FACULTADE DE FILOLOXÍA

Grao en Lingua e Literatura Galegas

Traballo de Fin de Grao

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

Helena Pérez Puente

Titor: Marcos Garcia González

Curso 2023/2024

Xuño 2024

Traballo de Fin de Grao presentado na Facultade de Filoloxía da Universidade de Santiago de Compostela para a obtención do Grao en Lingua e Literatura Galegas

FACULTADE DE FILOLOXÍA

Grao en Lingua e Literatura Galegas

Traballo de Fin de Grao

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

Helena Pérez Puente

Titor: Marcos Garcia González

Curso 2023/2024

Xuño 2024

Traballo de Fin de Grao presentado na Facultade de Filoloxía da Universidade de Santiago de Compostela para a obtención do Grao en Lingua e Literatura Galegas

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA FACULTADE DE FILOLOXÍA

Formulario de delimitación do título e resumo

Traballo de Fin de Grao curso 2023/2024

APELIDOS E NOME:	Pérez Puente Helena
GRAO EN:	Lingua e Literatura Galegas
(NO CASO DE MODERNAS) MENCIÓN EN:	
TITOR/A:	Marcos Garcia González
LIÑA TEMÁTICA ASIGNADA:	Lingüística e Procesamento das Linguas. Aplicacións

SOLICITO a aprobación do seguinte título e resumo:

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

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 de 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 enquisas a falantes destas linguas para comprobar a aceptabilidade das oracións en relación ás variables referidas. Cunha versión adaptada dos datasets, avaliáronse os modelos de lingua neurais máis avanzados para galego e portugués. Os resultados parecen indicar que os modelos de ambas as linguas non identifican adecuadamente a agramaticalidade das oracións nas que o principio de concordancia suxeito-verbo non se cumpre, ao contrario do que fan as persoas, que si son quen de distinguir as oracións gramaticais daquelas agramaticais.

Santiago de Compostela, 26 de abril de 2024.

Sinatura do/a interesado/a	Visto e prace (sinatura do/a titor/a)	Aprobado pola Comisión do Traballo de Fin de Grao coa data
	 Assinado digitalmente por Marcos Garcia González	- 6 MAI. 2024 Selo da Facultade de Filoloxía

SRA. PRESIDENTA DA COMISIÓN DO TRABALLO DE FIN DE GRAO

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

Título: El procesamiento de la estructura sintáctica en los modelos de lenguaje neuronales: la concordancia sujeto-verbo en gallego y portugués.

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

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 de 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 enquisas a falantes destas linguas para comprobar a aceptabilidade das oracións en relación ás variables referidas. Cunha versión adaptada dos datasets, avaliáronse os modelos de lingua neurais máis avanzados para galego e portugués. Os resultados parecen indicar que os modelos de ambas as linguas non identifican adecuadamente a agramaticalidade das oracións nas que o principio de concordancia suxeito-verbo non se cumpre, ao contrario do que fan as persoas, que si son quen de distinguir as oracións gramaticais daquelas agramaticais.

Resumen: Combinando conocimientos de las áreas de lingüística computacional y de psicolingüística, en este trabajo se explora el conocimiento sintáctico de los modelos de lenguaje neuronales, concretamente el principio de la concordancia sujeto-verbo en gallego y portugués. Con el fin de estudiar si los modelos procesan este fenómeno de forma semejante a los humanos, se elaboraron dos datasets (uno en gallego y otro en portugués) con 16 oraciones y 8 variantes para cada una de de ellas, que tienen en cuenta la gramaticalidad, la presencia o ausencia de un distractor y la distancia entre el sujeto y el verbo principal. Con estos elementos, se realizaron encuestas a hablantes de estas lenguas para comprobar la aceptabilidad de las oraciones en relación a las variables referidas. Con una versión adaptada de los datasets, se evaluaron los modelos de lenguaje neuronales más avanzados para gallego y portugués. Los resultados parecen indicar que los modelos de ambas lenguas no identifican adecuadamente la agramaticalidad de las oraciones en las que el principio de concordancia sujeto-verbo no se cumple, al contrario de lo que hacen las personas, que sí son capaces de distinguir las oraciones gramaticales de aquellas agramaticales.

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, the most advanced neural language models for Galician and Portuguese were evaluated. The results seem to indicate that the models for both languages do not adequately identify the ungrammaticality of sentences in which the subject-verb agreement principle is not met, unlike people, who are able to distinguish grammatical sentences from ungrammatical ones.

Palabras chave: modelos de linguaxe, procesamento da linguaxe natural, sintaxe, concordancia suxeito-verbo, galego, portugués.

Palabras clave: modelos de lenguaje, procesamiento del lenguaje natural, sintaxis, concordancia sujeto-verbo, gallego, portugués

Keywords: language models, natural language processing, syntax, subject-verb agreement, Galician, Portuguese.

Declaración de orixinalidade

Eu, Helena Pérez Puente, con DNI 53976965J, declaro que o presente Traballo de Fin de Grao é integramente orixinal e inédito. Polo tanto, non incorre en plaxio e as fontes de información están correctamente citadas.

Índice

Introdución

1. Lingüística computacional

A lingüística computacional é a disciplina que ten como obxectivo "a simulación da competencia comunicativa do home a nivel escrito e/ou a nivel oral, ou, ao menos, a simulación dalgunha subcompetencia desta"[1] (Tordera Yllescas, 2011, p. 8). O procesamento da linguaxe natural (PLN) constitúe a súa rama aplicada. Esta usa técnicas "orientadas á construción dunha representación do contido das producións lingüísticas en termos dunha determinada metalinguaxe" (Martí Antonín [ed.], 2003, p. 10, citado en Tordera Yllescas, 2011, p. 5). Ademais, ten carácter interdisciplinario, xa que nela conflúen coñecementos doutros campos como o da intelixencia artificial (Tordera Yllescas, 2011, p. 31). James F. Allen, profesor da Universidade de Toronto e membro da Association for the Advancement of Artificial Intelligence (AAAI), apunta que "o obxectivo último [da investigación nesta área] é poder deseñar modelos [computacionais da linguaxe] que se aproximen á actuación humana nas habilidades lingüísticas de ler, escribir, escoitar e falar" (1995, p. 1).

2. Modelos de linguaxe

Un modelo é unha "simplificação de um fenômeno complexo" (Paes et al., 2023, p. 318). Por extensión, os modelos de linguaxe (LMs, do inglés *Language Models*) son simplificacións de linguas naturais que se valen de ferramentas computacionais para representar con números a información simbólica destes sistemas lingüísticos. Isto é, o proceso de codificación non consiste en atribuír a cada unidade lingüística un número, senón que codifica tamén a información léxica, sintáctica e semántica de cada elemento (Paes et al., 2023, p. 318). Os modelos adéstranse cun amplo corpus de textos, que procesan e dos que extraen a información e patróns que lles permiten representar unha lingua (Paes et al., 2023, p. 341).

3. Tipos de modelos

3.1. Modelos de linguaxe probabilísticos

Existen varios tipos de modelos de linguaxe, entre eles os modelos de linguaxe probabilísticos ou modelos n-grama**,** que son os máis sinxelos. Toman o seu nome do n-

1 Esta e as restantes citas que se atopan ao longo do traballo son traducións propias, excepto aquelas extraídas de textos en portugués, que foron conservadas na súa lingua orixinal.

grama, unha secuencia de *n* número de palabras (Jurafsky e Martin, 2024, capítulo 3, p. 2). O seu funcionamento baséase no cálculo da probabilidade da próxima palabra nunha secuencia tendo en conta as *n* palabras inmediatamente anteriores. Cómpre recordar que o token (elementos nos que o modelo divide unha oración, que normalmente se corresponden con palabras, signos de puntuación, e eventualmente prefixos ou sufixos) que sucede ao n-grama non é xerado polo modelo, senón que se escolle segundo a distribución de probabilidade dentro do modelo. Esta está marcada polo corpus co que se adestra, pois varía segundo os tokens que coñeza o LM (Paes et al., 2023, p. 320).

O presuposto de Markov (Jurafsky e Martin, 2024, capítulo 3, p. 2) sostén que se pode predicir a probabilidade dun elemento sen mirar moito máis atrás. Así, se o modelo é de bigramas (2-grama) usa a palabra anterior para predicir a seguinte, se é de trigramas (3-grama) considera as dúas previas, e así sucesivamente.

3.2. Modelos de linguaxe neurais

Os modelos de linguaxe neurais, por outra banda, son os máis avanzados. Estes baséanse nas redes neurais artificiais, unidades computacionais que codifican palabras (*input*) como vectores (é dicir, convértenas en números), a partir dos que se calcula o *output* ou secuencia de saída. Este sistema tamén recibe o nome de *deep learning*, xa que ditas unidades se ordenan en varias capas sucesivas nas que se procesa a información. Os datos só poden ir cara a capa seguinte e así sucesivamente (Jurafsky e Martin, 2024, capítulo 7, p. 1). Cantas máis capas ten o modelo, maior é a súa complexidade.

Esta representación computacional da lingua dista moito daquelas que seguen os modelos simbólicos baseados na maior parte da lingüística teórica. Ademais, a forma na que estes modelos adquiren competencia gramatical difire tamén do tradicionalmente considerado necesario polos lingüistas (Linzen e Baroni, 2021, p. 196). De feito, o adestramento consiste en darlles grandes cantidades de texto, xa que necesitan millóns de palabras para adquirir a lingua, mentres que as persoas necesitan menos cantidade de input.

Os LMs son modelos distribucionais. É dicir, deducen o significado das palabras segundo o contexto no que ocorren, seguindo a hipótese distribucional, que afirma que "a similitude en significado resulta nunha similitude na distribución lingüística" (Boleda, 2020, p. 214). Isto quere dicir que asumen que dúas palabras que normalmente aparecen en contextos semellantes no corpus teñen un significado tamén semellante (Erk, 2012, p. 635) e que aquelas que aparecen recorrentemente próximas teñen algunha relación semántica (Paes et al., 2023, p. 318). Como resultado, os modelos usan estes datos para a predición de palabras. Ademais, esta información, codificada numericamente, pode representarse no espazo vectorial para analizar e comparar os contextos nos que aparecen as unidades lingüísticas e, polo tanto, o seu contido semántico (Paes et al., 2023, p. 318). Por exemplo, unha palabra polisémica está representada por vectores distintos dependendo do seu significado en cada contexto.

Como se comentou anteriormente, para procesar unha lingua os modelos de redes neurais converten as palabras en secuencias de números reais. Estas representacións

vectoriais denomínanse embeddings. A partir do 2017, os modelos de linguaxe neurais comezaron a construír os embeddings tendo en conta o contexto no que se presenta cada palabra, isto é, pasaron a utilizarse embeddings contextualizados (Paes et al., 2023, p. 326).

Para xerar embeddings contextualizados úsanse principalmente dúas arquitecturas: as redes neurais recorrentes (RNN), que xurdiron primeiro, e os transformers, que son posteriores (Paes et al., 2023, p. 327). Neste segundo tipo de arquitectura máis moderna é no que nos imos centrar no presente traballo, xa que supera nalgúns aspectos o procesamento insuficiente das RNN. Os transformers válense do mecanismo de *self-attention* ou auto-atención, que produce embeddings contextualizados integrando información das palabras que están ao seu redor (Jurafsky e Martin, 2024, capítulo 10, p. 2). Isto favorece a aprendizaxe das relacións que se establecen entre unidades lingüísticas nos textos. Ademais, a arquitectura de transformers conta cun codificador, que procesa a secuencia de texto e codifícaa como un vector; e pode contar cun decodificador, que procesa ese vector e o transforma na secuencia de saída (Paes et al. 2023, p. 332). Os LMs de transformers son os empregados nos modelos de linguaxe grandes (*large language models* ou LLMs), como BERT ou GPT (Paes et al., 2023, p. 327).

4. Motivación e obxectivos do traballo

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, p. 26)[2], "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 presente traballo, que busca revelar se os LMs adquiren realmente unha lingua coma nós.

Concretamente, o obxectivo deste estudo é comparar o procesamento da concordancia suxeito-verbo en falantes nativos de galego e portugués e 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 á dependencia suxeito-verbo afectan á capacidade de resolución das persoas e das máquinas. En concreto, consideraremos a distancia entre o núcleo do suxeito e o verbo principal e a presenza ou ausencia dun distractor. Por exemplo, unha das oracións utilizadas no test é "O cabalo que ten novas *ferraduras* atravesa|*atravesan o campo ao galope", que inclúe o distractor *ferraduras* entre o suxeito (*cabalo*) e o verbo principal (*atravesa/atravesan*). Esta palabra, que difire en número coa alternativa gramatical (*atravesa*) mais concorda coa agramatical (*atravesan*), axuda a verificar se o número do verbo que os modelos recoñecen como

[2] Liña de pensamento que segue, en xeral, toda a lingüística chomskyana.

máis probable está directamente vinculado coa palabra inmediatamente anterior a el. Isto é, os distractores son chave para comprobar se os modelos presentan unha concepción lineal da oración ou se, en cambio, identifican as relacións xerárquicas que existen na lingua.

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, p. 15). 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, p. 532) 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 e Baroni, 2021, p. 4) e, polo tanto, se son quen de procesar a relación sintáctica entre este e o verbo principal.

Estado da arte

No seu artigo publicado no 2016, Linzen et al. buscan descubrir se as redes neurais LSTM (Long Short-Term Memory) poden identificar as relacións sintácticas de dependencia en inglés. A través da tarefa de predición de número (*number prediction task*), que consiste en propor unha oración que non inclúe a forma verbal e que o modelo adiviñe o seu número (ex.: "As *chaves* do armario **está*|*están* enriba da mesa"), avalían LSTMs supervisados e modelos de linguaxe xenéricos no recoñecemento da concordancia entre suxeito e verbo. Os resultados apuntan a que, mentres os LSTMs adestrados si que identifican estas estruturas, os LMs xenéricos necesitan dunha supervisión centrada nas relacións xerárquicas da lingua para poder aprender estas dependencias, xa que o seu procesamento non é suficiente para acadar o obxectivo da tarefa con éxito. Ademais, os distractores parecen confundir estes modelos, cuxo índice de erro aumenta canto maior é o número destes elementos na oración (Linzen et al., 2016, p. 528).

