Proposta recibida en setembro 2025 e aceptada para publicación en novembro 2025.

O procesamento da estrutura sintáctica nos modelos de linguaxe neurais: a concordancia suxeito-verbo en galego e portugués

Processing of Syntactic Structure in Neural Language Models: Subject-Verb Agreement in Galician and Portuguese

Helena Pérez Puente
Facultade de Filoloxía
Universidade de Santiago de Compostela

Resumo

Combinando coñecementos das áreas de lingüística computacional e de psicolingüística, neste traballo explórase o coñecemento sintáctico dos modelos de linguaxe neurais, concretamente o principio da concordancia suxeito-verbo en galego e portugués. Co fin de estudar se os modelos procesan este fenómeno de forma semellante aos humanos, elaboráronse dous datasets (un en galego e outro en portugués) con 16 oracións e 8 variantes para cada unha delas, que teñen en conta a gramaticalidade, a presenza ou ausencia dun distractor e a distancia entre o suxeito e o verbo principal. Con estes elementos, realizáronse avaliacións de aceptabilidade a falantes destas linguas para comprobar a aceptabilidade das oracións en relación ás variables referidas. Cunha versión adaptada dos datasets, avaliáronse modelos de lingua neurais para galego e portugués. Os resultados parecen indicar que, aínda que tanto humanos coma modelos distinguen entre oracións gramaticais e agramaticais, os falantes amosan maior precisión e incluso se benefician dos distractores, mentres que estes confunden os LMs, especialmente en contextos longos.

Palabras chave

modelos de linguaxe, sintaxe, concordancia suxeito-verbo, galego, portugués

Abstract

Combining knowledge from the areas of computational linguistics and psycholinguistics, this paper explores the syntactic knowledge of neural language models, specifically the principle of subject-verb agreement in Galician and Portuguese. In order to study whether the models process this phenomenon in a similar way to humans, two datasets were created (one in Galician and one in Portuguese) with 16 sentences and 8 variants for each of them, taking into account grammaticality, the presence or absence of a distractor and the distance between the subject and the main verb. With these elements, surveys were carried out with speakers of these languages to check the acceptability of the sentences in relation to the variables referred to. Using an adapted version of the datasets, neural language models for Galician and Portuguese were evaluated. The results show that although both humans and models distinguish between grammatical and ungrammatical sentences, speakers exhibit greater accuracy and even benefit from distractors, while these confuse LMs, especially in long contexts.

Keywords

language models, syntax, subject-verb agreement, Galician, Portuguese

1. Introdución

Os modelos de linguaxe (LMs, do inglés *Language Models*) convertéronse nun dos eixos centrais do procesamento da linguaxe natural grazas á súa capacidade de xerar *embeddings contextualizados* ou representacións vectoriais que integran información léxica, semántica e sintáctica (Caseli & Nunes, 2024). A evolución das arquitecturas, especialmente co paso das redes neurais recorrentes (*RNN*) aos *transformers* (baseados no mecanismo de *self-attention*), permitiu mellorar a aprendizaxe de relacións de dependencia na lingua (Jurafsky & Martin, 2026). Modelos como BERT ou GPT exemplifican esta nova xeración de LMs de grande escala.

Entre as moitas dúbidas que rodean os modelos de linguaxe está a cuestión de se estes adquiren e comprenden a estrutura sintáctica das linguas como o fan as persoas ou se se limitan a emular o comportamento lingüístico humano baseándose na recompilación de texto sobre a que se adestran. De ser así, afirmacións como a de Raposo (1992), "as línguas naturais [...] são adquiridas e faladas espontaneamente apenas pelos membros da espécie humana, isto é, por organismos com um determinado tipo específico de estrutura e organização mental", poderían quedar obsoletas. Esta incógnita foi a que motivou o

DOI: 10.21814/lm.17.2.498

presente traballo, que busca revelar se os modelos de linguaxe adquiren realmente unha lingua coma nós.

Concretamente, o obxectivo deste estudo é comparar o procesamento da concordancia suxeito-verbo en falantes de galego e portugués e en modelos de linguaxe neurais destas linguas. Para isto, avaliaremos os humanos e os LMs utilizando un dataset elaborado de forma controlada para cada idioma coa finalidade de comprobar como os factores externos a esta dependencia afectan á capacidade de resolución das persoas e dos modelos. En concreto, consideramos dous factores: (i) a distancia entre o núcleo do suxeito e o verbo principal e (ii) a presenza ou ausencia dun distractor. Un exemplo ilustrativo é a oración:

> *O* ***cabalo*** *que ten novas ferraduras* ***atravesa*** | ****atravesan*** *o campo ao galope.*

Na secuencia, o distractor *ferraduras* concorda en número coa forma agramatical (*atravesan*) e permite comprobar se o modelo vincula o verbo coa palabra inmediatamente precedente ou se é quen de identificar a relación xerárquica co núcleo do suxeito (*cabalo*).

Dunha banda, esperamos que os resultados demostren que os humanos non teñen dificultades identificando a concordancia suxeito-verbo e a agramaticalidade daquelas oracións que non sigan este principio sintáctico, xa que como falantes teñen interiorizada a gramática da súa lingua (Raposo, 1992). Respecto aos modelos de linguaxe, agardamos que neles o procesamento deste fenómeno estea condicionado pola presenza de distractores e pola distancia entre o suxeito e o verbo. Esta hipótese baséase nos resultados de estudos sobre procesamento da concordancia dos LMs en inglés, os modelos máis avanzados e adestrados nunha lingua cun sistema morfolóxico menos complexo que o das romances. Linzen et al. (2016) apuntan que estes "poden aprender a realizar a tarefa de concordancia verbo-número na maioría dos casos, aínda que o seu índice de erro aumenta nas oracións particularmente complexas". Ademais, debido a que en inglés "o núcleo do suxeito resulta ser o nome que precede o verbo [...], xeralmente a precisión non aporta directamente información sobre se as redes neurais profundas son quen de identificar o núcleo do suxeito" (Linzen & Baroni, 2021) e, polo tanto, se son quen de procesar a relación sintáctica entre este e o verbo principal.

