what follows, we first introduce the original English SyntaxGym and then present in detail the novel SyntaxGymEs.

3.1 SyntaxGym for English

The tests in the SyntaxGym designed by Hu et al. (2020a) (henceforth also referred to as "English SyntaxGym") are based on the notion of *surprisal*. A sequence of words is given to a language model, which assigns a probability to each of the following candidate words. Given the syntactic properties of the considered language, some candidate words are less surprising than others, and so should be predicted by a language model. For instance, after the sequence *The cat*, the inflected word *sleeps* should be less surprising than *sleep*.

Each test consists of a list of ITEMS that vary in a controlled way according to a set of CONDITIONS determined by the experimental design. The other main component is a series of PREDICTIONS comparing surprisal values in specific regions of the items across conditions. If the relevant syntactic generalization has been learned by the model, the predictions should hold.

Moreover, some tests have versions with MODIFIERS, in which additional clauses or phrases have been embedded inside each item. These modifiers increase the linear distance between two co-varying items, making the task harder. Sometimes they also include a distractor word in the middle of a syntactic dependency, which can lead the models to misinterpret the dependency.

The test suites are arranged in terms of the following *circuits*:

• **Agreement**: Morphosyntactic phenomena that occur when the features of an item constrain another item to adopt a specific form. This is a marginal phenomenon in English, so the original circuit only includes 3 test suites on *Subject-verb number agreement*, all of them with modifiers (Marvin and Linzen, 2018).

• **Licensing**: A construction's need for the presence of a *licensor* to allow its occurrence in a sentence. The circuit consists of 4 suites on *Negative polarity items* (2 of them with modifiers) and 6 on *Reflexive pronouns* (all of them with modifiers), also from Marvin and Linzen (2018).

• **Center embedding**: Subordinate clauses that sit in the middle of their superordinate clause, creating nested dependencies. This circuit contains 2 test suites: *Center embedding* and *Center embedding with modifier*, from Wilcox et al. (2019a).

• **Long-distance dependencies (LDDs)**: LDDs occur when two constituents that are syntactically related do not appear adjacent to one another, but at a longer distance from one another. The circuit includes 6 suites on *Filler-gap dependencies* (2 with modifiers and 4 addressing extraction and hierarchy) from Wilcox et al. (2018) and Wilcox et al. (2019b), and 2 suites on *Cleft structure* that were first introduced in (Hu et al., 2020a).

• **Gross syntactic expectation**: Expectation for a large syntactic structure usually induced by subordinating adverbs or conjunctions. 4 test suites on *Subordination* (from Futrell et al. (2018), 3 of them with modifiers) constitute the circuit.

• **Garden path effects**: Effects that emerge when an incorrect but locally likely parse needs to be abandoned in favor of the correct one, once a specific word appears in the sentence. Two such effects are considered in this circuit: *Main verb/reduced relative clause (MVRR)* and *NP/Z garden paths*, with respectively 2 and 4 suites, all from Futrell et al. (2018).

3.2 SyntaxGymES: SyntaxGym for Spanish

For Spanish, we expand the tests in (Hu et al., 2020a) so as to cover language-specific phenomena. In this section, we detail which of the original tests we retained, which ones we modified, and which ones we added within each original circuit. A whole new circuit regarding the linear order of a sentence's basic constituents was also added, since flexibility in this respect is a characteristic that distinguishes Spanish (and other Romance languages) from English. For a more detailed description with examples and predictions, see the Supplementary Material; upon acceptance of the paper, Syntax-

[1] http://syntaxgym.org/

GymES will be published in the SyntaxGym platform http://syntaxgym.org/.

3.2.1 Notation

We follow the usual notations in linguistic literature. An asterisk '*' preceding an example signals that the sentence is ungrammatical, it violates some principle or constraint. A question mark '?' is used to indicate a marginal sentence, i.e., a sentence that is grammatical but very uncommon or that requires a non-straightforward interpretation. The exclamation mark '!' indicates a highly difficult sentence to process for the human mind.

3.2.2 Agreement

Unlike English, Spanish is a morphologically rich language, and as such it presents many morpho-syntactic phenomena related to agreement. For this reason, out of the six original circuits, **Agreement** was the one that underwent the most changes.

Regarding verbal agreement (constraints imposed on the verb by the subject), we adapted two existing test suites, **Subject-Verb Agreement with Object Relative Clause** and **Subject-Verb Agreement with Subject Relative Clause**, and created a new one, **Basic Subject-Verb Agreement**, in which both person and number features were taken into consideration.

(1) Tú cocinas
you.2SG cook.2SG

(2) * Tú cocinais/cocino/cocinan
you.2SG cook.2PL/1SG/3PL

As for nominal agreement (constraints that a noun's gender and number features can impose on the form of other words in the sentence), we also created several new test suites: **Determinant-Noun Agreement** simply pairs a noun with the four possible forms of the definite article (*el*, *la*, *los*, *las*), while **Adjective-Noun Agreement** pairs a noun with the four possible forms of an adjective that modifies it (we excluded articles to avoid providing extra information).