Un traballo posterior (Gulordava et al., 2018) retoma o estudo de Linzen et al. (2016), analizando ademais de inglés, italiano, hebreo e ruso, linguas morfoloxicamente máis complexas. Gulordava at al. inclúen algúns cambios nesta nova aproximación, como a avaliación doutras concordancias de número de contexto longo (refírese á distancia entre os dous elementos entre os que existe a relación; neste caso é considerada longa a partir de tres tokens), máis alá da de suxeito-verbo. Alén disto, tamén avalían datos sobre o procesamento destas estruturas en italiano por parte de humanos, que serven para comparar o comportamento dos modelos respecto ao dos falantes. Ademais, este estudo recoñece que factores como a información lexical, semántica ou a frecuencia dos tokens no corpus co que foi adestrado o modelo poden influír á hora de probar se as LSTMs teñen habilidades sintácticas. Así, as oracións que se usan neste experimento son gramaticais mais carecen de sentido (ex.: "The colorless green *ideas* I ate with the chair *sleep* furiously") (Gulordava et al., 2018, p. 1195). Neste caso, os modelos foron adestrados cun corpus controlado. Os resultados sinalan que os LMs procesan as relacións de concordancia de contexto longo incluso nas oracións sen sentido e en todas as linguas obxecto de estudo. Ademais, a diferenza do que se concluía en Linzen et al. (2016), os modelos LSTMs procesan as oracións medianamente ben a pesar da presenza de distractores, aínda que tanto humanos (falantes de italiano) como modelos mostran unha menor precisión canto maior é o número destes elementos.

En Marvin e Linzen (2018), búscase avaliar a gramaticalidade das predicións dos modelos de linguaxe en inglés, a través de datasets construídos de xeito controlado. Entre os fenómenos sintácticos incluídos neste estudo está a concordancia suxeito-verbo. Os

resultados da TSE (Targeted Syntactic Evaluation: test no que os LMs escollen entre as dúas alternativas para ocupar unha posición valeira na oración. Ex.: "O can * no parque"; "xoga" ou "xogan") mostran un non moi extraordinario rendemento. Tamén se recollen datos sobre a avaliación destes fenómenos por humanos, que superan a actuación dos modelos. O obxectivo desta comparación de base psicolingüística é confrontar os erros que cometen os LMs e as persoas, porque "se os erros son semellantes, os dous sistemas poden estar usando unha representación similar" (Marvin e Linzen, 2018, p. 1197).

Tras a introdución do modelo transformer BERT, Goldberg (2019) estuda se este LM identifica estruturas sintácticas en inglés, concretamente a concordancia suxeito-verbo. Para saber se recoñecen estas xerarquías, sérvese dalgúns dos estímulos usados por Linzen et al. (2016), Gulordava et al. (2018) e Marvin e Linzen (2018). Os resultados revelan un moi bo procesamento das relación de concordancia da sintaxe inglesa por parte do primeiro modelo de transformers BERT.

Mueller et al. (2020) van un paso máis alá e deciden levar a cabo un estudo para comprobar como esa habilidade dos modelos BERT para aprender sintaxe muda segundo a lingua. Así, avalíanse sintacticamente modelos de linguaxe LSTM e BERT multilingüe e monolingüe en inglés, francés, alemán, hebreo e ruso. Os resultados mostran que a precisión na identificación das relacións de concordancia é maior naquelas linguas morfoloxicamente máis ricas e que os modelos monolingües obteñen mellores resultados que os multilingües, revelando que non hai unha transferencia de información sintáctica dunha lingua ás outras. Ademais, BERT multilingüe presenta un elevado grao de precisión na avaliación destas estruturas en inglés, mais exhibe chamativas deficiencias nas outras linguas.

No traballo de Newman et al. (2021), proponse unha aproximación diferente á TSE das feitas ata ese momento en inglés para avaliar a sistematicidade do coñecemento sintáctico dos modelos de linguaxe neurais, así como as tendencias no seu comportamento. Os resultados obtidos evidencian que os modelos neurais son quen de completar as probas con éxito e que constitúen "potenciais modelos psicolingüísticos" (Newman et al., 2021, p. 3714), é dicir, que poden ser comparados con datos do procesamento sintáctico humano. Alén disto, tamén se mostra que a TSE sobreestima a sistematicidade dos modelos de linguaxe. Con todo, os LMs son un 40% máis precisos nos verbos que predín como máis probables no contexto.

Garcia e Crespo-Otero (2022) realizan unha avaliación sintáctica de modelos BERT monolingües e multilingües para galego-portugués. Con esta fin, crearon un dataset co que avaliaron o procesamento das concordancias de número (suxeito-verbo), de xénero (suxeito-atributo adxectivo) e persoa (suxeito-infinitivo flexionado) por parte dos modelos, controlando factores como a presenza de distractores e a distancia entre os elementos analizados. Os resultados mostran que estes modelos de transformers presentan un bo procesamento das estruturas sintácticas, mais a presenza de distractores en oracións con concordancias de contexto longo parecen confundilos. Ademais, no caso da concordancia de número, o distractor semella ter menor influencia nas oracións de contexto curto. Isto suxire que nestas "o modelo pode estar identificando a relación entre o suxeito e o verbo" (Garcia e Crespo-Otero, 2022, p. 52).

Nese mesmo ano, de-Dios-Flores e Garcia (2022) publican un traballo que verifica tamén a habilidade dos modelos de transformers BERT (monolingües e multilingües, de diferentes tamaños e adestrados con cantidades de datos e de vocabulario distintas) para identificar a concordancia suxeito-verbo e substantivo-adxectivo en galego. Nos tests comparárase a probabilidade que lle dan os modelos á alternativa correcta e á incorrecta en cada estrutura. Os resultados mostran que a precisión xeral dos modelos é máis alta nos casos de concordancia substantivo-adxectivo que nos de suxeito-verbo. Alén disto, o que parece exercer maior influencia no rendemento dos modelos testados é o tamaño do corpus de adestramento, sendo os resultados dos modelos Bertinho inferiores aos dos BERT, que foron adestrados con maior cantidade de textos (45 millóns no caso dos primeiros fronte os 550 millóns de tokens nos BERT).

Nestes dous estudos que teñen como lingua obxecto o galego e o portugués, os datasets non foron construídos de maneira moi sistemática e os resultados non foron contrastados con humanos. Así, este traballo busca contribuír coa elaboración controlada de datasets para a avaliación da concordancia suxeito-verbo, tendo en conta a frecuencia das unidades sintácticas valoradas nos modelos de linguaxe e a súa presenza no vocabulario dos modelos, así como a distancia entre os dous elementos analizados e a presenza de distractores. Ademais, confróntanse os resultados do procesamento das mesmas oracións por parte de modelos cos obtidos por falantes nativos desas linguas (ou con alto grao de competencia). Esta comparativa faise aplicando unha metodoloxía non usada anteriormente nos traballos para galego e portugués, a *surprisal*, que proporciona o valor aproximado de gramaticalidade que lle atribúen os modelos ás alternativas verbais para cada oración ("O cabalo que ten novas ferraduras atravesa|*atravesan o campo ao galope"). Deste xeito, alén de contribuír cun novo dataset construído de xeito sistemático, amplíase a investigación feita ata agora con modelos en galego e portugués (idiomas pouco estudados na área do PLN) empregando por vez primeira a metodoloxía *surprisal* nesta tarefa e comparando os datos de LMs sobre a concordancia suxeito-verbo nestas linguas cos de humanos.

Materiais e metodoloxía

A continuación, preséntanse os datos, modelos e experimentos usados neste traballo para contestar as preguntas de investigación expostas na introdución.

1. Materiais

Para alcanzar o noso obxectivo, comparar o procesamento da concordancia suxeito-verbo en falantes e en modelos de linguaxe neurais para galego e portugués, partimos de dous eixos fundamentais: un conxunto de datos (coñecido habitualmente polo seu nome en inglés: *dataset*) e os modelos que serán avaliados. Estes materiais son a base dos experimentos que se levan a cabo neste traballo e por iso o seu proceso de construción e selección é esencial para o bo desenvolvemento da pesquisa.

1.1. Datasets

Un dataset é "unha colección de diferentes conxuntos de información que son tratados como unha única unidade por un ordenador"[3]. Neste caso, construímos en Excel dous datasets, un en galego e outro en portugués, con 16 oracións. Cada unha conta con 8 variantes creadas tendo en conta os seguintes parámetros: a distancia entre o suxeito e o verbo principal (contexto), a presenza dun distractor e a gramaticalidade da oración (véxase o exemplo na Táboa 1). Os suxeitos foron escollidos de forma sistemática e balanceada dende o punto de vista morfolóxico: 4 oracións con substantivo masculino singular, 4 con substantivo masculino plural, 4 con substantivo feminino singular e 4 con substantivo feminino plural (16 oracións totais). Ademais, estes teñen trazos semánticos distintos, como [+ animal], [- concreto], [+ humano] e [- animado]. Así, hai diversidade semántica nos suxeitos analizados, aspecto que non se recolle nos anteriores traballos, onde todos eran [+humanos] (*vid.* Garcia e Crespo-Otero, 2022; e de-Dios-Flores e Garcia, 2022). En total, cada dataset está conformado por 128 oracións, que serán as avaliadas nos experimentos con humanos e LMs[4].

A pesar de compartir todas estas características, os dous datasets non son resultado dun proceso de tradución directa galego-portugués ou viceversa. Aínda que si se tentou conservar a equivalencia semántica das oracións de ambos os dous, isto non foi sempre

[3] Cambridge Dictionary, consultado o 26 de maio de 2024 https://dictionary.cambridge.org/dictionary/english/dataset

[4] Os datasets están no apartado "Anexos".

compatible co mantemento da estrutura necesaria para avaliar a concordancia suxeito-verbo.

Con.	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.

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

Un punto chave na construción dos datasets foi a selección das palabras para os suxeitos e os verbos principais. No dataset galego, estas foron extraídas dos vocabularios dos modelos monolingües Bertinho-base e BERT-gl-base, mentres que para o dataset portugués, foron tomadas do vocabulario do modelo BERTimbau-base (información ampliada na sección 1.2). Concretamente, escolléronse palabras cunha frecuencia maior a 0,5 por millón nun corpus de referencia, conformado pola Wikipedia, WikiBooks e un corpus de blogs baixados de internet. Evitáronse deste xeito aquelas máis frecuentes, porque potencialmente os modelos resolverían mellor estes casos e polo tanto os resultados de aceptabilidade non serían representativos da súa capacidade xeral de procesamento da dependencia sintáctica suxeito-verbo.

Un dos factores que se tiveron en conta á hora da elaboración das oracións foi a distancia entre o núcleo do suxeito e o verbo principal, ao que nos referiremos de aquí en diante como *contexto*. Diferéncianse dous tipos: o contexto curto (*short agreement dependency*) e o contexto longo (*long agreement dependency*). O primeiro é aquel no que entre os dous elementos analizados non hai máis de 5 palabras (ex.: “A nena cos zapatos vellos xogou co veciño.”). O segundo é o máis complexo e o que supón un maior nivel de dificultade para os modelos (*vid*. “Estado da arte”). Neste caso, a distancia entre suxeito e verbo é de entre 8 e 10 palabras (ex.: “A nena que acababa de chegar á vila cos seus pais xogou co veciño.”).

Por outro lado, a presenza ou ausencia dun distractor é considerado outro factor a ter en conta na avaliación do procesamento da dependencia suxeito-verbo, xa que traballos anteriores mostran que estes elementos poden confundir tanto humanos (Bock e Miller, 1991) como os modelos de linguaxe (Linzen et al., 2016; Gulordava et al., 2018; Garcia e Crespo-Otero, 2022). 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. Malia que se intentou evitar no máximo a presenza destas no deseño dos datasets, non sempre se conseguiu este obxectivo (ex.: "O cabalo que *sacode* a *crina* da *cor* das *amoras* atravesa o campo ao galope."). A análise destes casos pode ser unha potencial liña de investigación para traballos futuros.

Ademais, como se pode observar na Táboa 1, a gramaticalidade é o terceiro factor incluído na construción dos datasets. Así, para cada combinación contexto longo-con distractor, contexto longo-sen distractor, contexto curto-con distractor e contexto curto-sen distractor en cada oración hai dúas variantes (*vid*. Táboa 2): unha gramatical (concordancia suxeito-verbo) e outra agramatical (sen concordancia do suxeito e o verbo).

Contexto	Distractor	Gramaticalidade	Oración
Curto	Con	Gramatical	Eses paxaros de peteiro escuro fican á beira do estanque.
Curto	Con	Agramatical	Eses paxaros de peteiro escuro fica á beira do estanque.

Táboa 2. Parte do set de variantes da oración 8 do dataset para galego.

1.2. Modelos

Neste traballo avaliamos diferentes modelos de linguaxe de transformers. Para galego escollemos Bertinho-base e BERT-gl-base, mentres que para portugués eliximos BERTimbau-base[5]. Para todos os casos, seleccionouse a versión *base*, é dicir, de 12 capas, dos respectivos modelos, para que deste xeito os resultados dos experimentos sexan comparables.

O modelo Bertinho-base (Vilares et al., 2021) é un LM monolingüe en galego. Está preadestrado cun corpus extraído da versión galega da Wikipedia que conta con aproximadamente 45 millóns de palabras. O seu vocabulario é de 30.000 tokens.

O BERT-gl-base é un modelo monolingüe preadestrado para galego cun corpus de 500 millóns de palabras (Garcia, 2021). O seu vocabulario está conformado por 119.547 tokens (Garcia e Crespo-Otero, 2022).

O modelo avaliado para o portugués é BERTimbau-base. Este LM monolingüe foi adestrado con 2,68 billóns de tokens do *BrWaC* (*Brazilian Web as Corpus)*. Ten un vocabulario de 30.000 tokens (Souza et al., 2020).

2. Metodoloxía

Neste traballo realízanse tres experimentos, partindo todos eles da adaptación dos datasets deseñados. O primeiro, procedemento baseado en enquisas, busca comprobar o comportamento das persoas fronte a concordancia de suxeito e verbo; mentres que os outros dous, a TSE e a avaliación usando a *surprisal* (*vid*. sección 2.2.1) dos modelos,

[5] Cómpre ter en conta que BERTimbau-base foi adestrado na súa maioría con textos en portugués do Brasil, mais consideramos que isto non afecta a concordancia suxeito-verbo. Exploramos outros LMs da variedade de portugués europeo para o traballo, pero non puidemos avalialos xa que teñen como vocabulario base o inglés.

teñen como obxectivo avaliar o procesamento desta relación sintáctica nos modelos de linguaxe neurais. Os resultados obtidos de LMs e dos humanos serán posteriormente comparados.

