Proposta recibida en xuño 2025 e aceptada para publicación en outubro 2025.

Clasificación automática de textos por niveis de lecturabilidade: recursos e modelos para o galego

Automatic text readability classification: resources and models for Galician

Sandra Rodríguez Rey
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela

Marcos Garcia
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela

Resumo

A avaliación automática da lecturabilidade textual constitúe un campo en expansión no ámbito do procesamento das linguas naturais, cun impacto relevante en áreas como o ensino e a aprendizaxe de linguas e a accesibilidade. Neste contexto, este artigo presenta Corlega, o primeiro corpus de textos en galego clasificados por niveis de lecturabilidade, composto por 480 documentos dirixidos a persoas adultas. O corpus abrangue 11 categorías e 36 subcategorías con diferentes xéneros textuais, subxéneros e tipos de textos. O proceso de selección e compilación de documentos, así como de clasificación, segue os estándares do proxecto iRead4Skills, que desenvolve recursos e modelos computacionais para o portugués, o español e o francés. Para compilar Corlega, este traballo define seis niveis de lecturabilidade en galego, para os que propón un conxunto de descritores lingüísticos. Con base nesta taxonomía, describimos o proceso de compilación do corpus e a súa distribución actual —en catro dos seis niveis de lecturabilidade—, así como as principais características deste novo recurso. Adicionalmente, empregamos o corpus para adestrar e avaliar ferramentas de clasificación automática da lecturabilidade, mediante o axuste de modelos Transformers monolingües e multilingües e a implementación de modelos híbridos. Os resultados suxiren que, con corpus de adestramento de tamaño reducido, a extracción de características de modelos preadestrados permite obter resultados competitivos co axuste supervisado dos modelos. Con todo, a combinación de corpus de diferentes linguas permite axustar modelos multilingües con mellor desempeño. Tanto o corpus como os modelos están dispoñibles para a comunidade científica.

Palabras chave

corpus de lecturabilidade; avaliación automática da lecturabilidade; clasificación de textos; galego; fine-tuning; aprendizaxe para adultos

Abstract

The automatic readability assessment of texts is a growing field within Natural Language Processing, with significant implications in areas such as language teaching and learning and accessibility. In this context, this paper presents Corlega, the first corpus of Galician texts classified by readability level, consisting of 480 texts aimed at adult readers. The corpus covers 11 categories and 36 subcategories, including a variety of text types, genres and subgenres. The process of selection and compilation of documents, as well as classification, follows the standards of the iRead4Skills project, which develops resources and computational models for Portuguese, Spanish and French. To compile Corlega, this work defines six levels of readability in Galician and proposes a set of linguistic descriptors for each level. Using this taxonomy, we describe the compilation process of the corpus and its current distribution —across four of the six readability levels—, as well as the main features of this new resource. Additionally, we used the corpus to train and evaluate automatic readability classification tools by fitting monolingual and multilingual Transformer models, and the implementation of hybrid models. The results suggest that, with small training corpora, feature extraction from pre-trained models is an efficient method to achieve competitive results with supervised model fitting. However, combining corpora from different languages enables the fitting of multilingual models with better performance. Both the corpus and the models are available to the scientific community.

Keywords

readability corpus; automatic readability assessment; text classification; Galician; fine-tuning; adult learning

DOI: 10.21814/lm.17.2.488

1. Introdución

Os estudos sobre a complexidade dos textos e o seu impacto na comprensión lectora remóntanse ao século XIX, con investigacións que apuntan á lonxitude das oracións e ao tipo de léxico empregado como fenómenos relevantes na determinación da complexidade (Campos Saavedra et al., 2014). A partir destas primeiras aproximacións, xurdiron distintas propostas para medir o nivel de dificultade dun texto, o que dá lugar ao concepto de lecturabilidade, entendido como o grao de facilidade co que un texto pode ser lido e comprendido (Campos Saavedra et al., 2014), especialmente relevante en ámbitos como a aprendizaxe de linguas ou a accesibilidade textual (Vajjala, 2022). Tradicionalmente, a lecturabilidade dos textos estimábase mediante fórmulas baseadas principalmente na cantidade de oracións, palabras e sílabas dos textos (Flesch, 1948; McLaughlin, 1969). Estas fórmulas foron progresivamente substituídas por modelos baseados en características lingüísticas extraídas manualmente (Daoust et al., 1996; Crossley et al., 2007), por sistemas de aprendizaxe automática (Dell'Orletta et al., 2014) ou por novas arquitecturas de aprendizaxe profunda (Madrazo Azpiazu & Pera, 2019). Para implementar estes métodos, son necesarios corpus anotados segundo niveis de dificultade (Vajjala, 2022; Dell'Orletta et al., 2014) a través dos que conseguen integrar factores como a diversidade léxica, a complexidade sintáctica ou a coherencia textual (Wilkens et al., 2022).

Existen corpus e conxuntos de datos clasificados segundo a complexidade dos textos en moitas linguas, como o inglés (Schwarm & Ostendorf, 2005; Vajjala & Meurers, 2012), o castelán (Xu et al., 2015), o portugués (Ribeiro et al., 2024) ou o italiano (Dell'Orletta et al., 2014; Grego Bolli et al., 2017), así como ferramentas que miden a complexidade ou simplifican textos previamente identificados como complexos (Wilkens et al., 2022; Sheehan et al., 2013; Bengoetxea & Gonzalez-Dios, 2021).

Orientado especialmente para a poboación adulta, o proxecto iRead4Skills define unha nova taxonomía para a clasificación de textos segundo a súa lecturabilidade, e propón a compilación dun conxunto de datos multilingüe (en castelán, francés e portugués (Pintard et al., 2024a)), para o adestramento de clasificadores automáticos. Con todo, non coñecemos recursos textuais sobre lecturabilidade en galego dispoñibles, o que pode ter utilidade tanto para análises de carácter lingüístico como para o desenvolvemento de ferramentas computacionais.

Neste artigo presentamos unha proposta de seis niveis de lecturabilidade en galego e publicamos un novo corpus, Corlega, para o seu estudo e avaliación. O corpus componse de 480 textos dirixidos a adultos nativos pouco habituados a comunicarse en galego e a aprendices de galego nativos doutras linguas romances. A versión actual de Corlega inclúe textos en catro niveis de lecturabilidade, e en 11 categorías e 36 subcategorías que inclúen diversos xéneros textuais, subxéneros e tipos de texto. Corlega está principalmente deseñado para o desenvolvemento e avaliación de clasificadores automáticos.

Ademais de describir o propio proceso de creación do corpus e das súas características, neste artigo presentamos o uso deste novo recurso para adestrar modelos computacionais de clasificación automática de textos por niveis de lecturabilidade en galego. Exploramos tres estratexias: (1) modelos de base con clasificadores clásicos que empregan características textuais de superficie; (2) métodos híbridos que incorporan nos modelos de base información extraída de modelos Transformers (Vaswani et al., 2017) preadestrados (vectores de documento e *surprisal*); (3) axuste (*fine-tuning*) de modelos Transformers, tanto monolingües como multilingües, para os que adicionalmente avaliamos o impacto do uso de datos doutras linguas. Usando só recursos en galego, os modelos híbridos obtiveron mellores resultados, mais estes foron superados polo axuste de modelos multilingües que incorporaron corpus doutras linguas no adestramento.

En resumo, este traballo achega fundamentalmente as seguintes contribucións ao campo:

- Unha nova proposta de seis niveis de lecturabilidade en galego baseada na definida polo proxecto iRead4Skills, contrastada e ampliada cos descritores do Celga (Certificado de Estudos de Lingua Galega) e do MCER (Marco Común Europeo de Referencia para as Linguas).
- O primeiro corpus, até onde sabemos, de textos en galego clasificados por niveis de lecturabilidade.
- Unha análise das características dos textos en galego con correlación co nivel de lecturabilidade.
- Experimentos con clasificadores Random Forest adestrados tanto con características textuais como incorporando representacións de modelos Transformers.
- Experimentos de axuste de modelos Transformers monolingües en galego.

- Experimentos de axuste de modelos Transformers multilingües, tanto con corpus en galego como con estratexias de aumento de datos con recursos doutras linguas románicas.

Tanto Corlega como os resultados destes experimentos serven como unha referencia para traballo futuro sobre a clasificación automática da lecturabilidade en galego. O corpus está publicado na plataforma Zenodo[1], e distribúese para fins de reproducibilidade e investigación coa licenza CC BY-NC-ND 4.0. Unha selección dos mellores modelos está dispoñible en Hugging Face.[2]

A seguir a esta introdución, a Sección 2 presenta unha panorámica sobre corpus relacionados dispoñibles para outras linguas, así como ferramentas e modelos de clasificación automática. O proceso de creación do corpus e os seus resultados preséntanse na Sección 3, mentres que a Sección 4 mostra os experimentos de clasificación automática e os seus resultados. Por último, a Sección 5 conclúe este traballo e traza liñas para a investigación futura.

2. Traballos relacionados

Nesta sección realizamos unha revisión de traballos sobre a lecturabilidade e tarefas relacionadas, tanto do punto de vista dos recursos textuais como de ferramentas computacionais. Ao longo do texto facemos referencia a traballos sobre simplificación textual, tarefa cuxo obxectivo é modificar un texto para facelo máis doado de comprender, ben sexa mediante simplificación léxica ou sintáctica (Saggion, 2017).

2.1. Corpus e conxuntos de datos

No tocante aos corpus de complexidade textual que se poden usar para estudar a lecturabilidade, nesta sección centrarémonos principalmente nos traballos en inglés (por ser a lingua con máis recursos) e en corpus doutras linguas románicas próximas ao galego, con referencias ao euskera pola calidade dos seus recursos. Estes conxuntos de datos preséntanse agrupados pola tipoloxía de clasificación que conteñen.

Clasificación en dous ou tres niveis: Hai numerosos traballos que seguen este modelo de clasificación en textos fáciles e difíciles, empregando eventualmente unha terceira categoría intermedia. Moitos deles son empregados para tarefas de simplificación textual. O corpus CohMetrix-Esp (Quispesaravia et al., 2016) contén 100 textos en castelán: 50 fábulas para nenos (fáciles) e 50 historias para adultos (difíciles). Simplext (Saggion et al., 2011, 2015) está formado por 200 textos periodísticos en castelán e as súas adaptacións simplificadas feitas manualmente polo grupo de investigación DILES (UAM) (Saggion et al., 2011). O corpus VikiWiki é un corpus multilingüe que inclúe español, catalán, vasco, inglés, francés e italiano e está formado por 5400 textos de Vikidia — unha enciclopedia dirixida a persoas cun baixo nivel de comprensión lectora que adapta artigos da Wikipedia— e polos textos correspondentes da Wikipedia (Madrazo Azpiazu & Pera, 2020). Vikidia-En/Fr (Lee & Vajjala, 2022) é un corpus paralelo que tamén contén textos de Vikidia, cunha versión simple por texto e lingua. O corpus LBSPC (Leveled Basque Science Popularization Corpus, (Gonzalez-Dios et al., 2014)) é un corpus en vasco que inclúe 400 textos extraídos da revista científica Elhuyar Aldizkaria e da web de divulgación científica ZerNola, clasificados en dous niveis (Bengoetxea & Gonzalez-Dios, 2021). Outro corpus en vasco, o corpus de textos simplificados (ETSC), inclúe 227 frases orixinais do ámbito da divulgación científica e dúas versións simplificadas de cada unha (Gonzalez-Dios et al., 2018). READ-IT (Dell'Orletta et al., 2014) é un corpus formado por 638 textos periodísticos en italiano con dúas versións: orixinal e lectura fácil. Por último, o corpus OneStopEnglish (Vajjala & Lučić, 2018) contén 189 artigos de xornais de The Guardian, en inglés, rescritos por profesores para adaptalos a tres niveis.

