Lexical borrowing detection as a sequence labeling task

Data, modeling and evaluation methods for anglicism retrieval in Spanish

Elena Álvarez Mellado
Advisors: Julio Gonzalo (UNED), Constantine Lignos (Brandeis)

elDiario.es

RED

¿QUÉ ES...?

'Benching', estar en el banquillo de tu crush mientras otro juega de titular

Las redes sociales son el nuevo tablero en el que se deciden muchas de las relaciones de pareja. La invisibilidad que otorgan las pantallas facilita prácticas que pueden provocar inseguridad y estrés en quien las sufre

elDiario.es

RED

¿QUÉ ES...?

'Benching', estar en el banquillo de tu crush mientras otro juega de titular

Las redes sociales son el nuevo tablero en el que se deciden muchas de las relaciones de pareja. La invisibilidad que otorgan las pantallas facilita prácticas que pueden provocar inseguridad y estrés en quien las sufre

LA ERA DEL BIG DATA

Big data y machine learning: el mundo de los datos necesita especialistas

• Su aplicación permite a empresas y administraciones mejorar su funcionamiento, por eso la demanda de estos profesionales es cada vez mayor

¿QUÉ ES...?

canción Contir

Paraísos 'offshore' y riquezas ocultas de líderes mundiales y billonarios expuestas en una filtración sin precedentes

Los 'Papeles' de Pandora revelan el funcionamiento interno de una economía en las sombras que beneficia a los ricos y las élites a expensas de todos los demás

'Benching', est... crush mientras...

Las redes sociales son el nuevo tablero ...bilidad

La economía 'silver' el nuevo motor que impulsará 88 millones de empleos en la próxima década

Un informe de la Comisión Europea revela que los mayores de 55 años, con alto poder adquisitivo y tecnológicamente activos, transformarán el mercado

EL COMIDISTA

COCINA SENCILLA >

'Batch cooking' de abril: cocina una tarde y come toda la semana

¿Organizar el menú semanal te quita el sueño y de lunes a viernes no tienes tiempo? Aquí tienes muchas ideas de platos sencillos y sabrosos para cuatro personas que se preparan en una tarde.

Why are lexical borrowings interesting?

- Language change and language contact

- New words and meanings

- Sociolinguistics dynamics

Motivation

Propose automatic methods for retrieving lexical borrowings (focusing on anglicisms in Spanish)

Previous work

A similar task: codeswitching identification

Sometimes I'll start a sentence in Spanish y termino en español

Previous approaches to borrowing identification

- Lexicon lookup
- Pattern matching
- Character n-gram probability

Previous work: limitations

Previous work: limitations

Cada vez más personas optan por el van life y se van a vivir a una autocaravana o a una furgoneta camperizada.

Previous work: limitations


Cómo convertir tu piso en una autocaravana para tener una autocaravana grande en vez de un piso pequeño

Cada vez más personas optan por el van life y se van a vivir a una autocaravana o a una furgoneta camperizada.

EL MUNDO TODAY

Previous work: limitations

*Estoy sin **cash*** vs *El disco de Johnny Cash*

*Receta de **pie** de limón* vs *El pie de página*

Our proposal

Lexical borrowing identification should be framed as a sequence labeling task.

(≈ NER, MWE)

Our proposal

- Linguistic ambiguity (*pie*)
- Multiword borrowing (*machine learning*)
- Adjacency (*body beige*)

Our work from 2020

- A dataset annotated with anglicisms

- A CRF model for the task (F1 = 0.86)

Research questions

What type of data?

- Problem definition?
- Existing data?
- Annotation?

What type of models?

How should we evaluate?

- Blind spots in standard evaluation?
- Diagnostic? Actionable? Predictive?

Research questions

What type of data?

- Problem definition?
- Existing data?
- Annotation?

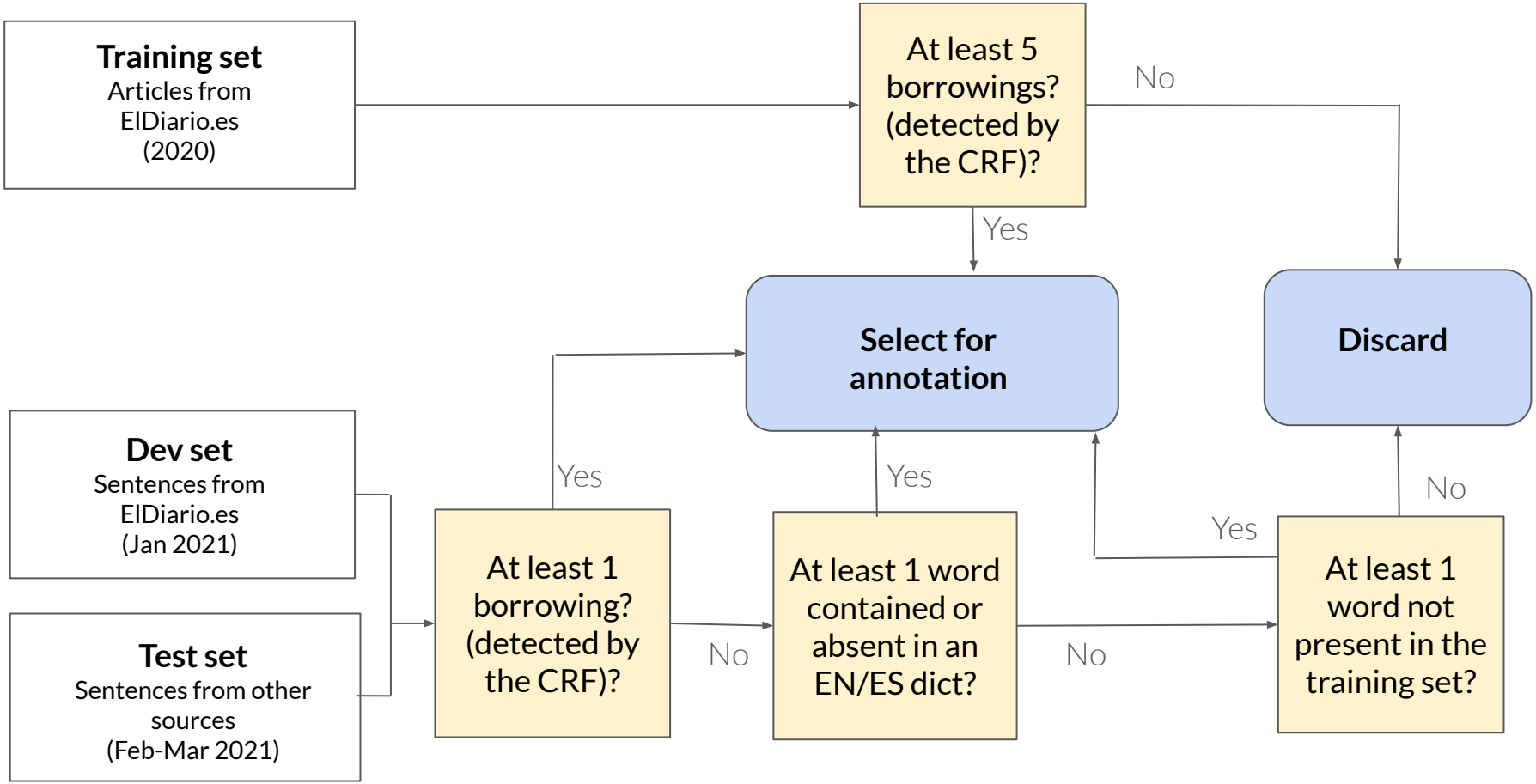
What type of models?

How should we evaluate?

- Blind spots in standard evaluation?
- Diagnostic? Actionable? Predictive?

Limitations of the 2020 dataset

- Headlines from *ElDiario.es*
- No IAA
- 54% of seen borrowings


Training set
Articles from ElDiario.es (2020)
At least 5 borrowings? (detected by the CRF)?
No
Yes
Select for annotation
Discard
Dev set
Sentences from ElDiario.es (Jan 2021)
Test set
Sentences from other sources (Feb-Mar 2021)
At least 1 borrowing? (detected by the CRF)?
Yes
No
At least 1 word contained or absent in an EN/ES dict?
Yes
No
At least 1 word not present in the training set?
Yes
No

Annotation: What is a lexical borrowing?

- Borrowing vs codeswitches, metalinguistic usage, etc.

- When does a word stop being a borrowing?

- Unassimilated lexical borrowings

COALAS: COrpus of AngLicisms in the spAnish PresS🐨

	2020	COALAS
# tokens	325.665	372.701
# journalistic sources	1	21
# ENG spans	1.304	3.038
# OTHER spans	102	123
% unseen spans (in test)	46%	92%
IAA	-	0.9

COALAS: COrpus of AngLicisms in the spAnish PresS🐨

	2020	COALAS
# tokens	325.665	372.701
# journalistic sources	1	21
# ENG spans	1.304	3.038
# OTHER spans	102	123
% unseen spans (in test)	46%	92%
IAA	-	0.9

Research questions

What type of data?

- Problem definition?
- Existing data?
- Annotation?

What type of models?

How should we evaluate?

- Blind spots in standard evaluation?
- Diagnostic? Actionable? Predictive?

Supervised models

- CRF

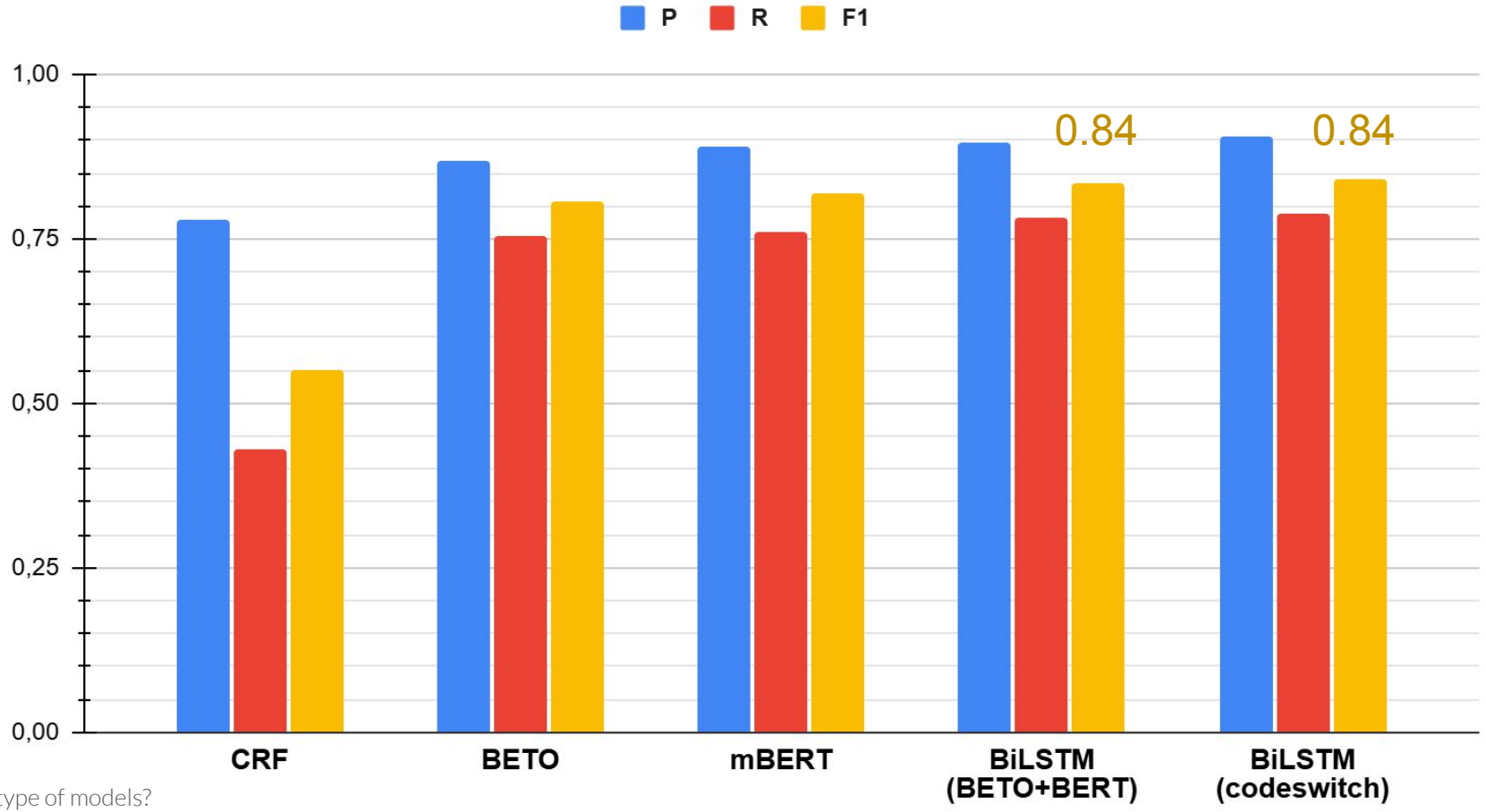
Supervised models

- CRF: F1= 0.55 [F1= 0.86, 2020]

Supervised models

- CRF
- Transformer-based models:
  - BETO
  - mBERT
- BiLSTM-CRF fed with:
  - Contextual embeddings (BETO, BERT) + subword embeddings
  - Fine-tuned embeddings (codeswitch) + subword embeddings