(3) La tienda vende discos usados
the store sells disc.M.PL used.M.PL

(4) * La tienda vende discos
the store sells disc.M.PL
usados/usado/usadas/usada
used.M.PL/M.SG/F.PL/F.SG

In addition to these two suites, we built similar ones for **Attribute Agreement** in copulative constructions, to which we added two versions with object or subject relative clauses as modifiers, and also for **Predicative Agreement** in constructions with subject or object predicative complement. The only difference here is that the two words that must agree are not adjacent anymore. In terms of predictions, the verb/noun with matching features should have a lower surprisal than the others, and the verb/noun that matches only one feature should have a lower surprisal that the one that doesn't match any.

3.2.3 Center Embedding

For this circuit, we adapted to Spanish the two existing test suites in English, creating **Center Embedding** and **Center Embedding with PP modifier**. In the basic suite, a relative clause is center embedded after the subject of the main clause. Verb transitivity and subject-verb plausibility are used to test if the models are capable of retaining the relevant information and predicting the verbs in the correct order.

3.2.4 Gross Syntactic Expectation

From the four original suites in this circuit, we adapted three of them: **Subordination**, and two of its versions with modifiers, **Subordination with Object Relative Clause** and **Subordination with Subject Relative Clause**. Given a sentence that starts with a typically subordinating adverb or conjunction, these suites test the models' ability to maintain the expectation for the onset of a matrix clause for as long as the subordinate one lasts.

3.2.5 Long-distance Dependencies

Filler-gap dependencies are an example of LDDs. They occur when a phrase (the filler) is realized somewhere in the sentence, but is semantically interpreted at some other point (the gap). For this circuit, we created a **Basic Filler-Gap Dependencies** test and adapted from the original English circuit a version that includes modifiers, **Filler-Gap Dependencies with Three Sentencial Embeddings**. Embedding three sentences between filler and gap makes the task more challenging. We also adapted to Spanish the novel **Pseudo-Cleft Structures** suite introduced in (Hu et al., 2020a).

3.2.6 Garden Path Effects

The Garden Path effect can be created by several syntactic ambiguities that differ cross-linguistically. The Main Verb/Reduced Relative garden path effect was the subject of two suites in the original

English circuit, but it does not translate to Spanish, so those suites were not included in Spanish SyntaxGym.

On the other hand, the ambiguity responsible for NP/Z also holds for Spanish. Here, an NP is initially interpreted as the object in a subordinate clause when it actually is the subject of the main clause (the subordinate clause having a Zero/null object). The ambiguity can be prevented with a comma, but also by placing an overt object in the subordinate clause, as is done in **NP/Z Garden Path Effect (with Overt Object)**, or by substituting its verb with a pure intransitive verb, as is done in **NP/Z Garden Path Effect (with Intransitive Verb)**. Both suites correspond to Spanish adaptations of the two original suites regarding this effect.

(5) !Mientras ella leía sus manuscritos se volaron por la ventana.

!'While she read her manuscripts went out the window.'

(6) Mientras ella [dormía]/[leía un libro]/[leía,] sus manuscritos se volaron por la ventana.

'While she [slept]/[read a book]/[read,] her manuscripts went out the window.'

3.2.7 Licensing

Negative polarity items (NPIs), like *any* or *ever* in English, are examples of words that need to be licensed by negation. Since Spanish NPIs do not function exactly in the same way, we took the original NPI Licensing test as inspiration and created two new suites: **Negative Polarity Items** and **NPIs and Polarity Agreement**.

Constructions with verbs in subjunctive mood also require the presence of a licensor. In Spanish, a verb expressing feelings (e.g. of joy, surprise, pleasantness) in the main clause, creates the expectation for subjunctive mood in the subordinate clause. This was the basis for a new test suite: **Subjunctive Mood and Verbs that Express Feeling**.

(7) Espero que mañana llueva/*lloverá.
(I)hope that tomorrow rain.SUB/will.rainIND

'I hope it [rains]/[will rain] tomorrow.'

The other new suite in this circuit, **Subjunctive Mood, Negation and Belief Verbs**, relies on the fact that belief verbs can also license subjunctive mood, but only when combined with negation:

(8) No creo que mañana
NEG (I)believe that tomorrow
llueva/*lloverá.
rain.SUB/will.rain.IND

'I don't think it [rain]/[will rain] tomorrow.'

(9) Creo que mañana no
(I)believe that tomorrow NEG
lloverá/*llueva.
will.rain.IND/rain.SUB

'I think it [won't]/[don't] rain tomorrow.'

