Assessing the Agreement Competence of Large Language Models

Alba Táboas García
NLP Group
Pompeu Fabra University
alba.taboas@upf.edu

Leo Wanner
Barcelona Supercomputing Center &
Catalan Institute for Research
and Advanced Studies (ICREA)
leo.wanner@bsc.es

Abstract

While the competence of LLMs to cope with agreement constraints has been widely tested for English, only a very limited number of works deals with morphologically rich(er) languages. In this work, we experiment with 25 mono- and multilingual LLMs, applying them to a collection of more than 5,000 test examples that cover the main agreement phenomena in three Romance languages (Italian, Portuguese, and Spanish) and one Slavic Language (Russian). We identify which of the agreement phenomena are most difficult for which models and challenge some common assumptions of what makes a good model. The test suites into which the test examples are organized are openly available and can be easily adapted to other agreement phenomena and other languages for further research.

1 Introduction

Agreement is one of the linguistic phenomena usually invoked to illustrate dependency in language (Mel'čuk, 2009). It reflects the fact that, within a sentence, the wordform $\mathbf{w}_2$ (the *target*) co-varies with a wordform $\mathbf{w}_1$ (the *controller*) with respect to selected morpho-syntactic features.[1] In English, agreement is most obvious in the number covariance in subject–verb constructions. Therefore, it is not surprising that compliance with subject–verb agreement restrictions has become a key diagnosis for the syntactic competence of neural language models. It allows researchers to evaluate whether a model truly captures hierarchical structure rather than merely learning surface-level patterns (Linzen et al., 2016; Goldberg, 2019; Nastase et al., 2024). However, in morphologically rich(er) languages, agreement is a considerably more prominent phenomenon than in English. *Canonical agreement features* include not only number, but also person, gender, and grammatical case; certain numerals as well as possessive and qualitative adjectives can also act as controllers; and among targets, in addition to verbs and adjectives as in English, we also find pronouns, numerals, adverbs, adpositions, nouns, etc. (Corbett, 2006). The goal of our work is to assess the competence of state-of-the-art neural models in handling a broader range of agreement features than encountered in English. To this end, we selected three Romance languages: Italian, Portuguese and Spanish, and Russian as a representative of Slavic languages. For both the Romance languages triple and for Russian, we define ten different agreement tests. These tests are applied to a number of monolingual and multilingual models. Our experiments show that despite a good overall performance, language models struggle with complex agreement constructions, with monolingual models outperforming multilingual ones. But model size and training data size, which are typically correlated with model performance, do not play a significant role in this case.

[1] Mel'čuk (1993) makes a distinction between the phenomena of *agreement*, *government*, and *congruence*. In our study, we refrain from such a detailed differentiation.

2 Related Work

Linzen et al. (2016) were among the first to propose using a model's assignment of higher probability to targets with the correct number grammeme (as opposed to the incorrect one) in subject–verb agreement in English as a benchmark for evaluating its syntactic competence. This test methodology, known as *targeted syntactic evaluation*, has since then been extended to test other syntactic structures in English, such as, e.g., reflexive anaphora agreement and licensing of negative polarity items (Marvin and Linzen, 2018), filler-gap constructions and island constraints (Wilcox et al., 2018, 2019), garden path effects (Futrell et al., 2018, 2019), or all of the above (Hu et al., 2020).

Early work on LSTMs includes studies on num-

Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025), pages 36–53

ber and case agreement in Basque (Ravfogel et al., 2018) and long-distance number agreement in Italian, English, Hebrew, Russian, covering both subject–verb agreement and noun–adjective agreement, where applicable (Gulordava et al., 2018). With transformer models, this line of research expanded to German (Zaczynska et al., 2020) as well as French, Hebrew and Russian (Mueller et al., 2020). Spanish has been the focus of broader agreement testing also beyond subject–verb (Pérez-Mayos et al., 2021), while targeted evaluations have also been developed for Galician and Portuguese (Garcia and Crespo-Otero, 2022; de Dios-Flores and Garcia, 2022). More recently, Basque auxiliary verb agreement with its complements and noun-class agreement in Swahili has been studied as well (Kryvosheieva and Levy, 2025). Overall, the studies have shown that the architecture of the tested model plays a role in its performance on agreement tasks. While LSTMs capture syntactic structure under certain conditions, their performance degrades in more complex configurations. Transformer-based models, on the other hand, exhibit a more robust agreement performance, especially in English. However, they are still sensitive to constructions involving agreement attractors or long-distance dependencies and perform worse on languages with a richer morphology, such, e.g., Basque, Hebrew, Russian, or Swahili.

Our work differs from previous works in two key ways. First, our test suites were manually curated by linguists to ensure that all examples are well-formed and semantically plausible, contrasting with other approaches that often rely on synthetically generated stimuli. Second, we provide a comparison across a wide range of monolingual and multilingual language models, which allows us to assess what elements have an impact on their agreement performance.

3 Agreement Test Suites

Following Hu et al. (2020); Pérez-Mayos et al. (2021) and others, we group the different tests into *test suites*. Each test suite focuses on a specific agreement rule and contains several *items*. Each item consists of a sentence sample adhering to the given rule and one or more samples that systematically vary from the first sample in the way they violate this rule. All test suites are based on the premise that a model should yield higher *surprisal* [2] values for a target whose features fail to match those of its controller than one with correctly matching features.

To assess model performance under more realistic conditions, some test suites include an adversarial sample featuring grammatical constructions that increase the linear distance between the target and the controller. These constructions also incorporate what is commonly referred to in the literature as an *agreement attractor*, i.e., an element that shares its part of speech with the controller but that differs from it in the values for some or all of the agreement features involved in the relation. We paid special attention to select, when possible, agreement attractors that remain semantically plausible in relation to the target. The following examples from one of our Spanish test suites illustrate a regular test sentence and its adversarial counterpart (the controller appears in bold, the correct target in blue, the incorrect target in red, and the attractor is underlined):

(1) Las **voluntarias** cayeron
the **volunteer.F.PL** fell
enfermas/*enfermos
ill.F.PL/*ill.M.PL
'The volunteers fell ill.'

(2) Las **voluntarias** [que ayudaron a los
the **volunteer.F.PL** who helped to the
refugiados] cayeron enfermas/*enfermos
refugee.M.PL fell ill.F.PL/*ill.M.PL
'The volunteers who helped the refugees fell ill.'

In (2), the construction increasing linear distance between the controller and target is the relative clause in brackets, where the agreement attractor is underlined. Note that the attractor semantically fits both the main verb and the target, and that its features match those of the incorrect target, making the test more difficult for the models.

Regarding the **building process**, every example in each test suite was hand-crafted by a linguist fluent in the specific language[3]. Starting from a grammatical sentence, ungrammatical variants were created by altering morphological features involved in

[2] Following the terminology in Information Theory, we use the term *surprisal* to denote the negative log probability of a token (Samson, 1953), and in line with its use in psycholinguistics (Hale, 2001; Hu et al., 2020). Note, however, that to better capture contrasts between matching and mismatching controller wordforms, we rely on a different scoring metric; cf., Section 4.2.

[3] The time required to create comparable test suites varies with the designer's linguistic expertise and creativity, but our examples can serve as a helpful starting point for future work.

agreement. To reduce frequency effects and bias, tests balanced all relevant feature values. Additionally, we deliberately included both stereotypical and non-stereotypical gender roles (e.g., female lawyers, male nurses) to ensure lexical diversity and further mitigate potential gender bias.

Below, we introduce the tested agreement phenomena, along with a list of all test suites created for them; for a more detailed description and additional examples, see Appendix B.

3.1 Italian, Portuguese, and Spanish

In Romance languages, controllers in agreement relations are nominal in nature, they are either nouns or pronouns; targets can be any words that can be inflected, such as finite verbs, participles, adjectives, and determiners; the features involved are gender (masculine and feminine), number (singular and plural) and person (first, second and third). As for the domain, we consider agreement within the noun phrase and agreement within the clause.