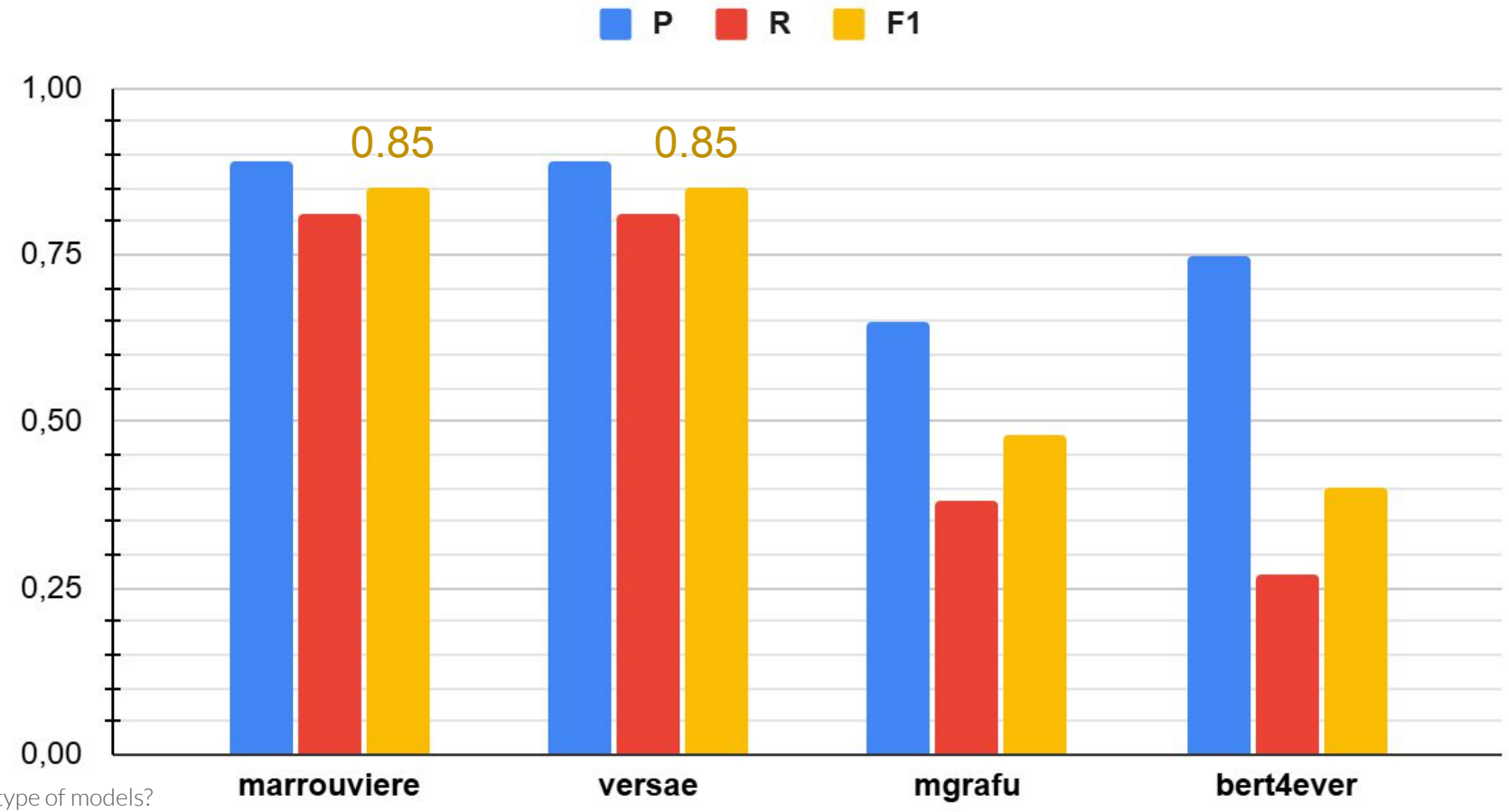
Model performance on COALAS



ADoBo 2021 (Automatic Detection of Borrowings)

- A shared task at Iberlef 2021
- Dataset: COALAS
- 9 systems, 4 teams. Best F1 = 0.85

Performance of participants of ADoBo 2021



An LLM for anglicism extraction

- Data scarce scenario
- Llama3 (**8B**)
- Few shot

Prompting methodology for 8B-Llama3

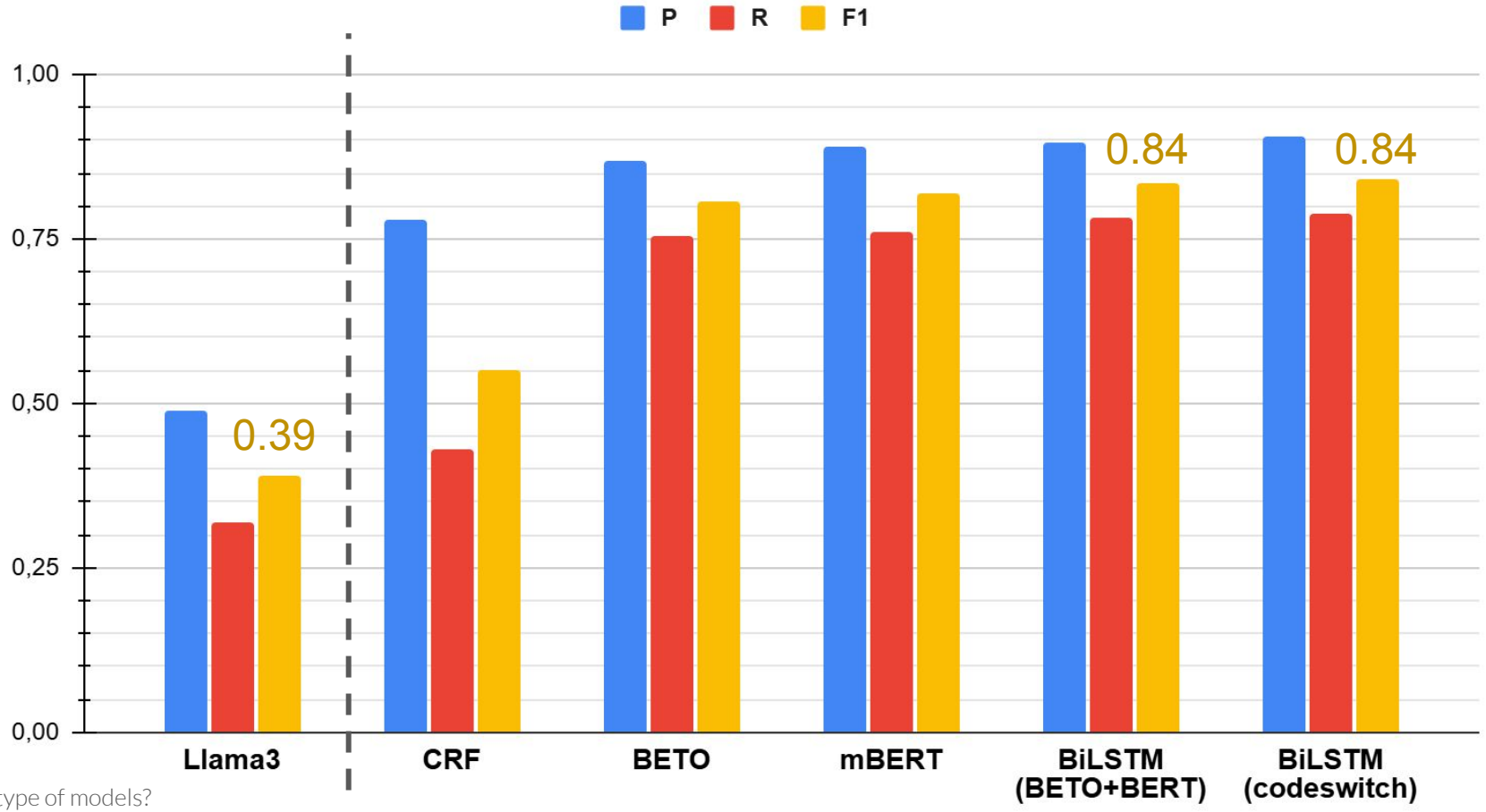
```
Un anglicismo es un préstamo crudo incorporado del inglés que se usa dentro de una frase en castellano sin adaptación, palabras
como 'online', 'machine learning', 'podcast' o 'blockchain'.  Un anglicismo puede estar formado por una única palabra (como
'streaming' en 'El evento se retransmitirá por streaming'), por varias palabras (como 'concept car' en 'La marca ha sacado a la
venta un nuevo concept car') o ser dos anglicismos independientes que aparecen uno al lado del otro (como 'smartphone' y
'online' en 'Compré mi smartphone online'), y pueden aparecer entrecomillados o no.  Los nombres propios (como los nombres de
personas, lugares o títulos de obras de ficción) no son anglicismos."

Voy a darte una frase en castellano, tu misión es identificar si la frase contiene algún anglicismo.  La frase puede contener un
anglicismo, varios o no contener ninguno.  Responde solo con el segmento que cumpla la condición de ser anglicismo (si es que lo
hay) y sin añadir ninguna consideración más.  Si la frase no contiene ningún anglicismo, responde 'None'.  Si la frase contiene
más de 1 anglicismo, devuelve todos los anglicismos, con un anglicismo por línea (es decir, añade un salto de línea entre
anglicismos).  No alucines ni añadas palabras que no estén en la frase original.  Esta es la frase:
El problema de la "fast fashion" es triple :  consume recursos desmesuradamente , produce toneladas de basura de difícil
reciclaje y fomenta la explotación laboral de la industria textil .

"fast fashion"
```

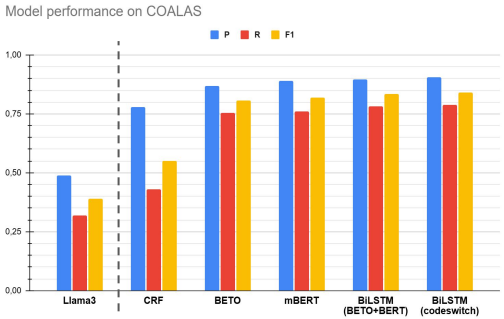
Prompt inspired by Ashok and Lipton (2023) template for NER: definition of the task with examples, question and response by Llama3.

Model performance on COALAS



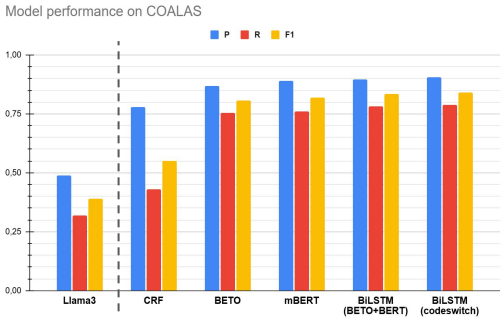
Standard evaluation

BiLSTM (cs) > 8B-Llama3



Standard evaluation

BiLSTM (cs) > 8B-Llama3



diagnostic? ?

actionable? ?

predictive? ?

Research questions

What type of data?

- Problem definition?
- Existing data?
- Annotation?

What type of models?

How should we evaluate?

- Blind spots in standard evaluation?
- Diagnostic? Actionable? Predictive?