3.2.8 Linearization

One of the main syntactic distinctions between languages is constituent order within the sentence. But, in addition to the canonical order in which these elements appear, languages also differ in their flexibility to alter that order. Spanish allows some flexibility, which was the basis for three new test suites.

For **Subject–Auxiliary Verb–Main Verb Linearization**, the possibility to postpone the subject is compared with the rigidity of the relation between main and auxiliary verb, which must be adjacent and do not allow inversion:

(10) Juan [ha comido]/*[comido ha].

'John [has eaten]/[eaten has].'

(11) Ha [comido Juan]/*[Juan comido].

*'Has [eaten John]/[John eaten].'

In the **Subject–Verb–Object Linearization** test, we compare the phenomenon in affirmative versus interrogative sentences. In Spanish, word order flexibility holds for affirmative sentences, but not for interrogative ones, where subject-verb inversion is compulsory:

(12) Ana compró un libro. / Compró un libro Ana.

'Ann bought a book. / Bought a book Ann.'

(13) ¿Qué compró Ana? / *¿Qué Ana compró?

'What did Ann buy? / *What Ana did buy?'

Word order variations also appear within the NP, as captured by the **Noun-Adjective and Noun-PP Linearization** test. Contrary to English, Spanish adjectives usually come after the noun. But again, the language allows for some flexibility and they can be swapped. This possibility, however, does not apply to other noun modifiers like prepositional phrases:

(14) Construyó una [mesa robusta]/[robusta mesa].

'He built a [sturdy table]/[table sturdy].'

(15) Construyó una [mesa de madera]/*[de madera mesa].

'He built a [wooden table]/*[table wooden].'

4 Experiments

We test the base cased versions of BERT and mBERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b) and XLM-R (Conneau et al., 2020) on the English SyntaxGym and BETO (Canete et al., 2020), mBERT and XLM-R on the Spanish SyntaxGym. To run the experiments, we use the SyntaxGym toolkit (Gauthier et al., 2020).

4.1 Experimental Setup

The SyntaxGym test suites are designed from the perspective of sentence generation, i.e., with the hypothesis that if a model has correctly learned some relevant syntactic generalization, it should assign higher probability to grammatical and natural continuations of sentences. This requires asking the models to predict the next token given a context of previous tokens, in a left-to-right generative fashion. However, BERT-based and RoBERTa-based families of models (in our case, BERT and mBERT on the one side, and RoBERTa and XLM-R on the other side) are bidirectional, they are trained with a masked language modeling objective to predict a word given its left and right context. In this work, we follow Wang and Cho (2019)'s sequential sampling procedure to evaluate the probability of a text sequence, encoding unidirectional context in the forward direction. To compute the probability distribution for a sentence with N tokens, we start with a sequence of $N+2$ tokens: a *begin_of_sentence* token plus $N+1$ *mask* tokens, where the last *mask* corresponds to the *end_of_sentence* token. For each token position i in $[1, N]$, we compute the probability distribution over the vocabulary given the left context of the original sequence, and select the probability assigned by the model to the original word.

For example, in an agreement test with the sentence '*The girls run fast.*', a model that has properly learned agreement should assign a higher probability to *run* than to *runs* for the third word. In order to test it, we feed the tokens sequence *[[bos] [The] [girls] [mask] [mask] [mask] [mask]]* to the model, and compare the probabilities assigned by the model to *run* and *runs* for position 4.

	Average SG performance	
Model	**English**	**Spanish**
BERT	77.80	—
RoBERTa	82.04	—
mBERT	77.55	72.31
XLM-R	71.84	78.50
BETO	—	67.92

Table 1: Average SG score by model class for the English and Spanish tests.

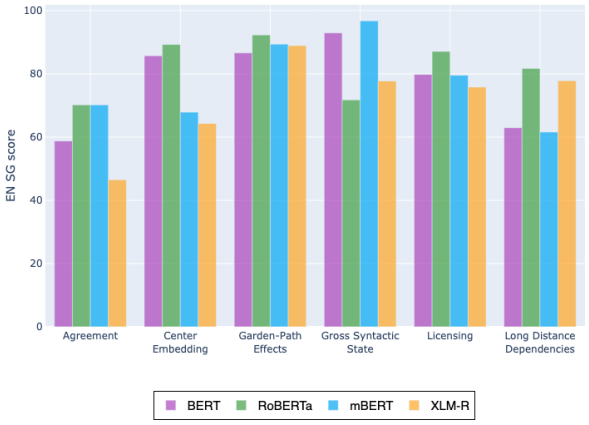

Figure 1: Performance accuracy across English circuits

4.2 Results of the experiments

This section summarizes the results of our experiments that aim to: (i) contrast the performance of monolingual and multilingual models on English and Spanish and (ii) provide insights on the performance of the multilingual models across languages.

Table 1 shows the average SyntaxGym (SG) performance of the evaluated monolingual and multilingual models on the English and Spanish SyntaxGyms. Figures 1 and 2 zoom in on the performance of the tested models with respect to specific circuits for English and Spanish respectively.