We created the test suites for Spanish, Italian and Portuguese based on these four variables, aiming to cover a representative range of their main agreement relations. Within the NP, we test nominal agreement (gender and number) between nouns and articles or possessives (both determiners) and between nouns and adjectives. Within the clause, we test nominal agreement between subject nouns and adjectives or predicate participles, as well as verbal agreement (person and number) between subject nouns or pronouns and the verb. In total, we experiment with ten different test suites for each of the three languages, some of which have already been introduced by Pérez-Mayos et al. (2021) for Spanish. Since these languages share many agreement-related properties, we present the test suites jointly, although separate instances of the test suites are used for each of the three languages, unless stated otherwise. Table 1 lists the different test suites and provides examples.

3.2 Russian

Similarly to Romance languages, in Slavic languages, agreement typically occurs between nominal controllers (nouns and pronouns) and targets that can be inflected (determiners, adjectives, participles and finite verbs), within a noun phrase and within a clause. Agreement features include number (singular and plural), gender (feminine, masculine, and neuter) and person (first, second, and third). The first difference to Romance languages is, however, that Russian has a more pronounced case system. Noun phrases have six different case markings: nominative, accusative, genitive, dative, prepositional, and instrumental. Although it can be argued that case is the morphological manifestation of government and not an agreement feature (Corbett, 2006), we have included it in our test suites for two reasons. First, in nominal agreement, controller and target can take different forms with the same values of their agreement features, depending on their case, so it is impossible to avoid the matter completely. Second, if we assume the grammar to be dependency-based, the noun in a NP may take a specific case due to it being governed by the verb, but other elements in the NP (for instance, determiners or adjectives) take the same case because they are targets in their agreement relation with the noun. Moreover, unlike in Romance languages, the behaviour of Russian verbs regarding agreement depends on tense. In Russian, the verb быть ('[to] be') is omitted in the present tense, leaving only the participle to carry past tense marking. As a result, Russian verbs show person and number agreement in present tense, but number and gender agreement in past tense.

Following the same approach as for the Romance languages, we designed for Russian ten different test suites, which aim to cover a range of fundamental agreement phenomena by manipulating the four key variables involved: controller, target, features, and domain. Within the NP, we test nominal agreement (gender and number) between nouns and articles or possessives (both determiners), and between nouns and adjectives. The test suites are grouped by syntactic structure, with all their test items sharing the same structure. Therefore, case, number and gender cannot be grouped together in one test for each controller–target combination. This means that for each combination, we can create six different, but partially repetitive, test suites. To avoid repetitions, we reduce the number of tests and explore noun–adjective agreement in nominative, accusative, and dative case, and noun-determiner agreement in genitive, prepositional, and instrumental case. Within the clause, we test nominal agreement between subject nouns and adjectives or participles in the predicate. Furthermore, we test verbal agreement (person and number) between subject nouns or pronouns and the verb. See Table 2 for the tests and examples.

Table 1: Test suites for Romance languages: agreement phenomena

Test suite	Languages	#Items**	Grammatical example***	Translation
Article–Noun	es, it, pt	32 × 4	(es) *El*.M.SG *gato*.(M).SG	'The cat'
Possessive–Noun	it, pt	32 × 4	(it) *Il*.M.SG *mio*.M.SG *lavoro*.(M).SG	'My job'
Adjective–Noun	es, it, pt	24 × 4	(es) *La tienda vende discos*.(M).PL *usados*.M.PL	'The store sells second-hand vinyls'
Predicative Attribute*	es, it, pt	32 × 4	(pt) *O apartamento*.(M).SG *está vazio*.M.SG	'The appartment is empty'
Predicative Complement*	es, it, pt	32 × 4	(it) *Le attrici*.(F).PL *ridevano spensierate*.F.PL (es) *El conserje dejó la puerta*.(F).SG *abierta*.F.SG	'The actresses laughed nonchalantly' 'The janitor left the door open'
Unaccusative Participle*	it	24 × 2	(it) *Il bambino*.M.SG *è andato*.M.SG *a scuola*	'The boy went to school'
Passive Participle*	es, pt	24 × 2	(pt) *Os livros*.(M).PL *tem sido publicados*.M.PL	'The books have been published'
Subject–Verb Basic	es, it, pt	24 × 4	(it) *Noi*.1.PL *cuciniamo*.1.PL	'We cook'
Subject–Verb with Subject Relative Clause	es, it, pt	22 × 4	(pt) *O encanador*.SG *[que ajudou os pedreiros] trabalha*.3.SG *de sábado*	'The plumber who helped the bricklayers works on Saturdays'
Subject–Verb with Object Relative Clause	es, it, pt	22 × 4	(es) *Los albañiles*.PL *[a los que ayudó el fontanero] trabajan*.3.PL *los sábados*	'The bricklayers who the plumber helped work on Saturdays'

* These test suites have an adversarial version (with approximately the same number of items, but only 2 examples per item).
** Number of items × number of examples per item.
*** Target and controller are underlined, with their agreement features in small capital letters. Features in brackets are inherent to the word.

Table 2: Russian test suites: agreement phenomena

Test suite	#Items**	Grammatical example***	Lit. translation
Determiner–Noun Genitive	21 × 6	Машина твоего.GEN.M.SG отца.GEN.(M).SG	'Car your father's'
Determiner–Noun Instrumental	21 × 6	Я обедал со своей.INS.F.SG сестрой.INS.(F).SG	'I had-lunch with my sister'
Determiner–Noun Prepositional	21 × 6	На том.PREP.M.SG столе.PREP.(M).SG	'On that table'
Adjective–Noun Nominative	21 × 6	красивая.NOM.F.SG женщина.NOM.(F).SG спит	'Beautiful woman sleeping'
Adjective–Noun Accusative	21 × 6	Президент примет серьезное.ACC.N.SG решение.ACC.(N).SG	'President will-make serious decision'
Adjective–Noun Dative	21 × 6	Работодатель позвонит лучшему.DAT.M.SG кандидату.DAT.(M).SG	'Employer will-call best candidate'
Predicative Attribute*	28 × 3	Квартира.(F).SG кажется пустой.F.SG	'Apartment seems empty'
Predicative Complement*	31 × 3	Учительница.(F).SG ушла сердитая.F.SG	'Teacher left angry'
Subject–Verb Present/Future	24 × 4	Я.1.SG читаю.1.SG книгу	'I am-reading book'
Subject–Verb Past*	27 × 3	Вода.(F).SG повредила.F.SG посевы	'Water damaged crops'

* These test suites have an adversarial version (with approximately the same number of items, but only 2 examples per item).
** Number of items × number of examples per item.
*** Target and controller are underlined with their agreement features in small capital letters. Features in brackets are inherent to the word.

4 Experimental Setup

We evaluated 25 different monolingual and multilingual models (cf. Table 6 in Appendix A), using the metric presented in Section 4.2 below and the *minicons* library (Misra, 2022), which allows for easy surprisal and probability computations. For bidirectional models, we applied the modified scoring technique proposed by Kauf and Ivanova (2023), which masks all tokens to the right of the target word (within the same word) to prevent overestimation in multi-token words. For causal models with tokenizers that mark the beginning of words (and not the continuation), we applied the correction suggested by Pimentel and Meister (2024).