Not all anglicisms are created equal

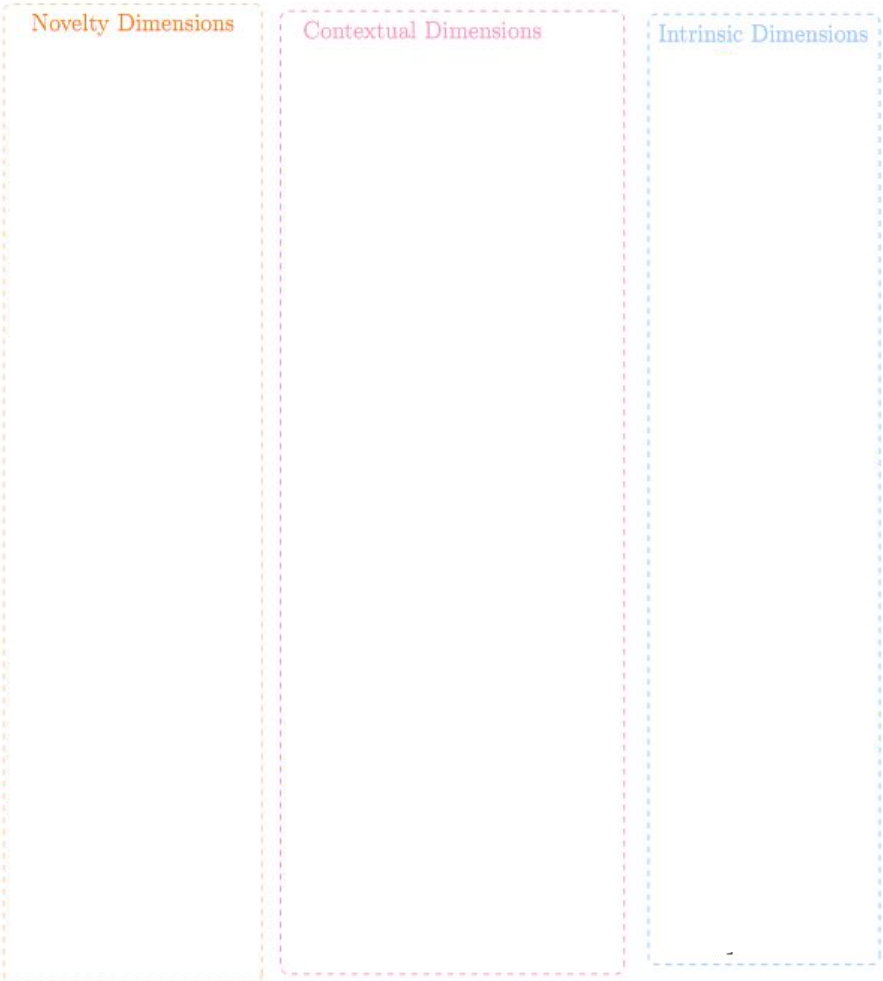
- *streaming*
- *online*
- *pie*

Not all anglicisms are created equal

- *streaming*
- *online*
- *pie*

- *Streaming del partido*

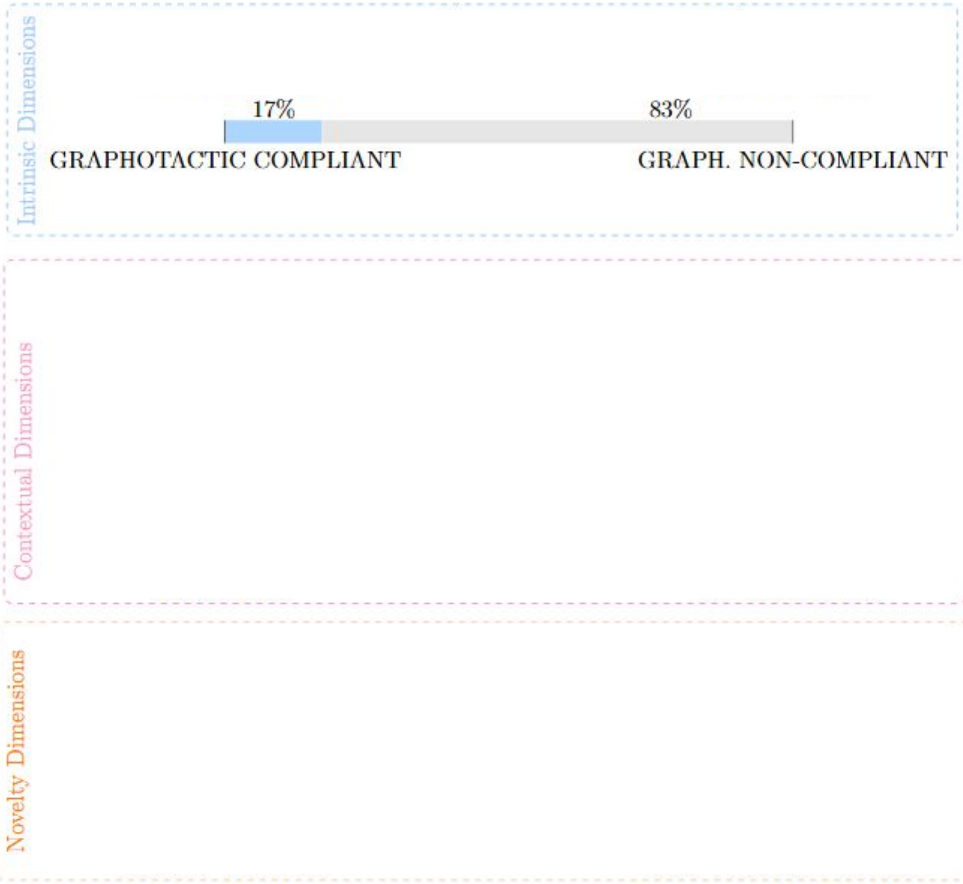
- *Receta de “pie” de limón*


Novelty Dimensions
Contextual Dimensions
Intrinsic Dimensions

podcast

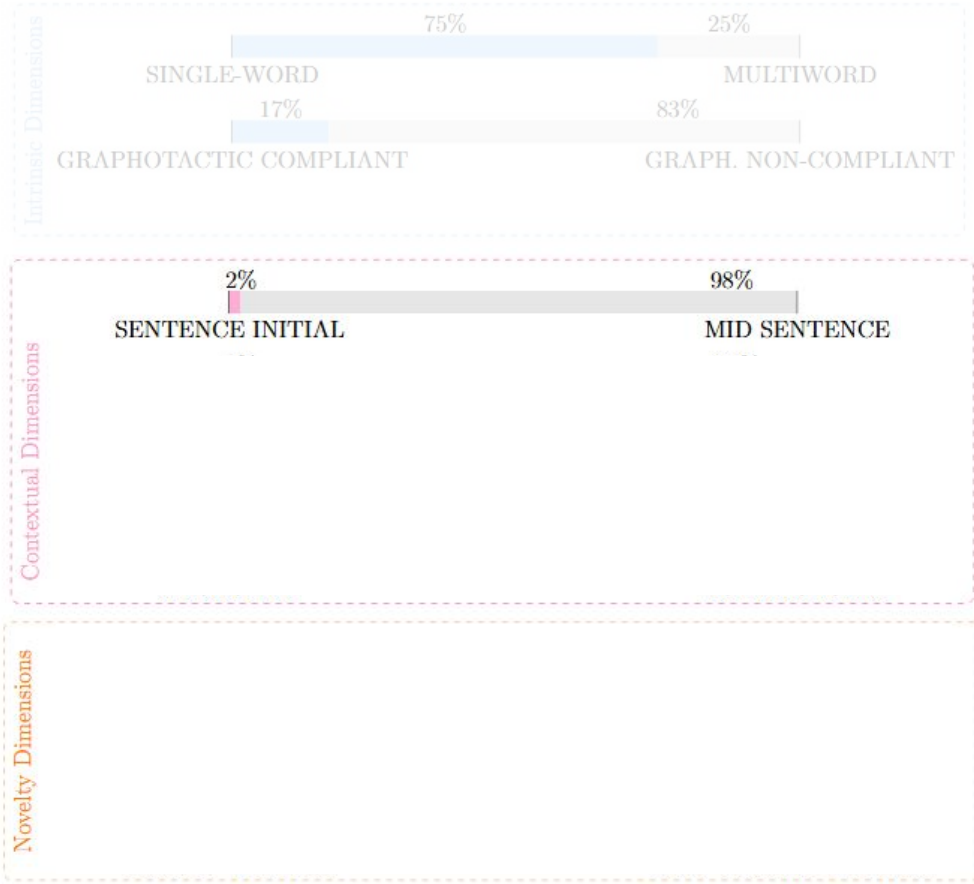
machine learning

online



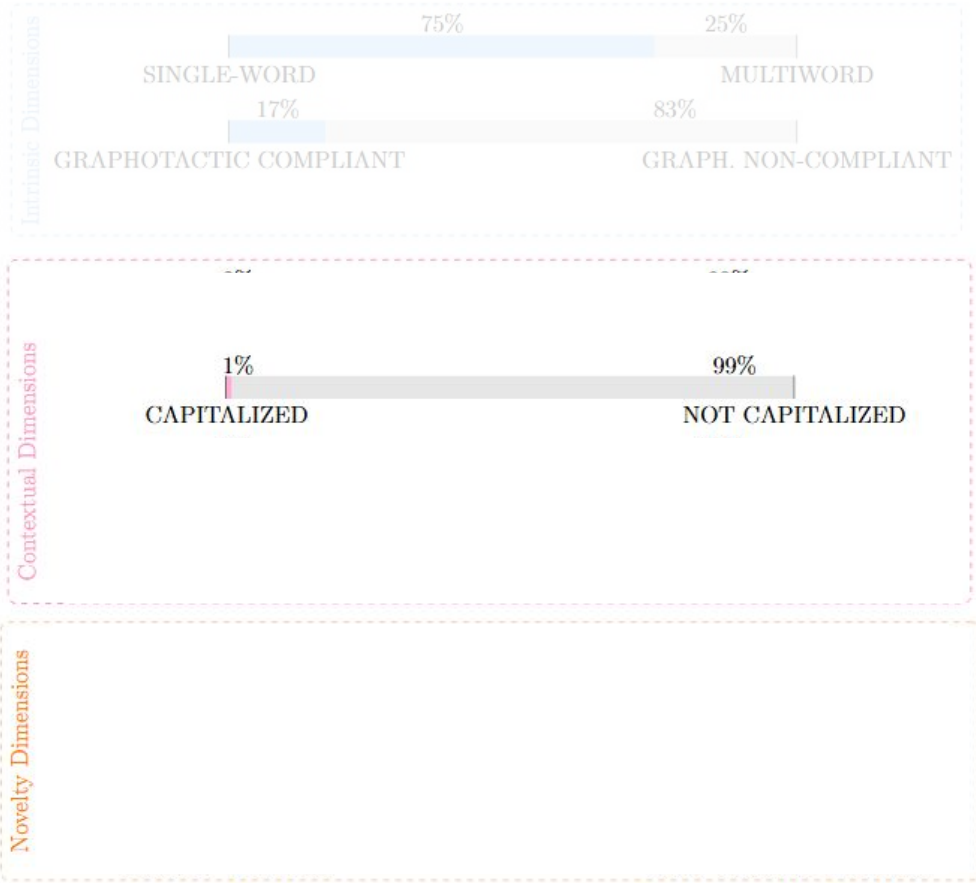
streaming

Streaming en TVE



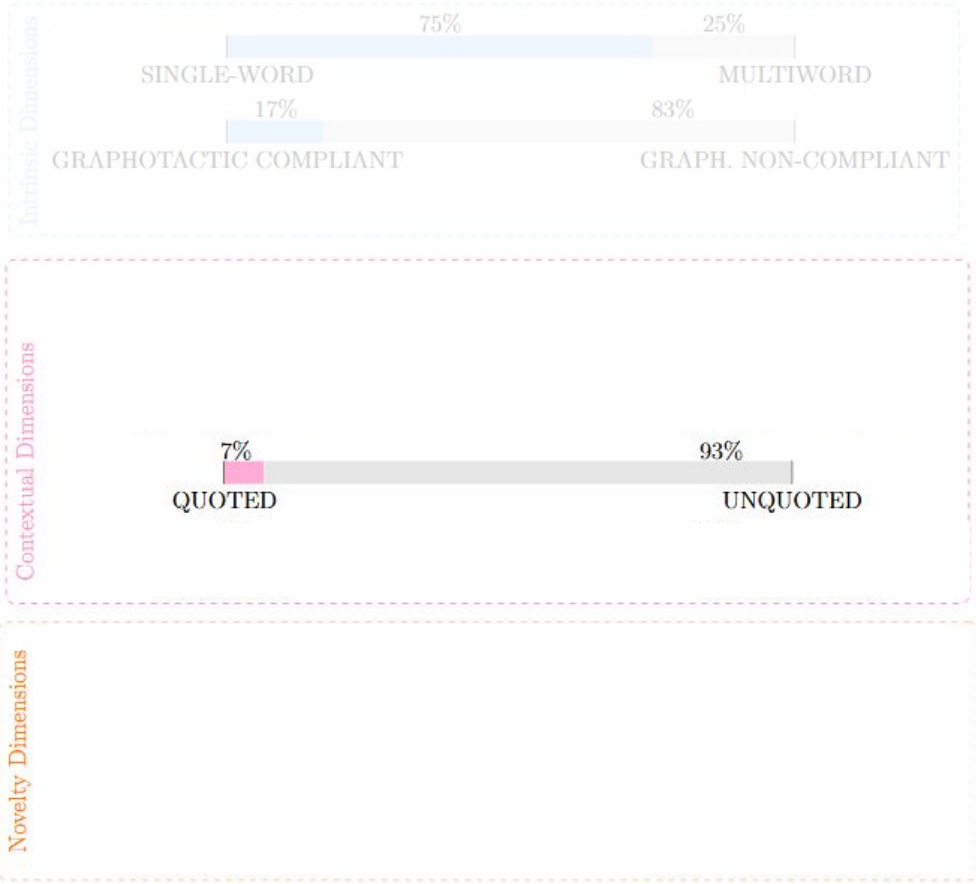
El streaming será en TVE

Big Data



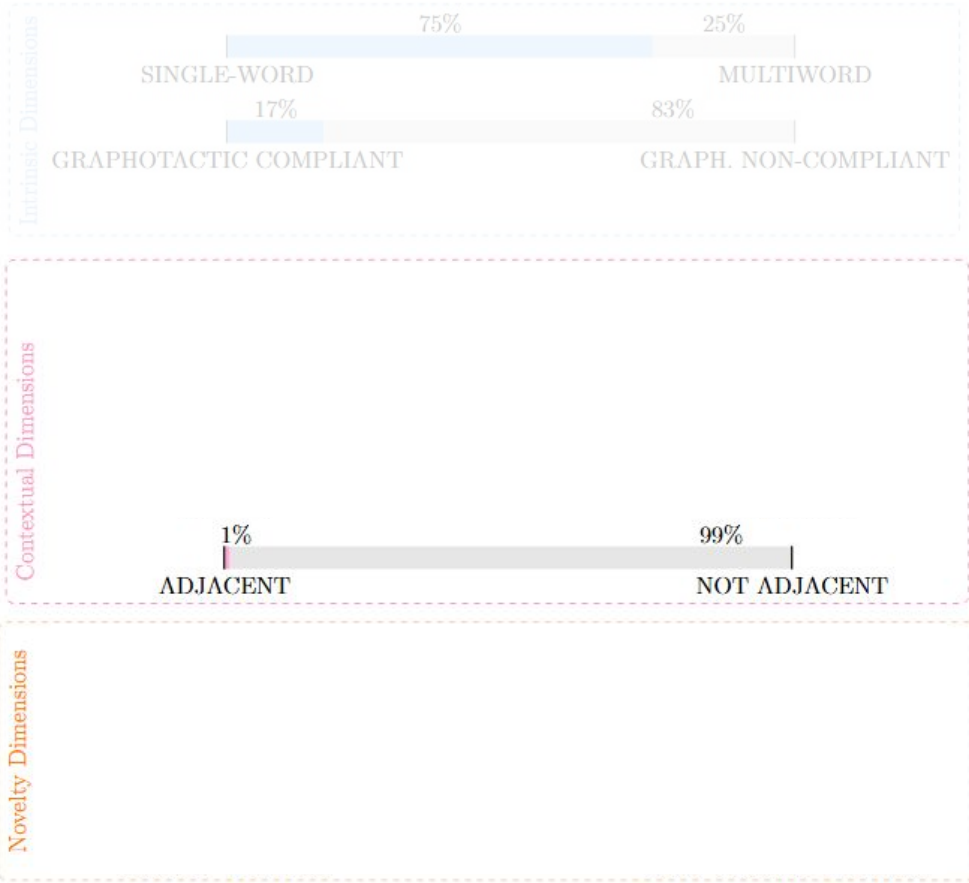
big data

“prime time”