2. Estado da arte

A avaliación da capacidade dos modelos de linguaxe neurais para procesar estruturas sintácticas, e en particular a concordancia suxeito-verbo, converteuse nunha das liñas de investigación máis activas no Procesamento da Linguaxe Natural (PLN).

O estudo pioneiro de Linzen et al. (2016) centrouse en comprobar se as redes LSTM (*Long Short-Term Memory*) podían identificar relacións de dependencia en inglés. Mediante a *number prediction task*, na que se lle presenta ao modelo unha oración incompleta para que prediga o número do verbo ausente, observouse que as LSTMs adestradas especificamente eran quen de recoñecer a concordancia, mais os modelos de linguaxe xenéricos precisaban dunha supervisión explícita. Ademais, a presenza de distractores incrementaba significativamente a taxa de erro.

Traballos posteriores ampliaron esta perspectiva. Gulordava et al. (2018) incorporaron linguas tipoloxicamente diversas (italiano, hebreo, ruso), introducindo tamén oracións gramaticais mais semanticamente anómalas, a fin de evitar que o modelo se apoiase en pistas léxicas. Ademais, avaliaron datos sobre o procesamento destas estruturas en italiano por parte de humanos. Os resultados amosaron que as LSTMs poden xestionar concordancias a longa distancia, tanto en oracións naturais como nas sen sentido, se ben a súa precisión diminúe coa presenza de distractores, de maneira similar ao que sucede cos falantes humanos.

Na mesma liña, Marvin & Linzen (2018) deseñaron un conxunto de probas controladas (*Targeted Syntactic Evaluation*) para avaliar a gramaticalidade das predicións. A comparación dos resultados dos LMs cos xuízos humanos puxo de manifesto que, malia o baixo rendemento relativo dos modelos, os erros de ambos sistemas amosan certas similitudes, o que suxire unha posible converxencia representacional.

O avance dos modelos baseados en *transformers* abriu un novo escenario. Goldberg (2019) mostrou que BERT procesa con grande precisión a concordancia suxeito-verbo en inglés, superando amplamente os resultados previos con LSTMs. Porén, Mueller et al. (2020), que experimentaron con inglés, francés, alemán, hebreo e ruso, evidenciaron que este comportamento varía segundo a lingua: os modelos monolingües superan os multilingües, e a riqueza morfolóxica influe na súa capacidade para identificar dependencias. BERT multilingüe, por exemplo, presenta altos niveis de acerto en inglés mais limitacións notables noutras linguas.

A cuestión da sistematicidade do coñecemento sintáctico foi retomada por Newman et al. (2021), quen observaron que os modelos poden actuar como "potenciais modelos psicolingüísticos", é dicir, que poden ser comparados con datos do procesamento sintáctico humano. Con todo, a TSE tende a sobreestimar as súas capacidades.

En relación co galego e o portugués, destacan dous estudos. Garcia & Crespo-Otero (2022) avaliaron a concordancia de número, xénero e persoa en modelos BERT, concluíndo que estes procesan ben as estruturas sintácticas, mais amosan dificultades cando interveñen distractores en contextos longos. Pola súa parte, de Dios-Flores & Garcia (2022) verificaron a capacidade de distintos BERT monolingües e multilingües para identificar concordancias suxeito-verbo e substantivo-adxectivo en galego. Así, constataron que o tamaño do corpus de adestramento é determinante: os modelos con máis datos (BERT) superan amplamente os de menor tamaño (Bertinho).

Cómpre salientar que, malia o valor destes traballos, os datasets empregados non sempre foron elaborados de maneira sistemática nin se compararon directamente cos xuízos de falantes humanos. A presente investigación busca avanzar nese punto, contribuíndo con datasets controlados para a avaliación da concordancia suxeito-verbo en galego e portugués, e incorporando por vez primeira nestas linguas a metodoloxía da *surprisal*, que permite aproximar o grao de gramaticalidade asignado polos modelos. Ao confrontar os resultados dos LMs cos dos falantes, aspírase non só a mellorar a avaliación dos modelos nestas dúas linguas, senón tamén a achegar novas evidencias sobre a súa capacidade para reflectir aspectos do procesamento humano da sintaxe.

3. Materiais e metodoloxía

Nesta sección descríbense os datos, modelos e metodoloxía empregados para dar resposta ás preguntas de investigación formuladas na introdución.

3.1. Materiais

A nosa análise baséase en dous eixos principais: os datasets e os modelos de linguaxe. Estes materiais constitúen a base dos experimentos realizados e a súa construción e escolla son esenciais para garantir a validez da investigación.

3.1.1. Datasets

Elaboráronse dous datasets, un en galego[1] e outro en portugués[2], cada un deles con 16 oracións, a partir das cales se xeraron 8 variantes considerando tres factores: (i) a distancia entre o suxeito e o verbo principal (contexto), (ii) a presenza dun distractor e (iii) a gramaticalidade da oración. Cada dataset está composto por 128 oracións, avaliadas tanto en experimentos con falantes humanos como con modelos de linguaxe.