Seguindo o Marco Común Europeo de Referencia para as Linguas (MCER): Seguindo os niveis propostos no MCER, destacamos o conxunto de datos en portugués creado por Ribeiro et al. (2024), con textos extraídos de exames do Instituto Camões. Con características similares encontramos recursos para o francés (FLE-CORP, Wilkens et al. (2022)) para o italiano (CELI, Grego Bolli et al. (2017)), ou para o castelán, extraídos de kwiziq e HablaCultura (Vásquez-Rodríguez et al., 2022) e que conteñen artículos libres destas webs etiquetados. Tamén para o castelán, o Corpus de Aprendices de Español (CAES, Rojo & Palacios Martínez (2016)) contén 6561 textos de diversas temáticas e xéneros textuais na versión actual (2.1), escritos por aprendices de español dos diferentes niveis e de diversas nacionalidades.

[1] https://zenodo.org/records/15441342
[2] https://huggingface.co/sandrarrey

Clasificación por idade recomendada de lectura: Esta categoría inclúe recursos recomendados para diferentes idades cuxa clasificación pode empregarse como indicador da complexidade lingüística. Aquí encontramos recursos para o inglés como o corpus Weekly Reader (Schwarm & Ostendorf, 2005), WeeBit (Vajjala & Meurers, 2012), ou os textos da ferramenta TextEvaluator (Sheehan et al., 2013), ou para o francés, como o corpus Bibebook (Hernandez et al., 2022).

Clasificación por cursos escolares: De maneira similar á categoría anterior, varios corpus clasifican os textos en función do curso escolar para o que serían recomendados. Nesta categoría existen traballos para o inglés (CLEAR: CommonLit Ease of Readability corpus, Heintz et al. (2022)), para o francés (FLM-CORP, Wilkens et al. (2022) e JeLisLibre, Hernandez et al. (2022)), ou para o castelán, onde destacamos o corpus Newsela, composto por 1130 novas rescritas por tradutores en catro niveis de complexidade (Xu et al., 2015). O corpus GRERLI (GRERLI Corpus) está formado por dous subcorpus —un en español de 320 textos e outro en catalán de 316— de textos de catro xéneros e catro niveis: 4º de primaria, 1º de secundaria, 1º de bacharelato e adultos maiores de 20 anos. Tamén para catalán, o corpus CesCa (El Català Escolar Escrit a Catalunya, Martí et al. (2012)) inclúe textos escritos por alumnos ata o último ano de ensino secundario, clasificados pola idade dos participantes e polo tipo de texto (definicións, narracións de películas, recomendacións argumentativas e explicacións dun chiste).

Recursos para o galego: Ata onde sabemos, non existen corpus especificamente deseñados para a avaliación da lecturabilidade en galego. Con todo, poden atoparse recursos útiles en tarefas próximas, como a simplificación textual e léxica, a complexidade ou a comprensión lectora (Ferrés et al., 2017). Entre estes destacan recursos do Proxecto Nós (Baucells et al., 2025), como PAWS-gl, corpus de paráfrases onde a lonxitude podería empregarse como indicador de dificultade, os corpus *mgsm-gl* e OpenBookQA_gl, orientados á avaliación da comprensión, ou *summarization_gl*, se asumimos que os resumos constitúen versións máis sinxelas dos textos orixinais e que polo tanto poden servir como base para tarefas de simplificación.

Como vimos, unha boa parte dos corpus existentes para as diferentes linguas foron deseñados para tarefas de simplificación textual ou están compostos en boa medida de textos para público infantil e xuvenil, o que non os fai axeitados para público adulto. Para dar resposta a esta necesidade, o proxecto iRead4Skills[3] desenvolveu un corpus multilingüe (Pintard et al., 2024a) orientado ao estudo da lecturabilidade para adultos.[4] Este conxunto de datos inclúe corpus de características similares en tres linguas: castelán, francés e portugués. Está formado por un conxunto de textos escritos (entre 2000 e 3000 por lingua) de once xéneros e numerosos subxéneros e catro niveis de complexidade, tres definidos por expertos (moi fácil, fácil e claro) e un cuarto nivel que inclúe textos máis complexos. O conxunto de datos é representativo dunha ampla variedade de xéneros e temáticas de lectura habitual e de interese para persoas adultas, e está especificamente deseñado para tarefas de clasificación de textos.

2.2. Modelos de clasificación e simplificación automática

Como se mencionou na introdución, a lecturabilidade calculouse durante décadas mediante fórmulas como Flesch (1948), SMOG (McLaughlin, 1969) ou Dale-Chall (Chall & Dale, 1995), que estiman a dificultade textual a partir de características superficiais como a lonxitude de palabras e oracións, ou a familiaridade do léxico (Campos Saavedra et al., 2014). Estas fórmulas aplican operacións matemáticas sinxelas sobre características observables do texto e devolven un valor numérico que representa a súa dificultade.

No entanto, estes enfoques foron progresivamente substituídos por modelos automáticos supervisados, que incorporan características lingüísticas. Algúns destes modelos extraen características definidas manualmente (Crossley et al., 2007; Daoust et al., 1996), mentres que outros aprenden representacións lingüísticas mediante técnicas de aprendizaxe automática (Dell'Orletta et al., 2014) e aprendizaxe profunda (Madrazo Azpiazu & Pera, 2019). A evolución máis recente da investigación en procesamento das linguas naturais (PLN) e intelixencia artificial deu lugar a técnicas que van desde o uso de enfoques clásicos baseados en características predefinidas (Wilkens et al., 2022), até modelos preadestrados que fan uso de representa-

[3] https://iread4skills.com/

[4] O proxecto iReadSkills (Intelligent Reading Improvement System for Fundamental and Transversal Skills Development) ten por obxectivo mellorar as competencias lectoras da poboación adulta cun nivel baixo de alfabetización creando un sistema intelixente que analice a complexidade textual e forneza lecturas axeitadas para este perfil de persoas, favorecendo así o seu acceso á información e á cultura.

cións distribuídas das palabras (Martinc et al., 2021), e modelos híbridos que combinan aprendizaxe automática clásica con técnicas de aprendizaxe profunda (Lee et al., 2021; Wilkens et al., 2024). Estes modelos precisan de corpus anotados segundo niveis de complexidade (Vajjala, 2022; Dell'Orletta et al., 2014) e ofrecen unha avaliación máis precisa da lecturabilidade que as fórmulas tradicionais.

Destes novos enfoques xurdiron traballos de investigación coa fin de desenvolver ferramentas capaces de medir o nivel de lecturabilidade dun texto, como por exemplo clasificadores automáticos, e, tamén, de ofrecer alternativas destes textos máis ou menos complexas, é dicir, ferramentas de simplificación textual. Entre estas ferramentas encóntranse o clasificador e avaliador TextEvaluator (Sheehan et al., 2013), o analizador híbrido InterpretARA (Zeng et al., 2024), o simplificador léxico para linguas ibéricas desenvolvido por Ferrés et al. (2017) ou o sistema de simplificación automática Simplext (Saggion et al., 2015). Outros traballos de investigación inclúen estudos diversos para o inglés, como por exemplo a influencia de novas características baseadas na fonética para a clasificación de textos por nivel educativo (Pinney et al., 2024), e para o francés, como por exemplo o léxico de francés como lingua estranxeira por niveis FLELEx (François et al., 2014) ou o analizador FABRA e traballos derivados (Wilkens et al., 2022, 2024).

Entre os traballos para o desenvolvemento e avaliación de modelos para o portugués encontramos modelos baseados en características lingüísticas (Curto et al., 2015) e en aprendizaxe profunda (Ribeiro et al., 2024). Outras ferramentas para linguas máis distantes son o clasificador ErreXail para o vasco (Gonzalez-Dios et al., 2014), ou MultiAzterTest, un clasificador multilingüe con múltiples niveis de lecturabilidade en inglés, castelán e vasco (Bengoetxea & Gonzalez-Dios, 2021). Por último, existen ferramentas deseñadas para dominios específicos, como por exemplo a ferramenta multilingüe orientada á clasificación de textos da Wikipedia en varias linguas, incluído o galego, presentada por Trokhymovych et al. (2024).

Tendo en conta a carencia tanto de recursos como de ferramentas para o galego, neste traballo presentamos un corpus clasificado en catro niveis de lecturabilidade e deseñado con base nas directrices do proxecto iRead4Skills, e avaliamos diferentes estratexias para o adestramento e avaliación de clasificadores automáticos para esta lingua.

3. Proceso de creación do Corlega

O Corlega é un corpus de lecturabilidade que contén 480 textos en galego. Está dirixido (i) a persoas adultas nativas que non se comuniquen en galego con frecuencia e que queiran mellorar as súas competencias, e (ii) a aprendices de galego nativos doutras linguas romances. Os textos están clasificados en niveis segundo o seu grao de lecturabilidade e están distribuídos en 11 categorías e 36 subcategorías, segundo o xénero ou subxénero textual, tipo de texto, área temática ou dominio. Os niveis están definidos a partir da clasificación proposta no proxecto iRead4Skills e adaptados ao galego seguindo as clasificacións existentes para o ensino e a aprendizaxe do galego.

3.1. Niveis

Para construír o corpus, definimos seis niveis de lecturabilidade que funcionan tamén como unha primeira proposta inicial na investigación nesta área en galego. Os criterios para a definición dos niveis baséanse en boa medida nos empregados polo proxecto iRead4Skills, tendo en consideración particularidades do galego. iRead4Skills combina diferentes propostas de avaliación da competencia lingüística para definir tres niveis de complexidade textual (para o portugués, o francés e o castelán):[5] *Very Easy*, *Easy*, e *Plain* (Monteiro et al., 2024)[6]. Adicionalmente, incorpora un cuarto nivel máis heteroxéneo (*+Complex*) onde son agrupados aqueles textos que conteñen elementos de maior complexidade que os definidos nos descritores dos tres niveis referidos.

Para definirmos os tres niveis iniciais, fixemos unha selección das características definitorias comúns ás tres linguas de iRead4Skills, así como das específicas dalgunha delas que aplican tamén ao galego (por exemplo, o uso do *infinitivo pessoal* en portugués e galego (chamado "infinitivo conxugado"). Estas características foron contrastadas e ampliadas coas especificacións, especialmente sobre comprensión lectora, dos niveis Celga, definidos para o galego polo Centro Ramón Piñeiro para a Investigación en Humanidades (CRPIH) (Fidalgo et al., 2006), e do MCER, valéndonos da implementación reali-

[5] MCER: https://www.coe.int/en/web/common-european-framework-reference-languages, Programme for the International Assessment of Adult Competencies (PIAAC): https://www.oecd.org/en/about/programmes/piaac.html e Association of Language Testers in Europe (ALTE): https://www.alte.org/

[6] https://zenodo.org/records/10459090

zada por unha Escola Oficial de Idiomas (EOI de Ponferrada, (Escuela Oficial de Idiomas de Ponferrada, 2023)), xa que presenta unha relación moi útil de contidos e competencias necesarias para a aprendizaxe do galego seguindo os niveis do MCER. Ademais, a adaptación dos niveis propostos neste artigo enriqueceuse cunha análise de contidos dos manuais *Aula de galego* (Chamorro et al., 2008) feita pola autora principal deste artigo. Os niveis propostos restantes (4, 5 e 6), foron deseñados empregando os mesmos recursos e metodoloxía, mais sen ter como guía ningunha análise previa.[7]

O Cadro 1 mostra as equivalencias entre os niveis propostos neste traballo e outras clasificacións. A equivalencia entre os niveis do MCER, Celga e ALTE vén dada polo CRPIH, e entre estes e os do iRead4Skills foi establecida polo propio proxecto (Monteiro et al., 2024).