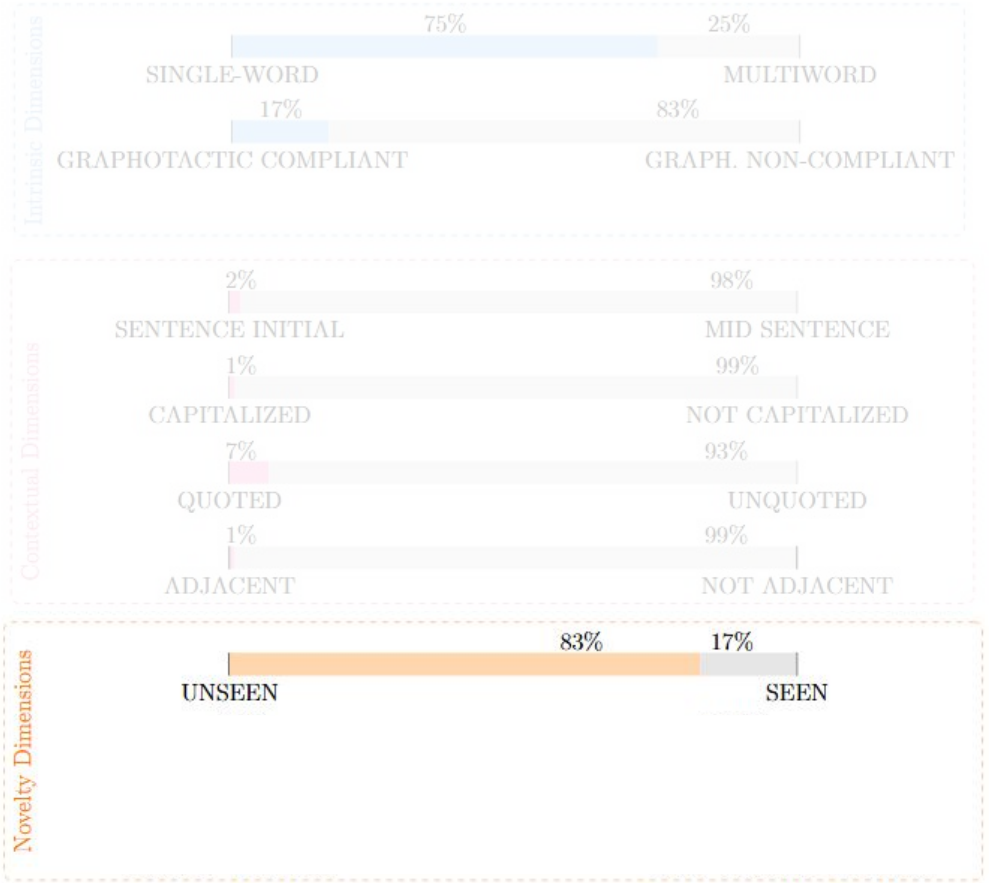
prime time

body beige



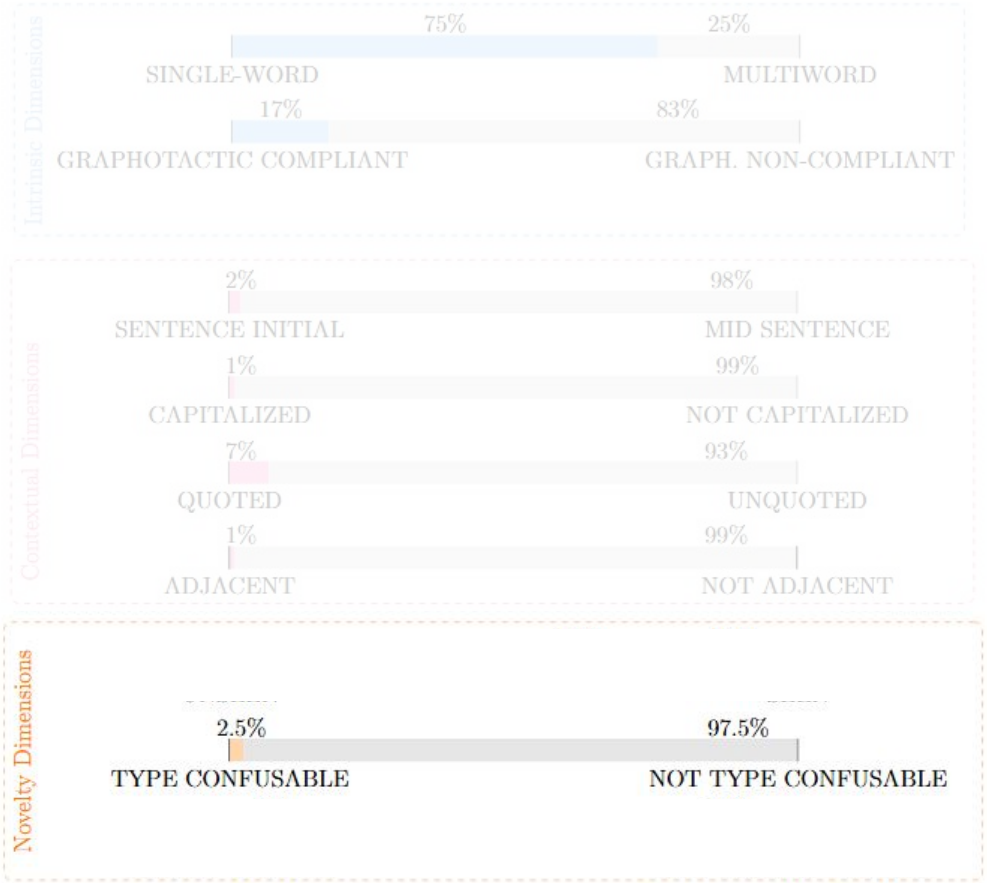
body blanco

benching

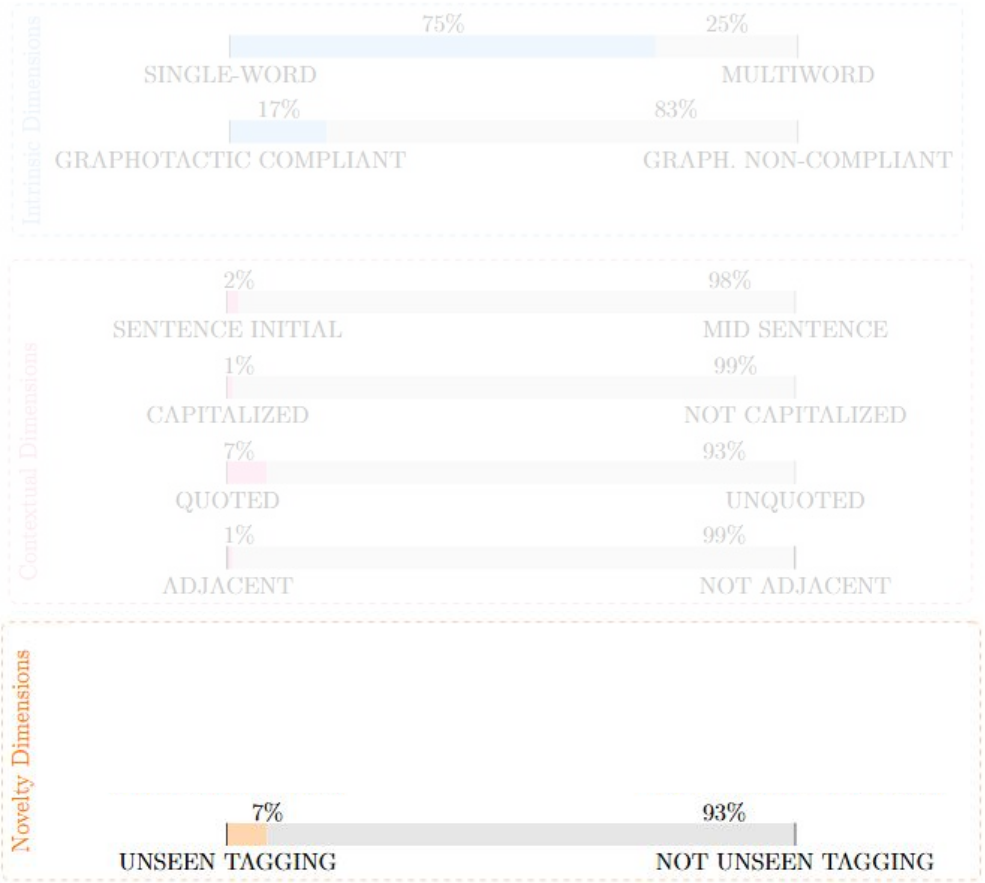


app

pie

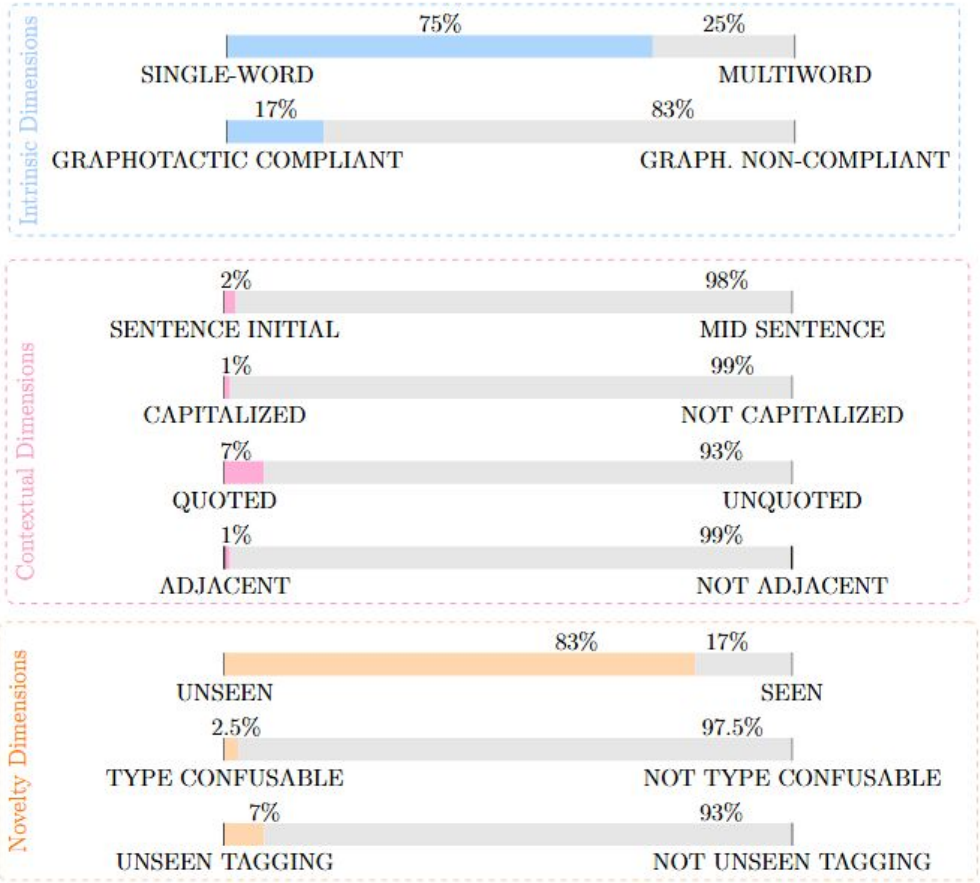

Intrinsic Dimensions
75%
25%
SINGLE-WORD
MULTIWORD
17%
83%
GRAPHOTACTIC COMPLIANT
GRAPH. NON-COMPLIANT
Contextual Dimensions
2%
98%
SENTENCE INITIAL
MID SENTENCE
1%
99%
CAPITALIZED
NOT CAPITALIZED
7%
93%
QUOTED
UNQUOTED
1%
99%
ADJACENT
NOT ADJACENT
Novelty Dimensions
2.5%
97.5%
TYPE CONFUSABLE
NOT TYPE CONFUSABLE

app



red carpet

app


Intrinsic Dimensions
75%
25%
SINGLE-WORD
MULTIWORD
17%
83%
GRAPHOTACTIC COMPLIANT
GRAPH. NON-COMPLIANT
Contextual Dimensions
2%
98%
SENTENCE INITIAL
MID SENTENCE
1%
99%
CAPITALIZED
NOT CAPITALIZED
7%
93%
QUOTED
UNQUOTED
1%
99%
ADJACENT
NOT ADJACENT
Novelty Dimensions
83%
17%
UNSEEN
SEEN
2.5%
97.5%
TYPE CONFUSABLE
NOT TYPE CONFUSABLE
7%
93%
UNSEEN TAGGING
NOT UNSEEN TAGGING

Model recall per dimension

Model	Overall recall	Unseen	Seen	Multi-token	Single-token	Type confus.	Unseen tag.	Sent. initial	Capital.	Quot.	Unquot.	Adjac.	Graph. n/ compl.	Graph. compl.
Llama3	0.32	0.32	0.32	0.44	0.29	0.35	0.38	**0.42**	0.31	0.76	0.28	0.19	0.33	0.26
CRF	0.43	0.35	0.82	0.36	0.45	0.52	0.25	0.08	0.12	0.75	0.40	0.25	0.46	0.26
BETO	0.81	0.79	0.90	0.84	0.80	**0.93**	0.67	**0.42**	0.25	**0.98**	0.80	0.19	0.84	**0.68**
mBERT	0.79	0.77	0.88	**0.84**	0.78	0.90	**0.74**	0.33	0.31	0.97	0.78	0.12	0.83	0.60
BiLSTM w/ BETO+BERT	0.80	0.78	**0.91**	0.83	0.79	0.74	0.62	**0.42**	0.19	0.96	0.79	**0.31**	0.83	0.63
BiLSTM w/ Codeswitch	**0.82**	0.80	0.91	0.84	**0.81**	0.71	0.63	0.33	**0.50**	0.94	**0.81**	0.12	**0.85**	0.62