4.1 The Models

For monolingual models, we tested for **Spanish** BETO (Canete et al., 2020), the base version of RoBERTa from the MarIA family of models (Gutiérrez-Fandiño et al., 2022), the open-source GPT2-Spanish model from DeepESP[4], and two lightweight models, alBETO and DistilBETO, introduced by Cañete et al. (2022). For **Italian**, we included an open-source Italian BERT model from the Bavarian State Library[5], and its distilled version BERTino[6] trained by indigo.ai. We also tested

[4] https://huggingface.co/DeepESP/gpt2-spanish
[5] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[6] https://huggingface.co/indigo-ai/BERTino

an Italian RoBERTa model from the Osiria project[7], UmBERTo, another RoBERTa-based model[8], and the GPT-based model GePpeTto (Mattei et al., 2020). For **Portuguese**, we considered BERTimbau (Souza et al., 2020) and its distilled version[9], as well as Tucano-160m, a LLaMA-based model (Corrêa et al., 2024), and GPorTuguese-2, a GPT-based model[10]. For **Russian**, we evaluated RuBERT (Kuratov and Arkhipov, 2019), which is the large version of RuRoBERTa, the small version of RuGPT3 (based on GPT2) from the family of models pre-trained by Zmitrovich et al. (2024), and DistilmBERTru, a model derived from mBERT via language reduction (Abdaoui et al., 2020).

For **multilingual models**, we tested the base versions of mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and the smallest version of XGLM (Lin et al., 2022), representing architectures based on BERT, RoBERTa and GPT3, respectively. Additionally, we included three larger and more recent decoder-only Transformer models: LLaMA-3.2-1B [11], Bloom-7B (Workshop, 2023) and Salamandra-7B (Gonzalez-Agirre et al., 2025).

Information about the size of the selected models can be found in Table 3. Further technical details about the models are provided in Appendix A.

4.2 Evaluation Metric

Traditional targeted syntactic evaluations (see the references in Section 2) assess a model's success in binary terms, i.e., whether it assigns a higher probability (or lower *surprisal*) to the correct word or sentence than to the incorrect one, without considering the magnitude of the difference.[12] This means that a model assigning nearly identical probabilities to both versions, with a slight preference for the correct one, would receive the same score as another model that strongly favors the correct choice. To address this limitation, we use a metric that accounts for the magnitude of the difference between the probabilities assigned by the model to each of the versions:

[7] https://huggingface.co/osiria/roberta-base-italian

[8] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1

[9] https://huggingface.co/adalbertojunior/distilbert-portuguese-cased

[10] https://huggingface.co/pierreguillou/gpt2-small-portuguese

[11] https://huggingface.co/meta-llama/Llama-3.2-1B

[12] Some works on targeted syntactic evaluation score single words while others score the full sentence.

Model	Number Params	Data Size	Model	Number Params	Data Size
Spanish			**Italian**		
BETO	110M	16GB	ItalianBERT	110M	81GB
RoBERTa-BNE	125M	570GB	UmBERTo	110M	70GB
GPT2-Spanish	125M	11.5GB	ItalianRoBERTa[1]	125M	30GB+
DistilBETO	67M	16GB	GePpeTto	117M	13GB
alBETO	12M	16GB	BERTino	68M	12GB
Portuguese			**Russian**		
BERTimbau	109	17.5GB	RuBERT[2]	178M	150GB+
Tucano-160M	160M	589GB	RuRoBERTa	355M	250GB
GPorTuguese-2[1]	124M	1GB+	RuGPT3	125M	450GB
DistilBERTimbau[3]	66M	??	DistilmBERTru[4]	55M	–
Multilingual models					
DistilmBERT	134M	∼75GB	LLaMA3.2	1.23B	∼40TB
mBERT	178M	∼75GB	Bloom	7.07B	1.5TB
XLM-R	270M	2.5TB	Salamandra	7.77B	∼20TB
XGLM	564M	9TB			

[1] ItalianRoBERTa and GPorTuguese-2 were not trained from scratch.
[2] RuBERT was adapted from mBERT by training a new tokenizer and replacing the embedding layer.
[3] No information was found about DistilBERTimbau's training data.
[4] DistilBERTru was created from mBERT via language reduction, there was no ulterior training.

Table 3: Number of parameters and dataset size of selected models.

$$Score(item) = \frac{1}{n} \sum_{x_i \in I} \frac{p(x_t|c)}{p(x_t|c) + p(x_i|c)}$$

where p represents the model's probability distribution, x_t is the target[13] word with matching morphological features, x_i is an incorrect word with all or some mismatching features, I is a set of n possible incorrect words for this item[14], and c represents the context (left context for causal models, and both left and right context for bidirectional ones).

This metric provides an estimation of the model's probability of choosing the correct word over an incorrect one, and then averages this probability across a set of incorrect alternatives. A value over 0.5 means the model assigned (on average) higher probability to the correct word than the incorrect ones, the higher the value, the bigger the difference.

5 Results and Discussion

Table 4 presents the average evaluation scores achieved by the individual models on our agreement test suites. Overall, it can be stated that the models have a reasonable agreement competence, although some significant differences can be ob-

[13] The term *target* is used here to refer to the expected or correct word, not to be confused with the grammatical concept of *target* in an agreement relation, as introduced in Section 1

[14] The number of possible incorrect words is one less than the number of examples per item, which is provided in Tables 1 and 2

Model	Agreement Score	Model	Agreement Score
Spanish		**Italian**	
BETO	0.9127	ItalianBERT	0.9009
RoBERTa-BNE	0.9167	UmBERTo	0.7581
GPT2-Spanish	**0.9223**	ItalianRoBERTa	0.8354
DistilBETO	0.7703	GePpeTto	0.8818
alBETO	0.7930	BERTino	**0.9112**
Portuguese		**Russian**	
BERTimbau	**0.9451**	RuBERT	0.8941
Tucano-160M	0.8967	RuRoBERTa	0.9078
GPorTuguese-2	0.8117	RuGPT3	**0.9159**
DistilBERTimbau	0.5126	DistilmBERTru	0.7220
Multilingual models			
DistilmBERT	0.7257	LLaMA3.2-1B	0.8568
mBERT	0.8036	Bloom-7B	0.8585
XLM-R	0.8402	Salamandra-7B	**0.9331**
XGLM	0.8664		

Table 4: Models' average agreement score.

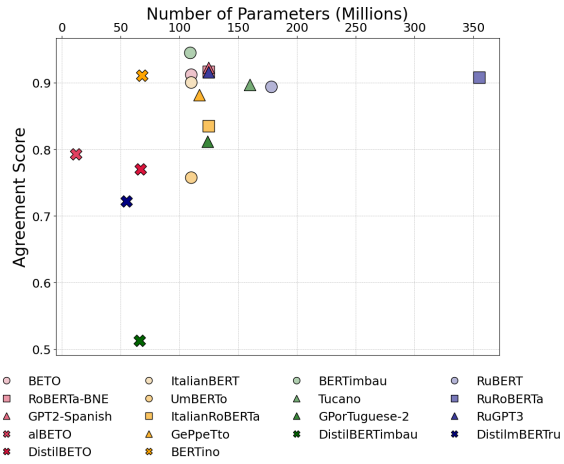

Figure 1: Average agreement score vs. model size (for monolingual models).

served. As expected, monolingual models generally outperform multilingual ones. However, the multilingual Salamandra achieves an average score across all four languages that is comparable to the score of the best monolingual models.

In what follows, we take a closer look at the performance of monolingual and multilingual models, emphasizing the most significant findings.

5.1 Monolingual Models

The **Spanish** models BETO, RoBERTa-BNE, and GPT2-Spanish achieve very similar scores, with GPT2-Spanish performing the best, reaching a 0.92 score. The two lightweight models perform worse, but still achieve reasonable scores over 0.77. For **Italian**, the distilled model BERTino outperforms all the rest at 0.91. GePpeTto and ItalianBERT perform similarly, with scores above 0.88. ItalianRoBERTa and UmBERTo score slightly lower, at 0.84 and 0.76, respectively. **Portuguese** models achieve the highest overall scores. BERTimbau is the strongest performer, reaching a 0.95 score, followed by Tucano-160m (0.90) and GPorTuguese-2 (0.81). The latter is particularly noteworthy, as it was fine-tuned from English GPT2-small using only 1GB of Portuguese data. The distilled version of BERTimbau performs significantly worse, with a score of 0.51. Finally, among **Russian** models, RuGPT3 leads with 0.92, closely followed by RuRoBERTa (0.91) and RuBERT (0.89). Once again, the lightweight model scores slightly lower at 0.72.