Proposta	**iR4S**	**MC**	**CG**	**ALTE**
Nivel 1	*very easy*	A1	-	-
Nivel 2	*easy*	A2	1	L. 1
Nivel 3	*plain*	B1	2	L. 2
Nivel 4	*+Complex*	B2	3	L. 3
Nivel 5	*+Complex*	C1	4	L. 4
Nivel 6	*+Complex*	C2	5	L. 5

Cadro 1: Equivalencia dos niveis do corpus con outras clasificacións. *iR4S* son os niveis definidos en iRead4Skills, *MC* no MCER, e *CG* no Celga.

Como amosa a táboa, non hai un nivel Celga equiparable ao básico 1 ou A1 do MCER, o que supuxo unha maior dificultade para definir as características dese nivel. Isto pode deberse a que a maioría das persoas que aprenden galego en centros de estudo regrados xa teñen certa competencia nalgunha lingua de orixe románica próxima ao galego. Este perfil de estudantes comezaría a estudar a lingua no nivel básico 2 ou A2, como sucede, por exemplo, con persoas galegas que deciden estudar portugués. Porén, para solventar esta falta de recursos, a definición deste nivel baseouse na presentada pola EOI de Ponferrada e na análise de materiais didácticos por parte da autora principal do artigo. Algúns destes materiais proceden de cursos introdutorios de galego, como pode ser a guía de acompañamento lingüístico da Xunta de Galicia *Falamos* (Conde et al., 2023) ou seccións do manual *Aula de galego 1* (Chamorro et al., 2008). O nivel desta última é equivalente a un nivel básico 2 (A2), que foron considerados de nivel básico 1 (A1) tanto por docentes que compartiron voluntariamente estes recursos coa autora do artigo como pola propia autora.

As descricións dos niveis resultantes da adaptación feita preséntanse nunha táboa coa seguinte estrutura: descrición de tipos de textos incluídos en cada nivel, características léxico-conceptuais, verbais, sintácticas e de cohesión. Para manter compatibilidade cos recursos producidos no proxecto iRead4Skills, nesta primeira versión do corpus clasificamos os textos en galego en catro niveis: Primeiro nivel (N1), segundo nivel (N2), terceiro nivel (N3) e nivel máis complexo (N4). Este último non se considera un nivel propiamente dito, senón que é unha agrupación de textos cun grao de dificultade maior que o do nivel 3.

Incluímos no Cadro 2 unha versión resumida cos aspectos máis relevantes dos niveis que empregamos no corpus. A taxonomía completa de niveis coas características detalladas de cada nivel pódese consultar no Anexo A.

3.2. Elaboración do corpus

Guía de elaboración: Para comezar, deseñouse unha guía de elaboración do corpus onde se definen as categorías e subcategorías nas que se clasifican os documentos e o número mínimo de textos por subcategoría e nivel, así como a súa lonxitude ideal.

Co obxectivo de incluírmos textos dunha ampla variedade de xéneros e dominios textuais, empregamos as categorías do proxecto iRead4Skills (Pintard et al., 2024a) cunha única variación, a supresión da categoría "Académico", xa que esta incluía textos de subxéneros propios dun nivel máis complexo, como son as memorias de proxectos, os artigos científicos ou os traballos de fin de grao, mestrado ou doutoramento. Para cada categoría foron adicionalmente definidas subcategorías, que se poden ver no Cadro 3. As subcategorías tamén son unha adaptación das presentes no iRead4Skills, de entre as que foron seleccionadas as máis frecuentes e as que máis se diferencian entre elas. As máis similares en xénero ou temática foron agrupadas. Ademais, foron adaptadas ao contexto galego: por exemplo, ao non existiren prospectos de medicamentos nesta lingua, definimos subcategorías similares, como "manual médico".

Para ter unha cantidade mínima de textos por subcategoría e nivel, valorando a cantidade de textos nos corpus de iRead4Skills e adaptándoo ao que podería compilar unha persoa nun perio-

[7] A definición de seis niveis —e non tres— ten como obxectivo a súa aplicación en traballo futuro, tanto do punto de vista da ampliación do Corlega, como da creación de novos recursos adaptados a estes niveis que permitan deseñar novas ferramentas.

	N1	**N2**	**N3**
Tipos de textos	Cotiáns, de actividades diarias.	Persoais, de traballo e ocio, informativos moi simples.	Informativos, de opinión e do ámbito académico e profesional propio.
Léxico-conceptual	Vocabulario cotiá, afixos máis frecuentes, adverbios máis comúns.	Vocabulario de actividades e de sentimentos, saúde, tempo, etc., interxeccións, adverbios de inclusión e dúbida.	Vocabulario laboral e de interese xeral (tecnoloxía, medio natural, economía, sociedade...)
Verbal	Verbos copulativos, regulares frecuentes, irregulares máis frecuentes, tempos de indicativo e perífrases máis habituais.	Regulares, irregulares frecuentes, infinitivo, xerundio e participio, presente de subxuntivo e perífrases habituais.	Voz pasiva, imperativo, perífrases verbais de diversos valores.
Estrutura sintáctica	Coordinadas e subordinadas simples, con conxuncións máis frecuentes, temporais con "cando" e comparativas e superlativas básicas.	Coordinadas copulativas, adversativas e disxuntivas, oracións impersoais simples e imperativas.	Coordinadas distributivas e explicativas, uso de "se" en impersoais, pasivas e verbos reflexivos, subordinadas con subxuntivo e infinitivo, adxectivas, substantivas e adverbiais, e estilo indirecto.
Cohesión	Preposicións máis frecuentes, pronomes con referente claro, elipse de elementos moi claros, marcadores de espazo e tempo máis frecuentes.	Conectores e enlaces máis frecuentes, marcadores "entón", "logo", "por outra banda"...	-

Cadro 2: Características principais dos tres niveis de lecturabilidade empregados no corpus. Cada nivel superior é acumulativo: mantén as características dos anteriores e incorpora outras adicionais, indicadas na táboa.

do de tempo similar, seleccionamos un mínimo de 12 textos por subcategoría para os tres niveis definidos (catro textos cada un), ampliándose ao incluírmos o nivel adicional (N4). Dentro do posible, intentouse incluír textos de todos os niveis en cada unha das subcategorías, aínda que hai casos onde isto non foi factible (por exemplo, non encontramos textos de medios de comunicación en galego de Nivel 1).

A lonxitude dos textos é, por regra xeral, de entre 100 e 500 palabras, agás nalgúns xéneros textuais nos que non sexa posible polas súas características, como por exemplo as mensaxes persoais ou os anuncios publicitarios. O obxectivo de establecer estes límites de extensión é que a lonxitude do texto non cree un nesgo na clasificación dos textos (é dicir, textos curtos en nivel baixo e textos longos en nivel alto), algo frecuente en tarefas de avaliación da complexidade textual. Por último, creouse unha base de datos para incluír os metadatos de cada documento, entre os que se encontran autor, tradutor, lingua, editor, lugar e data de publicación, identificador do documento, URL, Copyright, data de consulta, nivel de dificultade, xénero textual, temática e número de palabras.

Compilación: Os textos foron extraídos maiormente de fontes en liña de xeito manual, e foron revisados, corrixidos e compilados en formato de texto e en pdf. Unha vez compilados os textos de cada categoría e subcategoría, revisouse a cantidade de textos por nivel, resultando, como estaba previsto, nunha notable menor cantidade de textos de nivel 1. Para compensar esta carencia de recursos, solicitáronse textos deste nivel procedentes do ámbito educativo a diferentes institucións e docentes de galego. Os textos obtidos incluíronse nunha nova categoría chamada *didáctico* para diferencialos do resto de textos, por seren textos deseñados para a aprendizaxe do galego para non galegofalantes. Por último, clasificáronse como N4 textos de niveis de lecturabilidade superiores a N3. A cantidade de textos por

Categoría	Subcategoría
Comunicación persoal	Mensaxe Diario e blog
Comunicación profesional	Mensaxe Nota prensa Web
Medios de comunicación	Biografía Entrevista Horóscopo Nova Opinión Reportaxe Tempo Div. científica
Publicidade e difusión	Anuncio Etiqueta Instrucións Menú Manual médico
Literatura didáctica	Dicionario Autodidáctico
Literatura ficción	Novela Poesía Teatro
Literatura non ficción	Biográfico Crónicas Diario de viaxe Diario e cartas Guía turística
Política	Discurso Programa
Legal e administrativo	Legal Administrativo
Relixión	Ensinanzas Escrituras
Didáctico	-

Cadro 3: Listaxe de categorías e subcategorías do corpus.

nivel, categoría e subcategoría poden verse no cadro 4. A extensión do corpus en número de tokens tanto total coma por niveis aparece reflectida no cadro 5 (os detalles sobre como se tokenizou o corpus están na sección 3.3).

Validación por anotadoras externas: Co obxectivo de validar a clasificación dos documentos do corpus e de analizar a dificultade desta tarefa, realizouse unha anotación adicional dun subconxunto do corpus por especialistas, con formación superior en Lingüística e Ensino de Linguas, e con experiencia na creación de contidos para o ensino de galego.

	N1	N2	N3	N4	Total
Com. persoal	4	8	8	5	25
Com. profes.	8	12	12	5	37
Medios com.	0	32	32	5	69
Publicidade	20	22	23	5	70
Lit. didác.	8	8	8	5	29
Lit. ficción	0	12	12	5	29
Lit. non fic.	0	15	26	7	48
Política	0	8	8	8	24
Legal/admin.	4	4	5	5	21
Relixión	8	8	8	5	29
Didáctico	88	11	0	0	99
Total	**140**	**140**	**145**	**55**	**480**

Cadro 4: Textos do corpus por niveis e categorías. Cómpre destacar que o Nivel 4 contén menos textos (e tokens, Cadro 5) por ser un nivel non definido que sirve como referencia para agrupar textos máis complexos que os definidos.

	Tokens totais	Media tokens
Nivel 1	19085	136,32
Nivel 2	42147	301,05
Nivel 3	60322	416,01
Nivel 4	24300	441,82
Total	**145854**	**303,86**

Cadro 5: Extensión en tokens total e por niveis.

Seleccionáronse 80 textos do corpus, a cifra equivalente ao 20 % do corpus (de 420 textos cos tres niveis). Os textos foron elixidos de maneira aleatoria mais mantendo un número de documentos similar por cada nivel. Nesta selección excluímos documentos de nivel 4 (por non ter descritores definidos) e aqueles de nivel 1 que foran fornecidos por docentes e institucións (por estaren xa revisados por especialistas), e creáronse 8 formularios con 10 textos cada un. A pesar de incluírmos só textos dos niveis 1 ao 3, o formulario permitía clasificar o texto como de maior complexidade (nivel 4), o que dificulta máis a tarefa de validación.

Posteriormente, calculamos o acordo entre pares de anotadoras mediante Cohen's κ (con resultados de entre 0,142 e 0,411) e entre o conxunto de textos validados, empregando neste caso Krippendorff's α ordinal, cun resultado de $\alpha = 0,567$. Para obter estes resultados, eliminamos as respostas dos tres anotadores cun acordo notablemente inferior ($\kappa < 0,118$), pois consideramos que a súa interpretación das directrices da tarefa

foi potencialmente diferente e que iso repercutía nuns resultados pouco representativos. Os resultados obtidos indican unha concordancia relativamente baixa entre as anotadoras, mais similar a outros traballos (Pintard et al., 2024a,b). Estes resultados poden deberse, como noutros estudos, á dificultade intrínseca na clasificación de textos reais (non creados especificamente para o ensino de linguas), ademais de á propia configuración do experimento, onde se incluíron catro categorías para validar documentos de tres niveis.