Model recall per dimension

Model	Overall recall	Unseen	Seen	Multi-token	Single-token	Type confus.	Unseen tag.	Sent. initial	Capital.	Quot.	Unquot.	Adjac.	Graph. n/ compl.	Graph. compl.
Llama3	0.32	0.32	0.32	0.44	0.29	0.35	0.38	**0.42**	0.31	0.76	0.28	0.19	0.33	0.26
CRF	0.43	0.35	0.82	0.36	0.45	0.52	0.25	0.08	0.12	0.75	0.40	0.25	0.46	0.26
BETO	0.81	0.79	0.90	0.84	0.80	**0.93**	0.67	**0.42**	0.25	**0.98**	0.80	0.19	0.84	**0.68**
mBERT	0.79	0.77	0.88	**0.84**	0.78	0.90	**0.74**	0.33	0.31	0.97	0.78	0.12	0.83	0.60
BiLSTM w/ BETO+BERT	0.80	0.78	**0.91**	0.83	0.79	0.74	0.62	**0.42**	0.19	0.96	0.79	**0.31**	0.83	0.63
BiLSTM w/ Codeswitch	**0.82**	0.80	0.91	0.84	**0.81**	0.71	0.63	0.33	**0.50**	0.94	**0.81**	0.12	**0.85**	0.62

Model recall per dimension

Model	Overall recall	Unseen	Seen	Multi-token	Single-token	Type confus.	Unseen tag.	Sent. initial	Capital.	Quot.	Unquot.	Adjac.	Graph. n/ compl.	Graph. compl.
Llama3	0.32	0.32	0.32	0.44	0.29	0.35	0.38	**0.42**	0.31	0.76	0.28	0.19	0.33	0.26
CRF	0.43	0.35	0.82	0.36	0.45	0.52	0.25	0.08	0.12	0.75	0.40	0.25	0.46	0.26
BETO	0.81	0.79	0.90	0.84	0.80	**0.93**	0.67	**0.42**	0.25	**0.98**	0.80	0.19	0.84	**0.68**
mBERT	0.79	0.77	0.88	**0.84**	0.78	0.90	**0.74**	0.33	0.31	0.97	0.78	0.12	0.83	0.60
BiLSTM w/ BETO+BERT	0.80	0.78	**0.91**	0.83	0.79	0.74	0.62	**0.42**	0.19	0.96	0.79	**0.31**	0.83	0.63
BiLSTM w/ Codeswitch	**0.82**	0.80	0.91	0.84	**0.81**	0.71	0.63	0.33	**0.50**	0.94	**0.81**	0.12	**0.85**	0.62

pie

Model recall per dimension

Model	Overall recall	Unseen	Seen	Multi-token	Single-token	Type confus.	Unseen tag.	Sent. initial	Capital.	Quot.	Unquot.	Adjac.	Graph. n/ compl.	Graph. compl.
Llama3	0.32	0.32	0.32	0.44	0.29	0.35	0.38	**0.42**	0.31	0.76	0.28	0.19	0.33	0.26
CRF	0.43	0.35	0.82	0.36	0.45	0.52	0.25	0.08	0.12	0.75	0.40	0.25	0.46	0.26
BETO	0.81	0.79	0.90	0.84	0.80	**0.93**	0.67	**0.42**	0.25	**0.98**	0.80	0.19	0.84	**0.68**
mBERT	0.79	0.77	0.88	**0.84**	0.78	0.90	**0.74**	0.33	0.31	0.97	0.78	0.12	0.83	0.60
BiLSTM w/ BETO+BERT	0.80	0.78	**0.91**	0.83	0.79	0.74	0.62	**0.42**	0.19	0.96	0.79	**0.31**	0.83	0.63
BiLSTM w/ Codeswitch	**0.82**	0.80	0.91	0.84	**0.81**	0.71	0.63	0.33	**0.50**	0.94	**0.81**	0.12	**0.85**	0.62

Model recall per dimension

Model	Overall recall	Unseen	Seen	Multi-token	Single-token	Type confus.	Unseen tag.	Sent. initial	Capital	Quot.	Unquot	Adjac.	Graph. n/ compl.	Graph. compl.
Llama3	0.32	0.32	0.32	0.44	0.29	0.35	0.38	**0.42**	0.31	0.76	0.28	0.19	0.33	0.26
CRF	0.43	0.35	0.82	0.36	0.45	0.52	0.25	0.08	0.12	0.75	0.40	0.25	0.46	0.26
BETO	0.81	0.79	0.90	0.84	0.80	**0.93**	0.67	**0.42**	0.25	**0.98**	0.80	0.19	0.84	**0.68**
mBERT	0.79	0.77	0.88	**0.84**	0.78	0.90	**0.74**	0.33	0.31	0.97	0.78	0.12	0.83	0.60
BiLSTM w/ BETO+BERT	0.80	0.78	**0.91**	0.83	0.79	0.74	0.62	**0.42**	0.19	0.96	0.79	**0.31**	0.83	0.63
BiLSTM w/ Codeswitch	**0.82**	0.80	0.91	0.84	**0.81**	0.71	0.63	0.33	**0.50**	0.94	**0.81**	0.12	**0.85**	0.62

"streaming" vs streaming

Model recall per dimension

Model	Overall recall	Unseen	Seen	Multi-token	Single-token	Type confus.	Unseen tag.	Sent. initial	Capital.	Quot.	Unquot.	Adjac.	Graph. n/ comp.	Graph. compl.
Llama3	0.32	0.32	0.32	0.44	0.29	0.35	0.38	**0.42**	0.31	0.76	0.28	0.19	0.33	0.26
CRF	0.43	0.35	0.82	0.36	0.45	0.52	0.25	0.08	0.12	0.75	0.40	0.25	0.46	0.26
BETO	0.81	0.79	0.90	0.84	0.80	**0.93**	0.67	**0.42**	0.25	**0.98**	0.80	0.19	0.84	**0.68**
mBERT	0.79	0.77	0.88	**0.84**	0.78	0.90	**0.74**	0.33	0.31	0.97	0.78	0.12	0.83	0.60
BiLSTM w/ BETO+BERT	0.80	0.78	**0.91**	0.83	0.79	0.74	0.62	**0.42**	0.19	0.96	0.79	**0.31**	0.83	0.63
BiLSTM w/ Codeswitch	**0.82**	0.80	0.91	0.84	**0.81**	0.71	0.63	0.33	**0.50**	0.94	**0.81**	0.12	**0.85**	0.62

streaming vs *online*

Motivation for a benchmark

- Under-represented dimensions

- How do dimensions collide?

Motivation for a benchmark

- Under-represented dimensions

- How do dimensions collide?

  *El partido se retransmitirá por "**streaming**".*

Motivation for a benchmark

- Under-represented dimensions

- How do dimensions collide?

  *El partido se retransmitirá por "**streaming**".*

  *El partido se retransmitirá por **streaming**.*

Motivation for a benchmark

- Under-represented dimensions

- How do dimensions collide?

  *El partido se retransmitirá por "**streaming**".*

  *El partido se retransmitirá por **streaming**.*

  ***Streaming** del partido en TVE.*

Motivation for a benchmark

- Under-represented dimensions

- How do dimensions collide?

  *El partido se retransmitirá por "**streaming**".*

  *El partido se retransmitirá por **streaming**.*

  ***Streaming** del partido en TVE.*

  *"**Streaming**" del partido en TVE.*

Benchmark for Loanwords and Anglicisms in Spanish

- 1830 handcrafted sentences

- 1 anglicism per sentence

- Combination of dimensions exhaustively explored

Skyline

Skyline *de Nueva York en la nueva película de Woody Allen*

Benchmark for Loanwords and Anglicisms in Spanish

Skyline de Nueva York en la nueva película de Woody Allen

skyline de nueva york en la nueva película de woody allen

Skyline De Nueva York En La Nueva Película De Woody Allen

SKYLINE DE NUEVA YORK EN LA NUEVA PELÍCULA DE WOODY ALLEN

SKYLINE de Nueva York en la nueva película de Woody Allen

"Skyline" de Nueva York en la nueva película de Woody Allen

"skyline" de nueva york en la nueva película de woody allen

"Skyline" De Nueva York En La Nueva Película De Woody Allen

"SKYLINE" DE NUEVA YORK EN LA NUEVA PELÍCULA DE WOODY ALLEN

"SKYLINE" de Nueva York en la nueva película de Woody Allen

*El perfil de las cuatro torres caracteriza el **skyline** madrileño*

Benchmark for Loanwords and Anglicisms in Spanish

Skyline de Nueva York en la nueva película de Woody Allen

skyline de nueva york en la nueva película de woody allen

Skyline De Nueva York En La Nueva Película De Woody Allen

SKYLINE DE NUEVA YORK EN LA NUEVA PELÍCULA DE WOODY ALLEN

SKYLINE de Nueva York en la nueva película de Woody Allen

"Skyline" de Nueva York en la nueva película de Woody Allen

"skyline" de nueva york en la nueva película de woody allen

"Skyline" De Nueva York En La Nueva Película De Woody Allen

"SKYLINE" DE NUEVA YORK EN LA NUEVA PELÍCULA DE WOODY ALLEN

"SKYLINE" de Nueva York en la nueva película de Woody Allen

El perfil de las cuatro torres caracteriza el skyline madrileño

el perfil de las cuatro torres caracteriza el skyline madrileño

El Perfil De Las Cuatro Torres Caracteriza El Skyline Madrileño

EL PERFIL DE LAS CUATRO TORRES CARACTERIZA EL SKYLINE MADRILEÑO

El perfil de las cuatro torres caracteriza el Skyline madrileño

El perfil de las cuatro torres caracteriza el SKYLINE madrileño

El perfil de las cuatro torres caracteriza el "skyline" madrileño

el perfil de las cuatro torres caracteriza el "skyline" madrileño

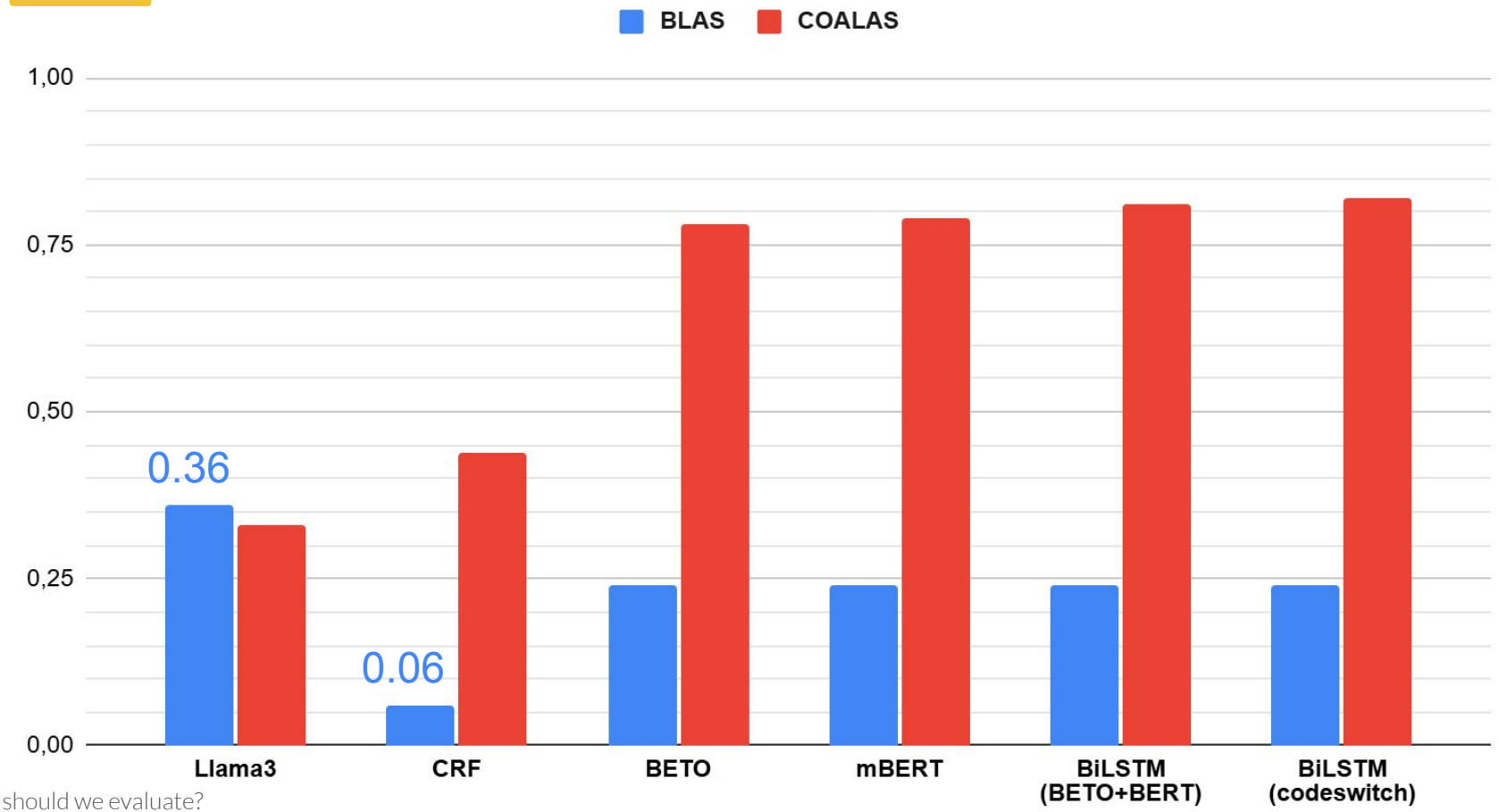
El Perfil De Las Cuatro Torres Caracteriza El "Skyline" Madrileño

EL PERFIL DE LAS CUATRO TORRES CARACTERIZA EL "SKYLINE" MADRILEÑO

El perfil de las cuatro torres caracteriza el "Skyline" madrileño

El perfil de las cuatro torres caracteriza el "SKYLINE" madrileño

Recall on BLAS vs COALAS



Recall across punctuation

Casing	CRF		BETO		mBERT		BiLSTM (un)		BiLSTM (cs)		Llama3	
	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Standard casing	0.22		**0.69**		0.61		0.66		0.60		0.66	
Text is lowercase	0.32		0.88		0.84		**0.90**		0.81		0.68	
Span is titlecase	0.00		0.05		0.05		0.06		0.06		**0.57**	
Text is titlecase	0.00		0.07		0.10		0.04		0.07		**0.33**	
Span is uppercase	0.00		0.00		0.00		0.00		0.01		**0.55**	
Text is uppercase	0.00		0.00		0.00		0.00		0.00		**0.22**	