Os suxeitos foron seleccionados de maneira sistemática e balanceada: 4 oracións con substantivo masculino singular, 4 con masculino plural, 4 con feminino singular e 4 con feminino plural. Engadiuse tamén diversidade semántica (ex.: [+*animal*], [- *concreto*], [+*humano*], [- *animado*]), aspecto ausente en estudos anteriores. O Cadro 1 exemplifica as oito variantes dunha mesma oración.

Os verbos e substantivos seleccionáronse dos vocabularios dos modelos a avaliar (*vid. infra*), excluíndo as palabras de maior frecuencia para evitar que os resultados reflectisen unicamente casos triviais de aprendizaxe estatística.

Distinguimos dous tipos de contexto: curto ou *short agreement dependecy* (ata 5 palabras entre núcleo de suxeito e verbo) e longo ou *long agreement dependency* (8–10 palabras). A presenza dun distractor[3] foi introducida como factor adicional, dado que pode interferir tanto en humanos como en modelos (Linzen et al., 2016; Gulordava et al., 2018; Garcia & Crespo-Otero, 2022; de Dios-Flores & Garcia, 2022). Finalmente, para cada combinación de contexto e distractor, construíronse versións gramaticais e agramaticais.

[1] https://github.com/helenaperezpuente/agreement-verbsubject-GL-PT/blob/main/dataset_GL.xlsx

[2] https://github.com/helenaperezpuente/agreement-verbsubject-GL-PT/blob/main/dataset_PT.xlsx

[3] Nestes datasets, consideramos como distractor unicamente a palabra inmediatamente anterior ao verbo principal que difire en número co núcleo do suxeito e polo tanto tamén coa alternativa verbal gramatical desa oración (ex.: "O cabalo que sacode a crina da cor das *amoras* atravesa o campo ao galope."). Con todo, hai que ter en conta que outras palabras próximas que non precedan inmediatamente ao verbo tamén poden actuar como distractores.

Contex.	Dis.	Gra.	Oración
Curto	Sen	Gra.	A propietaria que leu as páxinas detidamente asinou o contrato.
Curto	Sen	Agra.	A propietaria que leu as páxinas detidamente asinaron o contrato.
Curto	Con	Gra.	A propietaria que puxo as condicións asinou o contrato.
Curto	Con	Agra.	A propietaria que puxo as condicións asinaron o contrato.
Longo	Sen	Gra.	A propietaria que estaba encantada coas condicións propostas onte asinou o contrato.
Longo	Sen	Agra.	A propietaria que estaba encantada coas condicións propostas onte asinaron o contrato.
Longo	Con	Gra.	A propietaria que estaba encantada coas condicións propostas naquelas páxinas asinou o contrato.
Longo	Con	Agra.	A propietaria que estaba encantada coas condicións propostas naquelas páxinas asinaron o contrato.

Cadro 1: As oito variantes da oración número 9 do dataset para galego con variación de contexto (contex.), distractor (dis.) e gramaticalidade (gra.).

3.1.2. Modelos

Foron avaliados tres modelos baseados en *transformers*[4], sempre na súa versión *base* (12 capas):

- **Bertinho-base** (Vilares et al., 2021), modelo monolingüe en galego adestrado coa Wikipedia galega (45M palabras; vocabulario de 30.000 *tokens*).
- **BERT-gl-base** (Garcia, 2021), tamén monolingüe en galego, adestrado cun corpus de 500M palabras e vocabulario de 119.547 *tokens* (Garcia & Crespo-Otero, 2022).
- **BERTimbau-base** (Souza et al., 2020), modelo monolingüe en portugués adestrado con 2,68 billóns de *tokens* do BrWaC e vocabulario de 30.000 *tokens*.[5]

3.2. Metodoloxía

Deseñáronse tres experimentos: (i) avaliacións de aceptabilidade con falantes, (ii) *Targeted Syntactic Evaluation* (TSE) e (iii) análise baseada en *surprisal*. Os dous últimos céntranse nos modelos de linguaxe e permiten comparar o seu rendemento co dos falantes humanos.

3.2.1. Compilación de datos de falantes

A avaliación cos falantes levouse a cabo mediante avaliacións de aceptabilidade en Microsoft Forms. Deseñáronse 8 para galego e 8 para portugués, cada unha con 16 oracións (unha variante de cada ítem do dataset).[6] Así, cada participante avaliou un subconxunto reducido, o que permitiu minimizar a fatiga e garantir unha maior calidade nas respostas (Langsford et al., 2018).

Cada oración foi seguida dunha escala Likert do 1 ao 5, unha metodoloxía habitual e robusta para recoller graos de aceptabilidade mesmo con mostras pequenas (Langsford et al., 2018; Bross, 2019). Para evitar que a terminoloxía condicionase as respostas, empregouse a expresión "ben formadas" en lugar de "gramaticais" ou "correctas", seguindo recomendacións en psicolingüística e avaliación de xuízos de aceptabilidade (Bross, 2019; de Dios-Flores et al., 2023).

As avaliacións de aceptabilidade comezaban cunha breve introdución explicativa superficial, sen termos técnicos, que explicaba que se trataba dun estudo sobre como as persoas procesamos a lingua.[7] Ademais, recolléronse datos sobre a competencia lingüística dos participantes (preguntando pola súa condición de falantes nativos e o nivel autodeclarado en caso contrario), tanto en galego como en portugués.[8]

[4] Avaliáronse unicamente modelos de linguaxe bidireccionais, xa que un dos obxectivos principais deste proxecto é ampliar a investigación existente sobre diferentes versións de BERT para o galego e o portugués, sen entrar na comparativa con arquitecturas causais.