3.3. Características dos textos

O corpus foi convertido a formato CoNLL-U[8] empregando ferramentas de PLN e *scripts* propios: FreeLing (Padró, 2012) para identificar oracións, tokenizar e realizar análise morfosintáctica, e Stanza (Qi et al., 2020) para realizar análise de dependencias, empregando o modelo de Dependencias Universais *Galician-TreeGal* (Garcia et al., 2018). Do corpus procesado extraéronse estatísticas sobre características definitorias da lecturabilidade dos textos segundo estudos previos (Pilán, 2018; Wilkens et al., 2022; Curto et al., 2015; Vajjala & Lõo, 2014), como a lonxitude dos textos, oracións ou palabras, a porcentaxe de cada categoría morfosintáctica, etc.

Unha vez obtidos estes valores, computamos a correlación entre cada un deles e os niveis definidos no corpus, empregando Spearman ρ. Como era esperado (Bengoetxea & Gonzalez-Dios, 2021), as características con maior correlación son as que están relacionadas coa lonxitude, xa que os textos máis simples tenden a ser tamén máis curtos. Así, destácanse o número de lemas, tokens e palabras (con valores de $\rho = 0,66$, $\rho = 0{,}64$ e $\rho = 0{,}65$, respectivamente), ou a cantidade de tokens diferentes ($\rho = 0{,}64$) e vírgulas ($\rho = 0{,}55$)[9]. Porén, se non tivermos en conta valores absolutos mais si normalizados, as características con maior correlación cos niveis son a media de tokens e de palabras por frase ($\rho = 0{,}46$ nos dous casos), a distancia media de dependencias sintácticas ($\rho = 0{,}32$), e o tamaño medio de cada token ($\rho = 0{,}19$), característica xa empregada nas fórmulas tradicionais para a lecturabilidade e en traballos previos (Vajjala & Meurers, 2012; Bengoetxea & Gonzalez-Dios, 2021).

[8] https://universaldependencies.org/format.html

[9] O número de vírgulas pode indicar de maneira indirecta unha maior complexidade sintáctica, por usárense para indicar elementos coordinados, oracións parentéticas, alteración de elementos na oración, cláusulas de oracións compostas, etc.

4. Experimentación

Nesta sección presentamos un conxunto de experimentos para o adestramento de clasificadores automáticos de textos segundo o seu nivel de lecturabilidade. Primeiro, formulamos unha serie de preguntas de investigación e hipóteses. A seguir, describimos os experimentos realizados con diferentes modelos, así como os resultados de cada método.

4.1. Preguntas de investigación e hipóteses asociadas

Antes de proceder, definimos as seguintes preguntas de investigación (Ps), hipóteses (Hs), e experimentos (Es):

- P1: É posible axustar cun corpus de pequeno tamaño un modelo Transformers para clasificar textos en función do seu nivel de lecturabilidade en galego?

  H1: Con base en traballos similares (Vásquez-Rodríguez et al., 2022; Ribeiro et al., 2024) esperamos que o axuste dun modelo Transformers co corpus aquí presentado consiga resultados satisfactorios nesta tarefa, xa que pode aproveitar coñecemento adquirido na fase de preadestramento.

  E1: Para dar resposta a esta pregunta de investigación e validar ou refutar a H1, realizaremos múltiples experimentos de axuste de modelos monolingües en galego co corpus que presentamos neste traballo.

- P2: Para a tarefa anterior, terá mellor desempeño o axuste de modelos monolingües, ou de modelos multilingües que inclúan galego? Noutras palabras, o coñecemento lingüístico doutras linguas ten un impacto positivo ou negativo no modelado da lecturabilidade nun contexto monolingüe?

  H2: Tendo en conta que as características da complexidade textual son en boa medida independentes da lingua[10] (Monteiro et al., 2024; Imperial et al., 2025), é probable que os modelos multilingües funcionen mellor que os monolingües nesta tarefa. Pola contra, o feito de que os modelos monolingües teñan unha tokenización adaptada á lingua de destino pode favorecer o seu desempeño nesta tarefa.

  E2: Para responder a esta pregunta faremos diversos experimentos de axuste de modelos

[10] A independencia da lingua refírese ás características estruturais, variando os valores limiares destas en función de cada lingua.

multilingües co corpus de galego e os compararemos cos resultados obtidos con modelos monolingües.

- P3: Os modelos multilingües poden mellorar o seu desempeño cun maior volume de datos procedentes de linguas próximas?

  H3: Tendo en conta traballos previos (Lee & Vajjala, 2022; Madrazo Azpiazu & Pera, 2020), esperamos que os modelos multilingües funcionen mellor ao seren adestrados cun volume moito máis amplo de textos clasificados, aínda que estes sexan doutras linguas.

  E3: Para responder a esta pregunta faremos varios experimentos cos modelos multilingües avaliados, engadindo nos datos de adestramento varios corpus do proxecto iRead4Skills, xa que teñen niveis e características similares.

4.2. Experimentos

Como foi mencionado, realizamos varios conxuntos de experimentos con modelos Transformers. Empregamos a metodoloxía estándar de axuste, que consiste en seleccionar modelos preadestrados e axustar os seus parámetros para a clasificación de documentos con base nun corpus de adestramento. Para avaliar comparativamente o seu funcionamento, implementamos un modelo de base (*baseline*) simple empregando un pequeno conxunto de características textuais para adestrar clasificadores Random Forest. Alén diso, avaliamos modelos híbridos incorporando nestes clasificadores información extraída de representacións Transformers, o que nos permite observar o seu impacto como características adicionais e saber se a información codificada por modelos preadestrados é competitiva coas probabilidades asignadas por estes mesmos modelos axustados.

4.2.1. Conxuntos de datos

Para avaliar os diferentes modelos empregamos o Corlega. Adicionalmente, para realizar experimentos interlingüísticos (E3), usamos como datos de axuste os corpus de español, portugués e francés incluídos no Dataset 1 do proxecto iRead4Skills (Pintard et al., 2024a). Os textos foron procesados mediante un *script* propio que os converte en conxuntos de datos en formato CSV compatible coas librarías empregadas para o desenvolvemento dos modelos. O conxunto de datos de galego dividiuse de xeito automático en tres partes:[11] de adestramento (80 %), de validación (5 %), e de avaliación (15 %). Dado o número limitado de documentos de N4 no Corlega, os conxuntos de validación e avaliación foron balanceados en termos de número de documentos por clase, para asegurar unha avaliación máis informativa e equilibrada entre clases, especialmente en relación ao recoñecemento do N4.[12] A descompensación no conxunto de adestramento foi abordada mediante sobremostraxe, duplicando aleatoriamente documentos do N4 para mellorar a capacidade do modelo de recoñecer esta clase.

Dado que os datos das outras linguas só se empregan como corpus de adestramento adicional, non foi necesario dividilos. O tamaño final dos conxuntos de datos pode verse no Cadro 6.

		Nivel			
Ling.	**Parte**	**N1**	**N2**	**N3**	**N4**
GL	Adestr	112	112	116	40*
	Valid	7	7	7	7
	Aval	21	21	22	21
ES	Adestr	660	660	889	354
PT	Adestr	777	666	752	737
FR	Adestr	467	758	766	209

Cadro 6: Tamaño dos conxuntos de datos (en número de documentos) empregados para avaliar os modelos. *Indica documentos orixinais, incrementándose o número até 112 mediante sobremostraxe.

4.2.2. Modelos

Para levar a cabo os experimentos de clasificación mediante axuste de modelos Transformers empregamos os seguintes modelos preadestrados:

Codificadores: Tanto para realizar os axustes como para extraer características adicionais para os modelos de base, usamos estes modelos:

- **Monolingües:** Variantes *small* e *base* dos modelos BERT-gl (Garcia, 2021) e Bertinho (Vilares et al., 2021).
- **Multilingües:** Modelo multilingüe BERT-base (mBERT, Devlin et al. (2019)) e as variantes *base* e *large* de XLM-RoBERTa (Conneau et al., 2019).

[11] Tendo en conta o tamaño relativamente pequeno do corpus, decidimos reducir o tamaño dos datos de validación para asegurar exemplos suficientes no adestramento e na avaliación.

[12] En experimentos preliminares empregando a distribución orixinal por nivel, observáronse resultados globais superiores, mais diversas análises da precisión por cada nivel mostraban que os clasificadores raramente etiquetaban un documento como N4, polo que os modelos resultantes non terían aplicabilidade nun contexto real.

Xerativos: Adicionalmente, e só para obter características dos modelos de base, empregamos os seguintes modelos xerativos:

- **Monolingües:** Llama3.1-Carballo 8B[13], e Llama3.1-Carvalho-PT-GL 8B[14], ambos continuacións de adestramento de Llama 3.1 8B (Grattafiori et al., 2024), o primeiro nun grande corpus monolingüe galego (de Dios-Flores et al., 2024) e o segundo incluíndo tamén datos en portugués, español e inglés.[15]
- **Multilingües:** Variantes 1B e 3B da familia de modelos Llama 3.2.[16]

4.2.3. Métodos

Modelo de base: Como modelo de base contra o que comparar as adaptacións dos Transformers e as aproximacións híbridas, adestramos un clasificador Random Forest empregando características lingüísticas, seguindo traballos relacionados (Akef et al., 2024). Usamos a implementación Random Forest incluída na libraría `scikit-learn`.[17] Durante o adestramento, avaliamos o impacto de varios tamaños de *estimadores* (árbores de decisión individuais), cun mínimo de 50 e un máximo de 600.[18]

Para este modelo escollemos os dous seguintes tipos de características baseadas na lonxitude: (i) tamaño da oración e (ii) tamaño da palabra.[19] De cada unha delas seleccionamos catro valores (media, mediana, mínimo e máximo), obtendo clasificadores simples de 8 características.

Modelos híbridos: Adicionalmente, adestramos modelos híbridos incorporando ao modelo de base dous tipos de características extraídas de modelos Transformers:

- Vector do documento: como característica adicional para representar cada texto, incorporamos un vector de documento que é a media dos vectores de todas as oracións do documento. Os vectores das oracións foron previamente construídos facendo a media dos vectores de todos os subtokens de cada oración (Imperial, 2021). Este vector de documento concaténase ao vector das oito características definidas para o modelo de base.
- *Surprisal:*[20] Inspirándonos no uso da perplexidade de modelos de lingua para ordenar documentos segundo o nivel de complexidade (Martinc et al., 2021), exploramos o impacto da *surprisal* na clasificación. Neste caso obtemos a *surprisal* dun documento usando a libraría *minicons*[21] (Misra, 2022). Para calcular a *surprisal* dos modelos bidireccionais (codificadores), empregamos o método proposto por Kauf & Ivanova (2023), que ten en conta palabras descoñecidas e expresións multipalabra. En todos os casos obtemos catro valores de surprisal (media, mediana, mínima e máxima), que son concatenados ao vector de características de cada documento.