Recall across punctuation

Casing	CRF		BETO		mBERT		BiLSTM (un)		BiLSTM (cs)		Llama3	
	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Standard casing	0.22		**0.69**		0.61		0.66		0.60		0.66	
Text is lowercase	0.32		0.88		0.84		**0.90**		0.81		0.68	
Span is titlecase	0.00		0.05		0.05		0.06		0.06		**0.57**	
Text is titlecase	0.00		0.07		0.10		0.04		0.07		**0.33**	
Span is uppercase	0.00		0.00		0.00		0.00		0.01		**0.55**	
Text is uppercase	0.00		0.00		0.00		0.00		0.00		**0.22**	

*El Perfil De Las Cuatro Torres Caracteriza El "**Skyline**" Madrileño*

*EL PERFIL DE LAS CUATRO TORRES CARACTERIZA EL "**SKYLINE**" MADRILEÑO*

Recall across punctuation

Casing	CRF		BETO		mBERT		BiLSTM (un)		BiLSTM (cs)		Llama3	
	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Standard casing	0.22		**0.69**		0.61		0.66		0.60		0.66	
Text is lowercase	0.32		0.88		0.84		**0.90**		0.81		0.68	
Span is titlecase	0.00		0.05		0.05		0.06		0.06		**0.57**	
Text is titlecase	0.00		0.07		0.10		0.04		0.07		**0.33**	
Span is uppercase	0.00		0.00		0.00		0.00		0.01		**0.55**	
Text is uppercase	0.00		0.00		0.00		0.00		0.00		**0.22**	

*El Perfil De Las Cuatro Torres Caracteriza El "**Skyline**" Madrileño*

*EL PERFIL DE LAS CUATRO TORRES CARACTERIZA EL "**SKYLINE**" MADRILEÑO*

Recall across anglicism types

Type	Length	Position	CRF		BETO		mBERT		BiLSTM (unad.)		BiLSTM (codeswitch)		Llama3	
			W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Compliant	Multi	Ini	0.00		0.40		0.20		0.20		0.30		**0.90**	
		Mid	0.20		**1.00**		0.90		**1.00**		0.90		0.80	
	Single	Ini	0.00		0.10		0.00		0.20		0.10		**0.50**	
		Mid	0.10		0.70		0.80		0.90		**1.00**		0.40	
Non-compliant	Multi	Ini	0.10		**0.90**		0.60		0.30		0.20		**0.90**	
		Mid	0.60		**1.00**		**1.00**		**1.00**		**1.00**		0.90	
	Single	Ini	0.00		0.20		0.20		0.20		0.10		**0.80**	
		Mid	0.50		**1.00**		**1.00**		**1.00**		**1.00**		0.60	
Mixed-compliant	Multi	Ini	0.00		0.30		0.20		0.20		0.20		**0.80**	
		Mid	0.10		**1.00**		0.90		**1.00**		0.90		0.70	
Ambiguous	Multi	Mid	0.00		0.50		0.50		**0.70**		0.50		**0.70**	
	Single	Mid	0.00		**0.33**		**0.33**		**0.33**		0.00		**0.33**	
Mixed-ambiguous	Multi	Ini	0.00		0.30		0.30		0.20		0.00		**0.90**	
		Mid	0.10		0.70		0.70		**1.00**		0.60		0.80	
Adjacent	Multi	Mid	0.45		**0.90**		0.65		0.80		0.80		0.55	
	Single	Mid	0.60		0.95		0.95		0.95		**1.00**		0.30	

*El partido se retransmitirá por "**streaming**" en TVE.*
*La agencia de "**fact checking**" desmintió el bulo.*

Recall across anglicism types

Type	Length	Position	CRF		BETO		mBERT		BiLSTM (unad.)		BiLSTM (codeswitch)		Llama3	
			W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Compliant	Multi	Ini	0.00		0.40		0.20		0.20		0.30		**0.90**	
		Mid	0.20		**1.00**		0.90		**1.00**		0.90		0.80	
	Single	Ini	0.00		0.10		0.00		0.20		0.10		**0.50**	
		Mid	0.10		0.70		0.80		0.90		**1.00**		0.40	
Non-compliant	Multi	Ini	0.10		**0.90**		0.60		0.30		0.20		**0.90**	
		Mid	0.60		**1.00**		**1.00**		**1.00**		**1.00**		0.90	
	Single	Ini	0.00		0.20		0.20		0.20		0.10		**0.80**	
		Mid	0.50		**1.00**		**1.00**		**1.00**		**1.00**		0.60	
Mixed-compliant	Multi	Ini	0.00		0.30		0.20		0.20		0.20		**0.80**	
		Mid	0.10		**1.00**		0.90		**1.00**		0.90		0.70	
Ambiguous	Multi	Mid	0.00		0.50		0.50		**0.70**		0.50		**0.70**	
	Single	Mid	0.00		**0.33**		**0.33**		**0.33**		0.00		**0.33**	
Mixed-ambiguous	Multi	Ini	0.00		0.30		0.30		0.20		0.00		**0.90**	
		Mid	0.10		0.70		0.70		**1.00**		0.60		0.80	
Adjacent	Multi	Mid	0.45		**0.90**		0.65		0.80		0.80		0.55	
	Single	Mid	0.60		0.95		0.95		0.95		**1.00**		0.30	

*"**Streaming**" del partido en TVE.*
*"**Fact checking**" del debate electoral.*

Recall across anglicism types

Type	Length	Position	CRF		BETO		mBERT		BiLSTM (unad.)		BiLSTM (codeswitch)		Llama3	
			W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Compliant	Multi	Ini	0.00		0.40		0.20		0.20		0.30		**0.90**	
		Mid	0.20		**1.00**		0.90		**1.00**		0.90		0.80	
	Single	Ini	0.00		0.10		0.00		0.20		0.10		**0.50**	
		Mid	0.10		0.70		0.80		0.90		**1.00**		0.40	
Non-compliant	Multi	Ini	0.10		**0.90**		0.60		0.30		0.20		**0.90**	
		Mid	0.60		**1.00**		**1.00**		**1.00**		**1.00**		0.90	
	Single	Ini	0.00		0.20		0.20		0.20		0.10		**0.80**	
		Mid	0.50		**1.00**		**1.00**		**1.00**		**1.00**		0.60	
Mixed-compliant	Multi	Ini	0.00		0.30		0.20		0.20		0.20		**0.80**	
		Mid	0.10		**1.00**		0.90		**1.00**		0.90		0.70	
Ambiguous	Multi	Mid	0.00		0.50		0.50		**0.70**		0.50		**0.70**	
	Single	Mid	0.00		**0.33**		**0.33**		**0.33**		0.00		**0.33**	
Mixed-ambiguous	Multi	Ini	0.00		0.30		0.30		0.20		0.00		**0.90**	
		Mid	0.10		0.70		0.70		**1.00**		0.60		0.80	
Adjacent	Multi	Mid	0.45		**0.90**		0.65		0.80		0.80		0.55	
	Single	Mid	0.60		0.95		0.95		0.95		**1.00**		0.30	

*"**Streaming**" del partido en TVE.*
*"**Fact checking**" del debate electoral.*

Recall across anglicism types

Type	Length	Position	CRF		BETO		mBERT		BiLSTM (unad.)		BiLSTM (codeswitch)		Llama3	
			W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Compliant	Multi	Ini	0.00	0.00	0.40	**0.50**	0.20	0.40	0.20	0.40	0.30	0.30	**0.90**	0.20
		Mid	0.20	0.00	**1.00**	0.80	0.90	0.80	**1.00**	0.80	0.90	**0.90**	0.80	0.30
	Single	Ini	0.00	0.00	0.10	**0.20**	0.00	0.10	0.20	**0.20**	0.10	**0.20**	**0.50**	0.10
		Mid	0.10	0.00	0.70	0.50	0.80	0.80	0.90	0.80	**1.00**	**0.90**	0.40	0.10
Non-compliant	Multi	Ini	0.10	0.10	**0.90**	0.50	0.60	**0.70**	0.30	0.40	0.20	**0.70**	**0.90**	0.30
		Mid	0.60	0.10	**1.00**	0.80	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.90	0.40
	Single	Ini	0.00	0.00	0.20	0.20	0.20	0.30	0.20	0.10	0.10	**0.40**	**0.80**	**0.40**
		Mid	0.50	0.40	**1.00**	**1.00**	**1.00**	0.90	**1.00**	0.90	**1.00**	**1.00**	0.60	0.50
Mixed-compliant	Multi	Ini	0.00	0.00	0.30	**0.40**	0.20	0.30	0.20	0.30	0.20	0.20	**0.80**	0.20
		Mid	0.10	0.00	**1.00**	0.90	0.90	0.90	**1.00**	**1.00**	0.90	0.80	0.70	0.30
Ambiguous	Multi	Mid	0.00	0.00	0.50	0.30	0.50	**0.40**	**0.70**	0.20	0.50	0.10	**0.70**	0.10
	Single	Mid	0.00	0.00	**0.33**	**0.33**	**0.33**	0.00	**0.33**	0.00	0.00	0.00	**0.33**	0.00
Mixed-ambiguous	Multi	Ini	0.00	0.00	0.30	**0.40**	0.30	0.30	0.20	0.20	0.00	0.00	**0.90**	0.20
		Mid	0.10	0.00	0.70	0.50	0.70	0.50	**1.00**	**0.60**	0.60	0.50	0.80	0.40
Adjacent	Multi	Mid	0.45	0.10	**0.90**	0.20	0.65	**0.30**	0.80	0.10	0.80	0.15	0.55	0.10
	Single	Mid	0.60	0.25	0.95	0.15	0.95	0.25	0.95	**0.40**	**1.00**	0.35	0.30	0.15

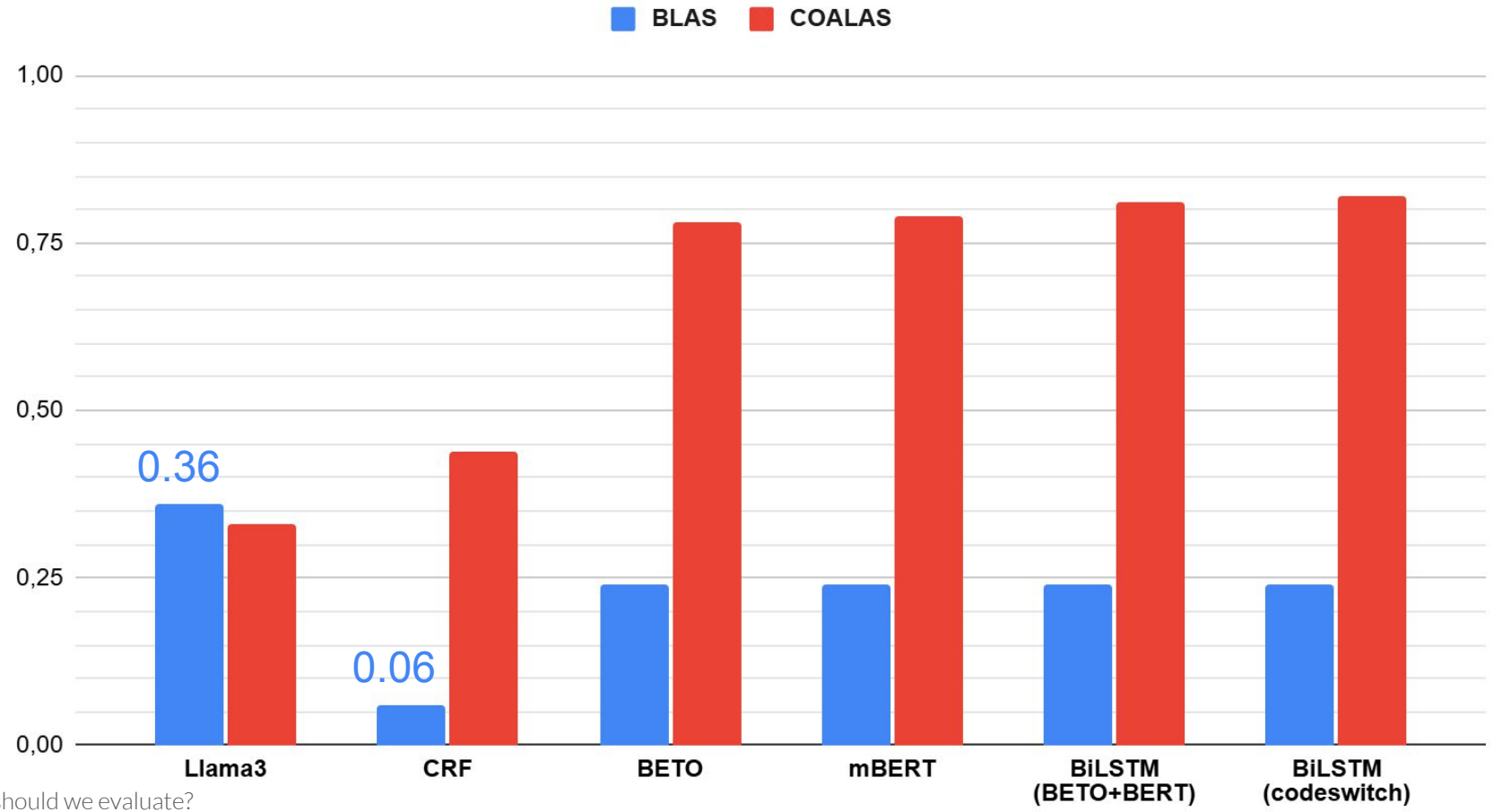
*El partido se retransmitirá por "**streaming**" en TVE.*
*El partido se retransmitirá por **streaming** en TVE.*