2.1. Compilación de datos de falantes

O método escollido para avaliar as persoas foi o das enquisas. Elaboráronse 8 para falantes de galego (GL) e 8 para falantes de portugués (PT) coa aplicación Microsoft Forms. Cada formulario inclúe unha variante de cada oración do dataset. Por exemplo, a primeira construción lingüística a avaliar no cuestionario 1 de portugués é a variante gramatical de contexto longo e con distractor da primeira oración do dataset: "O cavalo que sacode a crina da cor das amoras está no campo". Isto permítenos presentar ante os suxeitos un número suficientemente reducido de cuestións como para que a súa atención na proba non diminúa considerablemente e afecte os resultados. Cómpre salientar que a distribución das variantes foi feita sistematicamente para que os participantes non poidan identificar un padrón nas características das oracións presentadas. Cada unha destas está seguida dunha escala likert do 1 ao 5 para puntuar cada oración segundo a súa aceptabilidade (Langsford, 2018, p. 8). Este sistema foi escollido polo seu bo rendemento incluso en experimentos con mostras de tamaño reducido (Langsford, 2018, p. 1).

Cada enquisa está precedida por unha breve introdución explicativa. Nela exponse de forma superficial e informal o tema do traballo para o que se recollen os datos, de xeito que este non condicione a actitude do participante perante as cuestións: "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". Respecto ás instrucións que se dan para completar a enquisa, anímase a puntuar as oracións segundo "che pareza que están ben formadas ou non", evitando empregar termos como *gramaticais* ou *correctas* para que as respostas non se vexan condicionadas por unha asociación co prescritivismo ortográfico e gramatical do sistema educativo. Así, escolleuse a expresión "ben formadas", empregada en "Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies" (De-Dios-Flores et al., 2023) e en "Acceptability Ratings in Linguistics: A Practical Guide to Grammaticality Judgments, Data Collection, and Statistical Analysis" (Bross, 2019).

Ademais, precedendo as outras, inclúense dúas preguntas en cada enquisa para obter datos sobre o coñecemento lingüístico dos participantes: "Es falante nativa/o de galego?" e "En caso de que non sexas falante nativa/o, como cualificarías o teu nivel de galego?" (para galego); e "É falante nativo de português?" e "Em caso de não ser falante nativo, como qualificaria o seu nível de português?" para (portugués).

2.2. Compilación de datos de modelos de linguaxe

2.2.1. Surprisal

A metodoloxía de avaliación *surprisal* ("sorpresa") consiste na adaptación dos valores da probabilidade que ten unha palabra de ocorrer nunha oración segundo un modelo

de linguaxe, tendo en conta o contexto (Rezaii et al., 2023, p. 647). O valor de *surprisal* dunha palabra dada obtense mediante o logaritmo da probabilidade inversa desa palabra no contexto oracional. Os seus valores téñense relacionado cos tempos de lectura e o procesamento lingüístico (Hale, 2001). Canto máis baixo sexa o valor de *surprisal*, menos "sorprendido" está o LM pola presenza da palabra nese contexto e polo tanto considéraa máis probable. Para obter a *token surprisal*[6] ou probabilidade da palabra target no contexto, usamos a librería minicons[7], que "proporciona unha API estándar para investigadores interesados en levar a cabo análises conductuais e representacionais de modelos de linguaxe de transformers" (Misra, 2022, p.1).

Os datos de entrada ou *input* para este experimento consisten en pares de oracións completas (sen ningunha posición oculta): unha variante gramatical e outra agramatical. Así, a adaptación dos datasets para *surprisal* é menos complexa que a que levamos a cabo para a TSE. Neste proceso, creamos un arquivo CSV e incluímos o número da oración (1-16), as variantes de cada unha e, nunha terceira columna, a gramaticalidade (right ou gramatical e wrong ou agramatical). Deste xeito, o dataset resultante segue este modelo:

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
1	O cabalo que sacode a crina da cor do carbón intensamente atravesa o campo ao galope.	Right
1	O cabalo que sacode a crina da cor do carbón intensamente atravesan o campo ao galope.	Wrong
1	O cabalo que ten novas ferraduras atravesa o campo ao galope.	Right
1	O cabalo que ten novas ferraduras atravesan o campo ao galope.	Wrong
1	O cabalo desparasitado recentemente atravesa o campo ao galope.	Right
1	O cabalo desparasitado recentemente atravesan o campo ao galope.	Wrong

Táboa 4. Adaptación das variantes da oración 1 do dataset (GL) para *surprisal*.

Deseguido, executamos o script ou conxunto de comandos para calcular a *token surprisal* con Google Colab, que lanzaremos tres veces avaliando a Bertinho-base e BERT-gl-base co dataset para galego adaptado e o modelo BERTimbau co portugués. O primeiro paso do script consiste en instalar minicons e despois escoller o modelo de linguaxe a avaliar. A continuación, abrimos os datasets adaptados para TSE e *surprisal* na lingua do LM. O obxectivo deste paso é verificar se ambos os dous datasets usados nos experimentos deste traballo son idénticos, dando a oportunidade de corrixir posibles erros humanos no proceso de adaptación dos datos. Por último, calcúlase a *surprisal* e gárdanse os resultados nun ficheiro CSV.

[6] Tamén calculamos a *sentence surprisal*, que consiste na combinación da probabilidade de todas as palabras da oración para verificar se é máis probable neste caso a oración gramatical (concordancia suxeito-verbo) ou a agramatical (non concordancia suxeito-verbo), mais desbotamos este método porque difumina os resultados.

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

Unha vez que se obteñen os resultados do experimento, estes convértense á escala likert que se emprega nas enquisas. Deste xeito, os datos do procesamento da concordancia suxeito-verbo dos modelos poden ser confrontados cos dos humanos.

2.2.2. TSE

A TSE ou *Targeted Syntactic Evaluation* é un método que avalía o coñecemento sintáctico dun modelo de linguaxe usando unha oración cun elemento oculto para o que se ofrecen dúas opcións (Newman et al., 2021). Segundo a probabilidade que lle dea o LM á alternativa gramatical e á agramatical no contexto, obtense a súa precisión no procesamento do fenómeno sintáctico analizado. Neste caso, avaliaremos a concordancia suxeito-verbo en modelos para o galego e o portugués.

Para realizar este experimento, adaptamos os datasets (GL e PT) a un formato compatible coa TSE. Por unha parte, as variantes de cada oración redúcense á metade, porque a gramaticalidade deixa de ser un factor a considerar. Isto débese a que agora se oculta a posición do verbo (posición *masked*) e preséntanse como alternativas a forma verbal gramatical que concorda en número co suxeito e a agramatical que non concorda. Por exemplo, as variantes da oración 4 dispóñense do seguinte xeito para a TSE:

Or.	TSE: variante	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 este venres no litoral atlántico fortemente * que se superasen os 30 graos.	fixo	fixeron
4	O sol que luciu nas praias * que se superasen os 30 graos.	fixo	fixeron
4	O sol que luciu intensamente * que se superasen os 30 graos.	fixo	fixeron

Táboa 3. Adaptación das variantes da oración 4 do dataset (GL) para a TSE, onde "*" é unha posición oculta na cal se obteñen as probabilidade das dúas alternativas (gramatical e agramatical).

Despois, os datasets resultantes destas modificacións convértense en arquivos CSV, formato compatible co seguinte paso do experimento.

A continuación, empregamos un script ou conxunto de comandos para calcular a TSE. Executámolo con Google Colab testeando o dataset adaptado en galego con Bertinho-base e BERT-gl-base e o dataset adaptado en portugués co modelo BERTimbau. En primeiro lugar, cárgase o modelo de linguaxe que se quere avaliar e o documento CSV. Deseguido, tokenizamos as oracións e obtemos o índice da posición oculta e as predicións do modelo para cada unha destas. Por último, imprimimos nun ficheiro de saída o *output* ou saída do script. Aí atopamos o resultado da TSE: se o LM calcula que a alternativa correcta para unha oración é máis probable que a incorrecta, no ficheiro de saída aparece un *1*; mais se é o verbo incorrecto nese contexto o que o LM prevé como máis probable imprimirase un *0*. Para calcular o desempeño de cada modelo, usamos a métrica "precisión", obtida dividindo o número de oracións nas que acerta entre o total de oracións avaliadas.

Experimentos e resultados

1. Experimentos

Para a recolección de datos de humanos sobre o procesamento da concordancia suxeito-verbo, empregamos o método das enquisas (*vid*. "Materiais e metodoloxía", apartado 2.1), que foron distribuídas entre falantes nativos e persoas cunha alta competencia e desempeño na lingua obxecto de estudo. En total, participaron 93 individuos: 76 nas enquisas para galego e 17 para as de portugués. Nas primeiras (GL) hai ao menos tres valoracións de diferentes persoas por oración, mentres que para as segundas (PT) hai como mínimo unha resposta por oración. Respecto á autopercepción lingüística dos falantes non nativos participantes, os que cubriron a enquisa para o portugués puntuaron o seu nivel na lingua cun 4,87 de media sobre 5 e os que contestaron a de galego cualificáronse cunha media de 3,95 sobre 5, que debe ser considerada tendo en conta a situación sociolingüística do idioma.

Respecto aos modelos, levamos a cabo dous experimentos, *surprisal* e TSE, nos que avaliamos o procesamento da dependencia suxeito-verbo nos LMs Bertinho-base e BERT-gl-base para galego e BERTimbau para portugués (vid. "Materiais e metodoloxía", apartado 2.2). Os resultados de *surprisal* foron convertidos a escala likert (1-5) para a súa comparación cos datos recollidos de falantes.

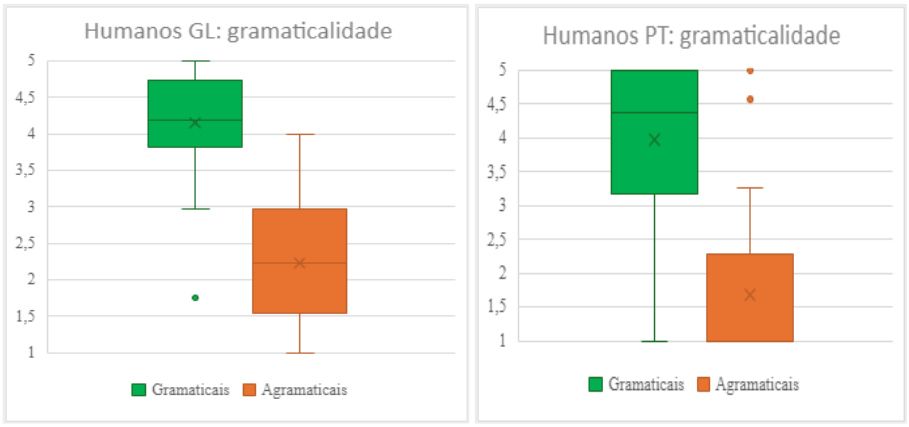
2. Resultados

Neste apartado, preséntanse os resultados tanto dos falantes como dos modelos de linguaxe. Concretamente, cando os comentamos usamos os valores medios obtidos, facendo referencia ao desvío padrón cando se considera necesario. Así, deixamos para un traballo futuro a realización de análises estatísticas que nos permitan coñecer a significación de cada subconxunto de resultados no procesamento da concordancia suxeito-verbo en LMs e humanos.

2.1. Compilación de datos de falantes: enquisas

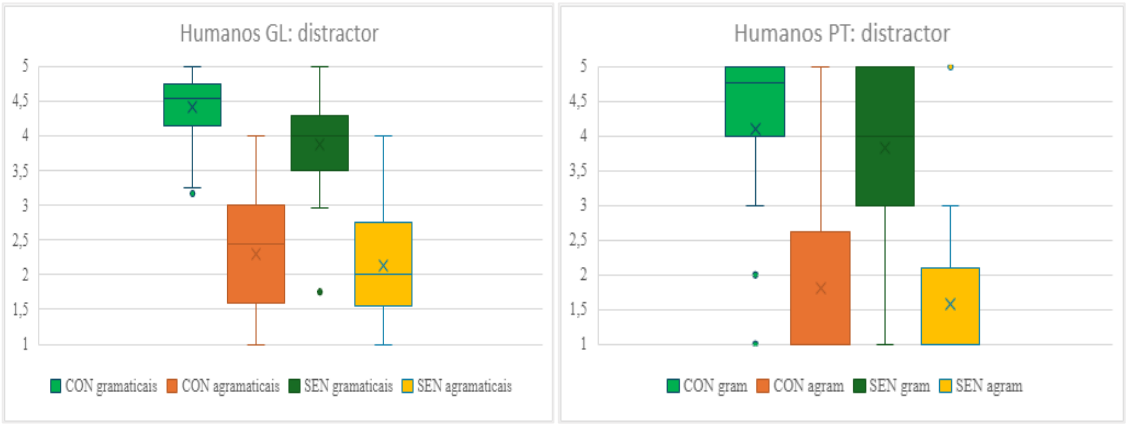
A continuación, expóñense os resultados das enquisas feitas con humanos segundo os diferentes factores considerados para a avaliación do procesamento da concordancia suxeito-verbo.

Gramaticalidade:



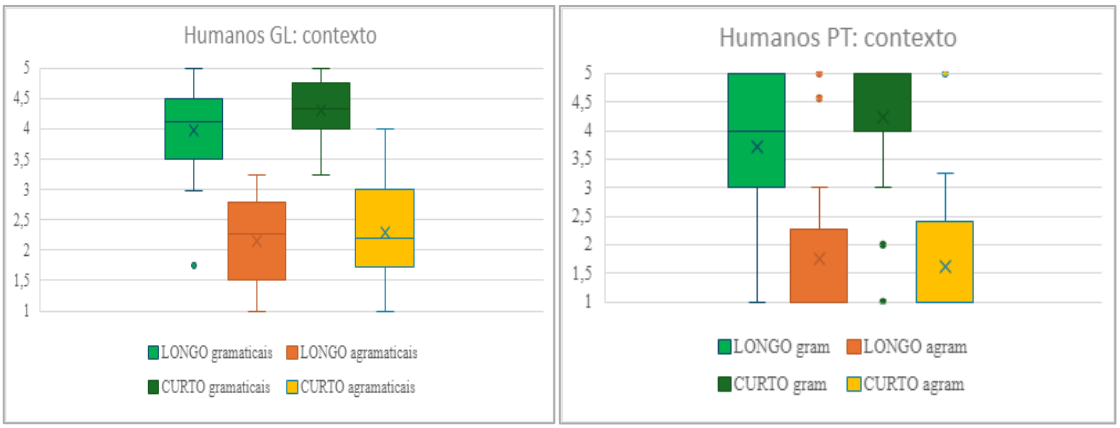
Os resultados mostran que os falantes dan ás oracións gramaticais (concordancia suxeito-verbo) unha puntuación media de 4,14 en galego e 3,97 en portugués, mentres que as agramaticais cualifícanas cun 2,22 e un 1,69 respectivamente segundo a súa gramaticalidade na escala likert (1-5). Aínda que existe variación nas valoracións outorgadas dentro do conxunto das oracións gramaticais e agramaticais (desvío padrón de 0,63 [GL] e 1,28 [PT], e 0,82 [GL] e 1,00 [PT] respectivamente), os humanos dan puntuacións ben diferenciadas ás correctas e ás incorrectas.