The performance of the models allows for some interesting conclusions that challenge common assumptions concerning model size, the size and quality of the training data, and model architecture. Thus, we find that model size (in terms of parameter count) has a weaker effect on agreement performance than one might expect (see Figure 1). For example, in Italian, BERTino, despite being half the size of GePpeTto and ItalianBERT, achieves a higher score. A similar pattern emerges in Spanish, where the lightweight alBETO model – despite being only 10% the size of GPT2-Spanish – still achieves 86% of its performance. In Russian, RuGPT3 outperforms RuRoBERTa despite having just over a third of its parameters; similarly, in Portuguese, BERTimbau surpasses Tucano-160m while using a third fewer parameters. **Training data size** also shows a limited impact on performance. In Spanish, GPT2-Spanish, trained on only 2% the amount of data used for RoBERTa-BNE, still outperforms it. Likewise, BERTimbau, trained on just 3% of the data used for Tucano-160m, achieves a considerably higher score in Portuguese. Italian follows the same pattern: GePpeTto, despite being trained on just 16% of the data used for ItalianBERT, performs at a comparable level. While it is not feasible to systematically assess within the scope of this study the linguistic **quality of the training data** for all models, we hypothesize that exposure to well-written, grammatically correct texts (e.g., Wikipedia, books, and news articles) likely contributes to better performance on linguistically demanding tasks such as ours. Testing this hypothesis directly would require a dedicated analysis beyond the current work.

Model architecture does not reveal a clear per-

formance trend either. Both encoder–decoder architectures like BERT and RoBERTa and decoder-only architectures like the GPT variants can achieve high agreement scores. In Spanish, BETO, RoBERTa-BNE, and GPT2-Spanish perform similarly, as do ItalianBERT, BERTino, and GePpeTto in Italian, and RuBERT, RuRoBERTa, and RuGPT3 in Russian. However, in Portuguese, the highest performance is achieved by BERTimbau, a BERT-based model. Neither the **tokenizer strategy** nor the **vocabulary size** appears to have a strong impact on agreement scores: the best models for Spanish and Russian use a BPE tokenizer with a 50K vocabulary, while those for Italian and Portuguese perform best with WordPiece and a 30K vocabulary. The top-performing multilingual model, meanwhile, uses SentencePiece.

5.2 Multilingual Models

Table 5 presents additional information on the performance of the multilingual models, including average scores for each language and the proportion of training data allocated to each language in each of the models. Overall, multilingual models perform well, but somewhat weaker than monolingual models, with the exception of Salamandra and Bloom for Spanish. Salamandra leads the rankings at an average score of 0.93, followed by XGLM at 0.87, and both Bloom and LLaMA-3.2 at 0.86. Looking at the results by language, Salamandra achieves the highest scores for Italian (0.92), Portuguese (0.94), and Russian (0.92), and Bloom stands out as the best-performing multilingual model for Spanish (0.96). Interestingly, Salamandra outperforms the best monolingual models for Spanish, Italian and Russian while remaining highly competitive for Portuguese. Similarly, Bloom surpasses the top Spanish monolingual model and nearly matches the best Portuguese one. Large multilingual models clearly benefit from **transfer learning** across languages, as evidenced by Bloom's results. Despite not being trained on any Italian or Russian data, it still performs reasonably well on these languages, most likely due to typological proximity.

Unlike for monolingual models, for multilingual models, **model size** appears to have a stronger impact on agreement competence, with larger models generally achieving better results, although it should be noted that we are now comparing much bigger sizes (see Figure 2). However, there are exceptions: XGLM, despite being half the size

	Our Score			
Model	**Spanish**	**Italian**	**Portuguese**	**Russian**
DistilmBERT	∼4.7% 0.691	∼4.7% 0.741	∼2.3% 0.744	∼4.7% 0.727
mBERT	∼.7% 0.795	∼4.7% 0.803	∼2.3% 0.806	∼4.7% 0.811
XLM-R	2.1% 0.834	1.2% 0.832	2.0% 0.839	11.1% 0.856
XGLM	4.3% 0.893	2.0% 0.809	1.8% 0.878	12.0% 0.885
LLaMA	?? 0.894	?? 0.827	?? 0.844	?? 0.862
Bloom	11.1% **0.960**	0% 0.751	5.2% 0.938	0% 0.785
Salamandra	16.1% 0.953	2.1% **0.919**	2.2% **0.944**	5.6% **0.917**

Table 5: Multilingual models' score by language

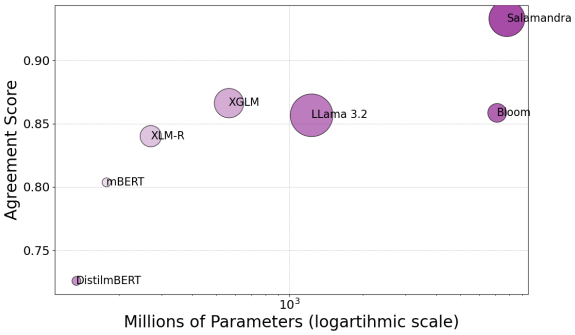

Figure 2: Average agreement score vs. model size (for multilingual models). Bubble size reflects training data size.

of LLaMA-3.2, outperforms it; and XLM-R with a fourth of LLaMA-3.2's size comes quite close. Something similar happens with **training data size**: models trained on more data tend to perform better. Yet again, there is an exception: although LLaMA-3.2 was trained on twice the amount of data as Salamandra (the second-largest model in terms of training data), it is still outperformed by it.

As far as **model architecture** is concerned, no definitive conclusions can be drawn since all larger models in the study are decoder-only Transformers, whereas the smaller ones follow an encoder–decoder architecture.

5.3 Results by Test Suite

Figure 3 presents the agreement scores across all individual test suites for all models and languages. As anticipated from the averages, the overall performance remains reasonably high. However, a few noteworthy observations stand out.

Although the **Article—Noun** test suite is relatively simple and should thus not be particularly