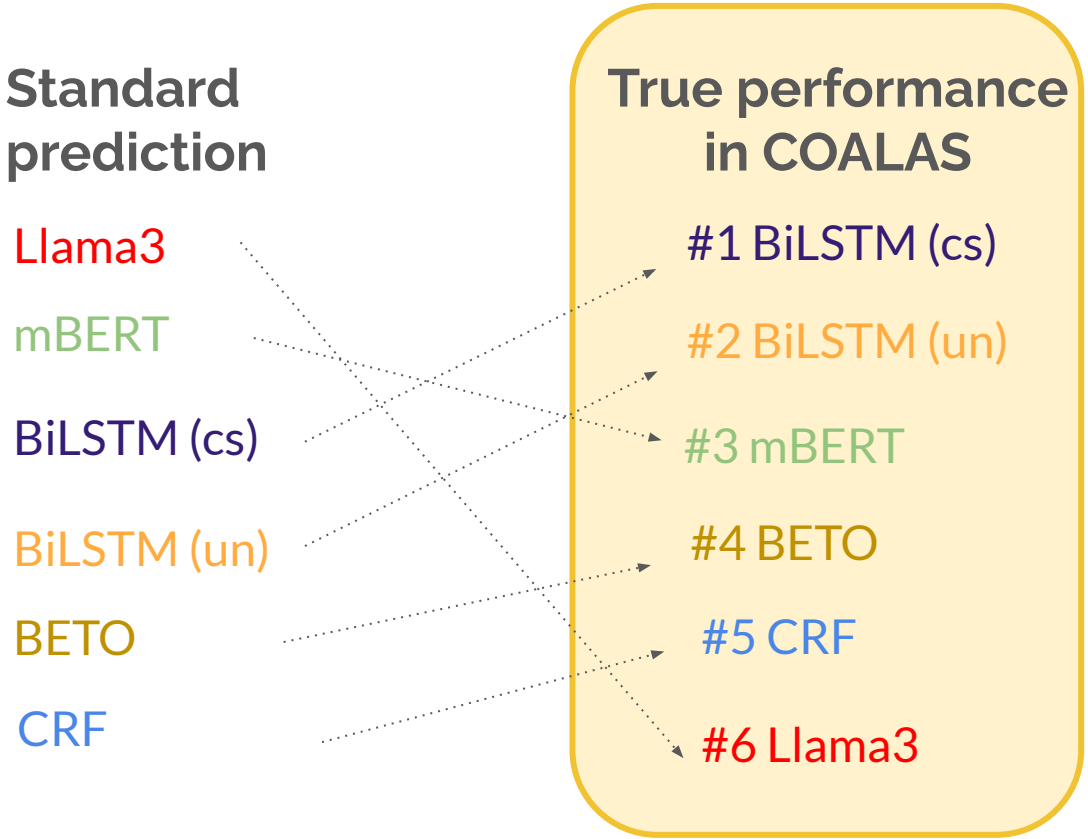
Recall across anglicism types

Type	Length	Position	CRF		BETO		mBERT		BiLSTM (unad.)		BiLSTM (codeswitch)		Llama3	
			W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Compliant	Multi	Ini	0.00	0.00	0.40	**0.50**	0.20	0.40	0.20	0.40	0.30	0.30	**0.90**	0.20
		Mid	0.20	0.00	**1.00**	0.80	0.90	0.80	**1.00**	0.80	0.90	**0.90**	0.80	0.30
	Single	Ini	0.00	0.00	0.10	**0.20**	0.00	0.10	0.20	**0.20**	0.10	**0.20**	**0.50**	0.10
		Mid	0.10	0.00	0.70	0.50	0.80	0.80	0.90	0.80	**1.00**	**0.90**	0.40	0.10
Non-compliant	Multi	Ini	0.10	0.10	**0.90**	0.50	0.60	**0.70**	0.30	0.40	0.20	**0.70**	**0.90**	0.30
		Mid	0.60	0.10	**1.00**	0.80	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.90	0.40
	Single	Ini	0.00	0.00	0.20	0.20	0.20	0.30	0.20	0.10	0.10	**0.40**	**0.80**	**0.40**
		Mid	0.50	0.40	**1.00**	**1.00**	**1.00**	0.90	**1.00**	0.90	**1.00**	**1.00**	0.60	0.50
Mixed-compliant	Multi	Ini	0.00	0.00	0.30	**0.40**	0.20	0.30	0.20	0.30	0.20	0.20	**0.80**	0.20
		Mid	0.10	0.00	**1.00**	0.90	0.90	0.90	**1.00**	**1.00**	0.90	0.80	0.70	0.30
Ambiguous	Multi	Mid	0.00	0.00	0.50	0.30	0.50	**0.40**	**0.70**	0.20	0.50	0.10	**0.70**	0.10
	Single	Mid	0.00	0.00	**0.33**	**0.33**	**0.33**	0.00	**0.33**	0.00	0.00	0.00	**0.33**	0.00
Mixed-ambiguous	Multi	Ini	0.00	0.00	0.30	**0.40**	0.30	0.30	0.20	0.20	0.00	0.00	**0.90**	0.20
		Mid	0.10	0.00	0.70	0.50	0.70	0.50	**1.00**	**0.60**	0.60	0.50	0.80	0.40
Adjacent	Multi	Mid	0.45	0.10	**0.90**	0.20	0.65	**0.30**	0.80	0.10	0.80	0.15	0.55	0.10
	Single	Mid	0.60	0.25	0.95	0.15	0.95	0.25	0.95	**0.40**	**1.00**	0.35	0.30	0.15

*El partido se retransmitirá por "**streaming**" en TVE.*
*El partido se retransmitirá por **streaming** en TVE.*

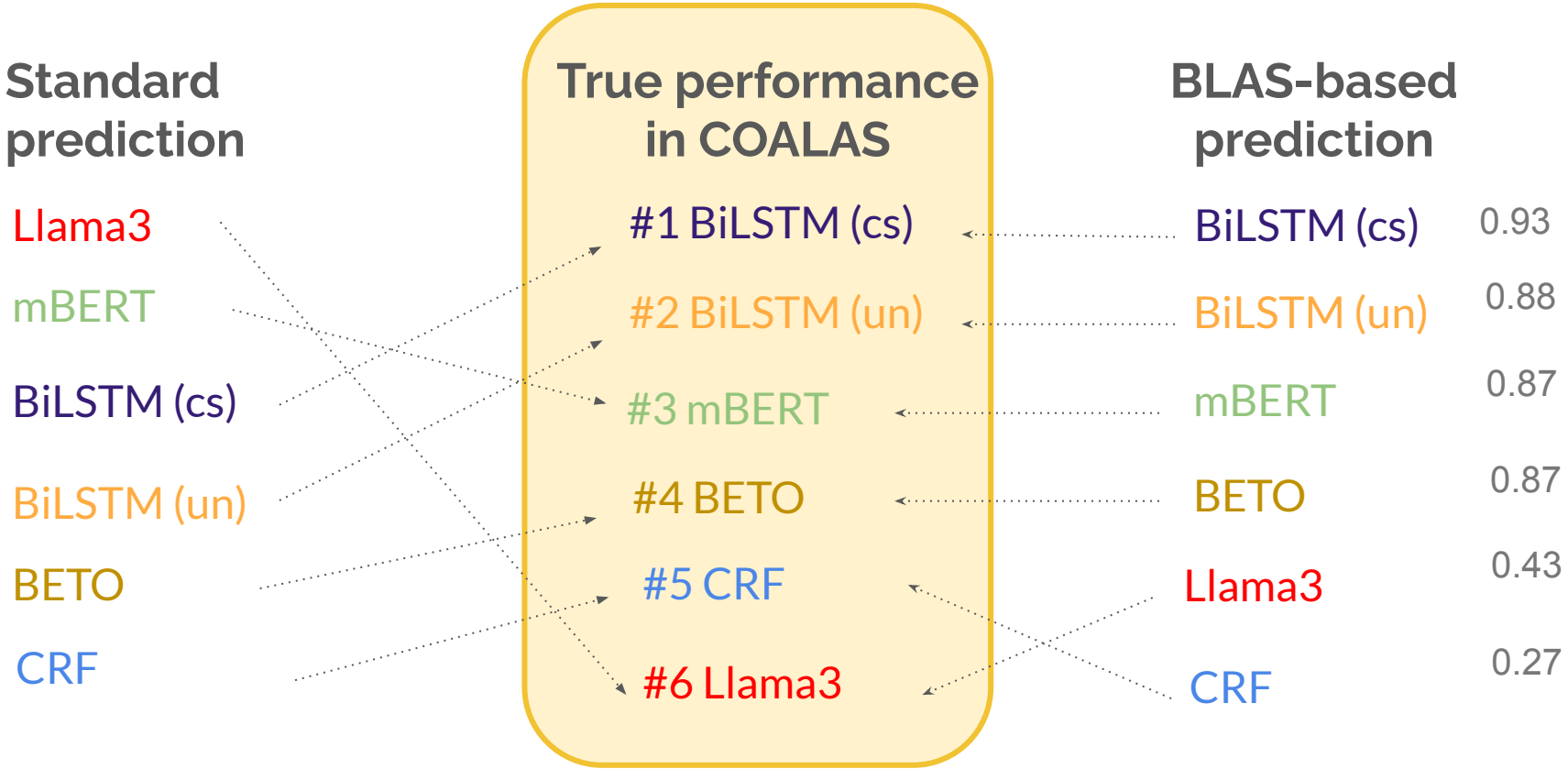
Recall on BLAS vs COALAS

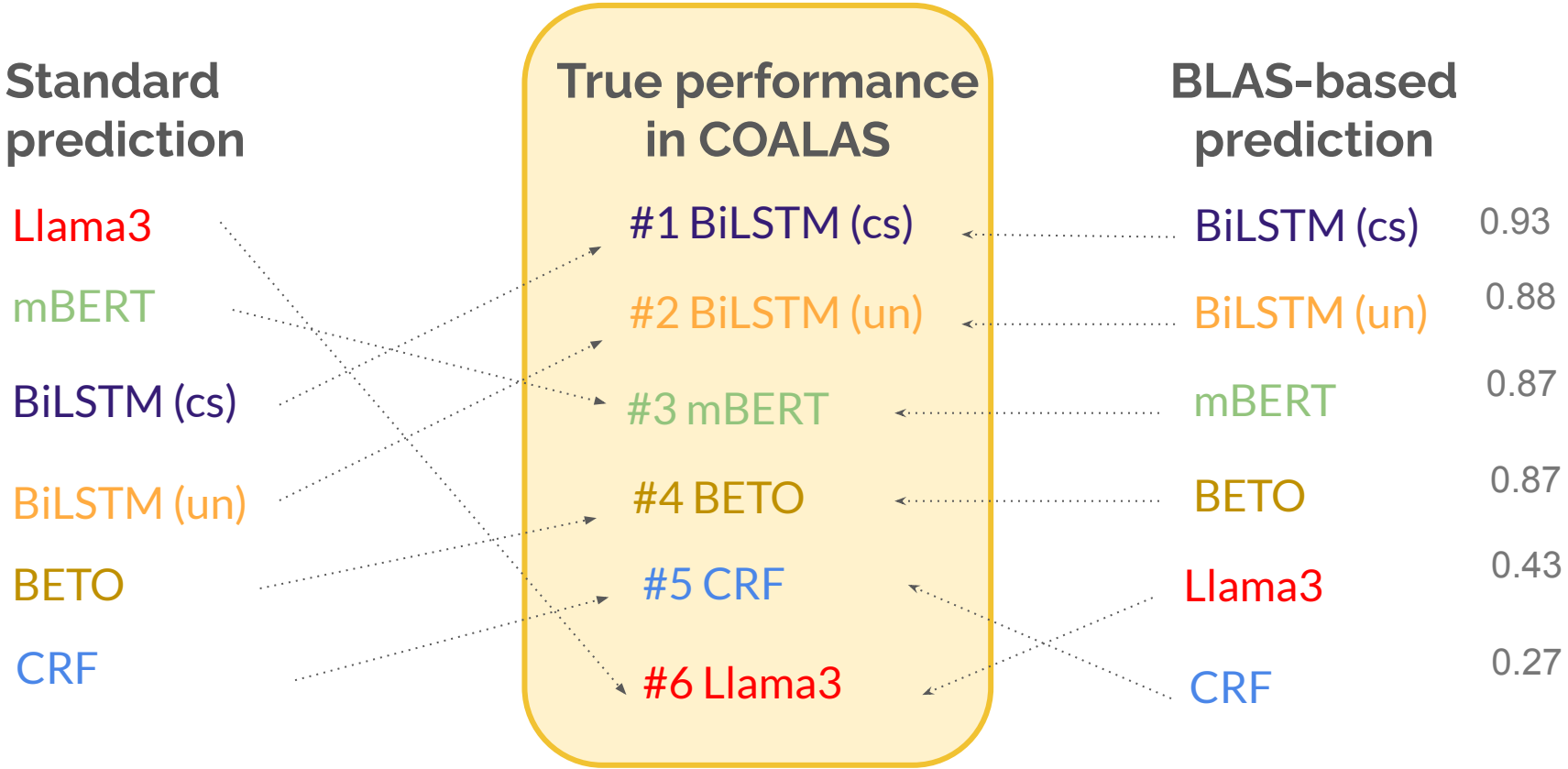



Standard prediction
Llama3
mBERT
BiLSTM (cs)
BiLSTM (un)
BETO
CRF
True performance in COALAS
#1 BiLSTM (cs)
#2 BiLSTM (un)
#3 mBERT
#4 BETO
#5 CRF
#6 Llama3