[5] Cómpre ter en conta que *BERTimbau-base* foi adestrado principalmente en portugués do Brasil, mais consideramos que isto non afecta á avaliación da concordancia suxeito-verbo. Exploráronse outros modelos para o portugués europeo, mais non se empregaron por teren o inglés como vocabulario base.

[6] https://github.com/helenaperezpuente/agreement-verbsubject-GL-PT/blob/main/surveys.md

[7] "Este cuestionario forma parte dun Traballo de Fin de Grao sobre linguas e tecnoloxías. En concreto, a finalidade desta breve enquisa (menos de 3 mins.) é observar como as persoas entendemos a lingua".

[8] Neste estudo, a categoría de "non nativos" refírese a falantes que non adquiriron a lingua como L1 estrita, pero que demostran unha competencia avanzada e un uso habitual da mesma. No contexto galego, este perfil correspóndese maioritariamente con neofalantes, mentres

Finalmente, a distribución das variantes foi realizada de forma sistemática mediante o uso do deseño en cadrado latino, garantindo así que cada participante recibise unha combinación equilibrada de factores (contexto, distractor, gramaticalidade) sen que se puidesen identificar padróns.

3.2.2. Compilación de datos de modelos

Surprisal. A metodoloxía de *surprisal* consiste en cuantificar a sorpresa que experimenta un modelo de linguaxe ante unha palabra nun contexto determinado. O valor de *token surprisal* calcúlase como o logaritmo da probabilidade inversa de que esa palabra apareza no contexto oracional (Hale, 2001; Rezaii et al., 2023). Valores máis baixos indican que o modelo considera a palabra máis probable no contexto e, polo tanto, "menos sorprendente". Esta medida correlaciónase co esforzo de procesamento humano e os tempos de lectura (Hale, 2001).

Para este experimento, os datos de entrada consisten en pares de oracións completas (sen posicións ocultas): unha variante gramatical e outra agramatical. A adaptación dos datasets é relativamente directa: creouse un arquivo CSV con tres columnas: número de oración, texto da oración e gramaticalidade (`Right` ou `Wrong`).

Item	Oración	Gram.
1	O cabalo que sacode a crina da cor das amoras atravesa o campo ao galope.	Right
1	O cabalo que sacode a crina da cor das amoras atravesan o campo ao galope.	Wrong

Cadro 2: Exemplo da adaptación das variantes da oración 1 do dataset galego para *surprisal*.

O cálculo realizouse mediante a libraría `minicons`[9] (Misra, 2022). Cada modelo (Bertinho-base, BERT-gl-base e BERTimbau-base) avaliouse sobre o dataset adaptado correspondente (GL ou PT). O output obtido converteuse á escala Likert (1–5), permitindo a comparación directa cos datos humanos.

TSE. A *Targeted Syntactic Evaluation* (TSE) é unha metodoloxía que avalía o coñecemento sintáctico dun LM mediante oracións nas que se oculta un elemento para o que se presentan dúas alternativas: a gramatical e a agramatical (Newman et al., 2021). A precisión do modelo calcúlase como a proporción de casos en que asigna maior probabilidade á alternativa correcta.

Para adaptar os datasets á TSE, reducíronse as variantes á metade, dado que a gramaticalidade xa se incorpora nas alternativas presentadas. Por exemplo, o Cadro 3 mostra a adaptación da oración 4 do dataset galego, onde o símbolo "*****" indica a posición do verbo a predicir.

A execución do experimento realizouse en Google Colab, onde se tokenizaron as oracións e se obtivo o índice da posición oculta e predicións do modelo. O output asigna 1 se a alternativa gramatical é máis probable, e 0 en caso contrario. Finalmente, a precisión calcúlase como a razón entre oracións correctamente preditas e o total de oracións avaliadas.

4. Experimentos e resultados

4.1. Experimentos

Para a recolección de datos humanos empregouse o método das avaliacións de aceptabilidade. Estas foron distribuídas entre falantes nativos e persoas cunha alta competencia na lingua obxecto de estudo. Participaron 93 individuos en total: 76 nas avaliacións de aceptabilidade para galego e 17 nas de portugués.

En galego (GL) recolléronse polo menos tres valoracións diferentes por oración, mentres que en portugués (PT) houbo como mínimo unha resposta por oración. A autopercepción lingüística dos participantes non nativos foi elevada: 4,87 de media sobre 5 no caso do portugués e 3,95 no caso do galego, valor este último que debe ser interpretado á luz da situación sociolingüística do idioma.[10]

No que atinxe aos modelos de linguaxe, realizáronse dous experimentos principais: *surprisal* e TSE (*Targeted Syntactic Evaluation*), aplicados aos LMs Bertinho-base e BERT-gl-base para galego e BERTimbau-base para portugués.

que no contexto portugués abrangue falantes de herdanza ou en contexto migratorio. A súa inclusión garante a representatividade da realidade sociolingüística destas comunidades.

[9] https://github.com/kanishkamisra/minicons

[10] No marco da complexa relación entre os galegofalantes e o estándar da lingua, é preciso matizar que na investigación non se tivo en conta o impacto da falta de familiaridade con certa terminoloxía estándar. En traballos futuros poderíase illar esta variable, xa que a reacción ante estes termos podería condicionar os seus xuízos lingüísticos.

Item	Variante TSE	Alt. correcta	Alt. incorrecta
4	O sol que luciu con forza este venres nas costas galegas * que se superasen os 30 graos.	fixo	fixeron
4	O sol que luciu intensamente * que se superasen os 30 graos.	fixo	fixeron

Cadro 3: Exemplo da adaptación das variantes da oración 4 do dataset galego para a TSE.