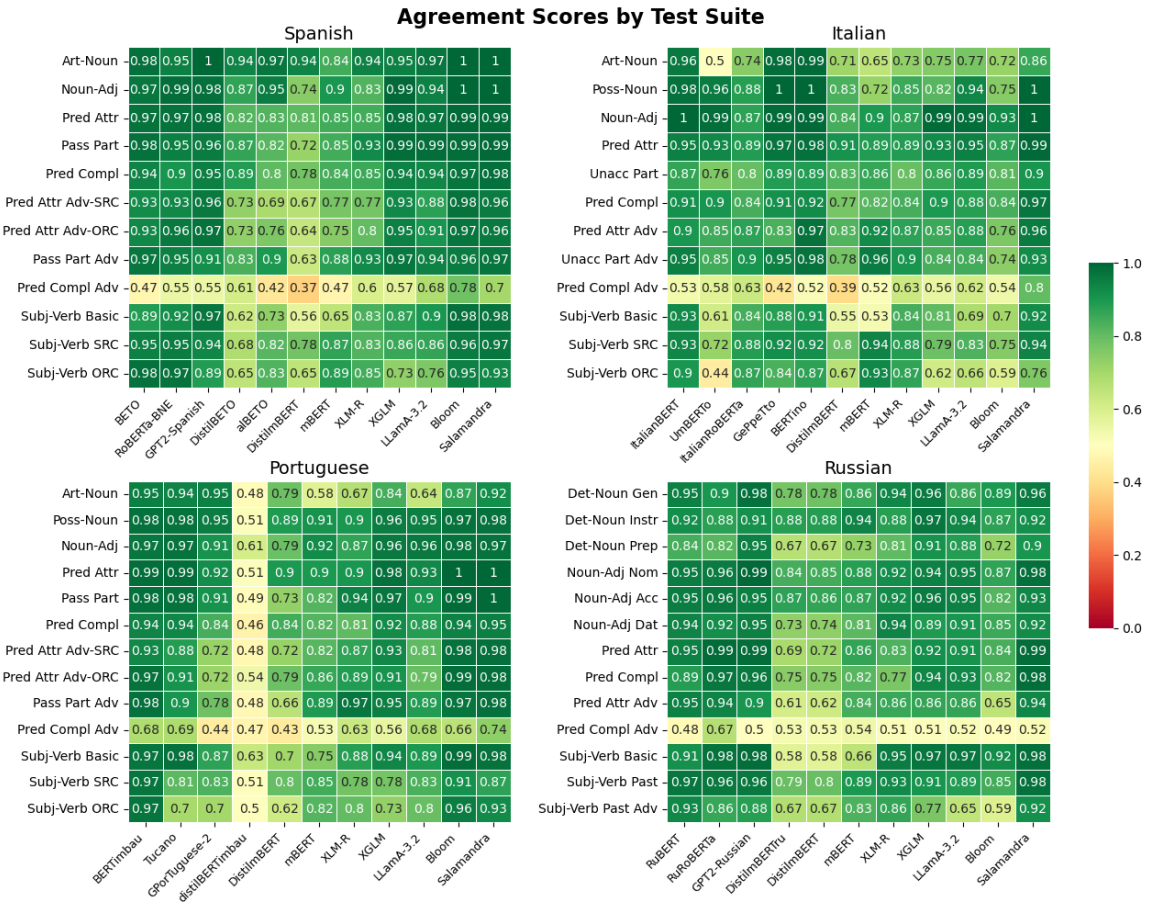

Figure 3: Agreement scores by test suite for all models and lnaguages.

demanding, multilingual models seem to struggle with it for Italian and Portuguese. This may be due to the brevity of articles in these languages, often made up of just one or two letters (e.g., *o* and *a* as singular definite articles in Portuguese for masculine and feminine nouns). The models might misinterpret them as the start of a word (perhaps in another language), leading to errors. Similarly, while the **Subject-–Verb Basic** test suite is also expected to be straightforward, it proved problematic for some models. Notably, mBERT exhibited a consistent difficulty across all four languages.

Surprisingly, most **adversarial versions** of the test suites were not radically more demanding than their standard counterparts. The only exception was the **Predicative Complement** test suite, which was already somewhat more complex than the other nominal agreement tests within the clause. In its adversarial form, it posed a substantial challenge, even for the strongest monolingual and multilingual models.

Both Subject—Verb with Relative Clause test suites should be considered inherently adversarial, despite lacking a standard equivalent. Adapted from English (Marvin and Linzen, 2018), they focus exclusively on number agreement with third-person subjects. These constructions proved difficult for many models, particularly multilingual ones, though even some monolingual models struggled, as can be observed in the results for Portuguese.

Looking at the lighter columns in Figure 3, which generally correspond to distilled models, we observe that **distillation** tends to negatively impact models' agreement performance, with the Italian model BERTino being a notable exception. In particular, the Spanish lightweight model alBETO, despite having only a fifth of DistilBETo's parameters, performs more robustly.

Although one might expect models to perform worse in **Russian** due to its more complex case system, the opposite could also be argued, as case markings provide additional grammatical cues. Moreover, relative clauses, used in adversarial test suites to increase the linear distance between the controller and the target, are typically enclosed by

commas in Russian, offering further structural hints that may increase the performance of the model.

Finally, **Salamandra** once again demonstrates exceptional robustness, facing real difficulty only in the Predicative Complement adversarial test suite for all languages, especially Russian, where it scored 0.52. It also encountered mild challenges in the Subject–Verb with Object Relative Clause for Italian: the only other instance where its score fell below 0.8.

5.4 Adversarial Test Suites

As our results show, the adversarial test suites were generally more difficult than their standard counterparts, though not as consistently as expected; some models even performed better on the adversarial versions.

Adversarial suites were carefully constructed to include semantically plausible agreement attractors, with incorrect targets deliberately chosen to agree with them rather than with the correct controller. Despite this, several models reliably assigned higher probabilities to the grammatically correct target, prompting for further investigation.

In the context of adversarial suites, we also developed alternative versions of the Subject–Verb with Relative Clauses test suites in Spanish. In the **subject** relative clause variant, we removed the preposition preceding the attractor[15], hypothesizing that its absence (potentially serving as a syntactic cue) would increase the difficulty. In contrast, for the **object** relative clause version, we repositioned the subject of the relative clause before the subordinate verb[16], expecting this to simplify the task by distancing the attractor from the main verb. Interestingly, the results contradicted our expectations: most models performed better on the revised subject relative clause suite and worse on the modified object relative clause suite. The improvement in the former may be due to the attractor's reduced semantic plausibility (being inanimate), while the decline in the latter might stem from the unusual surface structure, specifically, the sequence of two adjacent verbs confusing the model's internal representation of syntactic relations.

The key takeaway from these findings is that language models do not rely heavily on surface adjacency between controller and target. Instead, they seem to leverage other cues, including semantic compatibility. At the same time, the models' tendency to fail in the presence of non-canonical verb sequences suggests that their grasp of syntactic structure remains limited. Rather than encoding abstract syntactic relations robustly, they may depend more on frequent patterns and shallow heuristics.

6 Conclusions

Our experiments with 25 different Large Language Models, applied to a number of different test suites for Italian, Portuguese and Spanish on the one side and Russian on the other side, have shown that, in general, the models show a reasonable competence in agreement across languages, although monolingual models tend to perform better than multilingual ones. No significant difference has been noted between the Romance languages and Russian. However, all models still struggle to a certain extent with more complex syntactic constructions with attractors, for instance, in relative clauses. Interestingly, and contrary to common assumptions, model size, architecture, and training data volume had only a limited impact on agreement performance.

The test suites can be freely accessed and downloaded for research purposes[17]. Our approach to the creation of test suites can be extended to other languages, provided it remains focused on agreement phenomena. Creating test suites for a new language requires identifying potential *targets* and *controllers*, determining the relevant agreement *features*, and defining the *domain* where agreement occurs within the language's grammatical structure. The initial goal should be to cover core agreement phenomena by combining these four factors, which can then be expanded by incorporating more challenging or marginal cases, as well as semantic agreement and agreement resolution. This is precisely the direction we aim to take in the next phase of our work, broadening the range of languages covered and increasing the level of difficulty in our tests.

[15] For example, *La bióloga que colabora **con** los veterinarios tiene mucha experiencia* ('The biologist(f) who collaborates with the veterinarians has a lot of experience') became *La bióloga que comprueba los equipos tiene mucha experiencia* ('The biologist who tests the equipment has a lot of experience').

[16] We replaced *Los albañiles a los que ayudó el fontanero* with *Los albañiles a los que el fontanero ayudó* ('The bricklayers who the plumber helped').

[17] https://huggingface.co/datasets/albalbalba/SyntacticAgreement

Limitations

Despite a considerably broader coverage of agreement phenomena and different state-of-the-art neural language models than in previous work, our study has some obvious limitations. First, our study covers only four languages, three of which are closely related Romance languages. A broader typological coverage, particularly including languages from non-Indo-European families, languages with a rich inflectional morphology, or those with unique agreement systems, would provide a more comprehensive picture. Unfortunately, the development of high-quality test suites requires both linguistic expertise in the target languages and access to language models trained on them, both of which are often lacking for under-resourced or less commonly studied languages.

Second, the test suites used in this study primarily target core and relatively canonical agreement phenomena. While this allows for consistent and controlled evaluation, it may also underestimate the challenges that arise in more marginal or exceptional agreement cases–e.g., agreement across clause boundaries, or cases involving semantic factors. Future work should aim to incorporate such phenomena, both to test deeper syntactic and semantic understanding and to better represent the complexity found in natural language.