Presenza de distractor:



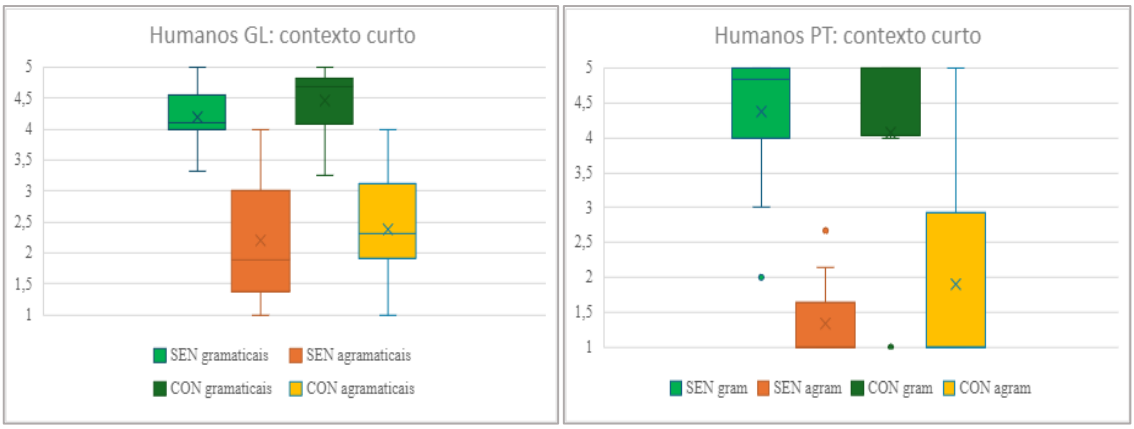
As oracións gramaticais con distractor obtiveron en ambas as dúas linguas puntuacións máis altas (cunha media de 4,42 en galego e 4,11 en portugués), que as gramaticais sen distractor (cunha media de 3,87 en galego e 3,82 en portugués). Aquelas agramaticais con distractor tamén foron valoradas de media 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:



Por outra banda, as oración gramaticais de contexto curto (4,32 GL e 4,22 PT) obteñen unha valoración media máis alta que as de contexto longo (3,97 GL e 3,71 PT). Os resultados das agramaticais en galego seguen a mesma tendencia: as de contexto longo teñen unha puntuación media de 2,15 e as de contexto curto de 2,29. Nas agramaticais en portugués ocorre o contrario, as de contexto longo teñen puntuacións máis altas que as de contexto curto: 1,76 fronte a 1,62.

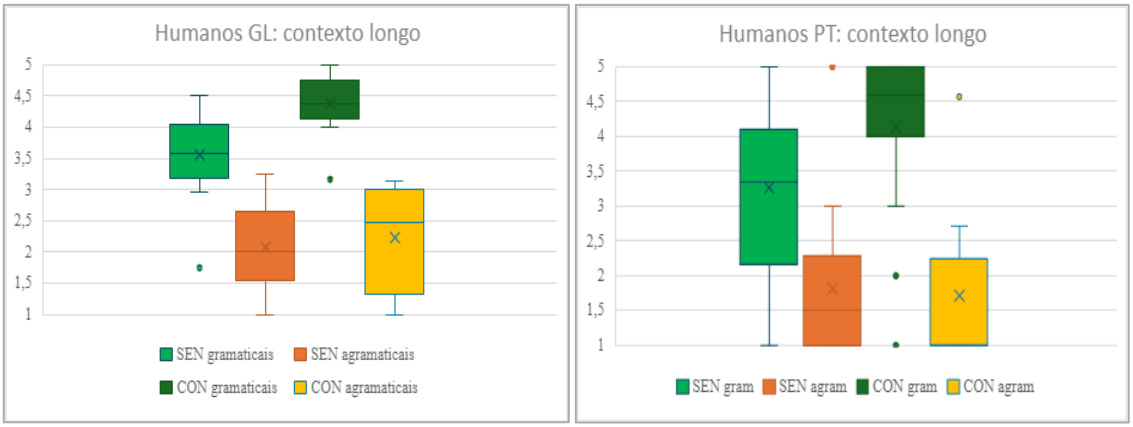
Contexto curto:



Nas enquisas en galego, as oracións gramaticais con distractor obteñen unha valoración media máis alta (4,45) que as que non o teñen (4,18), mentres que as agramaticais sen distractor teñen menor aceptabilidade (2,20) que as con distractor (2,38).

Os resultados das enquisas para portugués mostran que as oracións curtas agramaticais sen distractor son valoradas segundo os falantes como menos aceptables (puntuación media de 1,33) que as con distractor (1,90 de media). É salientable a ampla variación que se recolle nas valoracións das oracións agramaticais con distractor: desvío padrón de 1,51 nas oración con este elemento, fronte ao 0,84 das que non o teñen. Doutra banda, ás oración curtas gramaticais sen distractor atribúeselles unha puntuación máis alta de media (4,38) que ás con distractor (4,07).

Contexto longo:



En ambas as dúas linguas, os falantes valoran como máis aceptables aquelas oracións gramaticais de contexto longo que teñen distractor (4,38 GL e 4,15 PT), fronte ás que non teñen distractor (3,56 GL e 3,27 PT). Respecto ás agramaticais de contexto longo, en galego as sen distractor teñen de media unha valoración máis baixa (2,08) que as con distractor (2,23); mentres que en portugués se recolle o fenómeno contrario (1,81 fronte a 1,70).

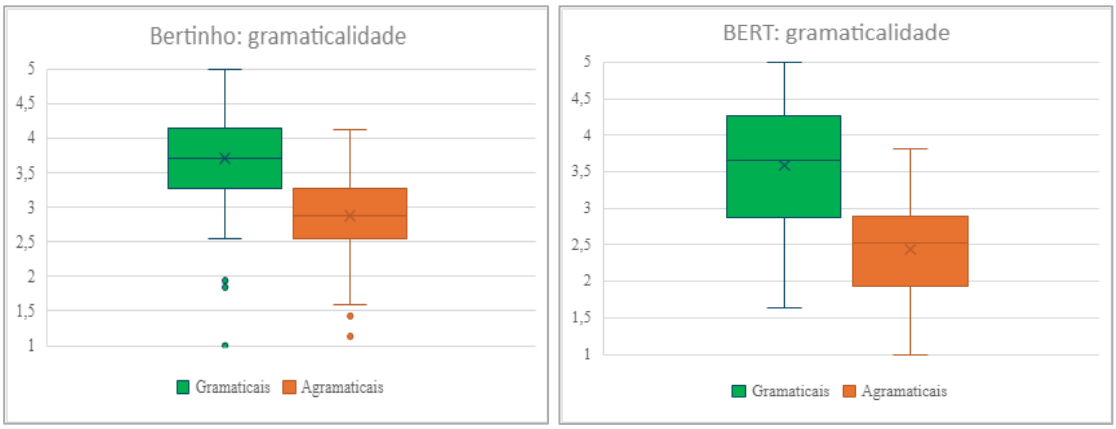
2.2. Compilación de datos de modelos de linguaxe

2.2.1. Surprisal

A continuación, expóñense os resultados de *surprisal*, metodoloxía empregada para a avaliación do procesamento da concordancia suxeito-verbo nos modelos de linguaxe. Os datos preséntanse ordenados segundo os factores considerados no deseño dos datasets.

2.2.1.1. Galego

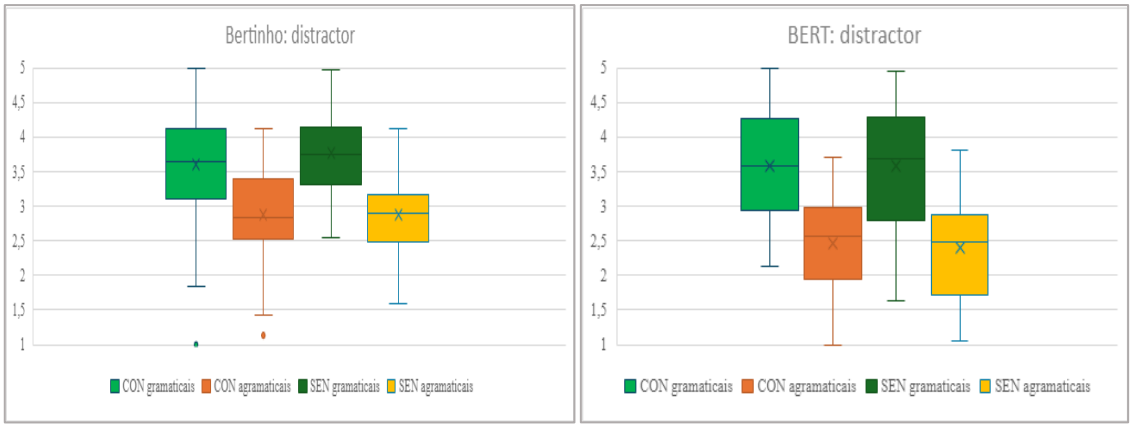
Gramaticalidade:



Os resultados mostran que ambos os dous modelos empregados para o galego dan ás oracións gramaticais (concordancia suxeito-verbo) unha aceptabilidade media máis alta que as agramaticais: 3,70 de Bertinho-base e 3,59 de BERT-gl-base, fronte a 2,87 en

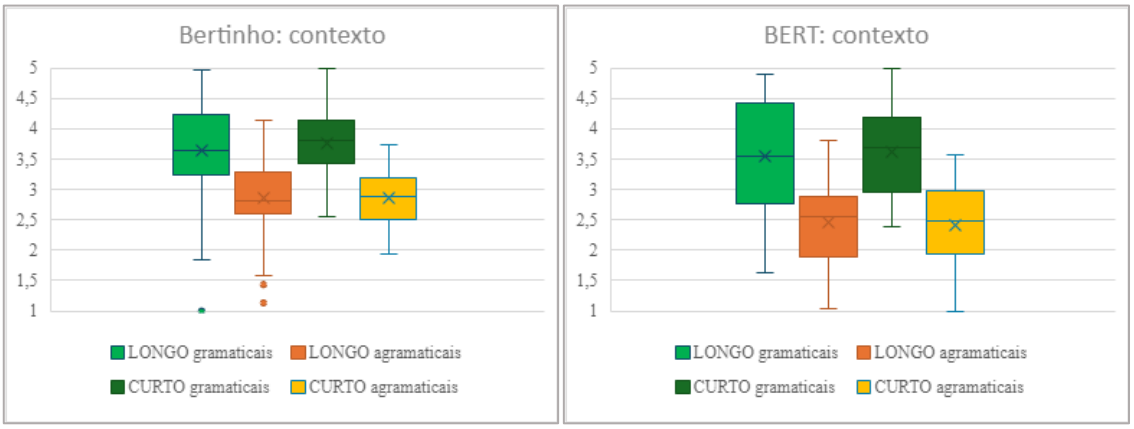
Bertinho-base e 2,43 en BERT-gl-base. Alén disto, os resultados de procesamento desta dependencia en BERT mostran unha diferenciación máis marcada entre as oracións non gramaticais e as gramaticais.

Presenza de distractor:



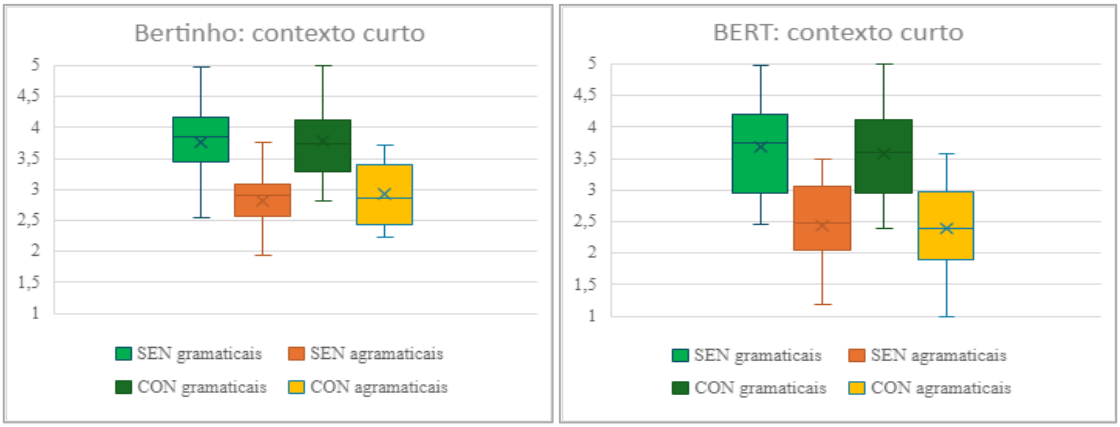
Mentres que nos resultados de Bertinho-base as oracións gramaticais sen distractor (3,78) foron valoradas como lixeiramente máis aceptables que as con distractor (3,63), nos resultados de BERT-gl-base esa diferenciación é ínfima (3,592 sen distractor, fronte con distractor 3,587). As agramaticais con distractor son máis aceptables segundo BERT-gl-base (2,70) que as sen distractor (2,40), mentres que a diferenza nos resultados de Bertinho-base é ínfima (2,873 fronte 2,866 de media respectivamente).