Six of the English test suites (Center Embedding, Cleft structure, MVRR, NPZ-Verb, NPZ-Object, Subordination) and five of the Spanish test suites (Attribute Agreement, Basic Subject-Verb Agreement, Subordination, Center Embedding, Basic Filler-Gap Dependencies) include tests with and without modifiers, i.e,. intervening content inserted before the critical region. Figures 3 and 4 show the models' average scores in these test suites, without modifiers (dark bars) and with modifiers (light bars), evaluating how robust each model is with

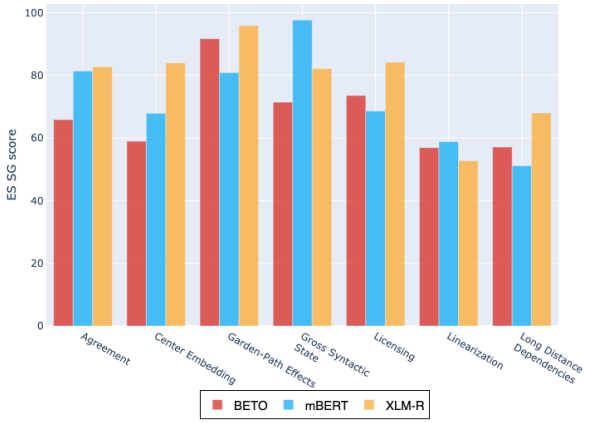

Figure 2: Performance accuracy across Spanish circuits

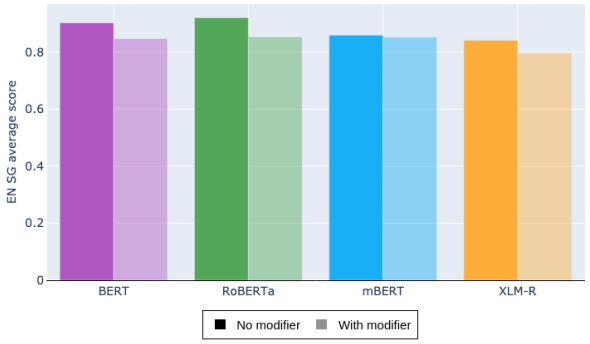

Figure 3: Models average English SG score in Center Embedding, Cleft structure, MVRR, NPZ-Verb, NPZ-Object and Subordination, with and without modifiers.

respect to the corresponding content.

5 Discussion

Let us assess in detail the results of the experiments from above. In what follows, we compare the performance of monolingual with the performance of multilingual models and analyze the cross-language performance of multilingual models, as well as the stability of the individual models with respect to modifiers.

5.1 Monolingual vs multilingual models

RoBERTa shows an overall higher performance than the other models for English (Table 1). This is not surprising since it is trained on 10 times more data than BERT, and it has been shown to improve over BERT in many NLU tasks. However, while mBERT does not seem to lose performance compared to BERT, XLM-R loses around 10 points compared to RoBERTa. As XLM-R is specifically designed to offer a more balanced performance across languages, with a special focus on low-resource languages, it appears natural that

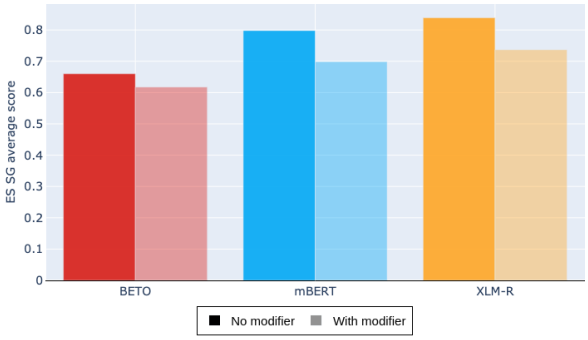

Figure 4: Models average Spanish SG score in Attribute Agreement, Subject-Verb Agreement, Subordination, Center Embedding and Filler-Gap Dependencies, with and without modifiers.

it loses some performance on high-resource languages such as English. For Spanish, the multilingual models clearly outperform the monolingual model. This is likely due to the fact that while BETO and mBERT are of comparable size and are trained with the same amount of data (16GB), BETO is only trained with a Masked Language Modeling (MLM) objective, and mBERT is trained on MLM and Next Sentence Prediction (NSP). On the other hand, XLM-R is also only trained on MLM, but it is trained on more than 2TB of data, 53 GB corresponding to Spanish data.

RoBERTa outperforms all other models in all the English circuits (cf. Figure 1), except in Gross Syntactic State, in which BERT-based models clearly outperform RoBERTa-based models, and the multilingual model outperforms the monolingual one in both families. Intuitively, we believe that the NSP training objective of BERT-based models helps them to better understand the relation between two sentences, and this knowledge can also be applied to the relation between two clauses (which is the basis of the Gross Syntactic State circuit). Comparing the BERT and RoBERTa model families, it is interesting to notice that while RoBERTa outperforms XLM-R in all circuits except Gross Syntactic State, BERT only outperforms mBERT in 3 of them.