Third, while our test suites were carefully constructed by expert linguists to ensure that grammatical and ungrammatical examples are well-formed and contrastive, we did not complement them with human acceptability judgments. Given the controlled design and linguistic motivation behind each example, we expect them to be generally reliable. However, incorporating human judgments in future work could provide additional insights into whether model predictions align with speaker intuitions, and help clarify to what extent observed model behavior reflects genuine linguistic competence.

Fourth, we observed some unexpected patterns, most notably, the strong performance of the lightweight Italian model BERTino and the generally lower scores of Italian models overall. We consider it unlikely that these differences arise from language-specific properties, given the close typological similarity among the Romance languages examined. Instead, the differences are more plausibly due to the variation in model pretraining or training data, which are not well documented for the Italian models. This highlights the need for more transparency in model development and warrants further investigation beyond the scope of the present study.

Finally, the range of language models we evaluated was constrained by the available computational resources. Many of the most recent state-of-the-art models are prohibitively large for independent researchers to use, even in a zero-shot evaluation setting. This limits the ability to fully explore the impact of scale and architecture. As model sizes continue to increase, the need for more equitable access to these technologies will become increasingly pressing, not just for training, but also for systematic evaluation.

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of mutililingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.

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

José Cañete, Sebastian Donoso, Felipe Bravo-Marquez, Andrés Carvallo, and Vladimir Araujo. 2022. ALBETO and DistilBETO: Lightweight Spanish language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4291–4298, Marseille, France. European Language Resources Association.

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.

Greville G. Corbett. 2006. *Agreement*. Cambridge University Press, Cambridge, UK.

Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2024. Tucano: Advancing neural text generation for portuguese. *Preprint*, arXiv:2411.07854.

Iria de Dios-Flores and Marcos Garcia. 2022. A computational psycholinguistic evaluation of the syntactic abilities of galician bert models at the interface of dependency resolution and training time. *Procesamiento del lenguaje natural*, 69:15–26.

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.

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.

Marcos Garcia and Alfredo Crespo-Otero. 2022. A targeted assessment of the syntactic abilities of transformer models for galician-portuguese. In *Computational Processing of the Portuguese Language*, pages 46–56, Cham. Springer International Publishing.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Oriol Pareras, Valle Ruiz-Fernández, and Marta Villegas. 2025. Salamandra technical report. *Preprint*, arXiv:2502.08489.

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.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodríguez Penagos, Aitor Gonzalez-Agirre, and Marta Villegas Montserrat. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68:39–60.

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

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. 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.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

Daria Kryvosheieva and Roger Levy. 2025. Controlled evaluation of syntactic knowledge in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 402–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *Preprint*, arXiv:1905.07213.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

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.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769. CEUR-WS.org. Italian Conference on Computational Linguistics 2020, CLiC-it 2020 ; Conference date: 01-03-2021 Through 03-03-2021.

Igor Mel'čuk. 1993. Agreement, governance and congruence. *Linguisticae investigationes*, 17(2):307–373.

Igor Mel'čuk. 2009. Dependency in natural language. In *Dependency in Linguistic Description*, pages 1–110. John Benjamins Publishing Company.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

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.

Vivi Nastase, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2024. Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 631–643, Pisa, Italy. CEUR Workshop Proceedings.

Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online. Association for Computational Linguistics.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium. Association for Computational Linguistics.

Edward W. Samson. 1953. Fundamental natural concepts of information theory. *ETC: A Review of General Semantics*, 10(4):283–297.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.

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. 2019. 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.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Karolina Zaczynska, Nils Feldhus, Robert Schwarzenberg, Aleksandra Gabryszak, and Sebastian Möller. 2020. Evaluating german transformer language models with syntactic agreement tests. *CoRR*, abs/2007.03765.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italia. ELRA and ICCL.

A Models' Size and Training Information

Table 6 provides detailed information about the language models selected for evaluation: model size, architecture, training dataset(s) and its size, tokenizer strategy and vocabulary size, as well as the average score obtained on our test suites.

B Description of test suites

This appendix is a description of all test suites created to assess the ability of LLMs to identify sentence samples that violate the rules of agreement in three Romance languages (Italian, Portuguese and Spanish) and one Slavic language (Russian). Each test suite contains sentence samples of a specific agreement rule and systematic variations of these samples violating that given rule. All of them are based on the premise that an element in an agreement relation whose features match those of the other element in the relation should yield higher

Table 6: Model information and average agreement score.

Model	Lgg	Architecture	Number Params	Training Data	Dataset Size	Tokenizer	Vocabulary Size	Our Score[1]
BETO	es	BERT	110M	Wikipedia, OPUS	16GB	WordPiece	31K	0.9127
RoBERTa-BNE	es	RoBERTa	125M	BNE BookCrawl	570GB	BPE	50K	0.9167
GPT2-Spanish	es	GPT2	125M	Wikipedia, Books	11.5GB	BPE	50K	**0.9223**
DistilBETO	es	DistilBERT	67M	Wikipedia, OPUS	16GB	SentencePiece	31K	0.7703
alBETO	es	alBERT	12M	Wikipedia, OPUS	16GB	SentencePiece	31K	0.7930
ItalianBERT	it	BERT	110M	Wikipedia OPUS, OSCAR	81GB	WordPiece	31K	0.9009
UmBERTo	it	RoBERTa	110M	OSCAR	70GB	SentencePiece	32K	0.7581
ItalianRoBERTa[2]	it	RoBERTa	125M	CommonCrawl	30GB+	BPE	50K	0.8354
GePpeTto	it	GPT2	117M	Wikipedia, ItWaC	13GB	BPE	30K	0.8818
BERTino	it	DistilBERT	68M	Paisa, ItWaC	12 GB	WordPiece	31K	**0.9112**
BERTimbau	pt	BERT	109M	brWaC	17.5GB	WordPiece	30K	**0.9451**
Tucano-160M	pt	LLaMA	160M	GigaVerbo	589GB	BPE	32K	0.8967
GPorTuguese-2[3]	pt	GPT2	124M	Wikipedia	1GB+	BPE	50K	0.8117
DistilBERTimbau	pt	DistilBERT	66M	??	??	WordPiece	30K	0.5126
RuBERT[4]	ru	BERT	178M	Wikipedia, News	150GB+	BPE	120K	0.8941
RuRoBERTa	ru	RoBERTa	355M	Wikipedia News(Corus), Books	250GB	BPE	50K	0.9078
RuGPT3	ru	GPT2	125M	Wikipedia(ru+en) News, Books, C4	450GB	BPE	50K	**0.9159**
DistilmBERTru[5]	ru	BERT	55M	-	-	WordPiece	14K	0.7220
DistilmBERT	multi	DistilBERT	134M	Wikipedia	∼50-100GB[6]	WordPiece	120K	0.7257
mBERT	multi	BERT	178M	Wikipedia	∼50-100GB[6]	WordPiece	120K	0.8036
XLM-R	multi	RoBERTa	270M	CommonCrawl	2.5TB	Unigram	150K	0.8402
XGLM	multi	GPT3	564M	CommonCrawl Books, Wikipedia	9TB	Unigram	250K	0.8664
LLaMA-3.2	multi	LLaMA	1.23B	??	∼40TB (9T tokens)[7]	Unigram	128K	0.8568
Bloom	multi	Bloom	7.07B	ROOTS Corpus	1.5TB	BPE	250K	0.8585
Salamandra	multi	Salmandra	7.77B	Colossal OSCAR FineWeb-Edu	∼20TB (13T tokens)[7]	SentencePiece	256K	**0.9331**

[1] The score for each monolingual model is the average calculated over all the test suites for each specific language. The score for multilingual models is the average score for the four languages evaluated. The best results for each language appears in bold.
[2] ItalianRoBERTa was not trained from scratch, but initialized with XLM-R weights.
[3] GPorTuguese-2 was not trained from scratch, but fine-tuned from the English GPT-2-small (Radford et al., 2019).
[4] RuBERT was adapted from mBERT by training a new tokenizer and replacing the embedding layer.
[5] DistilmBERTru is not really a distilled model, rather, it has been obtained by language reduction from mBERT as proposed by Abdaoui et al. (2020); there was no ulterior training performed.
[6] Dataset size was estimated considering the size of Wikipedia dumps at the time the model was released.
[7] Information about the dataset size was in number of tokens; size in TB was estimated by assuming an average token size of 5 bytes.
'?': No information was found about training data and/or its size.

probability values than one with mismatching features. In total, the test suites contain over 5,000 test samples.