Contexto:



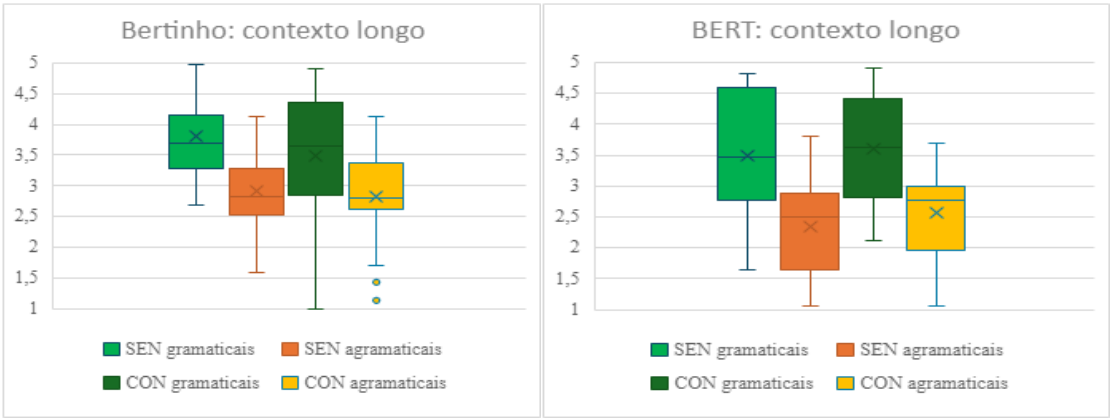
Por outra banda, as oración gramaticais de contexto curto (Bertinho-base 3,76 e BERT-gl-base 3,63) obteñen unha valoración media máis alta que as de contexto longo (Bertinho-base 3,64 e BERT-gl-base 3,55). En ambos os modelos, as agramaticais de contexto curto son identificadas como lixeiramente menos aceptables que as de contexto longo: 2,869 (Bertinho-base) e 2,41 (BERT-gl-base), fronte a 2,870 (Bertinho-base) e 2,45 (BERT-gl-base) de media.

Contexto curto:



Nas oracións gramaticais de contexto curto, Bertinho-base valora cun maior nivel de aceptabilidade aquelas con distractor (3,78) que as sen distractor (3,74), mentres que os resultados de BERT-gl-base apuntan como máis aceptables aquelas sen distractor (3,69 fronte 3,57). A mesma tendencia se manifesta nas oracións agramaticais de contexto curto: en Bertinho-base 2,82 de media (sen distractor) fronte 2,91 de media (con distractor), e en BERT-gl-base 2,44 de media (sen distractor) fronte 2,38 de media (con distractor).

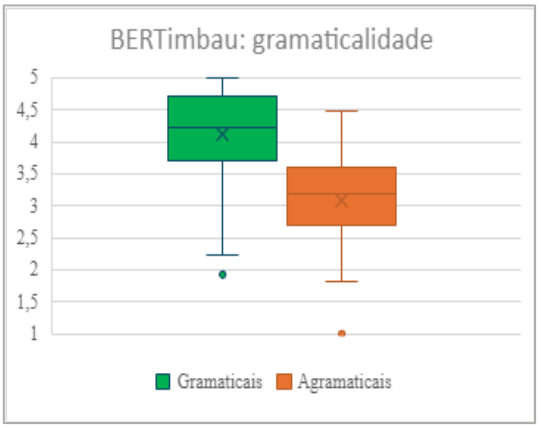
Contexto longo:



Nas oracións gramaticais de contexto longo, Bertinho-base valora cun maior nivel de aceptabilidade aquelas sen distractor (3,80) que as con distractor (3,48), mentres que os resultados de BERT-gl-base apuntan como máis aceptables aquelas con distractor (3,60 fronte o 3,50 de media das sen distractor). A mesma tendencia se manifesta nas oracións agramaticais de contexto curto: en Bertinho-base 2,90 de media (sen distractor) fronte 2,83 de media (con distractor), e en BERT-gl-base 2,35 de media (sen distractor) fronte 2,56 de media (con distractor).

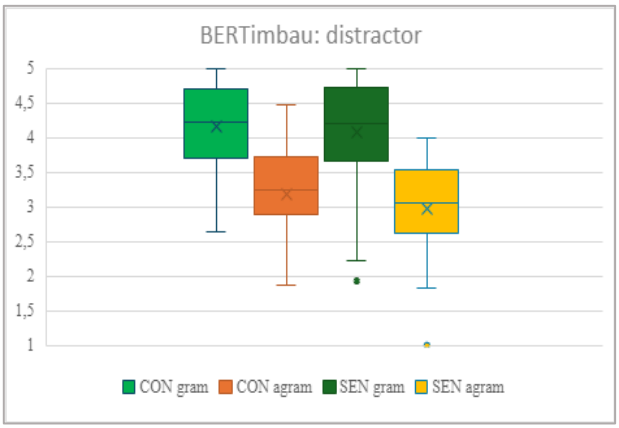
2.2.1.2. Portugués

Gramaticalidade:



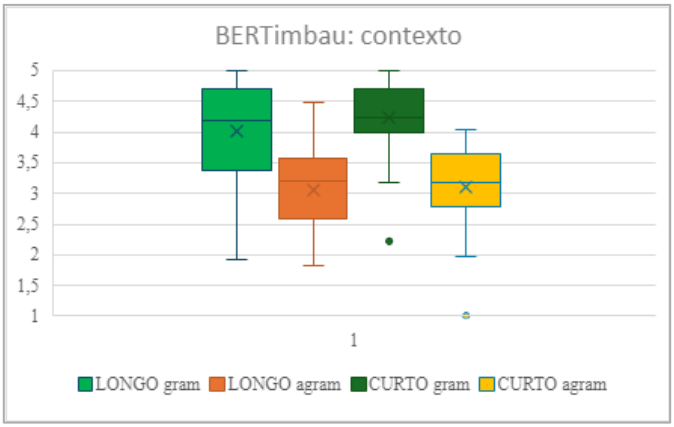
Os resultados de BERTimbau mostran unha diferenciación entre as oracións gramaticais e agramaticais: as primeiras valóraas de media cun 4,12 sobre 5, mentres que ás agramaticais atribúelle un 3,08 de media. O índice de aceptabilidade é, polo tanto, maior nas gramaticais.

Presenza de distractor:



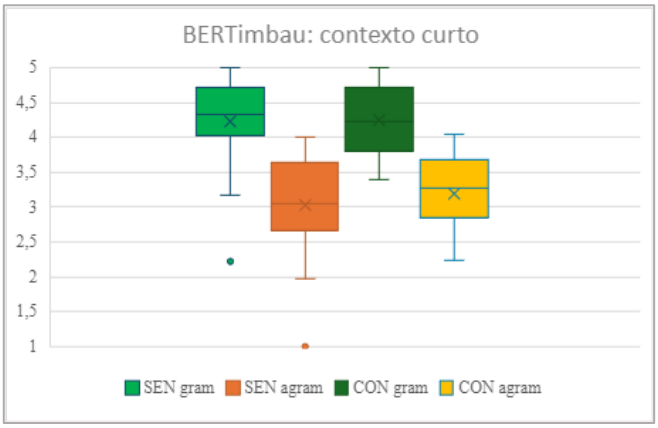
Este modelo considera lixeiramente máis aceptables de media as oración gramaticais con distractor (4,16) fronte ás sen distractor (4,08). As agramaticais sen distractor (2,96 de media) valóraas como menos aceptables que as con distractor (3,17 de media).

Contexto:



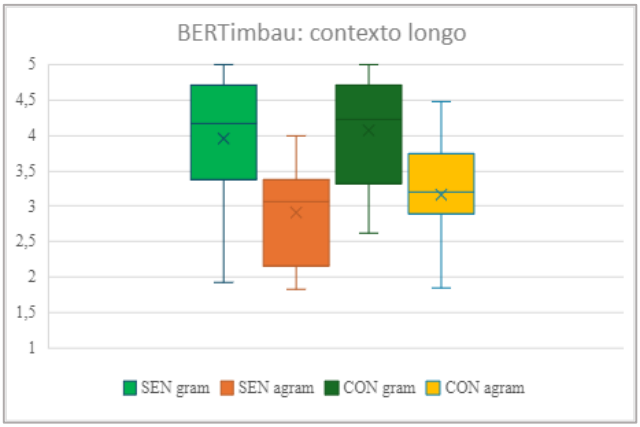
Segundo o contexto, BERTimbau-base valora as oracións gramaticais de contexto curto como máis aceptables (4,23 de media) que as de contexto longo (4,02 de media). Seguindo a mesma tendencia, as agramaticais de contexto longo (3,06 de media) son avaliadas como menos aceptables que as de contexto curto (3,11 de media).

Contexto curto:



Nas oracións de contexto curto, aquelas gramaticais con distractor obteñen de media unha valoración lixeiramente máis alta (4,25) pero rexístrase unha maior variación nas avaliacións, fronte as oracións que non teñen distractor (4,22). As agramaticais con distractor son consideradas máis aceptables (3,19 de media) que as sen distractor (3,03 de media).

Contexto longo:



Nas oracións de contexto longo, aquelas con distractor obteñen de media unha valoración lixeiramente máis alta (4,08), fronte as que non teñen este elemento (3,95). As agramaticais con distractor son consideradas máis aceptables (3,17 de media) que as sen distractor (2,92 de media).

2.2.2. TSE

A continuación expóñense os resultados de TSE, metodoloxía empregada para a avaliación da precisión no procesamento da concordancia suxeito-verbo nos modelos de linguaxe. Os datos preséntanse ordenados segundo os factores considerados no deseño das oracións. Cómpre apuntar que, malia que se mostran todos na mesma táboa, os datos obtidos por Bertinho-base e BERT-gl-base non son directamente comparables cos de BERTimbau-base, porque os datasets empregados son diferentes.

Variables		**Precisión Bertinho-base**	**Precisión BERT-gl-base**	**Precisión BERTimbau-base**
Gramaticalidade (precisión xeral)		0,34	0, 66	0,73
Contexto	Longo	0,31	0,50	0, 66
	Curto	0,38	0,81	0, 81
Distractor	Con	0,31	0,63	0,59
	Sen	0,38	0,69	0,88
Contexto curto	Con	0,31	0,81	0,69
	Sen	0,44	0,81	0,94
Contexto longo	Con	0,31	0,44	0,50
	Sen	0,31	0,56	0,81

Por un lado , a precisión de Bertinho-base é moi baixa. A súa precisión xeral é de 0,34. En ningún dos casos a súa precisión se aproxima ao 0,50, é dicir, non alcanza os resultados que por probabilidade se obterían no caso de facer o experimento de maneira aleatoria. O índice de precisión máis alto que obtén é naquelas oracións de contexto curto sen distractor (0,44).

Doutra banda, BERT-gl-base presenta unha precisión xeral de 0,66 sobre 1. Mostra unha clara dificultade procesando a gramaticalidade naquelas oracións de contexto longo, con só unha precisión de 0,50. Naquelas de contexto curto, en cambio, obtén bos resultados (0,81). Ademais, nas oracións con distractor a precisión decae lixeiramente (0,63) en comparación con aquelas que non teñen este elemento (0,69). Nas oracións de contexto curto a presenza do distractor non afecta a precisión do modelo, sendo esta nos casos con e sen distractor de 0,81. En cambio, nas oracións de contexto longo a precisión de BERT-gl-base descende considerablemente ata o 0,56 naquelas sen distractor e ata unha precisión de 0,44 nas que si hai distractor.

O modelo BERTimbau-base é o que mostra a máis alta precisión total de todos os LMs avaliados neste traballo. Este procesa notablemente mellor aquelas oracións de contexto curto (precisión de 0,81), fronte as de contexto longo (0,66). O distractor tamén inflúe negativamente nos resultados, obtendo un 0,59 nas oracións con este elemento, fronte ao 0,88 das que non o teñen. Ademais, a precisión nas oracións de contexto curto sen distractor obteñen a precisión máis alta de todo o dataset (0,94), mentres que as que si teñen distractor presentan un 0,69. BERTimbau-base mostra tamén un bo procesamento das oracións de contexto longo sen distractor, cunha precisión de 0,81, que cae ata 0,50 no caso daquelas nas que si hai distractor.

Discusión

Este capítulo comprende unha análise e discusión dos resultados obtidos nos tres experimentos realizados no traballo: enquisas, avaliación *surprisal* e TSE (*vid*. "Experimentos e resultados"). Primeiramente, expóñense as conclusións que se extraen dos datos dos falantes sobre o seu procesamento da concordancia de número de suxeito e verbo conseguidos coas enquisas. A continuación, analízanse os datos dos modelos de linguaxe neurais obtidos na avaliación *surprisal* e na TSE sobre o mesmo fenómeno sintáctico. Por último, confróntanse os resultados de humanos e modelos.

1. Resultados de humanos

Os resultados das enquisas confirman a hipótese exposta na "Introdución" deste traballo: as persoas identifican con facilidade a concordancia suxeito-verbo. Os datos presentados na sección 2.1 do capítulo "Experimentos e resultados" indican que tanto os falantes de galego como os de portugués que participaron no experimento danlle ás oración gramaticais e agramaticais (aquelas nas que o suxeito e o verbo non concordan en número) puntuacións ben diferenciadas, o que indica que son quen de identificar cando este principio sintáctico fundamental de ambas as dúas linguas é vulnerado.

Ademais, as persoas parecen procesar mellor as oracións de contexto curto que as de contexto longo. Os falantes de ambas as dúas linguas valoran con puntuacións de media máis altas as oración gramaticais de contexto curto que as de contexto longo, mentres que ás oracións agramaticais de contexto curto lles dan unha puntuación máis baixa de media que as de contexto longo. Así, confírmase que a distancia entre o núcleo do suxeito e o verbo principal condiciona o procesamento da concordancia suxeito-verbo en humanos, mais non coarta a capacidade das persoas para identificar a aceptabilidade das oracións.

Doutra banda, as oracións gramaticais con distractor obtiveron resultados máis altos na escala likert que aquelas sen este elemento, tanto nas oracións de contexto curto como nas de contexto longo. Estes sorprendentes datos, que coliden coa finalidade do distractor de confundir ao lector, lévannos a formular a seguinte hipótese: a presenza do distractor determina a atención dos falantes, así estes len as oracións que conteñen este elemento máis atentamente. Unha futura liña de traballo constitúea a exploración desta hipótese empregando técnicas como o *eye-tracking* e a análise dos tempos de lectura. As oracións agramaticais con distractor tamén recibiron valoracións máis altas que as sen distractor, o que apunta que nestes casos en cambio este elemento non axuda ao bo procesamento das oracións. Cómpre, así, continuar a investigación deste

fenómeno para unha maior comprensión do procesamento sintáctico humano e comprobar se son outros factores oracionais os que condicionan estes resultados.