Interestingly, all models seem to struggle with Agreement in English. This observation is aligned with Mueller et al. (2020)'s hypothesis that language models learn better hierarchical syntactic generalizations in morphologically complex languages (such as, e.g., Spanish), which frequently provide overt cues to syntactic structure, than in morphologically simpler languages (such as, e.g., English). Indeed, the fact that XLM-R offers the

lowest performance may be related to the fact that the model has been more exposed to more complex languages than the others. For Long Distance Dependencies, BERT-based models show a low performance compared to RoBERTa-based models. This might be due to the different training procedures adopted in both model families (i.e., that RoBERTa does not include the Next Sentence Prediction task (as BERT does) and introduces dynamic masking).

On the other hand, in specific circuits for Spanish (cf. Figure 2) XLM-R outperforms the other two models in 5 out of 7 circuits. As observed for English, the BERT-based models struggle with the Long Distance Dependencies tests, and mBERT offers an outstanding performance in Gross Syntactic State. The monolingual model, BETO, is outperformed by mBERT in 4 out of 7 tests, and by XLM-R in all 6 out of 7 tests. As mentioned before, these differences may be related to the fact that, unlike BERT, BETO is not trained with the NSP objective; but also to the difference in training data size: 16GB for BETO vs. more than 2TB (of which 53GB of Spanish data) for XLM-R.

All models offer a low performance in the new Linearization test for Spanish. A more in-depth investigation is necessary to explain this. The test has been designed with literary Peninsular Spanish in mind, and it is possible that the training data may not contain enough samples that show the targeted word order varieties, or may contain data from American Spanish sources, which may show differences in canonical word order with respect to Peninsular Spanish.

5.2 Cross-language multilingual models performance

As shown in Table 1, multilingual models do not syntactically generalize equally well in both languages. While mBERT offers a better generalization in English, outperforming XLM-R by almost 6 points, XLM-R generalizes better in Spanish, outperforming mBERT by 6 points. This observation corroborates our intuition that XLM-R sacrifices performance in high-resource languages (e.g., English, with 300GB of training data) to be able to offer a more balanced performance across languages (e.g., Spanish, with 53GB of training data).

Comparing Figures 1 and 2, we observe improvements in the Spanish tests for XLM-R in 4 out of 6 circuits, particularly noticeable in Agreement and Center Embedding, while it loses around 10 points in Long Distance Dependencies. On the other hand, mBERT also shows a big improvement in the Spanish tests in Agreement, while it loses performance in Garden Path Effects, Licensing and Long Distance Dependencies.

5.3 Model stability with respect to modifiers

Since modifiers increase the linear distance between the elements in a dependency structure, thus making the task more demanding, stability in this respect indicates that models have robustly learnt the appropriate syntactic generalization and do not depend that much on adjacency. Figures 3 and 4 show the models' average scores in those test suites that have two versions: without modifiers (dark bars) and with modifiers (light bars). As was intuitively expected, all the models offer a higher performance in the tests without modifiers. While for English the multilingual models are the less affected, for Spanish BETO seems to be more robust than the multilingual models, even though it offers a lower performance.

6 Conclusions

In this paper, we assessed the syntactic generalization potential of selected transformer-based language models on English and Spanish. We have shown that multilingual models do not generalize equally well across languages: mBERT generalizes better for phenomena in English, while XLM-R does it better for phenomena in Spanish. We have also shown that the answer to the question whether monolingual or multilingual models generalize better is equally language-specific: the monolingual RoBERTa generalizes better on English, while the multilingual XLM-R generalizes better on Spanish. While it is possible that the multilingual abstractions captured by XLM-R become useful for morphologically rich languages such as Spanish, this difference may also be related to the difference in the amount of training data used to train BETO and XLM-R, and therefore it is possible that a monolingual model trained with a comparable amount of data could outperform the multilingual models.

The performance of all models is affected by the presence of modifiers, which shows that the complexity of the syntactic structure is still a challenge. In general, each syntactic phenomenon deserves attention. For instance, Agreement in English is hard to learn, given the scarcity of cues (especially if compared to a morphologically rich language),

and so is Linearization in Spanish.

As far as the nature of the training procedures of the models is concerned, the lack of Next Sentence Prediction (NSP) objective in the RoBERTa model family seems to harm BETO, but not XLM-R; this suggests that the performance of BETO may be improved with (much) more training data. It also seems to harm in the case of the Gross Syntactic State circuit, suggesting that RoBERTa-based models may also benefit from complementary training objectives in their pretraining procedure.