In all test suites, the element that is systematically modified to violate the rule is the *target* of the agreement relation. The probabilities used to calculate the metric described in Section 4.2 are the ones assigned by the model to the correct and incorrect versions of this element. However, to be able to apply all our test suites to both bidirectional and causal LLMs, there are a couple of exceptions in which the element that varies is the target, but the probability is measured for the *controller*. This happens with the determiner–noun agreement relation, in which the target is to the left of the controller. Note that although the controller is the same, the probabilities are not, because the target in its left context has been modified.

B.1 Italian, Portuguese, and Spanish

Since Italian, Portuguese and Spanish share many agreement-related properties, we present most of the test suites for them jointly, with an example in one of the three languages. Note, however, that in these cases, each test also has a version in all three of them, unless explicitly stated otherwise.

In total, we experiment with ten different test suites for these languages. We explore nominal agreement–gender and number–within the noun phrase (three suites) and within the clause (four suites), and verbal agreement–person and number–within the clause (three suites).

• **Article–Noun Agreement**

To test article–noun agreement, the four possible forms of the definite article ('sg.' vs. 'pl.' × 'fem.' vs. 'masc.') are paired with different nouns to capture all four forms. Cf. the following example in Spanish, with the masculine noun *gato* 'cat':

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

• **Possessive–Noun Agreement**

Unlike in Spanish, in Italian and Portuguese, the possessive determiner has the four possible forms 'sg.' vs. 'pl.' × 'fem.' vs. 'masc.', as the definite article. This test pairs them with nouns that reflect these forms; cf. an example in Italian:

(4) Il mio lavoro
the.M.SG my.M.SG job
'My job'

(5) *[La mia] / *[I miei] / *[the.F.SG my.F.SG] / *[the.M.PL my.M.PL] / *[Le mie] lavoro
*[the.F.PL my.F.PL] job
'My job'

• **Adjective–Noun Agreement** This test pairs a noun with the 'sg.' vs. 'pl.' × 'fem.' vs. 'masc.' forms of an adjective that modifies it. To avoid providing extra information to the model, the test uses constructions without a determiner; cf., an example in Spanish:

(6) 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'

• **Predicative Attribute Agreement**

In Romance languages, the predicative attribute in copulative constructions must agree with the grammatical subject in gender and number. Consider an example in Portuguese:

(7) O apartamento está
the.M.SG apartment is
vazio/*vazios/*vazia/*vazias
empty.M.SG/*M.PL/*F.SG/*F.PL
'The apartment is empty'

For Spanish and Portuguese, this suite has two adversarial versions one with object and one with subject relative clauses. For Italian, there is only one adversarial version in which the intervening material can be a relative clause or a prepositional phrase. Since the role of the intervening material is to increase the linear distance between the covarying words and to include an agreement attractor, we decided to condense the two versions into one, as long as it maintains these two important factors. Here is an example in Italian:

(8) L' appartamento che guarda
the.M.SG apartment that look.3.SG
verso la spiaggia è
towards the.F.SG beach is
vuoto/*vuoti/*vuota/*vuote
empty.M.SG/*M.PL/*F.SG/*F.PL
'The apartment facing the beach is empty'

• **Predicative Complement Agreement**

In Italian, Portuguese, and Spanish, an attribute that functions as a predicative complement to the grammatical subject or the object must agree with it in gender and number; cf. an example in Spanish for a complement to the object:

(9) El tenista dejó la raqueta
the tennis-player left the.F.SG racket
destrozada/*destrozadas/*destrozado/*destrozados.
destroyed.F.SG/*F.PL/*M.SG/*M.PL
'The tennis player left his racket destroyed.'

This suite has an adversarial version with intervening material in the form of a relative clause or a prepositional phrase; cf. this example for a complement to the subject:

(10) Las voluntarias que ayudaron a
the.F.PL volunteer.F.PL who helped to
los refugiados cayeron
the.M.PL refugee.M.PL fell
enfermas/*enferma/*enfermos/*enfermo.
ill.F.PL/*F.SG/*M.PL/*M.SG
'The volunteers who helped the refugees fell ill.'

• **Participle Agreement**

In contrast to Portuguese and Spanish, in Italian, unaccusative verbs in past tense are conjugated with the auxiliary verb *essere* ('[to] be') and their past participle form. The past participle must agree in gender and number with the grammatical subject; cf.:

(11) Il bambino è andato/*andata a scuola.
the child is gone.M.SG/*F.SG to school
'The child has gone to school.'

This suite has an adversarial version as well, with relative clauses or prepositional phrases as intervening material.

(12) Il bambino che ha litigato con sua
the child who has fought with his.F.SG
sorella è andato/*andata a scuola.
sister is gone.M.SG/*F.SG to school
'The child who had a fight with his sister has gone to school.'

• **Passive Participle Agreement**

In passive constructions of Portuguese and Spanish, the past participle must agree in gender and number with the grammatical subject; cf. a Portuguese example:

(13) Os encontros serão
the.M.PL matches will.be
transmitidos/*transmitidas ao vivo
broadcast.M.SG/*F.SG to.the live
'The matches will be broadcast live.

This suite also has an adversarial version for both languages.

(14) Os encontros de qualificação para
the.M.PL matches of qualification for
as semifinais serão
the.F.SG semifinal will.be
transmitidos/*transmitidas ao vivo
broadcasted.M.SG/*F.SG to.the live
'The qualifying matches for the semifinals will be broadcast live.

• **Basic Subject–Verb Agreement**

In Italian, Portuguese and Spanish, the finite verb tense and mood forms must agree in person and number with the grammatical subject, as in the Italian example below:

(15) Tu cucini
you.2SG cook.2SG

(16) * Tu cucinate/cucino/cucinano
you.2SG cook.2PL/1SG/3PL
'You cook'

• **Subject–Verb Agreement with Subject Relative Clause**

This test suite, which has been adapted from the English test introduced by Marvin and Linzen (2018), focuses on number agreement. The subject relative clause includes an *agreement attractor* differing in number with the subject. The model is expected to assign higher probability to the verb agreeing with the subject (instead of the attractor), in both singular and in plural; cf. an example from Portuguese:

(17) O encanador que ajudou os
the.SG plumber that helped.3SG thePL
pedreiros trabalha/*trabalham de sábado.
bricklayers work.3SG/3PL of saturday.
'The plumber who helped the bricklayers works/*work on saturdays.'

(18) Os encanadores que ajudaram o
the.PL plumbers that helped.3SG thePL
pedreiro *trabalha/trabalham de sábado.
bricklayer work.3PL/3SG of saturday.
'The plumbers who helped the bricklayer *works/work on saturdays.'

• **Subject–Verb Agreement with Object Relative Clause**

As the previous test suite, this test is also on number agreement, only that it contains an object instead of a subject relative clause. Furthermore, in view of the stricter subject–verb order in Brazilian Portuguese, we introduce two versions of this test, one for Brazilian Portuguese and one for Italian and Spanish. The one for Portuguese follows the same pattern as the English version:

(19) Os pedreiros que o encanador
the.PL bricklayers that theSG plumber
ajudou *trabalha/trabalham de sábado.
helped.3SG work.3SG/3PL of saturday.
'The bricklayers who the plumber helped *works/work on saturdays.'