Axuste de modelos: Fixemos diversos experimentos de axuste de modelos codificadores (Sección 4.2.2) empregando a libraría *Transformers* de Hugging Face (Wolf et al., 2020). Realizamos tres tipos de experimentos (véxase a Sección 4.1): (i) axuste de modelos monolingües en galego co corpus de galego; (ii) axuste de modelos multilingües só co corpus de galego, e (iii) axuste de modelos multilingües aumentando os datos con corpus doutras linguas (Sección 4.2.1). Neste último caso, comparamos o desempeño dos modelos incluíndo unha lingua adicional (portugués, español ou francés), dúas (español e portugués, por seren as máis próximas do galego), e co conxunto completo dos datos.

No axuste dos modelos exploramos os seguintes hiperparámetros: tres tamaños de *batch* (8, 16 e 32) e 5 valores de *learning rate* (de 1e-5 a 5e-5) durante 5 épocas, até escoller a mellor configuración en cada caso.[22] Tanto nos modelos de base como no axuste dos modelos Transformers, o adestramento foi realizado usando as particións de adestramento e avaliando o impacto de dife-

[13] https://huggingface.co/proxectonos/Llama-3.1-Carballo

[14] https://huggingface.co/Nos-PT/Llama-Carvalho-PT-GL

[15] Estes modelos non son portanto estritamente monolingües, mais dado que a maior parte do seu corpus de adestramento é galego diferenciámolos aquí dos multilingües que conteñen unha variedade moito máis ampla e balanceada de linguas.

[16] https://huggingface.co/collections/meta-llama/llama-32-66f448ffc8c32f949b04c8cf.

[17] https://scikit-learn.org

[18] En xeral, os mellores resultados déronse con 400 ou 500 estimadores.

[19] Seleccionamos estas características por seren as dúas características de superficie con maior correlación cos niveis do corpus (véxase a Sección 3.3) para as que non é necesario o uso de ferramentas de análise automática.

[20] A *surprisal* é o logaritmo negativo da probabilidade que un modelo de lingua lle asigna a unha palabra en contexto (Levy, 2008), e empregouse tanto para predicir tempos de lectura a nivel de palabra (Wilcox et al., 2023) como como medida de aceptabilidade gramatical a nivel de oracións (Lee & Vu, 2024).

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

[22] Por limitacións da GPU, os experimentos co modelo XLM-large con *batches* de 16 e 32 non se puideron realizar co conxunto completo dos datos.

rentes configuracións de hiperparámetros no conxunto de validación. Para cada modelo, seleccionamos a configuración con mellor desempeño na validación para obter os resultados no conxunto de avaliación.[23]

4.3. Resultados e discusión

O Cadro 7 mostra os resultados do modelo de base e dos modelos híbridos, mentres que o Cadro 8 os do axuste de modelos Transformers. O modelo de base, un clasificador simple con só 8 características de superficie, obtivo unha exactitude de 0,42. Nos métodos híbridos, a incorporación de información extraída de modelos Transformers tivo sistematicamente impacto positivo (salvo nun caso, adicionando *surprisal* de BERT-base), e a combinación das características de superficie cos vectores de documento de BERT-small obtivo os mellores resultados (0,55).

Cando comparamos o efecto da incorporación dos vectores de documento e da *surprisal* observamos que os primeiros teñen un maior impacto no desempeño dos modelos, aumentando a exactitude media de todos os modelos en 0,08 (frente a un 0,02 da *surprisal*). Por último, a combinación de todas as características produciu resultados diverxentes: en 5 modelos a exactitude con *Todas* foi menor que as variantes con vectores, en 3 casos foi ao revés, e noutros 3 os resultados foron equivalentes. Polo tanto, a incorporación da *surprisal* a modelos con vectores de documento non ten un impacto positivo de maneira sistemática.

A propósito das propiedades dos modelos, os resultados suxiren que o tamaño non parece ter un efecto claro no desempeño, xa que a comparación entre modelos *small* e *base* (BERT e Bertinho), ou entre *base* e *large* (XLM-RoBERTa) non segue as mesmas tendencias, sendo os resultados moi similares e con variacións non correlacionadas coa complexidade do modelo. Porén, se compararmos os codificadores e xerativos, os resultados son lixeiramente máis claros, e indican que os modelos xerativos (tamén de moito maior tamaño) obteñen, en xeral, resultados máis baixos que os codificadores.[24] Así, o valor máis alto

Modelo	**Caract.**	**Exact.**
Base	Tamaño (4+4)	0,42
BERT-small	Vectores	**0,55**
	Surp. (4)	0,44
	Todas	*0,54*
BERT-base	Vectores	0,52
	Surp. (4)	0,40
	Todas	0,53
Bertinho-small	Vectores	0,53
	Surp. (4)	0,44
	Todas	*0,51*
Bertinho-base	Vectores	0,54
	Surp. (4)	0,44
	Todas	*0,51*
mBERT	Vectores	0,49
	Surp. (4)	0,44
	Todas	0,49
XLM-base	Vectores	0,48
	Surp. (4)	0,45
	Todas	*0,47*
XLM-large	Vectores	0,46
	Surp. (4)	0,44
	Todas	0,47
Llama-Carballo	Vectores	0,48
	Surp. (4)	0,47
	Todas	0,48
Llama-Carvalho	Vectores	0,47
	Surp. (4)	0,46
	Todas	0,51
Llama 3.2 1B	Vectores	0,49
	Surp. (4)	0,42
	Todas	*0,46*
Llama 3.2 3B	Vectores	0,48
	Surp. (4)	0,47
	Todas	0,48

Cadro 7: Exactitude (*Exact.*) do modelo de base (liña superior) e dos modelos híbridos (inferiores) no corpus galego. O modelo *base* emprega as 8 características baseadas no tamaño das oracións (4) e das palabras (4). Cada un dos restantes clasificadores híbridos incorpora as características adicionais dos modelos transformers: *Vectores*, *Surprisal*, e a combinación de todas as anteriores (*Todas*). Os 7 modelos centrais son codificadores, mentres que os 4 modelos da parte inferior son xerativos. Os valores subliñados indican desempeños inferiores ao modelo de base, en itálica aqueles que obteñen menor exactitude con *Todas* que só con algunhas das características, e en letra grosa os mellores resultados.

[23] Para seleccionarmos os mellores modelos baseámonos na taxa de 'exactitude' (*accuracy*), que se obtén dividindo o número de documentos correctamente clasificados entre o total de documentos. O desempeño dos modelos cunha maior exactitude medímolo posteriormente usando precisión, cobertura (*recall*) e medida F (*F1-score*) en cada un dos 4 niveis.

[24] A este respecto, traballos recentes no ámbito da psicolingüística computacional mostran que as representacións obtidas dos actuais LLMs teñen unha menor correlación co procesamento lingüístico humano que modelos de menor tamaño (Oh & Linzen, 2025).

(Llama-Carvalho, con 0.51) é máis baixo que o obtido por BERT-small (0,55), a pesar de ser de menor tamaño e adestrado cun corpus de galego moito máis pequeno. En relación coas linguas de cada modelo, parece haber unha tendencia lixeiramente máis favorable para os modelos monolingües, dado que os mellores resultados tanto das variantes de BERT coma das de Bertinho superan mBERT e as dúas versións de XLM. Esta tendencia é máis clara nos codificadores que nos xerativos, onde os mellores resultados son do Llama-Carvalho, mais en xeral as catro variantes de Llama conseguen taxas de exactitude similares.

Aínda que o desempeño dos clasificadores híbridos nos adianta algunha información sobre as preguntas de investigación formuladas (Sección 4.1), estas son mellor respondidas analizando os resultados do axuste de modelos Transformers:

- P1: Os resultados máximos de 0,49, en comparación con outros traballos con desempeños moito superiores (0,78 en castelán (Vásquez-Rodríguez et al., 2022), 0,79 en estonio (Vajjala & Lõo, 2014) e entre 0,81 (Ribeiro et al., 2024) e 0,75 (Curto et al., 2015) en portugués), suxiren que o axuste de modelos monolingües cun corpus pequeno, no caso do galego, non é satisfactorio. Estes resultados poden deberse á cantidade de datos empregada, inferior aos dos experimentos mencionados. Aínda que os resultados obtidos do axuste de modelos monolingües son superiores aos do modelo de base, estes son inferiores aos dos modelos híbridos. Por tanto, parece máis axeitado extraer directamente as características dos modelos preadestrados (vectores) que facer un axuste do modelo.
- P2: Non hai tendencias claras respecto do desempeño dos modelos multilingües frente aos monolingües adestrados co corpus de galego. XLM-large obtén o mellor resultado, mais cun valor moi próximo a BERT-base (e ao resto de monolingües), de tamaño inferior. Por tanto, os modelos multilingües adestrados só cun corpus pequeno de galego tampouco parecen competitivos frente aos modelos híbridos, e obteñen resultados que distan moito doutros traballos para outras linguas, por exemplo, dun 0,78 de precisión obtido do axuste de mBERT para a mesma tarefa en castelán (Vásquez-Rodríguez et al., 2022). Os resultados poden deberse, unha vez máis, á falta de datos en galego para o axuste.
- P3: O aumento de datos de adestramento doutras linguas ten un impacto positivo para a clasificación en galego (salvo en tres casos con XLM-large). Para mBERT, a mellor combinación é coas catro linguas, e para XLM-base e XLM-large, a mellor opción é engadir portugués. Con todo, os resultados distan moito dos obtidos noutros traballos, como por exemplo os obtidos por Lee & Vajjala (2022) con precisións superiores a un 80 % en castelán e francés axustando un modelo mBERT con datos en inglés, cunha maior cantidade de datos.

Modelo	**Corpus**	**Exact.**
BERT-small	GL	0,42
BERT-base	GL	0,49
Bertinho-small	GL	0,46
Bertinho-base	GL	0,43
mBERT	GL	0,45
	GL+ES	0,54
	GL+FR	0,49
	GL+PT	0,49
	GL+ES+PT	0,50
	GL+ES+PT+FR	**0,62**
XLM-base	GL	0,39
	GL+ES	0,50
	GL+FR	0,41
	GL+PT	0,56
	GL+ES+PT	0,48
	GL+ES+PT+FR	0,55
XLM-large	GL	0,50
	GL+ES	0,53
	GL+FR	0,38
	GL+PT	0,54
	GL+ES+PT	0,47
	GL+ES+PT+FR	0,41

Cadro 8: Exactitude no conxunto de avaliación dos modelos axustados. Os catro primeiros modelos son monolingües e o resto son multilingües. Os valores subliñados indican desempeños superiores co axuste con corpus externos que só co de galego, e en letra grosa os mellores resultados.

Unha vez analizados os resultados xerais de todos os modelos no conxunto de avaliación, construímos unha matriz de confusión para observar con máis detalle como se comporta o modelo con maior taxa de exactitude. Na Figura 1 podemos ver que o modelo parece predicir de maneira robusta os documentos de nivel 1 e 2, cun desempeño lixeiramente menor nos niveis 3 e 4. Neste sentido, a principal confusión deuse entre estes dous niveis superiores (por exemplo, dez documentos de nivel 3 foron incorrectamente clasificados como N4). Estes resultados poden deberse

a que N4 non foi definido como un nivel en si mesmo, mais como unha clase que engloba documentos de maior complexidade que o nivel 3.