Overall, our experiments have also shown the importance of testing models on a wider range of languages, in particular, morphologically rich ones. As part of our future work, we plan to expand further SyntaxGymES and develop SyntaxGyms for a number of other selected languages. Also, careful examination of a wider range of material is necessary to ensure that important phenomena are not left out, so as to assess the actual coverage of the test suites.

Acknowledgments

This work has been partially funded by the European Commission via its H2020 Research Program under the contract numbers 779962, 786731, 825079, and 870930.

References

José Canete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. *PML4DC at ICLR*, 2020.

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, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Edmiston. 2020. A systematic analysis of morphological content in BERT models for multiple languages. *arXiv preprint arXiv:2004.03032v1*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2020. When Bert forgets how to POS: Amnesic probing of linguistic properties and MLM predictions. *arXiv preprint arXiv:2006.00995*.

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

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020a. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020b. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2020. Contextualized representations using textual encyclopedic knowledge. *arXiv preprint arXiv:2004.12006*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

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.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019a. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 181–190, Florence, Italy. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019b. Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3302–3312, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.

A Spanish SyntaxGym: Description of Test Suites

This appendix lists and describes all the test suites compiled for Spanish SyntaxGym. Each test consists of a list of ITEMS that vary in a controlled way according to a set of CONDITIONS determined by the experimental design. A series of PREDICTIONS compare surprisal values at specific regions of the items across conditions. Some tests have versions with MODIFIERS that increase the linear distance between two co-varying items, making the task more demanding.

The test suites are arranged in terms of *circuits* of related syntactic phenomena. Each of the following sections corresponds to one of these circuits.

Notation. An asterisk * signals an ungrammatical sentence, a question mark ? indicates a marginal sentence (grammatical but very uncommon or requiring a difficult interpretation), an exclamation point ! denotes high processing difficulty.

A.1 Agreement

Agreement is a morpho-syntactic phenomenon that occurs when the features of an item constrain another item to adopt a specific form.

• **Basic Subject-Verb Agreement**. New suite. Spanish finite verbs in any tense/mood have six inflected forms according to person and number features. The verb's features the subject's, otherwise the result is ungrammatical.

(16) Tú cocinas
you.2SG cook.2SG

(17) * Tú cocinais/cocino/cocinan
you.2SG cook.2PL/1SG/3PL

Predictions: The surprisal at the verb region is expected to be lower when it matches the subject than in any other condition. It is also expected to be lower when at least one of the features (person or number) agrees than when both disagree.

• **Subject-Verb Agreement with Subject Relative Clause**. Adapted from English. This test focuses on number agreement. The subject relative clause includes a *distractor* NP differing in number with the subject.

(18) El fontanero que ayudó a los
the.SG plumber that helped.3SG to thePL
albañiles trabaja/*trabajan los sábados.
bricklayers work.3SG/3PL the saturdays.
'The plumber who helped the bricklayers works/*work on saturdays.'

(19) Los fontaneros que ayudaron al
the.PL plumbers that helped.3SG to.thePL
albañil *trabaja/trabajan los sábados.
bricklayer work.3PL/3SG the saturdays.
'The plumbers who helped the bricklayer *works/work on saturdays.'

Predictions: A successful model should place higher probability to the verb agreeing with the subject (instead of the distractor) both in singular and in plural.

• **Subject-Verb Agreement with Object Relative Clause**. Adapted from English. Equal to the previous one, but with an object relative clause.

Nominal agreement was the basis for the following 6 new test suites. All of them share the same predictions: the surprisals should be lower when both gender and number features in the second word of the agreement relation match those in the first word. They should also be lower when only one of the features agrees than when both disagree.

• **Determiner-Noun Agreement**. New suite. The four possible forms of the definite article are paired with different nouns.

(20) El/*La/*Los/*Las gato
the.M.SG/*F.SG/*M.PL/*F.PL cat

• **Adjective-Noun Agreement**. New suite. The test pairs a noun with the four possible forms of an adjective that modifies it (we used constructions without determiner to avoid providing the models with extra information).

(21) La tienda vende discos
the store sells discs
usados/*usado/*usadas/*usada
used.M.PL/M.SG/F.PL/F.SG
'The store sells second-hand discs.'

• **Attribute Agreement**. New suite. Here, a noun is paired with and adjective through a copulative construction. This suite has 2 versions with object or subject relative clauses as modifiers.

(22) El piso está vacío/*vacía/*vacíos/*vacías
the flat is empty.M.SG/*F.SG/*M.PL/*F.PL

• **Predicative Agreement**. New suite. The subject or the object is paired with an adjective functioning as a predicative complement.

(23) Los niños llegaron
the children arrived
cansados/*cansado/*cansadas/*cansada
tired.M.PL/*M.SG/*F.PL/*F.SG
'The children arrived tired.'

A.2 Center Embedding

A center embedded clause is a subordinate clause that sits in the middle of its superordinate clause, creating nested dependencies that may be challenging for the models.