4.2. Resultados

A continuación preséntanse os resultados das avaliacións de aceptabilidade e dos experimentos con LMs. Nas descricións facemos referencia a valores medios.

4.2.1. Resultados das avaliacións de aceptabilidade con humanos

Os falantes valoraron claramente de forma diferente as oracións gramaticais e as agramaticais. A continuación presentamos polo miúdo os resultados.

Distractor. As oracións con distractor recibiron puntuacións máis altas que as que non o teñen, tanto en exemplos gramaticais como en agramaticais. Por exemplo, en galego as oracións gramaticais con distractor acadaron unha media de 4,42 fronte a 3,87 das sen distractor, mentres que en portugués recibiron un 4,11 e un 3,82 respectivamente. Aquelas agramaticais con distractor tamén foron valoradas de media cunha cifra máis alta que as que non presentan este elemento: 2,31 (GL) e 1,80 (PT), fronte a 2,14 (GL) e 1,57 (PT).

Contexto. As oracións de contexto curto foron mellor valoradas que as de contexto longo en ambas as linguas. En galego, as gramaticais de contexto curto recibiron 4,32 puntos de media fronte a 3,97 das de contexto longo. No portugués observouse unha tendencia semellante, aínda que nas oracións agramaticais houbo unha inversión lixeira (1,76ñas de contexto longo fronte a 1,62 nas de contexto curto).

As combinacións de contexto e distractor amosaron efectos complexos. Por exemplo, en galego as oracións de contexto longo gramaticais con distractor foron mellor valoradas (4,38) que as sen distractor (3,56). No portugués, en cambio, as oracións agramaticais de contexto curto sen distractor recibiron a menor puntuación media (1,33).

	Condición	Galego	Portugués
Gramaticalidade	Gram.	4,14	3,97
	Agram.	2,22	1,69
Distractor	Con (gr.)	4,42	4,11
	Sen (gr.)	3,87	3,82
	Con (agr.)	2,31	1,80
	Sen (agr.)	2,14	1,57
Contexto	Curto (gr.)	4,32	4,22
	Longo (gr.)	3,97	3,71
	Curto (agr.)	2,29	1,62
	Longo (agr.)	2,15	1,76

Cadro 4: Resultados das avaliacións de aceptabilidade con humanos.

4.2.2. Resultados dos LMs: surprisal

A avaliación con *surprisal* mostrou diferenzas claras entre oracións gramaticais e agramaticais, aínda que menos marcadas que nos humanos.

Galego. En primeiro lugar, ambos os modelos distinguen entre oracións gramaticais e agramaticais, aínda que BERT-gl-base mostra unha separación máis marcada. No caso de Bertinho-base, as oracións gramaticais acadaron unha media de 3,70 fronte a 2,87 nas agramaticais, mentres que BERT-gl-base obtivo 3,59 fronte a 2,43.

A presenza de distractores tivo un efecto desigual: en Bertinho-base, as oracións gramaticais sen distractor (3,78) foron lixeiramente máis aceptables que con distractor (3,63); en cambio, BERT-gl-base apenas amosou diferenzas (3,592 vs. 3,587). Curiosamente, este último modelo considera as agramaticais con distractor (2,70) máis aceptables ca as sen distractor (2,40).

No relativo ao contexto, ambos os modelos atribúen maior aceptabilidade ás oracións gramaticais de contexto curto (Bertinho-base: 3,76; BERT-gl-base: 3,63) que ás de contexto longo (3,64 e 3,55 respectivamente). Para as oracións agramaticais, as diferenzas entre contextos son mínimas.

Portugués. No caso de BERTimbau-base, obsérvase unha clara diferenciación entre oracións gramaticais (4,12) e agramaticais (3,08). A presenza de distractor incrementa lixeiramente a aceptabilidade tanto en oracións gramaticais (4,16 vs. 4,08) coma en agramaticais (3,17 vs. 2,96).

	Condición	Bertinho	BERT-gl
Gramaticalidade	Gram.	3,70	3,59
	Agram.	2,87	2,43
Distractor	Con (gr.)	3,63	3,59
	Sen (gr.)	3,78	3,59
	Con (agr.)	2,87	2,70
	Sen (agr.)	2,87	2,40
Contexto	Curto (gr.)	3,76	3,63
	Longo (gr.)	3,64	3,55
	Curto (agr.)	2,87	2,41
	Longo (agr.)	2,87	2,45

Cadro 5: Resultados de *surprisal* en galego.

A nivel de contexto, as oracións gramaticais de contexto curto (4,23) superan ás de contexto longo (4,02). Do mesmo xeito, as oracións agramaticais de contexto curto (3,11) resultan lixeiramente menos aceptables cás de contexto longo (3,06).

	Condición	BERTimbau-base
Gramaticalidade	Gram.	4,12
	Agram.	3,08
Distractor	Con (gr.)	4,16
	Sen (gr.)	4,08
	Con (agr.)	3,17
	Sen (agr.)	2,96
Contexto	Curto (gr.)	4,23
	Longo (gr.)	4,02
	Curto (agr.)	3,11
	Longo (agr.)	3,06

Cadro 6: Resultados de *surprisal* en portugués.

4.2.3. Resultados dos modelos de linguaxe: TSE

Os resultados da TSE (*vid.* Figura 1) amosan unha grande disparidade entre os modelos: Bertinho-base non supera a probabilidade aleatoria (0,34); BERT-gl-base acada unha precisión sólida (0,66) pero cae en contextos longos con distractor; mentres que BERTimbau-base ofrece os mellores resultados (0,73), especialmente en contextos curtos sen distractor (0,94).