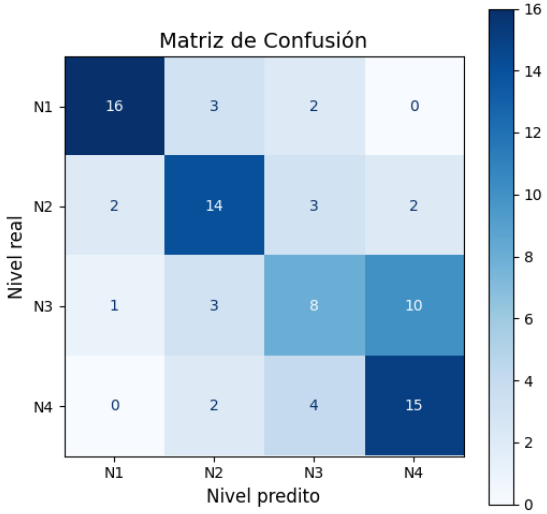

Figura 1: Distribución das predicións do modelo mBERT axustado con todos os corpus externos (GL+ES+PT+FR).

A distribución das predicións do modelo permítenos tamén obter resultados máis pormenorizados, incluíndo precisión, cobertura, e medida F por cada nivel.[25] O Cadro 9 mostra os resultados por nivel do mellor modelo (axustado) e do mellor modelo híbrido (BERT-small con vectores de documento e características superficiais). Observando a medida F, vemos que o modelo axustado ten valores superiores en dous niveis (2 e 4), inferiores no N3 (debido á alta cobertura do modelo híbrido), e iguais ao modelo híbrido para o nivel 1, onde este ten maior cobertura, mais menor precisión.

N	Híbrido			Axustado		
	P	Cob.	F1	P	Cob.	F1
1	0,69	0,95	**0,80**	0,84	0,76	**0,80**
2	0,69	0,43	0,53	0,64	0,67	**0,65**
3	0,39	0,73	**0,51**	0,47	0,36	0,41
4	1,00	0,10	0,17	0,56	0,71	**0,62**

Cadro 9: Resultados por nivel do mellor modelo axustado (mBERT con todos os corpus) e híbrido (BERT-small con vectores de documento).

En suma, este conxunto de experimentos posibilitou, por un lado, establecer un modelo de base para a clasificación automática da lecturabilidade en galego. Por outro lado, permitiunos comparar estratexias de clasificación híbridas (construídas unicamente con características de superficie e modelos preadestrados) con modelos Transformers de diferentes características e axustados con corpus de diversas linguas, onde os mellores resultados foron obtidos por modelos multilingües adestrados con estratexias de aumento de datos.

5. Conclusións e traballo futuro

Este artigo presenta Corlega, o que é, ata onde sabemos, o primeiro corpus de textos clasificados por niveis de lecturabilidade en galego. Ademais, introduce unha proposta de taxonomía que inclúe seis niveis de lecturabilidade cun conxunto de características descritivas.

Por unha banda, o corpus empregouse nunha primeira etapa para analizar a correlación dun conxunto de características lingüísticas e textuais cos niveis de complexidade definidos. Os resultados confirman estudos previos que indican que as características relacionadas coa lonxitude do texto son as que presentan maior correlación co nivel de lecturabilidade do texto, como son o tamaño das oracións e das palabras. No entanto, fenómenos como a lonxitude media das oracións en tokens ou palabras, da distancia entre dependencias e de caracteres por token amosan tamén unha correlación significativa que convén explorar en futuros traballos.

Por outra banda, o Corlega demostrou ser útil para o adestramento e avaliación de clasificadores automáticos. O adestramento de clasificadores híbridos, que incorporan características extraídas de modelos neuronais en algoritmos clásicos como Random Forest, obtivo taxas de exactitude de ata 0,55 (incrementando en 0,12 o modelo de base). Estas aproximacións híbridas obtiveron mellores resultados que os modelos Transformers monolingües axustados con datos en galego, mais foron superados ao usarmos modelos multilingües e estratexias de aumento de datos. Así, o modelo mBERT atinxiu unha taxa de exactitude de 0,62 ao adestralo con corpus de catro linguas. Unha análise do desempeño por nivel tamén forneceu información detallada sobre o comportamento do modelo, o que pode ser útil para desenvolvementos futuros que melloren este tipo de clasificación de textos en galego.

Así, deste traballo xurden novas liñas de investigación tanto na análise das características que definen a lecturabilidade dos textos como na exploración máis a fondo do axuste de modelos

[25] Onde a precisión é o número de verdadeiros positivos dividido polo total de verdadeiros e falsos positivos por clase; a cobertura é o número de verdadeiros positivos dividido polo total de verdadeiros e falsos negativos; e a medida F é a media harmónica da precisión e a cobertura.

para crear clasificadores automáticos de textos, na experimentación con novos métodos ou mesmo na ampliación do corpus. Como traballo futuro, sería interesante ampliar o conxunto de datos de adestramento con corpus de características similares traducidos automaticamente, asegurando que se manteña o nivel de lecturabilidade na tradución, e tamén ampliar o corpus con textos do resto de niveis definidos no apéndice A e observar como funcionan os mesmos clasificadores con máis niveis.

Esperamos que este traballo contribúa ao desenvolvemento de recursos e ferramentas para o ensino do galego, e que resulte útil para analizar aquelas características lingüísticas que poden influír nos procesos de aprendizaxe do galego, especialmente por parte de persoas adultas con menor hábito lector nesta lingua. Entre as posibles aplicacións, estaría o uso dos textos do corpus como materiais de lectura axeitados segundo o nivel ou o uso dalgún destes clasificadores ou outros futuros para a selección de materiais de lectura axeitados, xa sexa por parte de docentes ou de maneira autodidacta, tanto para persoas adultas que aprenden o galego coma lingua estranxeira como para persoas nativas que non se comunican en galego con frecuencia e queren mellorar as súas competencias. Ademais, no futuro espérase que estes clasificadores se complementen cun analizador que indique onde reside a complexidade do texto e un asistente de escritura con recomendacións de adaptación, que poderían ser usados para o deseño de exercicios.

Agradecementos

Un agradecemento especial a Laura Formoso Cuesta (CLM da USC) e a Marisol Ríos Noya (Servizo de Normalización Lingüística da UDC) por contribuír á creación do corpus fornecendo materiais textuais usados para o ensino do galego para adultos. Tamén a André Bernárdez Braña pola axuda técnica para a extracción de estatísticas sobre o corpus.

Agracedemos ademais a participación a todas as lingüistas que axudaron no proceso de avaliación da calidade do corpus aquí presentado: Marta Vázquez Abuín, Albina Sarymsakova, Laura Castro Sánchez, Laura Formoso Cuesta, Adina Ioana Vladu, Silvia Paniagua Suárez, Susana Sotelo Docío e Noémia Tato Abella.

Este traballo foi financiado pola Comisión Europea (Proxecto: iRead4Skills, referencia: 1010094837, convocatoria: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837), pola Consellería de Cultura, Educación, Formación Profesional e Universidades da Xunta de Galicia (Centro de investigación de Galicia acreditación 2024-2027 ED431G 2023/04 e GPC ED431B 2025/16) e a Unión Europea (Fondo Europeo de Desarrollo Regional - FEDER), polo MCIU/AEI/10.13039/501100011033 (proxectos con referencias PID2021-128811OA-I00 e CNS2024-154902) e por unha axuda *Ramón y Cajal* (RYC2019-028473-I).

Referencias

Akef, Soroosh, Amália Mendes, Detmar Meurers & Patrick Rebuschat. 2024. Investigating the generalizability of Portuguese readability assessment models trained using linguistic complexity features. En *16th International Conference on Computational Processing of Portuguese (PROPOR)*, 332–341.

Baucells, Irene, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau & Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. En *31st International Conference on Computational Linguistics*, 10491–10519.

Bengoetxea, Kepa & Itziar Gonzalez-Dios. 2021. MultiAzterTest: a multilingual analyzer on multiple levels of language for readability assessment. arXiv [cs.CL/cs.AI]. 10.48550/arXiv.2109.04870

Campos Saavedra, Daniela, Paula Contreras Carmona, Bernardo Riffo Ocares, Mónica Véliz & Alejandro Reyes Reyes. 2014. Complejidad textual, lecturabilidad y rendimiento lector en una prueba de comprensión en escolares adolescentes. *Universitas Psychologica* 13(3). 10.11144/Javeriana.UPSY13-3.ctlr

Chall, Jeanne S. & Edgar Dale. 1995. *Readability revisited: The new dale-chall readability formula*. Brookline Books

Chamorro, Margarita, Ivonete da Silva & Xaquín Núñez. 2008. *Aula de galego 1*. Xunta de Galicia

Conde, Nancy Domínguez, Martín Ramos Insua & Iria Fernández Silva. 2023. *Falamos! guía de acompañamento lingüístico*. Xunta de Galicia, Dirección Xeral de Xuventude e Voluntariado, Consellería de Traballo e Benestar

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. arXiv [cs.CL]. 10.48550/arXiv.1911.02116

Crossley, Scott Andrew, David F. Dufty, Philip M. McCarthy & Danielle S. McNamara. 2007. Toward a new readability: A mixed model approach. En *Annual Meeting of the Cognitive Science Society*, 197–202.

Curto, Pedro, Nuno Mamede & Jorge Baptista. 2015. Automatic text difficulty classifier - assisting the selection of adequate reading materials for european Portuguese teaching. En *7th International Conference on Computer Supported Education (CSEDU)*, 36–44. 10.5220/0005428300360044

Daoust, François, Léo Laroche & Lise Ouellet. 1996. SATO-CALIBRAGE: présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique* 25(1). 205–234. 10.7202/603132ar

Dell'Orletta, Felice, Simonetta Montemagni & Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *ITL - International Journal of Applied Linguistics* 165. 163–193. 10.1075/itl.165.2.03del

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. En *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186. 10.18653/v1/N19-1423

de Dios-Flores, Iria, Silvia Paniagua Suárez, Cristina Carbajal Pérez, Daniel Bardanca Outeiriño, Marcos Garcia & Pablo Gamallo. 2024. CorpusNÓS: A massive Galician corpus for training large language models. En *16th International Conference on Computational Processing of Portuguese (PROPOR)*, 593–599.

Escuela Oficial de Idiomas de Ponferrada. 2023. *Programación didáctica del departamento de gallego: Curso 2023-2024 alumnado libre*. Escuela Oficial de Idiomas de Ponferrada.

Ferrés, Daniel, Horacio Saggion & Xavier Gómez Guinovart. 2017. An adaptable lexical simplification architecture for major Ibero-Romance languages. En *1st Workshop on Building Linguistically Generalizable NLP Systems*, 40–47. 10.18653/v1/W17-5406

Fidalgo, Elvira, Margarita Chamorro, Bieito Silva, Xulio C. Sousa, ManuelÑúñez Singala, Pilar Concheiro, Yolanda López & Olga Iglesias. 2006. *Niveis de competencia en lingua galega: Descrición de habilidades e de contidos adaptados ao marco europeo común de referencia para as linguas*. Xunta de Galicia, Secretaría Xeral de Política Lingüística, Centro Ramón Piñeiro para a Investigación en Humanidades

Flesch, Rudolf. 1948. A readability formula in practice. *Elementary English* 25(6). 344–351.

François, Thomas, Nùria Gala, Patrick Watrin & Cédrick Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. En *9th International Conference on Language Resources and Evaluation (LREC)*, 3766–3773.

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

Garcia, Marcos, Carlos Gómez-Rodríguez & Miguel A Alonso. 2018. New treebank or repurposed? on the feasibility of cross-lingual parsing of romance languages with universal dependencies. *Natural Language Engineering* 24(1). 91–122. 10.1017/S1351324917000377

Gonzalez-Dios, Itziar, María Jesús Aranzabe, Arantza Díaz de Ilarraza & Haritz Salaberri. 2014. Simple or complex? assessing the readability of Basque texts. En *25th International Conference on Computational Linguistics (COLING)*, 334–344.