• **Center Embedding**. Adapted from English. A relative clause is center embedded after the subject of the main clause. Verb transitivity and subject-verb plausibility are used to test if the models are capable of retaining the relevant information and predicting the verbs in the correct order.

(24) La tormenta que el capitán [capeó amainó]/?[amainó capeó].

'The storm the captain [weathered abated]/?[abated weathered].'

Prediction: The surprisal of the combination of verbs should be smaller when their relative order creates a plausible sentence than when it creates an implausible one.

• **Center Embedding with modifier**. In the version with modifier, a prepositional phrase is inserted after the subject of the subordinate clause.

A.3 Gross Syntactic State

Expectation for a large syntactic structure at some point within the sentence.

• **Subordination**. Adapted from English. A sentence starting with a subordinate clause creates the expectation for the onset of a matrix clause for as long as the subordinate one lasts.

(25) ?(Mientras) ella miraba los resultados, el doctor entró en la habitación.

'While she looked at the results, the doctor entered the room.'

(26) (*Mientras) ella miraba los resultados.

'(*While) she looked at the results.'

Predictions: The surprisal for the lack of a second clause should be higher when there is a subordinating conjunction or adverb than where there is not. But having two clauses joined by a conjunction/adverb should be less surprising than their juxtaposition.

• **Subordination with Object Relative Clause** and **Subordination with Subject Relative Clause**. Adapted from English. Versions of the previous suite but with a modifier.

A.4 Long-distance Dependencies

LDDs occur when two syntactically related groups do not appear adjacent to one another but at a longer distance from one another.

• **Basic Filler-Gap Dependencies**. New suite, a simplified version of the existing FGD tests for English. FGDs occur when a phrase (the filler) is realized somewhere in the sentence but is semantically interpreted at some other point (the gap).

(27) Yo sé [lo que]/*que tu amigo tiró _ al suelo.

'I know what/*that your friend threw _.'

(28) Yo sé *[lo que]/que tu amigo tiró una colilla al suelo.

'I know *what/that your friend threw a cigarette butt.'

Predictions: The overt object should be more surprising when there is a filler when there is not. We also expect lower surprisal when the sentence has a filler later followed by gap than when it has a conjunction instead but the gap remains.

• **Filler-Gap Dependencies with Three Sentencial Embeddings**. Adapted from English. It is a version of the previous test that includes a modifier (three sentential embeddings) between filler and gap. This makes the task more challenging. The predictions, though, remain the same.

• **Pseudo-Cleft Structures** Adapted from English. A pseudo-cleft or wh-cleft is formed by a wh-element extracting content from a relative clause joined by a copula to a constituent that provides the content requested by the wh-element. The extracted constituent can be a NP or a VP. In the VP case, the verb in the relative clause must be an inflected form of 'hacer' ('to do').

(29) Lo que tú difundiste/?hiciste fue un rumor.

'What you spread/*did was a rumor.'

(30) Lo que tú *difundiste/hiciste fue confirmar un rumor.

'What you *spread/did was confirm a rumor.'

Predictions: The surprisal should be lower for the extracted VP when the verb in the relative clause is a light verb (*hacer* – 'to do') than when it is not, but it should be higher for the extracted NP when the verb is light than when it is semantically heavier and matches the NP. In addition, the difference in the first case should be more important than in the

second one. This happens because the light verb admits a wider range of objects, whereas in the first case, one of the options is syntactically incorrect.

A.5 Garden Path Effects

Garden-path effects emerge when an incorrect but locally likely parse needs to be abandoned in favor of the correct one. In the NP/Z garden path, an NP is initially interpreted as the object in a subordinate clause, but when the main verb appears, this NP should be reinterpreted as its subject. The effect can be prevented by adding a comma, but also by placing an overt object in the subordinate clause, or by substituting its verb with a purely intransitive one. These are the basis for the next two suites.

• **NP/Z Garden Path Effect (Overt Object)**.

• **NP/Z Garden Path Effect (Intransitive Verb)**. Both adapted from English.

(31) !Mientras ella leía sus manuscritos se volaron por la ventana.

!'While she read her manuscripts went out the window.'

(32) Mientras ella [dormía]/[leía un libro]/[leía,] sus manuscritos se volaron por la ventana.

'While she [slept]/[read a book]/[read,] her manuscripts went out the window.'

Predictions: The main verb should be more surprising in the garden path condition than when the effect has been prevented either by the comma or by interfering with the verb. Moreover, the difference in surprisal should be bigger when the comma is essential to solve the garden path effect than when it is not.

A.6 Licensing

In natural language, some words or constructions need the presence of a licensor to allow their occurrence in a sentence. This happens with NPIs (Negative polarity items) and subjunctive mood, for instance.