2. Resultados de modelos de linguaxe

Nas seguintes liñas preséntase a análise dos resultados da avaliación de *surprisal* e de TSE, experimentos levados a cabo para avaliar os tres modelos de linguaxe neurais escollidos neste traballo.

2.1. Surprisal

A continuación, expóñense os resultados de *surprisal*. Primeiramente, discútense os datos obtidos na avaliación dos modelos de linguaxe neurais para galego (Bertinho-base e BERT-gl-base). Deseguido, analízanse os referentes ao modelo para portugués BERTimbau-base.

2.1.1. Modelos de linguaxe para galego

Os resultados dos modelos de transformers para galego Bertinho-base e BERT-gl-base amosan que estes diferencian as oracións gramaticais e as agramaticais sen dificultade (*vid*. sección 2.2.1.1. de "Experimentos e resultados"). Con todo, BERT-gl-base presenta unha maior eficacia á hora de distinguir as cadeas sintacticamente correctas daquelas incorrectas, pois atribúelles valores máis diferenciados dos que asigna Bertinho-base. Estes resultados seguen as mesma tendencias que traballos anteriores como o de Garcia e Crespo-Otero (2022), o de-Dios-Flores e Garcia (2022) ou o de-Dios-Flores et al. (2023) xa anunciaban: BERT-gl supera ao primeiro LM de transformers para galego, Bertinho.

Ademais, as cifras obtidas no experimento sinalan que estes modelos procesan mellor as oracións de contexto curto que as de contexto longo. Ambos os dous LMs dan unha puntuación máis alta de media ás oracións gramaticais de contexto curto que ás de contexto longo e unha valoración máis baixa ás agramaticais de contexto curto que ás de contexto longo.

Do mesmo xeito, os resultados indican que o distractor confunde o modelo, xa que aquelas oracións gramaticais que presentan este elemento obteñen unha puntuación menor que as que non o teñen. No caso das agramaticais, a presenza do distractor parece ser a causa da atribución dunha maior aceptabilidade a aquelas que teñen este elemento, fronte as que non teñen distractor.

2.1.2. Modelo de linguaxe para portugués

Os resultados de BERTimbau-base mostran que este modelo fai unha boa diferenciación entre oracións gramaticais e agramaticais. Os valores dados ás sintacticamente correctas son máis altos que os rexistrados polos modelos para galego, mais as oracións incorrectas obteñen tamén unha puntuación máis alta de media que as destes outros modelos. Aínda así, BERTimbau-base presenta a máis clara distinción entre gramaticais e agramaticais.

Respecto ao contexto, os resultados da avaliación *surprisal* indican que o modelo procesa mellor as oracións gramaticais de contexto curto en comparación con aquelas

de contexto longo. Con todo, a aceptabilidade das oracións agramaticais de contexto curto é maior que aquelas de contexto longo, polo que a presenza do distractor neste caso non axuda ao mellor procesamento das oracións. A exploración destes resultados constitúe unha potencial liña de traballo futuro, na que a análise doutros factores que entran en contacto coa variable de contexto no dataset empregado pode explicar os datos obtidos.

Ademais, BERTimbau-base presenta un mellor procesamento das oracións gramaticais con distractor que de aquelas que non teñen este elemento. Porén, nas agramaticais ocorre o contrario: aquelas con distractor son consideradas máis aceptables que as sen distractor. Neste caso, como ocorre coa variable do contexto, cómpre examinar outros factores oracionais que poidan explicar estes resultados.

2.2. TSE

Os resultados de TSE amosan que todos os modelos procesan máis doadamente as oracións de contexto curto sen distractor que outras variantes. Nas oracións con estas variables tanto Bertinho-base como BERT-gl-base e BERTimbau-base rexistran a precisión máis alta de todos os seus resultados. Isto confirma a hipótese coa que comezamos este traballo: a distancia entre o núcleo do suxeito e o verbo principal así como a presenza do distractor inflúen no procesamento das oracións por parte dos modelos, sendo aqueles casos menos complexos (contexto curto e sen distractor) os que rexistran unha maior aceptabilidade.

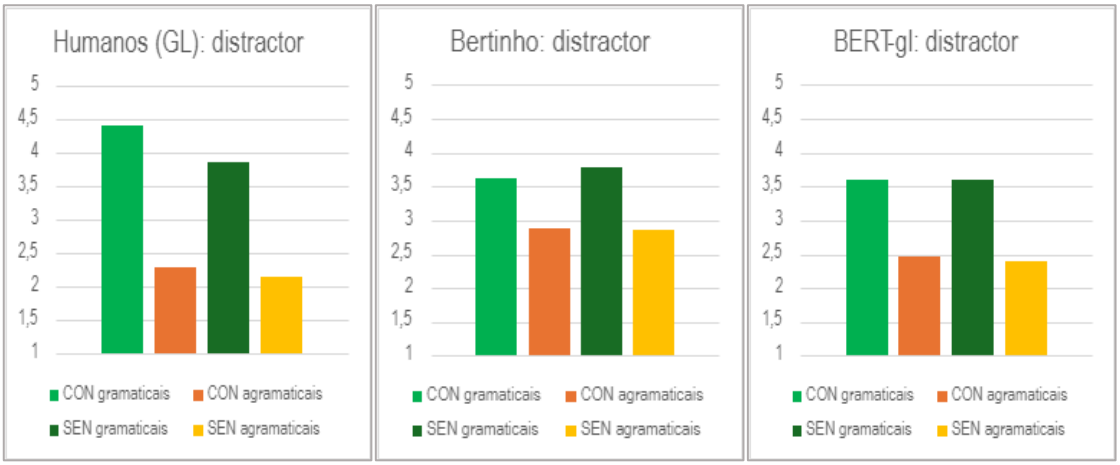
No caso de Bertinho-base, como xa comentamos, o seu mellor resultado é nas oracións de contexto curto e sen distractor, pero a súa precisión no procesamento destas variantes é aínda así menor ao resultado que se obtería ao realizar o experimento de forma aleatoria (< 0,50). Así, pode concluírse que Bertinho-base non identifica a estrutura xerárquica fundamental do galego e do portugués e non é quen de procesar a dependencia suxeito-verbo. A avaliación TSE permite afondar na análise das capacidades dos LMs, porque, aínda que en termos de *surprisal* Bertinho parecía diferenciar gramatical de agramatical, a avaliación TSE mostra o seu procesamento das oracións é deficiente.

Respecto aos outros modelos, en BERT-gl-base parece ter maior impacto o contexto, é dicir, a distancia entre o núcleo do suxeito e o verbo principal, que en BERTimbau-base. O contexto longo confunde notablemente a BERT-gl-base, xa que, malia presentar bos resultados xerais no procesamento das oracións, naquelas de contexto longo a súa precisión cae ata situarse no 0,50. De feito, naquelas oracións de contexto longo que teñen distractor a precisión é aínda menor, fronte ás que non teñen este elemento, que rexistran un resultado lixeiramente mellor. No entanto, BERT-gl-base presenta un moi bo procesamento nas oracións de contexto curto, cuxa precisión non varía nos casos con e sen distractor. Estes resultados confirman o exposto por Garcia e Crespo-Otero (2022): na concordancia de número suxeito-verbo o distractor parece ter menos influencia naquelas oracións de contexto curto, o que parece indicar que os modelos identifican a relación existente entre as dúas unidades *targeted* do traballo. Non obstante, nos resultados de BERTimbau-base, a presenza do distractor semella ter maior impacto que a distancia entre suxeito e verbo (contexto), pois naquelas oracións con distractor tanto

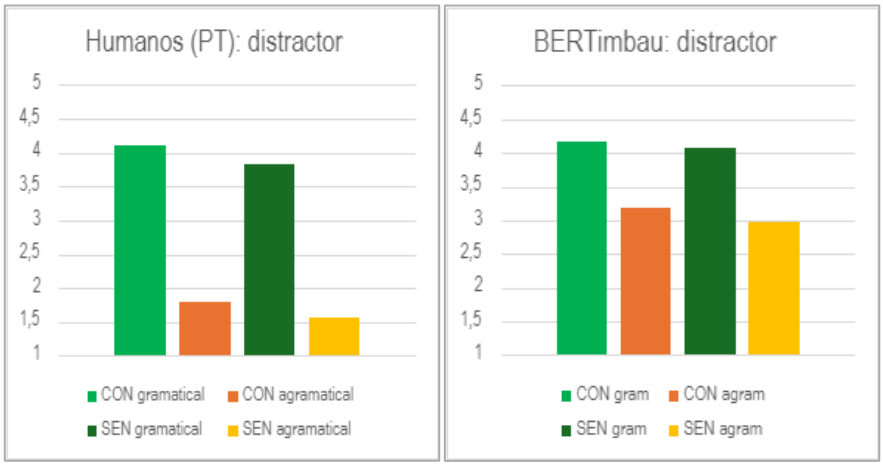
de contexto curto como de contexto longo a precisión do modelo cae considerablemente respecto ás oracións que non teñen distractor (*vid*. sección 2.2.2. de "Experimentos e resultados"). Mentres tanto, a súa precisión nas oracións de contexto curto e nas de contexto longo non é tan distante. Ademais, no procesamento das oracións de contexto longo sen distractor, variantes problemáticas para os outros modelos avaliados pola distancia entre suxeito e verbo, BERTimbau-base mostra unha moi alta precisión.

3. Comparación dos resultados dos humanos e dos modelos

Para facilitar a comparación, colocamos a continuación figuras que enfrontan os datos obtidos por humanos e por modelos de linguaxe no procesamento das oracións dos datasets.



Gráfica 1. Resultados da avaliación *surprisal* de falantes de galego e dos modelos de linguaxe Bertinho-base e BERT-gl-base, presentados segundo as variables presenza de distractor e gramaticalidade.



Gráfica 2. Resultados da avaliación *surprisal* de falantes de portugués e do modelo de linguaxe BERTimbau-base, presentados segundo as variables presenza de distractor e gramaticalidade.

Como era esperado, os resultados suxiren que os humanos identifican e procesan mellor que os modelos de linguaxe neurais a relación sintáctica de suxeito e verbo expresada a

través da concordancia de número. Tanto os falantes como os LMs distinguen as oracións gramaticais das agramaticais (aquelas nas que suxeito e verbo non concordan), mais as persoas outórganlles valores máis diferenciados a unhas e a outras que os modelos.

Ademais, mentres que o distractor nas oracións confunde os LMs, este parece axudar aos humanos na avaliación da aceptabilidade das cadeas lingüísticas. Con todo, falantes e modelos comparten un peor procesamento das oracións de contexto longo, fronte ás de contexto curto. 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 (con pequenas variacións en cada lingua).

Conclusións e traballo futuro

Nas páxinas deste traballo presentouse unha avaliación das habilidades sintácticas para procesar a concordancia suxeito-verbo de modelos de linguaxe neurais para galego e portugués e de falantes nativos e "case nativos" de ambas as dúas linguas, cunha conseguinte comparación dos datos duns e outros. A finalidade deste estudo era verificar se os modelos codifican a estrutura sintáctica da lingua e, polo tanto, se existe unha adquisición de coñecementos lingüísticos comparable coa dos humanos. Para isto, construíronse dous datasets controlando as variables de gramaticalidade, presenza de distractor, distancia entre o núcleo do suxeito e o verbo e frecuencia destes elementos nun corpus de referencia. Os datasets foron adaptados para realizar tres experimentos: por un lado, enquisas para explorar a resolución da dependencia suxeito-verbo por parte dos falantes; e por outro, a metodoloxía *surprisal* e a TSE para testar os modelos Bertinho-base (GL), BERT-gl-base (GL) e BERTimbau-base (PT).

Os resultados indican que tanto os humanos como os modelos teñen a capacidade de distinguir oracións gramaticais de agramaticais, aínda que as persoas establecen unha diferenciación máis acentuada entre unhas e outras. Os datos tamén revelan que a presenza do distractor na oración facilita, en contra do esperado, o procesamento da dependencia albo nos falantes, mentres que este elemento confunde os modelos. Isto suxire que a representación da estrutura sintáctica nos modelos de linguaxe vese afectada pola presenza de distractores, ao contrario do que acontece coas persoas. Ademais, a capacidade de resolución de humanos e LMs vese afectada negativamente nas oracións de contexto longo. Consecuentemente, os modelos presentan os seus mellores resultados no procesamento das oracións formalmente máis sinxelas: aquelas de contexto curto sen distractor. Estes datos parecen indicar que, ao menos nestes casos, os LMs si identifican e comprenden a estrutura sintáctica xerárquica que relaciona o núcleo do suxeito e o verbo principal da oración. Con todo, non son quen de recoñecer esta dependencia nas oracións máis complexas, especialmente se inclúen distractor. Isto suxire que os modelos seguen a ampararse na concepción lineal da oración naqueles casos onde existe unha grande distancia entre suxeito e verbo ou un distractor. Así pois, estes factores externos á dependencia suxeito-verbo si afectan á capacidade de resolución dos LMs.

Respecto aos modelos, este estudo revela un moi diferente rendemento entre eles. Por unha banda, o LM para portugués BERTimbau-base presenta a precisión xeral máis alta de todos os avaliados, 73%, que supera o 80% e 90% en varios casos. Doutro lado, o modelo BERT-gl-base ten unha precisión xeral do 66%, pero supera tamén o 80% nalgúns escenarios. Con todo, o outro modelo para galego, Bertinho-base, ten unha precisión

xeral do 34%, que só supera o 40% nos casos de contexto curto sen distractor. Isto indica que os modelos adestrados cun corpus de tamaño suficiente (BERTimbau-base e BERT-gl-base) acertan a maioría dos casos, mentres que os adestrados con poucos datos teñen *below-chance performance* ou un rendemento por debaixo da media probable.

Alén disto, nesta pesquisa trazáronse potenciais liñas de investigación que esperamos explorar en traballos futuros. Por unha banda, planeamos afondar no impacto do distractor no procesamento da dependencia suxeito-verbo en humanos, incluíndo na avaliación do procesamento das oracións métodos do campo da psicolingüística, como o *eye-tracking* ou o control do tempo de lectura. Ademais, respecto á valoración da capacidade de resolución tanto dos modelos como dos humanos, planeamos realizar análises estatísticas máis complexas, como tests de significancia ou correlacións entre os diferentes resultados. Pensamos que tamén sería útil para o avance da investigación avaliar outros fenómenos sintácticos (podendo ser adaptados para galego e portugués os incluídos nos datasets en inglés de Warstadt et al. [2020]) aplicando a metodoloxía empregada no presente traballo e explorar como os modelos multilingües resolven estas dependencias en varias linguas.