4.3. Discusión

Resultados de humanos. Os datos confirman que os falantes identifican con claridade a dependencia suxeito-verbo, incluso en oracións con distractores. Naquelas agramaticais, obsérvase un aumento na puntuación de aceptación cando o distractor está presente. Este achado é consistente co efecto de atracción de concordancia (*agreement attraction effect*) ben documentado na área da psicolingüística (Bock & Miller, 1991; Wagers et al., 2009). Neste contexto, a presenza

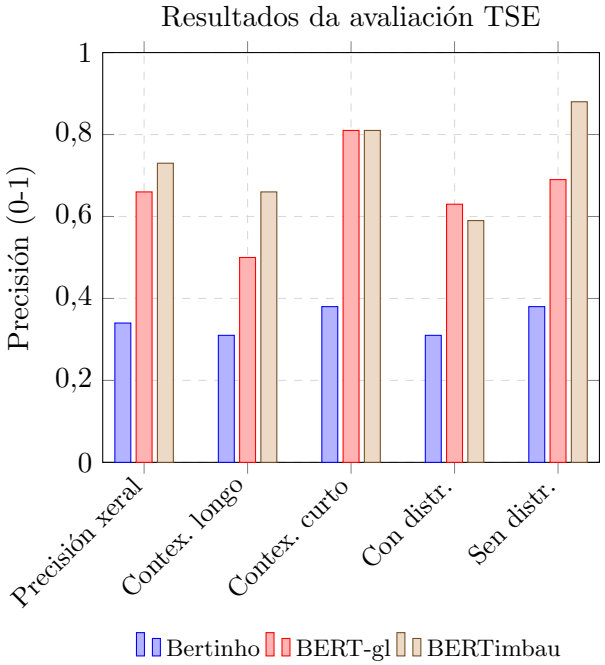

Figura 1: Resultados da avaliación TSE.

dun distractor que concorda en número co verbo xera unha "ilusión de gramaticalidade", o que dificulta a detección do erro por parte do lector (Wagers et al., 2009). En segundo lugar, nas oracións gramaticais, a presenza do distractor tamén incrementou significativamente as puntuacións. Dado que a teoría de Wagers predí interferencia nas condicións gramaticais, esta mellora pode ser resultado da carga informativa ou da probabilidade semántica das construcións avaliadas. Unha liña de investigación futura sería explorar os mecanismos de procesamento nestes casos con técnicas de *eye-tracking* e análise de tempos de lectura.

Resultados de modelos. BERT-gl-base supera a Bertinho-base na diferenciación entre oracións gramaticais e agramaticais, coincidindo con traballos previos (Garcia & Crespo-Otero, 2022; de Dios-Flores & Garcia, 2022; de Dios-Flores et al., 2023). O contexto curto é máis doadamente procesado ca o longo, e a presenza de distractor tende a confundir os modelos, atribuíndo maior aceptabilidade ás oracións incorrectas.

En portugués, BERTimbau destaca pola súa maior precisión global. Con todo, tamén mostra unha certa vulnerabilidade ante os distractores, o que confirma que estes elementos representan un reto para os LMs (efecto de atracción de concordancia).