Gonzalez-Dios, Itziar, María Jesús Aranzabe & Arantza Díaz de Ilarraza. 2018. The corpus of basque simplified texts (CBST). *Language Resources and Evaluation* 52(1). 217–247. 10.1007/s10579-017-9407-6

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri et al. 2024. The Llama 3 Herd of Models. arXiv [cs.AI/cs.CL/cs.CV]. 10.48550/arXiv.2407.21783

Grego Bolli, Giuliana, Danilo Rini & Stefania Spina. 2017. Predicting readability of texts for italian l2 students: A preliminary study.

En *Learning and Assessment: Making the Connections – ALTE 6th International Conference*, 272–278.

GRERLI Corpus. 1998. Corpus compiled by the international project "Developing Literacy in Different Contexts and in Different Languages". Spencer Foundation, Tel Aviv University, and University of Barcelona.

Heintz, Aron, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi & Agnes Malatinszky. 2022. A large-scaled corpus for assessing text readability. *Behavior Research Methods* 55. 491–507. 10.3758/s13428-022-01802-x

Hernandez, Nicolas, Nabil Oulbaz & Tristan Faine. 2022. Open corpora and toolkit for assessing text readability in French. En *2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, 54–61.

Imperial, Joseph Marvin. 2021. BERT embeddings for automatic readability assessment. En *International Conference on Recent Advances in Natural Language Processing (RANLP)*, 611–618.

Imperial, Joseph Marvin, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego & Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. arXiv [cs.CL]. 10.48550/arXiv.2506.01419

Kauf, Carina & Anna Ivanova. 2023. A better way to do masked language model scoring. En *61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 925–935. 10.18653/v1/2023.acl-short.80

Lee, Bruce W., Yoo Sung Jang & Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 10669–10686. 10.18653/v1/2021.emnlp-main.834

Lee, Justin & Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. arXiv [cs.CL/cs.AI/cs.LG]. 10.48550/arXiv.2203.07450

Lee, So Young & Mai Ha Vu. 2024. The effects of distance on NPI illusive effects in BERT. En *Empirical Methods in Natural Language Processing (EMNLP)*, 9443–9457. 10.18653/v1/2024.emnlp-main.530

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177. 10.1016/j.cognition.2007.05.006

Madrazo Azpiazu, Ion & Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics* 7. 421–436. 10.1162/tacl_a_00278

Madrazo Azpiazu, Ion & Maria Soledad Pera. 2020. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology* 71(6). 644–656. 10.1002/asi.24293

Martinc, Matej, Senja Pollak & Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics* 47(1). 141–179. 10.1162/coli_a_00398

Martí, Antonia, Liliana Tolchinsky & Anna Llaurado. 2012. Corpus cesca compiling a corpus of written catalan produced by school children. *International Journal of Corpus Linguistics* 17. 421–448. 10.1075/ijcl.17.3.06lla

McLaughlin, G. Harry. 1969. SMOG grading - a new readability formula. *The Journal of Reading* 639–646.

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

Monteiro, Ricardo, Raquel Amaro, Susana Correia, Alice Pintard, Roser Gauchola, Michell Moutinho & Xavier Blanco Escoda. 2024. iRead4Skills - Complexity Levels. 10.5281/zenodo.10459090

Oh, Byung-Doh & Tal Linzen. 2025. To model human linguistic prediction, make LLMs less superhuman. arXiv [cs.CL]. 10.48550/arXiv.2510.05141

Padró, Lluís. 2012. Analizadores multilingües en FreeLing. *Linguamática* 3(2). 13–20.

Pilán, Ildikó. 2018. *Automatic proficiency level prediction for intelligent computer-assisted language learning*: University of Gothenburg. Faculty of Arts. Tese de Doutoramento

Pinney, Christine, Casey Kennington, Maria Soledad Pera, Katherine Landau Wright & Jerry Alan Fails. 2024. Incorporating word-level phonemic decoding into readability assessment. En *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 8998–9009.

Pintard, Alice, Thomas François, Justine Nagant de Deuxchaisnes, Sílvia Barbosa, Maria Leonor Reis, Michell Moutinho, Ricardo Monteiro, Raquel Amaro, Susana Correia, Sandra Rodríguez Rey, Marcos Garcia González, Keran Mu & Xavier Blanco Escoda. 2024a. iRead4Skills Dataset 1: corpora by complexity level for FR, PT and SP. 10.5281/zenodo.10889888

Pintard, Alice, Thomas François, Nagant de Deuxchaisnes Justine, Sílvia Barbosa, Maria Leonor Reis, Michell Moutinho, Ricardo Monteiro, Raquel Amaro, Susana Correia, Sandra Rodríguez Rey, Keran Mu, Marcos Garcia González, André Bernárdez Braña & Xavier Blanco Escoda. 2024b. iRead4Skills Dataset 2: annotated corpora by level of complexity for FR, PT and SP. 10.5281/zenodo.13127674

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. En *58th Annual Meeting of the Association for Computational Linguistics*, 101–108. 10.18653/v1/2020.acl-demos.14

Quispesaravia, Andre, Walter Perez, Marco Sobrevilla Cabezudo & Fernando Alva-Manchego. 2016. Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish. En *10th International Conference on Language Resources and Evaluation (LREC)*, 4694–4698.

Ribeiro, Eugénio, Nuno Mamede & Jorge Baptista. 2024. Automatic text readability assessment in European Portuguese. En *16th International Conference on Computational Processing of Portuguese (PROPOR)*, 97–107.

Rojo, Guillermo & Ignacio M. Palacios Martínez. 2016. Learner Spanish on computer: The CAES 'corpus de aprendices de español' project. En *Spanish Learner Corpus Research*, 55–87. John Benjamins Publishing Company. 10.1075/scl.78.03roj

Saggion, Horacio. 2017. *Automatic text simplification*. Morgan & Claypool Publishers.

Saggion, Horacio, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula & Lorena Bourg. 2011. Text simplification in Simplext. making text more accessible. *Procesamiento del Lenguaje Natural* 47. 341–342.

Saggion, Horacio, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello & Biljana Drndarevic. 2015. Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing* 6(4). 10.1145/2738046

Schwarm, Sarah & Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. En *43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 523–530. 10.3115/1219840.1219905

Sheehan, Kathleen M., Michael Flor & Diane Napolitano. 2013. A two-stage approach for generating unbiased estimates of text complexity. En *Workshop on Natural Language Processing for Improving Textual Accessibility*, 49–58.

Trokhymovych, Mykola, Indira Sen & Martin Gerlach. 2024. An open multilingual system for scoring readability of wikipedia. arXiv [cs.CL/cs.AI]. 10.48550/arXiv.2406.01835

Vajjala, Sowmya. 2022. Trends, limitations and open challenges in automatic readability assessment research. En *13thLanguage Resources and Evaluation Conference (LREC)*, 5366–5377.

Vajjala, Sowmya & Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. En *3rd workshop on NLP for computer-assisted language learning*, 113–127.

Vajjala, Sowmya & Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. En *13th Workshop on Innovative Use of NLP for Building Educational Applications*, 297–304. 10.18653/v1/W18-0535

Vajjala, Sowmya & Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. En *7th Workshop on Building Educational Applications using NLP*, 163—-173.

Vásquez-Rodríguez, Laura, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel & Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in

Spanish. En *Workshop on Text Simplification, Accessibility, and Readability (TSAR)*, 188–198. 10.18653/v1/2022.tsar-1.18

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*

Vilares, David, Marcos Garcia & Carlos Gómez-Rodríguez. 2021. Bertinho: Galician BERT representations. *Procesamiento del Lenguaje Natural* 66. 13–26.

Wilcox, Ethan G., Tiago Pimentel, Clara Meister, Ryan Cotterell & Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics* 11. 1451–1470. 10.1162/tacl_a_00612

Wilkens, Rodrigo, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P. Yancey & Thomas François. 2022. FABRA: French aggregator-based readability assessment toolkit. En *13th Language Resources and Evaluation Conference (LREC)*, 1217–1233.

Wilkens, Rodrigo, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont & Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. En *Findings of the Association for Computational Linguistics (EACL)*, 2316–2331.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 38–45. 10.18653/v1/2020.emnlp-demos.6

Xu, Wei, Chris Callison-Burch & Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics* 3. 283–297. 10.1162/tacl_a_00139

Zeng, Jinshan, Xianchao Tong, Xianglong Yu, Wenyan Xiao & Qing Huang. 2024. InterpretARA: Enhancing hybrid automatic readability assessment with linguistic feature interpreter and contrastive learning. *AAAI Conference on Artificial Intelligence* 38(17). 19497–19505. 10.1609/aaai.v38i17.29921

A. Características dos niveis de lecturabilidade en galego

<table>
<tr><th colspan="4">**Primeiro nivel** $\simeq A1$ ($\simeq Pre-Celga1$)</th></tr>
<tr><td colspan="4">Textos cotiáns, como mensaxes curtas de correo ou relacionados coas actividades diarias, letreiros, listaxes, manuais, propaganda, formularios, materiais relacionados cos seus intereses, documentos auténticos breves (billetes, entradas, cartas de restaurante, facturas, etiquetas, planos, embalaxes, horarios, mapas...) de uso moi frecuente, instrucións sinxelas, relatos fáciles, letras de cancións e poemas sinxelos.</td></tr>
<tr><td colspan="4">**Léxico-conceptual**</td></tr>
<tr><td>Nomes comúns e propios.</td><td>Sufixo -iño/a. Valor dos afixos máis frecuentes. Superlativo en -ísimo/a.</td><td>Adverbios máis comúns de cantidade (moito, pouco), modo (ben, mellor, amodo), exclusión (só, soamente), afirmación e negación.</td><td>Repertorio de vocabulario limitado a situacións cotiás e de especial interese.</td></tr>
<tr><td colspan="4">**Verbal**</td></tr>
<tr><td>Verbos copulativos</td><td>Verbos regulares e semirregulares de uso frecuente (durmir, servir, espir). Irregulares máis frecuentes (ir, facer, estar). Reflexivos e pronominais máis comúns (chamarse, esquecerse, queixarse).</td><td>Tempos de indicativo en presente, pasado e futuro. Futuro hipotético.</td><td>Perífrases máis habituais: ir + inf, estar/andar + xerundio, volver + inf, comezar a + inf, ter que + inf</td></tr>
<tr><td colspan="4">**Estrutura sintáctica**</td></tr>
<tr><td>Coordinadas simples con e, mais, ou, pero.</td><td>Oracións temporais (cando) e comparativas e superlativas básicas.</td><td>Afirmativas, negativas, interrogativas con partículas e exclamativas básicas con partículas.</td><td>Subordinadas simples con conxuncións máis frecuentes: así que, porque, cando...</td></tr>
<tr><td colspan="4">**Cohesión**</td></tr>
<tr><td>Preposicións e locucións preposicionais máis frecuentes.</td><td>Uso de pronomes só con referente claro. Elipse só de elementos coñecidos e moi claros.</td><td colspan="2">Marcadores de orde do discurso, de espazo e de tempo máis frecuentes (hoxe, agora, cedo, lonxe, preto, enriba...)</td></tr>
</table>

<table>
<tr><th>**Segundo nivel** $\simeq A2$ ($\simeq Celga1$)</th></tr>
<tr><td>Mensaxes de carácter persoal (SMS, correos, cartas) ou de carácter social con frases rutinarias, cuestionarios sinxelos, notas e mensaxes relacionadas coas súas actividades de traballo, estudo e ocio e textos sociais breves tipificados (para felicitar, invitar, aceptar/refusar, agradecer, solicitar un servizo, pedir desculpas).
Xéneros textuais do primeiro nivel e tamén páxinas web, receitas, periódicos e revistas (titulares, novas e anuncios), horóscopos, contos e novelas curtos, anuncios de traballo... Instrucións sobre seguridade, aparellos de uso frecuente, aloxamento, menú de restaurante, vida académica, sanidade, posoloxía de medicamentos... Información sinxela de formularios e documentos administrativos (Correos, Administración Pública, bancos, entidades académicas...)
Rexistro formal e informal.</td></tr>
<tr><td>**Léxico-conceptual**</td></tr>
</table>