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, Magariños et al., 2022, p. 52).

Agradecementos

Este traballo non tería sido posible sen a colaboración de Lisa Ferreira, estudante de Línguas, Literaturas e Culturas da Universidade de Lisboa, que respondeu as miñas dúbidas co portugués sempre cun sorriso e de todas as persoas que participaron nas enquisas e as compartiron alén da fronteira. Grazas tamén á miña familia, que sempre fixo todo o posible para que hoxe eu poida estar aquí.

Bibliografía

Allen, J. (1995 [2ª ed.]). *Natural Language Understanding*. Benjamin/Cummings.

Bock, K. e Miller, C. A. (1991). Broken Agreement. *Cognitive Psychology*, 23(1), 45-93. https://doi.org/10.1016/0010-0285(91)90003-7.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, 6, 213-34.

Bross, F. (2019). *Acceptability Ratings in Linguistics: A Practical Guide to Grammaticality Judgments, Data Collection, and Statistical Analysis*. Version 1.02. Mimeo. www.fabianbross.de/acceptabilityratings.pdf

De-Dios-Flores, I. e Garcia, M. (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.

De-Dios-Flores, I.; Garcia Amboage, J. e Garcia, M. (2023). Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1 (Long Papers).

De-Dios-Flores, I.; Magariños, C.; Vladu, A. I.; Ortega, J. E.; Pichel, J. R.; Garcia, M.; Gamallo, M.; Fernández Rei, E.; Bugarín, A; González González, M.; Barro, S.; Regueira, X. L. (2022). The Nós Project: Opening routes for the Galician language in the field of language technologies. *Proceedings of the Towards Digital Language Equality Workshop*, 52–61.

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6 (10), 635-653.

Gamallo, P.; Garcia, M; de-Dios-Flores, I. (2022). Evaluating Contextualized Vectors from both Large Language Models and Compositional Strategies. *Procesamiento del Lenguaje Natural*, 69, 153-164.

Garcia, M. (2021). Exploring the representation of word meanings in context: A case study on homonymy and synonymy. *Proceedings of the 59th Annual Meeting of the*

Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 1 (Long Papers), 3625– 3640.

Garcia, M. e Crespo-Otero, A. (2022). A Targeted Assessment of the Syntactic Abilities of Transformer Models for Galician-Portuguese. *Computational Processing of the Portuguese Language. PROPOR 2022*, 13208. https://doi.org/10.1007/978-3-030-98305-5_5

Goldberg, Y. (2019). *Assessing BERT's Syntactic Abilities*. arXiv preprint arXiv:1901.05287.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 1195–1205.

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics,* 1-8. https://doi.org/10.3115/1073336.1073357

Jurafsky, D. e Martin, J. H. (2024). *Speech and Language Processing*. Stanford University.

Langsford, S.; Perfors, A.; Hendrickson, A.; Kennedy, L. e Navarro, D. (2018). Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a journal of general linguistics*, 3(1): 37, 1–34. https://doi.org/10.5334/gjgl.396

Linzen, T. e Baroni, M. (2021). Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, 7, 195-212.

Lizen, T.; Dupoux, E. E Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics,* 4, 521–535.

Marvin, R., Linzen, T. (2018). Targeted syntactic evaluation of language models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1192–1202.

Misra, K. (2022). *minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models*. arXiv:2203.13112v1.

Mueller, A., Nicolai, G., Petrou-Zeniou, P., Talmina, N., Linzen, T. (2020). Cross-linguistic syntactic evaluation of word prediction models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5523–5539.

Newman, B.; Ang, K.S.; Gong, J. e Hewitt, J. (2021). Refining targeted syntactic evaluation of language models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3710–3723.

Paes, A.; Vianna, D. e Rodrigues, J. (2023). Modelos de Linguagem. En Caseli, H.M. e Nunes, M.G.V. (org.), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português* (2ª ed., pp. 317-366). BPLN. https://brasileiraspln.com/livro-pln/2a-edicao.

Raposo, E. P. (1992). *Teoria da Gramática: A Faculdade da Linguagem*. Caminho.

Rezaii, N.; Michaelov, J.; Josephy-Hernandez, S.; Ren, B.; Hochberg, D.; Quimby, M.; Dickerson, B. C. (2023). Measuring Sentence Information via Surprisal: Theoretical and Clinical Implications in Nonfluent Aphasia. *Annals of Neurology*, 94 (4), 647-657. https://doi.org/10.1002/ana.26744

Souza, F.C.; Nogueira, R. F.; Lotufo, R.A. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020*, 403-417.

Tordera Yllescas, J.C. (2011). Sobre la Lingüística Computacional: Fundamentos. *Lingüística Computacional. Tecnologías del Habla,* anejo n.º 74, Revista *Quaderns de Filologia*.

Vilares, D., Garcia, M., Gómez-Rodríguez, C. (2021). Bertinho: Galician BERT Representations. *Procesamiento del Lenguaje Natural*, 66, 13–26.

Warstadt, A.; Parrish, A.; Liu, H.; Mohananey, A.; Peng, W.; Wang, S. e Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics* 2020, 8, 377–392. https://doi.org/10.1162/tacl_a_00321

Anexos

1. Dataset (GL)

Item	Contexto	Distractor	Gramaticalidade	Oración
1	Longo	Con	Gramatical	O cabalo que sacode a crina da cor das amoras atravesa o campo ao galope.
1	Longo	Con	Agramatical	O cabalo que sacode a crina da cor das amoras atravesan o campo ao galope.
1	Curto	Con	Gramatical	O cabalo que ten novas ferraduras atravesa o campo ao galope.
1	Curto	Con	Agramatical	O cabalo que ten novas ferraduras atravesan o campo ao galope.
2	Longo	Con	Gramatical	O amor que é coidado por ambas as partes con mimo durante os tempos difíciles sobrevive a todas as adversidades.
2	Longo	Con	Agramatical	O amor que é coidado por ambas as partes con mimo durante os tempos difíciles sobreviven a todas as adversidades.
2	Curto	Con	Gramatical	O amor que coidan ambas as partes sobrevive a todas as adversidades.
2	Curto	Con	Agramatical	O amor que coidan ambas as partes sobreviven a todas as adversidades.
3	Longo	Con	Gramatical	O directivo que impulsou a introdución no mercado da nova gama de electrodomésticos figurou na lista de convidados.
3	Longo	Con	Agramatical	O directivo que impulsou a introdución no mercado da nova gama de electrodomésticos figuraron na lista de convidados.
3	Curto	Con	Gramatical	O directivo que impulsou as reformas figurou na lista de convidados.
3	Curto	Con	Agramatical	O directivo que impulsou as reformas figuraron na lista de convidados.
4	Longo	Con	Gramatical	O sol que luciu con forza este venres nas costas galegas fixo que se superasen os 30 graos.

4	Longo	Con	Agramatical	O sol que luciu con forza este venres nas costas galegas fixeron que se superasen os 30 graos.
4	Curto	Con	Gramatical	O sol que luciu nas praias fixo que se superasen os 30 graos.
4	Curto	Con	Agramatical	O sol que luciu nas praias fixeron que se superasen os 30 graos.
5	Longo	Con	Gramatical	Os alumnos que gozan de réxime de pensión completa todo o ano residen na parte esquerda do edificio.
5	Longo	Con	Agramatical	Os alumnos que gozan de réxime de pensión completa todo o ano reside na parte esquerda do edificio.
5	Curto	Con	Gramatical	Os alumnos de primeiro ano residen na parte esquerda do edificio.
5	Curto	Con	Agramatical	Os alumnos de primeiro ano reside na parte esquerda do edificio.
6	Longo	Con	Gramatical	Os celos que Brais sente polo seu irmán dende que é cativo controlan a súa vida.
6	Longo	Con	Agramatical	Os celos que Brais sente polo seu irmán dende que é cativo controla a súa vida.
6	Curto	Con	Gramatical	Os celos polo irmán pequeno controlan a súa vida.
6	Curto	Con	Agramatical	Os celos polo irmán pequeno controla a súa vida.
7	Longo	Con	Gramatical	Os pais da rapaza que vive na casa co muro de granito rosa casaron na Arxentina.
7	Longo	Con	Agramatical	Os pais da rapaza que vive na casa co muro de granito rosa casou na Arxentina.
7	Curto	Con	Gramatical	Os pais da rapaza con coleta casaron na Arxentina.
7	Curto	Con	Agramatical	Os pais da rapaza con coleta casou na Arxentina.
8	Longo	Con	Gramatical	Eses paxaros que teñen as plumas pardas e o peteiro encarnado fican á beira do estanque.
8	Longo	Con	Agramatical	Eses paxaros que teñen as plumas pardas e o peteiro encarnado fica á beira do estanque.
8	Curto	Con	Gramatical	Eses paxaros de peteiro escuro fican á beira do estanque.
8	Curto	Con	Agramatical	Eses paxaros de peteiro escuro fica á beira do estanque.
9	Longo	Con	Gramatical	A propietaria que estaba encantada coas condicións propostas naquelas páxinas asinou o contrato.

9	Longo	Con	Agramatical	A propietaria que estaba encantada coas condicións propostas naquelas páxinas asinaron o contrato.
9	Curto	Con	Gramatical	A propietaria que puxo as condicións asinou o contrato.
9	Curto	Con	Agramatical	A propietaria que puxo as condicións asinaron o contrato.
10	Longo	Con	Gramatical	A lagoa que esconde as casas da antiga vila baixo as augas semella un espello.
10	Longo	Con	Agramatical	A lagoa que esconde as casas da antiga vila baixo as augas semellan un espello.
10	Curto	Con	Gramatical	A lagoa de augas calmas semella un espello.
10	Curto	Con	Agramatical	A lagoa de augas calmas semellan un espello.
11	Longo	Con	Gramatical	A mirada que a testemuña do crime lle dirixiu aos acusados revelou un medo atroz.
11	Longo	Con	Agramatical	A mirada que a testemuña do crime lle dirixiu aos acusados revelaron un medo atroz.
11	Curto	Con	Gramatical	A mirada das vítimas revelou un medo atroz.
11	Curto	Con	Agramatical	A mirada das vítimas revelaron un medo atroz.
12	Longo	Con	Gramatical	A nena que acababa de chegar á vila cos seus pais xogou co veciño.
12	Longo	Con	Agramatical	A nena que acababa de chegar á vila cos seus pais xogaron co veciño.
12	Curto	Con	Gramatical	A nena cos zapatos vellos xogou co veciño.
12	Curto	Con	Agramatical	A nena cos zapatos vellos xogaron co veciño.
13	Longo	Con	Gramatical	As xornalistas que cubrían a fronte oriental do conflito bélico publicaron unha nova reportaxe.
13	Longo	Con	Agramatical	As xornalistas que cubrían a fronte oriental do conflito bélico publicou unha nova reportaxe.
13	Curto	Con	Gramatical	As xornalistas da renovada redacción publicaron unha nova reportaxe.
13	Curto	Con	Agramatical	As xornalistas da renovada redacción publicou unha nova reportaxe.
14	Longo	Con	Gramatical	As bolboretas que estaba a estudar a nova investigadora asociada buscaban pole nas flores.

14	Longo	Con	Agramatical	As bolboretas que estaba a estudar a nova investigadora asociada buscaba pole nas flores.
14	Curto	Con	Gramatical	As bolboretas do xardín buscaban pole nas flores.
14	Curto	Con	Agramatical	As bolboretas do xardín buscaba pole nas flores.
15	Longo	Con	Gramatical	As ondas que vai traer consigo a fronte que entra esta noite poden superar os sete metros de altura.
15	Longo	Con	Agramatical	As ondas que vai traer consigo a fronte que entra esta noite pode superar os sete metros de altura.
15	Curto	Con	Gramatical	As ondas que trae o temporal poden superar os sete metros de altura.
15	Curto	Con	Agramatical	As ondas que trae o temporal pode superar os sete metros de altura.
16	Longo	Con	Gramatical	Esas poetas que se acaban de estrear na produción cinematográfica sobresaen no panorama literario internacional.
16	Longo	Con	Agramatical	Esas poetas que se acaban de estrear na produción cinematográfica sobresae no panorama literario internacional.
16	Curto	Con	Gramatical	Esas poetas que participaron na curtametraxe sobresaen no panorama literario internacional.
16	Curto	Con	Agramatical	Esas poetas que participaron na curtametraxe sobresae no panorama literario internacional.

2. Dataset (PT)

Item	Contexto	Distractor	Gramaticalidade	Oración
1	Longo	Con	Gramatical	O cavalo que sacode a crina da cor das amoras está no campo.
1	Longo	Con	Agramatical	O cavalo que sacode a crina da cor das amoras estão no campo.
1	Longo	Sen	Gramatical	O cavalo que sacode a crina da cor da hulha intensamente está no campo.
1	Longo	Sen	Agramatical	O cavalo que sacode a crina da cor da hulha intensamente estão no campo.
1	Curto	Con	Gramatical	O cavalo que tem novas ferraduras está no campo.
1	Curto	Con	Agramatical	O cavalo que tem novas ferraduras estão no campo.
1	Curto	Sen	Gramatical	O cavalo desparasitado recentemente está no campo.