• **Negative Polarity Items and Polarity Agreement**. New suite. In Spanish, NPIs that follow the verb (such as *nunca* 'never', *nadie* 'nobody', and *nada* 'nothing') need to be licensed by negation. This 'double negative' does not result in an affirmative, it is a sort of polarity agreement.

(33) Yo no bebo nunca/?siempre.
I NEG drink never/always
'I never drink./I don't drink always.'

(34) Yo bebo *nunca/siempre.
'I *ever/always drink.'

Predictions: We expect the surprisals in both agreeing conditions (negative-NPI, positive-PPI) to be lower than in any of the non-agreeing conditions (negative-PPI, positive-NPI).

• **Negative Polarity Items**. New suite. NPIs also need to be in the scope of the negation to be licensed by it. This suite compares between a negative particle that "commands" the NPI and one that doesn't.

(35) Tú, como no mirabas por la ventana,
You, as NEG looked by the window,
*(no) has visto a nadie.
NEG have seen at nobody
'As you weren't looking through the window, you have *(not) seen anybody.'

(36) Tú, como mirabas por la ventana, *(no)
You, as looked by the window, NEG
has visto a nadie.
have seen at nobody
'As you were looking through the window, you have *(not) seen anybody.'

Predictions: The NPI should be more surprising when there isn't a negative particle that commands it, independently of the presence of another one that does not command it.

• **Subjunctive Mood and Verbs that Express Feeling**. New suite. Feeling verbs that introduce a subordinate clause serve as licensors for subjunctive mood, whereas other type of verbs do not.

(37) Espero que mañana llueva/*lloverá.
(I)hope that tomorrow rain.SUB/will.rain.IND
'I hope it rains/*[will rain] tomorrow.'

(38) Sé que mañana
(I)know that tomorrow
*llueva/lloverá.
rain.SUB/will.rain.IND
'I know it [will rain]/rains tomorrow.'

Predictions: Subjunctive mood should be less surprising than indicative mood when the verb in the main clause expresses feelings. But when it doesn't, subjunctive should be more surprising than indicative mood. Moreover, subjunctive mood should also be more surprising with a feeling verb than with a non-feeling verb.

• **Subjunctive Mood, Negation and Belief Verbs**. New suite. Belief verbs can also license subjunctive mood, but only when combined with negation.

(39) No creo que mañana
NEG believe that tomorrow
llueva/*lloverá.
rain.SUB/will.rain.IND
'I don't think it rains/[will rain] tomorrow.'

(40) Creo que mañana no
(I)believe that tomorrow NEG
*llueva/lloverá.
rain.SUB/will.rain.IND
'I think it rains/[won't rain] tomorrow.'

Predictions: The subordinate verb should be less surprising in subjunctive than in indicative mood when the main clause is negated. However, the contrary should hold when the subordinate clause is negated but the main one is not. In addition, subjunctive mood should be less surprising when the negation is in the main clause than when it is in the subordinate clause.

A.7 Linearization

Constituent order is commonly used in linguistics as a way to classify languages. But, in addition to the canonical order in which elements appear, languages also differ in their flexibility to alter that order.

• **Subject – Auxiliary Verb – Main Verb Linearization**. New suite. Subject-verb order admits inversion in Spanish but main and auxiliary verb do not and they must be adjacent.

(41) Juan ha comido. / Ha comido Juan
'John has eaten. / Has eaten John.'

(42) *Juan comido ha. / *Ha Juan comido.
'John eaten has. / Has John eaten.'

Predictions: The postposed subject should be less surprising than any of the alterations involving auxiliary and main verb. The canonical SV order, however, should be less surprising than postposing the subject, and the difference in this case should be less important than the differences in the first two cases.

• **Subject – Verb – Object Linearization**. New test. In Spanish, word order flexibility holds for affirmative sentences but not for interrogative ones, where subject-verb inversion is compulsory.

(43) Ana compró un libro/Compró un libro Ana.
'Ann bought a book. / Bought a book Ann.'

(44) ¿Qué compró Ana? / ¿Qué Ana compró?
'What did Ana buy? / 'What Ana did buy?'

Predictions: A postposed subject in an affirmative sentence should be less surprising than lack of SV inversion in an interrogative one. The canonical SV order in the affirmative sentence, however, should be less surprising than postposing the subject, and the difference in this case should be less important than the difference in the first one.

• **Noun-Adjective and Noun-PP Linearization**. New suite. Spanish adjectives usually come after the noun, but this order can be inverted. Other noun modifiers like prepositional phrases cannot.

(45) Construyó una [mesa robusta]/[robusta mesa].
'He built a [sturdy table]/[table sturdy].'

(46) Construyó una [mesa de madera]/*[de madera mesa].
'He built a [wooden table]/*[table wooden].'

Predictions: A PP preceding the noun should be more surprising than one following it. An adjective preceding the noun should also be more surprising than one following it, but the difference in this case should be less important than in the first one.