Nomes comúns e propios.	Interxeccións máis frecuentes.	Vocabulario de situacións cotiás e actividades habituais (por especial interese) e un vocabulario receptivo máis amplo que permita comprender textos (sentimentos, meteoroloxía, fauna e flora, alimentación, vida académica, saúde, Administración, tempo libre, actividades profesionais máis habituais).	
Sufixo -iño/a . Afixos comúns relativamente frecuentes. Superlativo absoluto e relativo. Comparativos analíticos e sintéticos máis frecuentes, forma meirande.		Adverbios máis comúns de cantidade (moito, pouco), modo (ben, mellor, amodo), exclusión (só, soamente), inclusión (até, mesmo, incluso), afirmación, negación e dúbida.	Usos de ti, vós / vostede, vostedes.
Verbal			
Verbos copulativos	Verbos regulares e semirregulares (durmir, servir, espir) e irregulares frecuentes (ser, estar, facer, saber, querer, ir, vir, haber, poder, ter, deber, etc.). Reflexivos e pronominais máis comúns (chamarse, esquecerse, queixarse, lavarse, vestirse).		Tempos de indicativo en presente, pasado e futuro. Futuro hipotético. Infinitivo, xerundio e participio.
Presente de subxuntivo para expresión de desexos, sentimentos e reaccións (Oxalá veñas!/Que teñas sorte!/Sinto que marches). Imperativo para consellos, instrucións e ordes.		Perífrases habituais: estar, andar + xerundio, volver, comezar a, ter que, estar a/para, estar a/andar a + infinitivo, ter + participio, haber (de) + infinitivo.	
Estrutura sintáctica			
Coordinadas copulativas (e, mais, a mais, nin), adversativas (mais, pero, senón) e disxuntivas (ou...ou, nin...nin)	Oracións impersoais sinxelas.	Afirmativas, negativas (incluíndo dobre negación), interrogativas con partículas, exclamativas e imperativas.	Subordinadas simples con conxuncións máis frecuentes: así que, porque, cando, ademais...
Subordinadas temporais (cando, ao + inf), causais, finais, condicionais e comparativas (máis...ca, meirande...ca)			
Cohesión			
Preposicións, locucións preposicionais, conectores e enlaces máis frecuentes.	Uso de pronomes só con referente claro. Elipse só de elementos coñecidos e moi claros.	Marcadores de orde do discurso, de espazo e de tempo máis frecuentes (hoxe, agora, cedo, lonxe, preto, enriba...) e tamén: entón, logo, por outra banda...	

Terceiro nivel $\simeq B1$ ($\simeq Celga2$)	
Textos breves en lingua estándar relacionados coa vida cotiá que informen sobre acontecementos ou transmitan opinións, desexos ou ordes. Tamén textos sinxelos sobre temas do seu ámbito académico ou profesional. Entre outros: cartas de restaurantes, formularios, anuncios, folletos, visitas turísticas, hoteis, alugamento, cartas e correos, relatos, anuncios ou artigos da prensa sobre temas de actualidade, prospectos de medicamentos, folletos de divulgación e publicitarios...	
Léxico-conceptual	
Nomes comúns e propios.	Vocabulario de situacións cotiás, actividades de lecer, sentimentos, do ámbito laboral e de temas de interese xeral, como saúde, accidentes, tecnoloxía, medio natural, economía, sociedade, xeografía, servizos...
Verbal	

Voz pasiva	Tempos de indicativo en presente, pasado e futuro, futuro hipotético e presente de subxuntivo dos verbos regulares e dos irregulares máis comúns.	Imperativo (tamén negativo) para consellos, instrucións e ordes.	Perífrases verbais de diversos valores: poder + inf. (obriga), deber (de) + infinitivo (obriga), haber + inf. (obriga), ter + part. (aspecto), andar + xer., ir + xer, estar para + inf., poñerse a + inf.), volver + inf, empezar a + inf, acabar de + inf, deixar de + inf, levar + xer, seguir + xer, dar + part...
Estrutura sintáctica			
Coordinadas copulativas (e, mais, a mais, nin), adversativas (mais, pero, senón) e disxuntivas (ou...ou, nin...nin), distributivas (ben... ben, ora... ora) e explicativas (é dicir, ou sexa, isto é)		Usos de se: oracións impersoais, pasivas e verbos reflexivos.	Afirmativas, negativas (incluíndo dobre negación), interrogativas con partículas (tamén indirectas), exclamativas e imperativas.
Subordinadas con subxuntivo e infinitivo.	Subordinadas adxectivas e substantivas, adverbiais (causais, finais, consecutivas, condicionais, temporais)		Estilo indirecto
Cohesión			
Preposicións, locucións preposicionais, conectores e enlaces máis frecuentes.	Uso de pronomes só con referente claro. Elipse só de elementos coñecidos e moi claros.	Marcadores de orde do discurso, de espazo e de tempo máis frecuentes e tamén entón, logo, por outra banda...	

Cuarto nivel $\simeq B2$ **(** $\simeq Celga3$**)**
Dentro do ámbito de interese ou especialización: instrucións, información técnica, informes e artigos, textos oficiais, institucionais ou corporativos, comunicación formal, textos periodísticos (incluídos artigos e reportaxes) Temas xerais: notas, mensaxes, correos, foros e blogs, textos literarios sinxelos (ficción e non ficción) contemporáneos en prosa. Textos sobre a vida cotiá, laboral e académica como: cartas persoais ou profesionais, textos diversos dos medios de comunicación, normas, formularios administrativos, artigos de temáticas descoñecidas, materiais de cursos académicos. Textos extensos e complexos. Relacións lóxicas de conxunción, disxunción, oposición, contraste, concesión, comparación, condición, causa, finalidade, resultado e correlación. O aspecto puntual, perfectivo/imperfectivo, durativo, progresivo, habitual, prospectivo, incoativo, terminativo, iterativo e causativo. Expresión da capacidade, necesidade, posibilidade, probabilidade, vontade, permiso, obriga, prohibición.
Léxico-conceptual

Interxeccións habituais.	Anteposición do adxectivo e do superlativo (diferente valor)	Valor enfático dos interrogativos e exclamativos pospostos.	Vocabulario común e vocabulario especializado dos ámbitos de interese persoal, público, educativo e profesional propios. Outras temáticas: servizos, lingua e comunicación intercultural, ciencia, tecnoloxía, historia e cultura.
Expresións idiomáticas moi habituais.	Derivación.	Siglas frecuentes.	Algúns modismos e expresións coloquiais moi frecuentes.
Verbal			
Condicional simple (cortesía, desexo, hipótese).	Perífrases verbais: as anteriores e haber de + inf. (expresar posibilidade de futuro, necesidade ou intención).	Voz pasiva.	Imperativo (formas usuais) e imperativos lexicalizados comúns.
Pretérito imperfecto de subxuntivo para hipóteses.	Infinitivo conxugado.	Futuro imperfecto de subxuntivo en oracións temporais (cando), condicionais (se) e adxectivas.	
Estrutura sintáctica			
Coordinadas afirmativas e negativas copulativas, adversativas, explicativas, disxuntivas (p. Ex. Quer... quer)	Alteración da orde habitual dos elementos na oración con valor enfático ou focalizador.	Todas as afirmativas, negativas, interrogativas, exclamativas e imperativas.	Pasivas.
Usos de se: oracións impersoais, pasivas, verbos reflexivos, pronominais.	Subordinadas con indicativo e subxuntivo.	Substantivas con uso do pretérito perfecto de subxuntivo.	Circunstanciais con indicativo, subxuntivo e infinitivo. Condicionais con subxuntivo (presente, futuro, imperfecto) ou infinitivo, e causais.
Cohesión			
Expresión do espazo e do tempo, e das relacións espaciais e temporais.		Conectores de inicio, explicación, reformulación, exemplificación, conclusión, etc.	

Quinto nivel $\simeq C1$ ($\simeq Celga4$)
Textos extensos e complexos de calquera ámbito de especialidade: instrucións, normativas, advertencias e informacións diversas técnicas. Formato pouco habitual, linguaxe coloquial, humor. Comunicación formal, profesional ou institucional. Inclúe artigos, informes, actas ou memorias do ámbito profesional ou académico. Tamén textos periodísticos e artigos sobre temas de actualidade e textos literarios contemporáneos cun nivel intermedio de complexidade (lingüística, estrutural e temática).
Léxico-conceptual

Vocabulario común e vocabulario especializado dos ámbitos de interese persoal, público, educativo e profesional propios.	Modismos, coloquialismos, rexionalismos e argot.	Léxico especializado da saúde, léxico dos negocios, mundo laboral, económico, medio natural, organismos oficiais públicos, sistema político, arte, literatura, música, cine, ciencia, tecnoloxía, festividades...	Expresións idiomáticas e frases feitas.
Verbal			
Perífrases verbais con varios valores.	Todos os tempos verbais.	Infinitivo conxugado: usos e equivalencias en oracións temporais, condicionais, e inf conxugado / subxuntivo.	
Estrutura sintáctica			
Expresión da certeza, crenza, conxectura, dúbida, capacidade, posibilidade, probabilidade, necesidade, obriga, prohibición, permiso, vontade, intencionalidade, etc.		O aspecto puntual, perfectivo/imperfectivo, durativo, progresivo, habitual, prospectivo, incoativo, terminativo, iterativo e causativo.	Todas as afirmativas, negativas, interrogativas, exclamativas e exhortativas.
Relacións lóxicas de conxunción, disxunción, oposición, contraste, concesión, comparación, condición, causa, finalidade, resultado e correlación.		Oracións subordinadas substantivas, finais, modais, condicionais, adxectivas, concesivas, adverbiais con xerundio, condicionais e causais.	
Cohesión			
Expresión do espazo e do tempo, e das relacións espaciais e temporais.	Conectores discursivos de comezo, resumo, repetición, explicación, etc.	Outras expresións: seica, disque...	

Sexto nivel $\simeq C2$ ($\simeq Celga5$)			
Comunicación persoal (cartas, correos, notas...) en rexistro coloquial, con modismos, fraseoloxía e vocabulario específicos dun ámbito. Anuncios e información con dobre sentido ou que haxa que interpretar. Textos con trazos non normativos que pertencen a unha variedade do galego. Textos académicos, profesionais, de opinión ou ensaio e literarios, tanto de épocas pasadas, con variacións na lingua, coma de actualidade. Textos complexos e mesmo ambiguos ou pouco claros, que é necesario interpretar.			
Léxico-conceptual			
Pronomes de solidariedade e interese.	Derivación e composición.	Fraseoloxía en contextos específicos. Connotación, intención e polisemia en textos coloquiais ou literarios.	Vocabulario de todos os ámbitos, entre eles: relacións, vivenda, paisaxe, servizos públicos, sociedade, tempo libre, actividades comerciais, saúde, educación...
Verbal			
Formas verbais simples e perífrases para expresar a temporalidade.		Todas as formas verbais e usos.	
Estrutura sintáctica			
Alteracións da orde dos elementos da oración.		Todas as estruturas sintácticas e usos.	
Cohesión			
Todos os conectores e usos.			