1	Curto	Sen	Agramatical	O cavalo desparasitado recentemente estão no campo.
2	Longo	Con	Gramatical	O amor que foi cultivado por ambas as partes com carinho durante os tempos difíceis sobreviveu a todas as adversidades.
2	Longo	Con	Agramatical	O amor que foi cultivado por ambas as partes com carinho durante os tempos difíceis sobreviveram a todas as adversidades.
2	Longo	Sen	Gramatical	O amor que foi cultivado por ambas as partes com carinho durante o conflito finalmente sobreviveu a todas as adversidades.
2	Longo	Sen	Agramatical	O amor que foi cultivado por ambas as partes com carinho durante o conflito finalmente sobreviveram a todas as adversidades.
2	Curto	Con	Gramatical	O amor cultivado por ambas as partes sobreviveu a todas as adversidades.
2	Curto	Con	Agramatical	O amor cultivado por ambas as partes sobreviveram a todas as adversidades.
2	Curto	Sen	Gramatical	O amor que foi cultivado cuidadosamente sobreviveu a todas as adversidades.
2	Curto	Sen	Agramatical	O amor que foi cultivado cuidadosamente sobreviveram a todas as adversidades.
3	Longo	Con	Gramatical	O diretor que promoveu no mercado a nova gama de eletrodomésticos apareceu na lista dos convidados.
3	Longo	Con	Agramatical	O diretor que promoveu no mercado a nova gama de eletrodomésticos apareceram na lista dos convidados.
3	Longo	Sen	Gramatical	O diretor que promoveu a introdução no mercado das novas secadoras possivelmente apareceu na lista dos convidados.
3	Longo	Sen	Agramatical	O diretor que promoveu a introdução no mercado das novas secadoras possivelmente apareceram na lista dos convidados.
3	Curto	Con	Gramatical	O diretor que fomentou as mudanças apareceu na lista dos convidados.
3	Curto	Con	Agramatical	O diretor que fomentou as mudanças apareceram na lista dos convidados.

3	Curto	Sen	Gramatical	O diretor que fomentou as mudanças ontem apareceu na lista dos convidados.
3	Curto	Sen	Agramatical	O diretor que fomentou as mudanças ontem apareceram na lista dos convidados.
4	Longo	Con	Gramatical	O sol que brilhou bastante nesta sexta nas costas galegas fez com que os 30 graus fossem ultrapassados.
4	Longo	Con	Agramatical	O sol que brilhou bastante nesta sexta nas costas galegas fizeram com que os 30 graus fossem ultrapassados.
4	Longo	Sen	Gramatical	O sol que brilhou nesta sexta no litoral atlântico intensamente fez com que os 30 graus fossem ultrapassados.
4	Longo	Sen	Agramatical	O sol que brilhou nesta sexta no litoral atlântico intensamente fizeram com que os 30 graus fossem ultrapassados.
4	Curto	Con	Gramatical	O sol que brilhou nas praias fez com que os 30 graus fossem ultrapassados.
4	Curto	Con	Agramatical	O sol que brilhou nas praias fizeram com que os 30 graus fossem ultrapassados.
4	Curto	Sen	Gramatical	O sol que brilhou no interior intensamente fez com que os 30 graus fossem ultrapassados.
4	Curto	Sen	Agramatical	O sol que brilhou no interior intensamente fizeram com que os 30 graus fossem ultrapassados.
5	Longo	Con	Gramatical	Os alunos que desfrutam do regime de pensão completa durante todo o ano letivo vivem na ala esquerda do edifício.
5	Longo	Con	Agramatical	Os alunos que desfrutam do regime de pensão completa durante todo o ano letivo vive na ala esquerda do edifício.
5	Longo	Sen	Gramatical	Os alunos que desfrutam do regime de pensão completa anualmente vivem na ala esquerda do edifício.
5	Longo	Sen	Agramatical	Os alunos que desfrutam do regime de pensão completa anualmente vive na ala esquerda do edifício.
5	Curto	Con	Gramatical	Os alunos do primeiro ano vivem na ala esquerda do edifício.
5	Curto	Con	Agramatical	Os alunos do primeiro ano vive na ala esquerda do edifício.
5	Curto	Sen	Gramatical	Os alunos do primeiro ano geralmente vivem na ala esquerda do edifício.

5	Curto	Sen	Agramatical	Os alunos do primeiro ano geralmente vive na ala esquerda do edifício.
6	Longo	Con	Gramatical	Os sentimentos que o Brais tem pela sua primeira namorada deveriam permanecer.
6	Longo	Con	Agramatical	Os sentimentos que o Brais tem pela sua primeira namorada deveria permanecer.
6	Longo	Sen	Gramatical	Os sentimentos que o Brais tem pela sua namorada atualmente deveriam permanecer.
6	Longo	Sen	Agramatical	Os sentimentos que o Brais tem pela sua namorada atualmente deveria permanecer.
6	Curto	Con	Gramatical	Os sentimentos que tem pela namorada deveriam permanecer.
6	Curto	Con	Agramatical	Os sentimentos que tem pela namorada deveria permanecer.
6	Curto	Sen	Gramatical	Os sentimentos que tem atualmente deveriam permanecer.
6	Curto	Sen	Agramatical	Os sentimentos que tem atualmente deveria permanecer.
7	Longo	Con	Gramatical	Os avós da menina que vive na casa com o muro de granito casaram na Argentina.
7	Longo	Con	Agramatical	Os avós da menina que vive na casa com o muro de granito casou na Argentina.
7	Longo	Sen	Gramatical	Os avós da menina que vive na casa com o muro de granito habitualmente casaram na Argentina.
7	Longo	Sen	Agramatical	Os avós da menina que vive na casa com o muro de granito habitualmente casou na Argentina.
7	Curto	Con	Gramatical	Os avós da menina com gorro casaram na Argentina.
7	Curto	Con	Agramatical	Os avós da menina com gorro casou na Argentina.
7	Curto	Sen	Gramatical	Os avós da menina que compete hoje casaram na Argentina.
7	Curto	Sen	Agramatical	Os avós da menina que compete hoje casou na Argentina.
8	Longo	Con	Gramatical	Esses pássaros que têm as penas pardas e o bico encarnado ficam na beira da lagoa.

8	Longo	Con	Agramatical	Esses pássaros que têm as penas pardas e o bico encarnado fica na beira da lagoa.
8	Longo	Sen	Gramatical	Esses pássaros que têm as penas e o bico encarnados parcialmente ficam na beira da lagoa.
8	Longo	Sen	Agramatical	Esses pássaros que têm as penas e o bico encarnados parcialmente fica na beira da lagoa.
8	Curto	Con	Gramatical	Esses pássaros de bico escuro ficam na beira da lagoa.
8	Curto	Con	Agramatical	Esses pássaros de bico escuro fica na beira da lagoa.
8	Curto	Sen	Gramatical	Esses pássaros de penas encarnadas agora ficam na beira da lagoa.
8	Curto	Sen	Agramatical	Esses pássaros de penas encarnadas agora fica na beira da lagoa.
9	Longo	Con	Gramatical	A proprietária que ficou encantada com as condições propostas naquelas páginas assinou o contrato.
9	Longo	Con	Agramatical	A proprietária que ficou encantada com as condições propostas naquelas páginas assinaram o contrato.
9	Longo	Sen	Gramatical	A proprietária que ficou encantada com as condições propostas ontem assinou o contrato.
9	Longo	Sen	Agramatical	A proprietária que ficou encantada com as condições propostas ontem assinaram o contrato.
9	Curto	Con	Gramatical	A proprietária que propôs as condições assinou o contrato.
9	Curto	Con	Agramatical	A proprietária que propôs as condições assinaram o contrato.
9	Curto	Sen	Gramatical	A proprietária que leu as páginas detidamente assinou o contrato.
9	Curto	Sen	Agramatical	A proprietária que leu as páginas detidamente assinaram o contrato.
10	Longo	Con	Gramatical	A lagoa que esconde as casas da antiga vila sob as águas parece um espelho.
10	Longo	Con	Agramatical	A lagoa que esconde as casas da antiga vila sob as águas parecem um espelho.
10	Longo	Sen	Gramatical	A lagoa que escondia as casas da antiga vila medieval antes parece um espelho.
10	Longo	Sen	Agramatical	A lagoa que escondia as casas da antiga vila medieval antes parecem um espelho.

10	Curto	Con	Gramatical	A lagoa de águas calmas parece um espelho.
10	Curto	Con	Agramatical	A lagoa de águas calmas parecem um espelho.
10	Curto	Sen	Gramatical	A lagoa que está aqui perto parece um espelho.
10	Curto	Sen	Agramatical	A lagoa que está aqui perto parecem um espelho.
11	Longo	Con	Gramatical	A cara que a testemunha do crime tinha quando os réus entraram mostrou que ela tinha medo.
11	Longo	Con	Agramatical	A cara que a testemunha do crime tinha quando os réus entraram mostraram que ela tinha medo.
11	Longo	Sen	Gramatical	A cara que a testemunha do crime tinha quando os réus entraram ontem mostrou que ela tinha medo.
11	Longo	Sen	Agramatical	A cara que a testemunha do crime tinha quando os réus entraram ontem mostraram que ela tinha medo.
11	Curto	Con	Gramatical	A cara das vítimas mostrou que tinham medo.
11	Curto	Con	Agramatical	A cara das vítimas mostraram que tinham medo.
11	Curto	Sen	Gramatical	A cara das vítimas indubitavelmente mostrou que tinham medo.
11	Curto	Sen	Agramatical	A cara das vítimas indubitavelmente mostraram que tinham medo.
12	Longo	Con	Gramatical	A menina que acabava de chegar à vila com os seus pais jogou com o vizinho.
12	Longo	Con	Agramatical	A menina que acabava de chegar à vila com os seus pais jogaram com o vizinho.
12	Longo	Sen	Gramatical	A menina que chegara à vila com os seus pais recentemente jogou com o vizinho.
12	Longo	Sen	Agramatical	A menina que chegara à vila com os seus pais recentemente jogaram com o vizinho.
12	Curto	Con	Gramatical	A menina com os sapatos velhos jogou com o vizinho.
12	Curto	Con	Agramatical	A menina com os sapatos velhos jogaram com o vizinho.
12	Curto	Sen	Gramatical	A menina que corre livremente jogou com o vizinho.
12	Curto	Sen	Agramatical	A menina que corre livremente jogaram com o vizinho.

13	Longo	Con	Gramatical	As jornalistas que cobriram a fronte oriental do conflito bélico escreveram uma nova reportagem.
13	Longo	Con	Agramatical	As jornalistas que cobriram a fronte oriental do conflito bélico escreveu uma nova reportagem.
13	Longo	Sen	Gramatical	As jornalistas que cobriram a fronte oriental durante as revoltas recentemente escreveram uma nova reportagem.
13	Longo	Sen	Agramatical	As jornalistas que cobriram a fronte oriental durante as revoltas recentemente escreveu uma nova reportagem.
13	Curto	Con	Gramatical	As jornalistas da nova redação escreveram uma nova reportagem.
13	Curto	Con	Agramatical	As jornalistas da nova redação escreveu uma nova reportagem.
13	Curto	Sen	Gramatical	As jornalistas rapidamente escreveram uma nova reportagem.
13	Curto	Sen	Agramatical	As jornalistas rapidamente escreveu uma nova reportagem.
14	Longo	Con	Gramatical	As abelhas que estava a estudar a nova pesquisadora associada buscam pólen nas flores.
14	Longo	Con	Agramatical	As abelhas que estava a estudar a nova pesquisadora associada busca pólen nas flores.
14	Longo	Sen	Gramatical	As abelhas que estava a estudar a pesquisadora nos jardins ontem buscam pólen nas flores.
14	Longo	Sen	Agramatical	As abelhas que estava a estudar a pesquisadora nos jardins ontem busca pólen nas flores.
14	Curto	Con	Gramatical	As abelhas do jardim buscam pólen nas flores.
14	Curto	Con	Agramatical	As abelhas do jardim busca pólen nas flores.
14	Curto	Sen	Gramatical	As abelhas que estão ali buscam pólen nas flores.
14	Curto	Sen	Agramatical	As abelhas que estão ali busca pólen nas flores.
15	Longo	Con	Gramatical	As ondas que trará consigo a tormenta que entra esta noite poderão ultrapassar os sete metros de altura.

15	Longo	Con	Agramatical	As ondas que trará consigo a tormenta que entra esta noite poderá ultrapassar os sete metros de altura.
15	Longo	Sen	Gramatical	As ondas que vai trazer consigo a tormenta que entra hoje poderão ultrapassar os sete metros de altura.
15	Longo	Sen	Agramatical	As ondas que vai trazer consigo a tormenta que entra hoje poderá ultrapassar os sete metros de altura.
15	Curto	Con	Gramatical	As ondas que traz a tormenta poderão ultrapassar os sete metros de altura.
15	Curto	Con	Agramatical	As ondas que traz a tormenta poderá ultrapassar os sete metros de altura.
15	Curto	Sen	Gramatical	As ondas que chegarão hoje poderão ultrapassar os sete metros de altura.
15	Curto	Sen	Agramatical	As ondas que chegarão hoje poderá ultrapassar os sete metros de altura.
16	Longo	Con	Gramatical	Estas artistas que acabam de estrear-se na produção cinematográfica destacam-se no panorama literário internacional.
16	Longo	Con	Agramatical	Essas artistas que acabam de estrear-se na produção cinematográfica destaca-se no panorama literário internacional.
16	Longo	Sen	Gramatical	Estas artistas que se estrearam dirigindo e produzindo longa-metragens recentemente destacam-se no panorama literário internacional.
16	Longo	Sen	Agramatical	Estas artistas que se estrearam dirigindo e produzindo longa-metragens recentemente destaca-se no cenário literário internacional.
16	Curto	Con	Gramatical	Estas artistas que participaram na curta-metragem destacam-se no panorama literário internacional.
16	Curto	Con	Agramatical	Estas artistas que participaram na curta-metragem destaca-se no panorama literário internacional.
16	Curto	Sen	Gramatical	Estas artistas que publicam obras anualmente destacam-se no panorama literário internacional.
16	Curto	Sen	Agramatical	Estas artistas que publicam obras anualmente destaca-se no panorama literário internacional.

3. Enquisas (GL)

    1. https://forms.office.com/e/VQ1SJnEp1s
    2. https://forms.office.com/e/nmeLA6FtaG
    3. https://forms.office.com/e/cFzY0v38xh
    4. https://forms.office.com/e/rqDEg6U27a
    5. https://forms.office.com/e/VsPfPDsCLb
    6. https://forms.office.com/e/e3JJWK2uzs
    7. https://forms.office.com/e/9TLihLgdQA
    8. https://forms.office.com/e/8i4DkhTPPE

4. Enquisas (PT)

    1. https://forms.office.com/e/eCDZ1Ddung
    2. https://forms.office.com/e/6TDf7g0AFJ
    3. https://forms.office.com/e/NiQjajpZ3e
    4. https://forms.office.com/e/bZaqrPL16J
    5. https://forms.office.com/e/wq7V9DuTSM
    6. https://forms.office.com/e/9Z2ejysGa7
    7. https://forms.office.com/e/wLkFQUXFTk
    8. https://forms.office.com/e/z73tUUv9gL