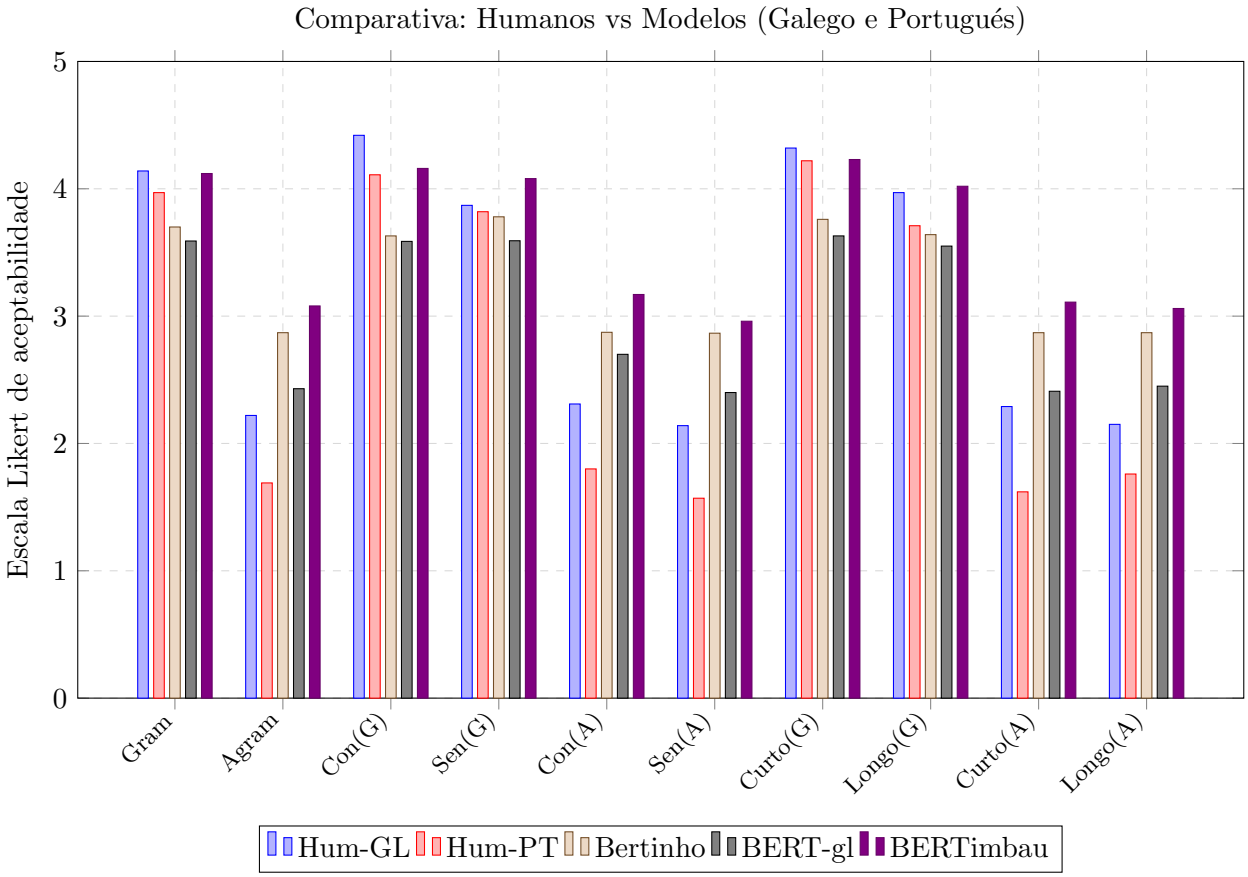

Figura 2: Comparación de xuízos humanos e surprisal dos modelos. As abreviaturas G/A refírense a Gramatical/Agramatical, Con/Sen a con/sen distractor, e Curto/Longo a contexto curto/longo.

Comparación entre humanos e modelos. Os humanos procesan mellor que os LMs a concordancia suxeito-verbo, distinguindo con máis claridade entre oracións gramaticais e agramaticais, como mostra a Figura 2. Ademais, mentres que para os modelos os distractores introducen ruído, para os falantes parecen ter un efecto positivo inesperado na avaliación da aceptabilidade das cadeas lingüísticas gramaticais. Tanto humanos como modelos comparten a dificultade no procesamento de oracións con contexto longo, aínda que a intensidade do efecto é maior nos LMs. Así, as variantes que os modelos de linguaxe procesan mellor son as oracións de contexto curto sen distractor, mentres que os falantes, en xeral, procesan mellor aquelas de contexto curto con distractor.

Finalmente, cómpre matizar o alcance do termo "procesamento" empregado neste estudo. As métricas utilizadas céntranse no resultado final da comprensión sintáctica e non nos mecanismos cognitivos ou computacionais que operan en tempo real. Tanto os xuízos emitidos polos falantes como as probabilidades xeradas polos modelos constitúen medidas *a posteriori*: reflicten a competencia gramatical a través dunha resposta posterior á análise da oración, e non mediante a observación directa dos procesos internos que teñen lugar durante a lectura.

5. Conclusións e traballo futuro

Neste traballo presentouse unha avaliación das habilidades sintácticas para procesar a concordancia suxeito-verbo en modelos de linguaxe neurais para galego e portugués, así como en falantes nativos e altamente competentes de ambas as dúas linguas. O obxectivo principal era comprobar se os LMs son quen de codificar a estrutura sintáctica da lingua e se esta codificación é equiparable á adquisición de coñecementos lingüísticos en humanos.

Para tal fin construíronse dous datasets (GL e PT), controlando variables de interese como a gramaticalidade, a presenza de distractores, a distancia entre o núcleo do suxeito e o verbo, e a frecuencia destes elementos nun corpus de referencia. Estes materiais permitiron levar a cabo tres experimentos: (i) avaliacións de aceptabilidade a falantes, (ii) avaliación por *surprisal* e (iii) avaliación mediante *Targeted Syntactic Evaluation* (TSE). Testeáronse os modelos Bertinho-base (GL), BERT-gl-base (GL) e BERTimbau-base (PT).

Os resultados mostran que tanto falantes como modelos discriminan entre oracións gramaticais e agramaticais, mais os humanos establecen diferenzas máis acusadas. De feito, a presenza dun distractor nas oracións gramaticais amosouse facilitadora para os falantes, en contra do esperado, mentres que tivo un efecto negativo para os modelos. Isto suxire que a representación sintáctica dos LMs se ve perturbada por factores lineais en consonancia co efecto de atracción de concordancia, ao contrario que nos humanos. Ademais, tanto en persoas como en modelos a distancia longa entre suxeito e verbo dificulta o procesamento, o que confirma a complexidade deste tipo de dependencias.

En canto ao rendemento dos modelos, atopamos diferenzas salientables: o LM portugués BERTimbau-base acadou unha precisión xeral do 73 %, con valores superiores ao 80 % e mesmo ao 90 % en determinados contextos; BERT-gl-base chegou ao 66 % e superou tamén o 80 % nalgúns casos; mentres que Bertinho-base, adestrado cun corpus máis reducido, ficou nun 34 % de precisión xeral, situándose mesmo por debaixo da media probable (*below-chance performance*). Estes datos parecen confirmar que o tamaño do corpus de adestramento é chave para a robustez sintáctica dos LMs.

Como liñas de traballo futuro, propoñemos afondar no impacto dos distractores no procesamento humano da dependencia suxeito-verbo gramaticais mediante técnicas da psicolingüística, como *eye-tracking* ou rexistro dos tempos de lectura. Así mesmo, resulta pertinente aplicar análises estatísticas máis complexas (tests de significancia, correlacións, etc.) que permitan matizar a comparación entre resultados de modelos e humanos. Tamén consideramos de interese ampliar a avaliación a outros fenómenos sintácticos, empregando a metodoloxía descrita e adaptando para galego e portugués os datasets dispoñibles en inglés (Warstadt et al., 2020). Unha vía adicional é avaliar modelos multilingües, explorando se a aprendizaxe cruzada favorece a xeneralización destas dependencias en linguas próximas.