(20) O pedreiro que os encanadores
the.SG bricklayer that the.PL plumbers
ajudaram trabalha/*trabalham de sábado.
helped.3PL work.3SG/3PL of saturday.
'The bricklayer who the plumbers helped works/*work on saturdays.'

In the one for Italian and Spanish, the agreement attractor within the relative clause is adjacent to the critical region where the main verb is located. In this case, the agreement attractor is the subject of the relative clause, and thanks to these languages' flexibility, it can appear post-posed to the subordinate verb and hence adjacent to the main one, as shown in the Spanish example below.

(21) Los albañiles a los que aydudó
the.PL bricklayers to the.PL that helped.3SG
el fontanero *trabaja/trabajan los
theSG plumber work.3SG/3PL the
sábados.
saturdays.
'The bricklayers who the plumber helped *works/work on saturdays.'

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

B.2 Russian

For Russian, we experiment with ten different test suites. We explore nominal agreement–gender and number–within the noun phrase (six suites) and within the clause (two suites), and verbal agreement–person and number–within the clause (two suites).

• **Determiner–Noun Agreement in Genitive**

In this test, a noun in a genitive construction is paired with a possessive or demonstrative determiner that modifies it:

(23) Машина твоего отца
car your.GEN.M.SG father.GEN
'Your father's car'

(24) * Машина
car
твоих/твоей/твой
your.GEN.PL/GEN.F.SG/NOM.M.SG
отца
father.GEN

(25) * Машина твоя/твоими
car your.GEN.PL/NOM.F.SG/INS.PL
отца
father.GEN

• Determiner-Noun Agreement in Instrumental

Analogous to the previous test suite. Here, a noun preceded by a specific preposition or verb that requires it to appear in instrumental case is paired with a determiner (a possessive or a demonstrative) that modifies it.

(26) Я обедал со своей
I had.lunch with my.INS.F.SG
сестрой
sister.INS.(F).SG
'I had lunch with my sister.'

(27) * Я обедал со
I had.lunch with
своими/своим/своя
my.INS.PL/INS.M.PL/NOM.F.SG
сестрой
sister.INS.(F).SG

(28) * Я обедал со своём/своих
I had.lunch with my.PREP.M.SG/ACC.PL
сестрой
sister.INS.(F).SG

• Determiner-Noun Agreement in Prepositional

The same as before, but with a noun preceded by a preposition that requires prepositional case and a determiner (a possessive or a demonstrative) that modifies it.

(29) На том столе
on that.PREP.M.SG table.PREP.(M).SG
'On that table'

(30) * На тех/той/тому
on that.PREP.PL/PREP.F.SG/DAT.M.SG
столе
table.PREP.(M).SG

(31) * На то/те
on that.ACC.N.SG/NOM.PL
столе
table.PREP.(M).SG

• Adjective–Noun Agreement in Nominative

This test suite pairs a noun in the subject position (hence in nominative case) with an adjective that modifies it:

(32) красивая женщина спит.
beautiful.NOM.F.SG woman.NOM sleeps.
'A beautiful woman is sleeping.'

(33) * красивые/красивый/красивую
beautiful.NOM.PL/NOM.M.SG/ACC.F.SG/
женщина спит.
woman sleeps.

(34) * красивого/красивым женщина
beautiful.GEN.M.SG/DAT.PL woman
спит.
sleeps.

• Adjective–Noun Agreement in Accusative

This suite is analogous to the previous one, but with the noun in the object position (hence in accusative case):

(35) Медсестра держала маленькое
nurse held small.ACC.N.SG
чадо.
child.ACC.(N).
'The nurse was holding a small child.'

(36) * Медсестра держала
nurse held
маленьких/маленькую/маленькому
small.ACC.PL/ACC.F.SG/DAT.N.SG
чадо.
child.ACC.(N).

(37) * Медсестра держала
nurse held
маленькой/маленькими чадо.
small.GEN.F.SG/INS.PL child.ACC.

• Adjective–Noun Agreement in Dative

This test suite is analogous to the previous two, but here the noun occupies a position (e.g., indirect object) in the sentence that requires dative case.

(38) Старик радуется солнечному
old-man enjoys sunny.DAT.N.SG
утру.
morning.DAT.(N)
'The old man enjoys the sunny morning.'

(39) * Старик радуется
old-man enjoys
солнечным/солнечной/солнечном
sunny.DAT.PL/DAT.F.SG/PREP.N.SG
утру.
morning.DAT.(N)

(40) * Старик радуется
old-man enjoys
солнечная/солнечными
sunny.NOM.F.SG/INS.PL
утру.
morning.DAT.(N)

• Predicative Attribute Agreement

This test suite is similar to the corresponding test suite for Spanish, Italian and Portuguese. A noun is paired with an adjective through a copulative

construction. The main difference comes from the fact that in Russian the gender feature is neutralized in plural. This means that to be able to capture mismatches in gender, only singular subjects are to be used:

(41) Квартира кажется
apartment(F).SG seems
старой/*старыми/*старым.
empty.F.SG/*PL/*M.SG
'The apartment seems empty'.

Note that gender and number cannot disagree at once (as it happened with Spanish, Italian and Portuguese), since gender is not apparent in plural.

This suite has an adversarial version, with a relative clause (sometimes a reduced one) serving as modifier for the grammatical subject. The modifier includes an agreement attractor differing in gender or number with the subject:

(42) Квартира, которая была
apartment.(F).SG which was
обставлена моим братом, кажется
furnished my brother.(M).SG seems
пустой/*пустым/*пустыми.
empty.F.SG/*M.SG/*PL
'The apartment that my brother furnished seems empty.'

• Predicative Complement Agreement

This test suite is also similar to the corresponding test suite for Italian, Portuguese, and Spanish. The subject is paired with an adjective functioning as a predicative complement. Again, the main difference is that to be able to capture mismatches in gender, only subjects in singular are used.

(43) Ребенок приехал
kid arrived
счастливый/*счастливая/*счастливые.
happy.M.SG/*F.SG/*PL
'The kid arrived happy.'

As the previous one, this suite also has an adversarial version.

(44) Ребенок, которого похвалила
kid who was.praised
воспитательница, приехал
teacher arrived
счастливый/*счастливая/*счастливые.
happy.M.SG/*F.SG/*PL
'The kid who was praised by the teacher arrived happy.'

• Basic Subject–Verb Agreement in Present/Future Tense

Finite verbs in present/future tense and indicative mood have six inflected forms according to person and number features. The verb's features must agree with the subject's:

(45) Я читаю книгу.
I.1SG read.1SG book
'I am reading a book.'

(46) * Я читаем/читаешь/читают
I.1SG read.1PL/2SG/3PL
книгу.
book

• Subject–Verb Agreement in Past Tense

In contrast, finite verbs in past tense and indicative mood have four inflected forms according to gender and number features (person is not involved). This applies to any person, but personal pronouns without context do not provide gender information, so the test only includes subjects in the third person singular (recall that gender feature is not apparent in plural).

(47) Учитель прочитал поэму в классе.
teacher.(M).SG read.M.SG poem in class
'The teacher read a poem in class.'

(48) * Учитель прочитала/прочитали
teacher.(M).SG read.F.SG/PL
поэму в классе.
poem in class

There is also an adversarial version of this test suite, as shown below:

(49) Учитель, которого ненавидели
teacher.(M).SG who was.hated
девочки, прочитал поэму в классе.
girl.PL read.M.SG poem in class
'The teacher who the girls hated read a poem in class.'

(50) * Учитель, которого ненавидели
teacher.(M).SG who was.hated
девочки, прочитала/прочитали
girl.(PL read.F.SG/PL
поэму в классе.
book in class