Ademais de achegar dous novos datasets construídos de forma sistemática para avaliar modelos en galego e en portugués, contribuímos ao estado da arte coa primeira comparación nestas linguas do procesamento da dependencia suxeito-verbo en modelos e falantes, sendo usada por vez primeira a metodoloxía *surprisal* nesta tarefa en galego e portugués. Así, con este traballo colabórase na ampliación da investigación para linguas minoritarias e economicamente menos potentes en relación ao inglés na área das tecnoloxías da linguaxe, o que "favorece o seu prestixio (un factor decisivo na normalización lingüística), mais tamén garante os dereitos lingüísticos dos cidadáns, reduce a desigualdade social e acurta a fenda dixital" (de Dios-Flores et al., 2022).

Agradecementos

Gustaríame agradecer a Marcos Garcia pola súa orientación durante a elaboración do meu Traballo de Fin de Grao, que deu lugar ao presente artigo. Tamén quero agradecer a Lisa Ferreira pola súa axuda na resolución das dúbidas relacionadas co portugués, así como a todas as persoas que colaboraron nas avaliacións de aceptabilidade e fixeron posible esta investigación. Finalmente, agradezo aos revisores os seus comentarios, que contribuíron a mellorar a calidade deste traballo.

Referencias

Bock, Kathryn & Carol A Miller. 1991. Broken agreement. *Cognitive Psychology* 23(1). 45–93. 10.1016/0010-0285(91)90003-7

Bross, Fabian. 2019. Acceptability ratings in linguistics: A practical guide to grammaticality judgments, data collection, and statistical analysis. Version 1.0. Mimeo.

Caseli, Helena M. & Maria G. V. Nunes (eds.). 2024. *Processamento de linguagem natural: Conceitos, técnicas e aplicações em Português*. BPLN.

de Dios-Flores, Iria & 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. 10.26342/2022-69-1

de Dios-Flores, Iria, Juan Garcia Amboage & Marcos Garcia. 2023. Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies. En *61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 203–222. 10.18653/v1/2023.acl-long.12

de Dios-Flores, Iria, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, José Ramom Pichel, Marcos García, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín-Diz, Manuel González González, Senén Barro & Xosé Luis Regueira. 2022. The nós project: Opening routes for the Galician language in the field of lan-

guage technologies. En *Workshop Towards Digital Language Equality*, 52–61.

Garcia, Marcos. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. En *59th Annual Meeting of the Association for Computational Linguistics (ACL) and the 11th International Joint Conference on Natural Language Processing (IJCNLP)*, 3625–3640. 10.18653/v1/2021.acl-long.281

Garcia, Marcos & Alfredo Crespo-Otero. 2022. A targeted assessment of the syntactic abilities of transformer models for galician-portuguese. En *15th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 46–56. 10.1007/978-3-030-98305-5_5

Goldberg, Yoav. 2019. Assessing BERT's syntactic abilities. arXiv [cs.CL]. 10.48550/arXiv.1901.05287

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen & Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. En *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 1195–1205. 10.18653/v1/N18-1108

Hale, John. 2001. A probabilistic earley parser as a psycholinguistic model. En *2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 1—8. 10.3115/1073336.1073357

Jurafsky, Daniel & James H. Martin. 2026. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, with language models*. Online manuscript.

Langsford, Steven, Andrew Perfors, Andrew Hendrickson, Lauren Kennedy & Danielle Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a journal of general linguistics* 3(1). 37. 10.5334/gjgl.396

Linzen, Tal & Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics* 7. 195–212. 10.1016/j.csl.2023.101520

Linzen, Tal, Emmanuel Dupoux & Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4. 521–535. 10.1162/tacl_a_00115

Marvin, Rebecca & Tal Linzen. 2018. Targeted syntactic evaluation of language models. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1192–1202. 10.18653/v1/D18-1151

Misra, Kanishka. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. arXiv [cs.CL]. 10.48550/arXiv.2203.13112

Mueller, Aaron, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina & Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. En *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 5523–5539. 10.18653/v1/2020.acl-main.490

Newman, Benjamin, Kai-Siang Ang, Julia Gong & John Hewitt. 2021. Refining targeted syntactic evaluation of language models. En *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 3710–3723. 10.18653/v1/2021.naacl-main.290

Raposo, Eduardo Paiva. 1992. *Teoria da gramática: a faculdade da linguagem*. Caminho

Rezaii, Neguine, James Michaelov, Sylvia Josephy Hernandez, Boyu Ren, Daisy Hochberg, Megan Quimby & Bradford Dickerson. 2023. Measuring sentence information via surprisal: Theoretical and clinical implications in nonfluent aphasia. *Annals of Neurology* 94(4). 647–657. 10.1002/ana.26744

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. En *9th Brazilian Conference on Intelligent Systems (BRACIS)*, 403–417. 10.1007/978-3-030-61377-8_28

Vilares, David, Marcos Garcia & Carlos Gómez-Rodríguez. 2021. Bertinho: Galician BERT representations. arXiv [cs.CL]. 10.26342/2021-66-1

Wagers, Matt, Ellen Lau & Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61(2). 206–237. 10.1016/j.jml.2009.04.002

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang & Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics* 8. 377–392. 10.1162/tacl_a_00321