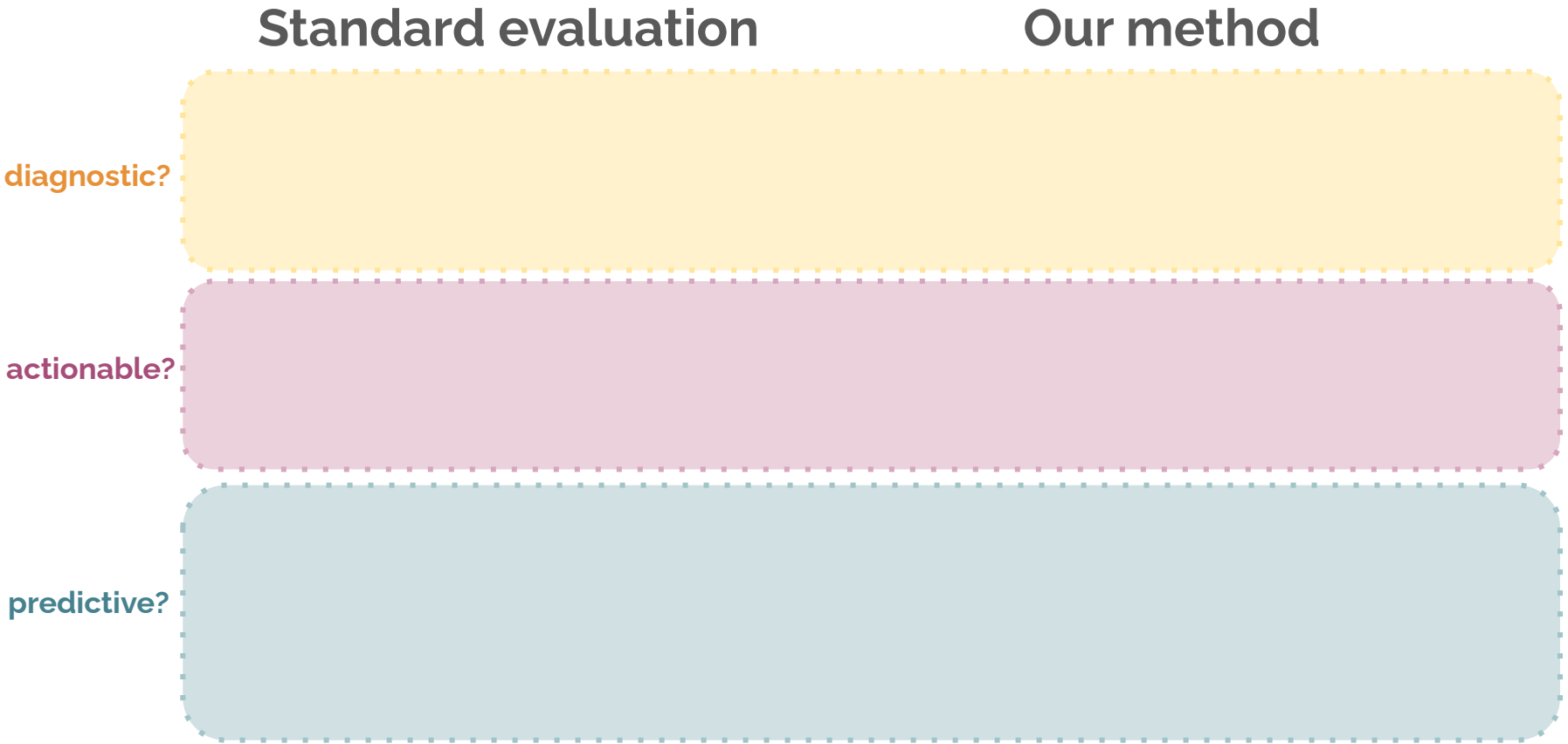
Recall across anglicism types

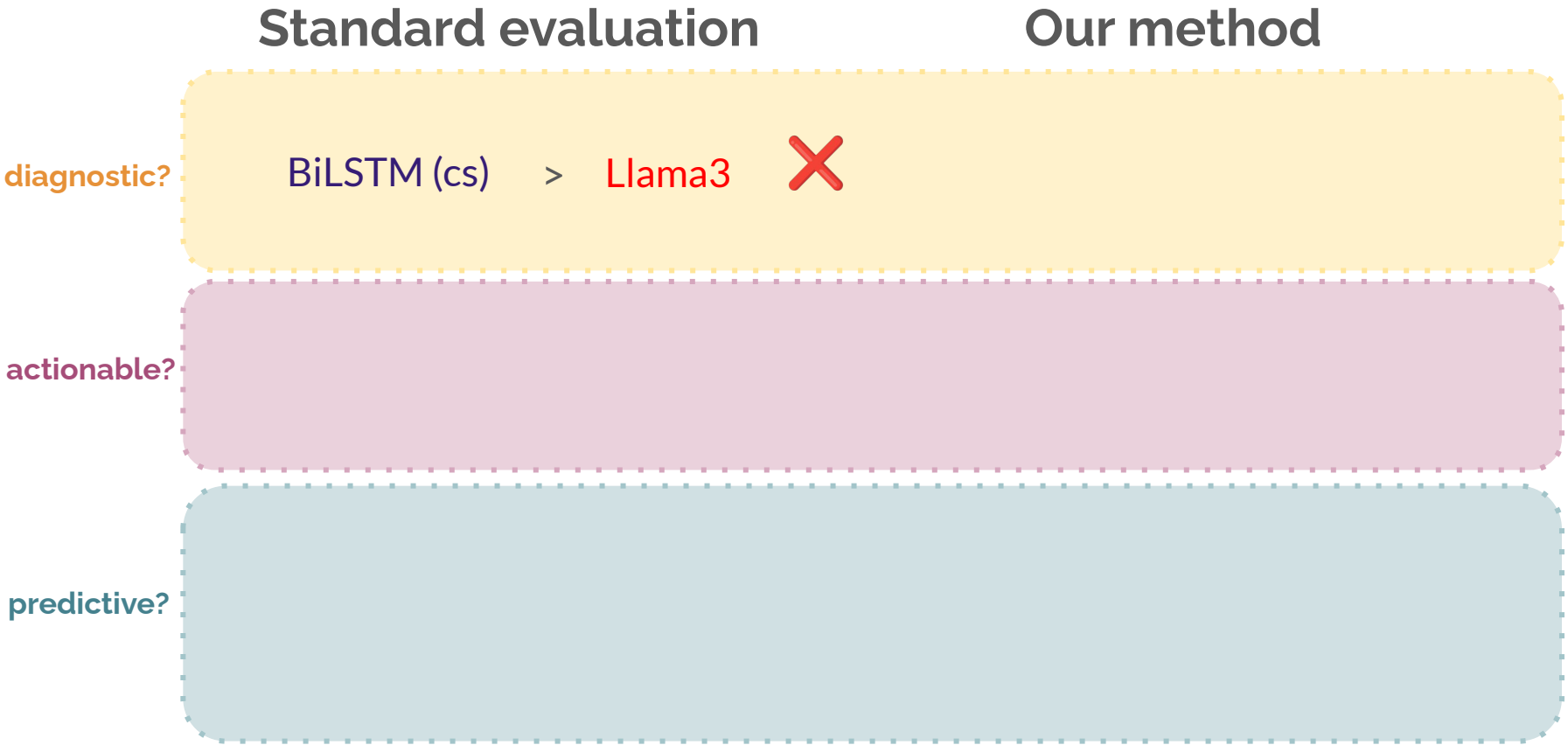
Type	Length	Position	CRF		BETO		mBERT		BiLSTM (unad.)		BiLSTM (codeswitch)		Llama3	
			W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Compliant	Multi	Ini	0.00	0.00	0.40	**0.50**	0.20	0.40	0.20	0.40	0.30	0.30	**0.90**	0.20
		Mid	0.20	0.00	**1.00**	0.80	0.90	0.80	**1.00**	0.80	0.90	**0.90**	0.80	0.30
	Single	Ini	0.00	0.00	0.10	**0.20**	0.00	0.10	0.20	**0.20**	0.10	**0.20**	**0.50**	0.10
		Mid	0.10	0.00	0.70	0.50	0.80	0.80	0.90	0.80	**1.00**	**0.90**	0.40	0.10
Non-compliant	Multi	Ini	0.10	0.10	**0.90**	0.50	0.60	**0.70**	0.30	0.40	0.20	**0.70**	**0.90**	0.30
		Mid	0.60	0.10	**1.00**	0.80	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.90	0.40
	Single	Ini	0.00	0.00	0.20	0.20	0.20	0.30	0.20	0.10	0.10	**0.40**	**0.80**	**0.40**
		Mid	0.50	0.40	**1.00**	**1.00**	**1.00**	0.90	**1.00**	0.90	**1.00**	**1.00**	0.60	0.50
Mixed-compliant	Multi	Ini	0.00	0.00	0.30	**0.40**	0.20	0.30	0.20	0.30	0.20	0.20	**0.80**	0.20
		Mid	0.10	0.00	**1.00**	0.90	0.90	0.90	**1.00**	**1.00**	0.90	0.80	0.70	0.30
Ambiguous	Multi	Mid	0.00	0.00	0.50	0.30	0.50	**0.40**	**0.70**	0.20	0.50	0.10	**0.70**	0.10
	Single	Mid	0.00	0.00	**0.33**	**0.33**	**0.33**	0.00	**0.33**	0.00	0.00	0.00	**0.33**	0.00
Mixed-ambiguous	Multi	Ini	0.00	0.00	0.30	**0.40**	0.30	0.30	0.20	0.20	0.00	0.00	**0.90**	0.20
		Mid	0.10	0.00	0.70	0.50	0.70	0.50	**1.00**	**0.60**	0.60	0.50	0.80	0.40
Adjacent	Multi	Mid	0.45	0.10	**0.90**	0.20	0.65	**0.30**	0.80	0.10	0.80	0.15	0.55	0.10
	Single	Mid	0.60	0.25	0.95	0.15	0.95	0.25	0.95	**0.40**	**1.00**	0.35	0.30	0.15

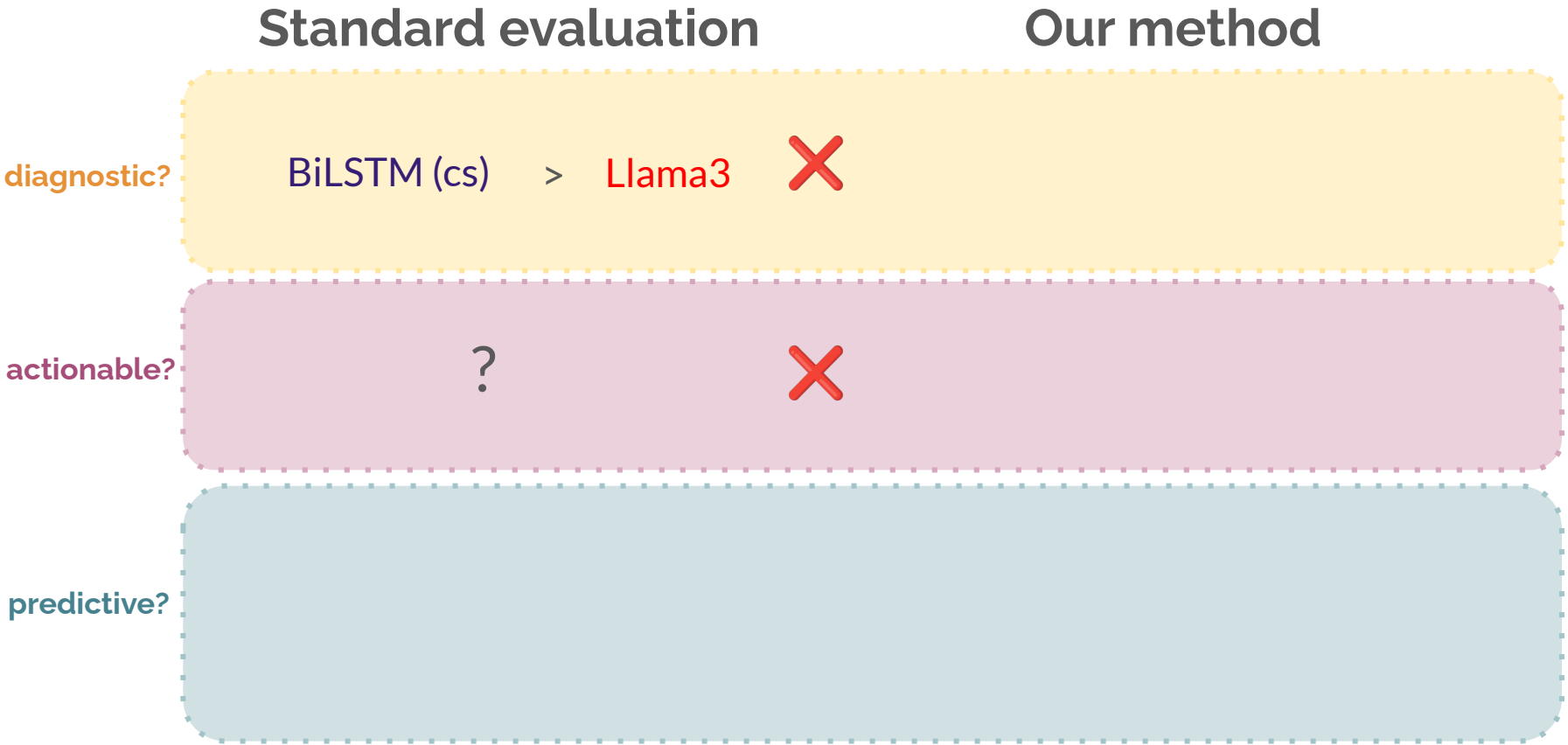

Standard prediction
Llama3
mBERT
BiLSTM (cs)
BiLSTM (un)
BETO
CRF
True performance in COALAS
#1 BiLSTM (cs)
#2 BiLSTM (un)
#3 mBERT
#4 BETO
#5 CRF
#6 Llama3
BLAS-based prediction
BiLSTM (cs)
0.93
BiLSTM (un)
0.88
mBERT
0.87
BETO
0.87
Llama3
0.43
CRF
0.27

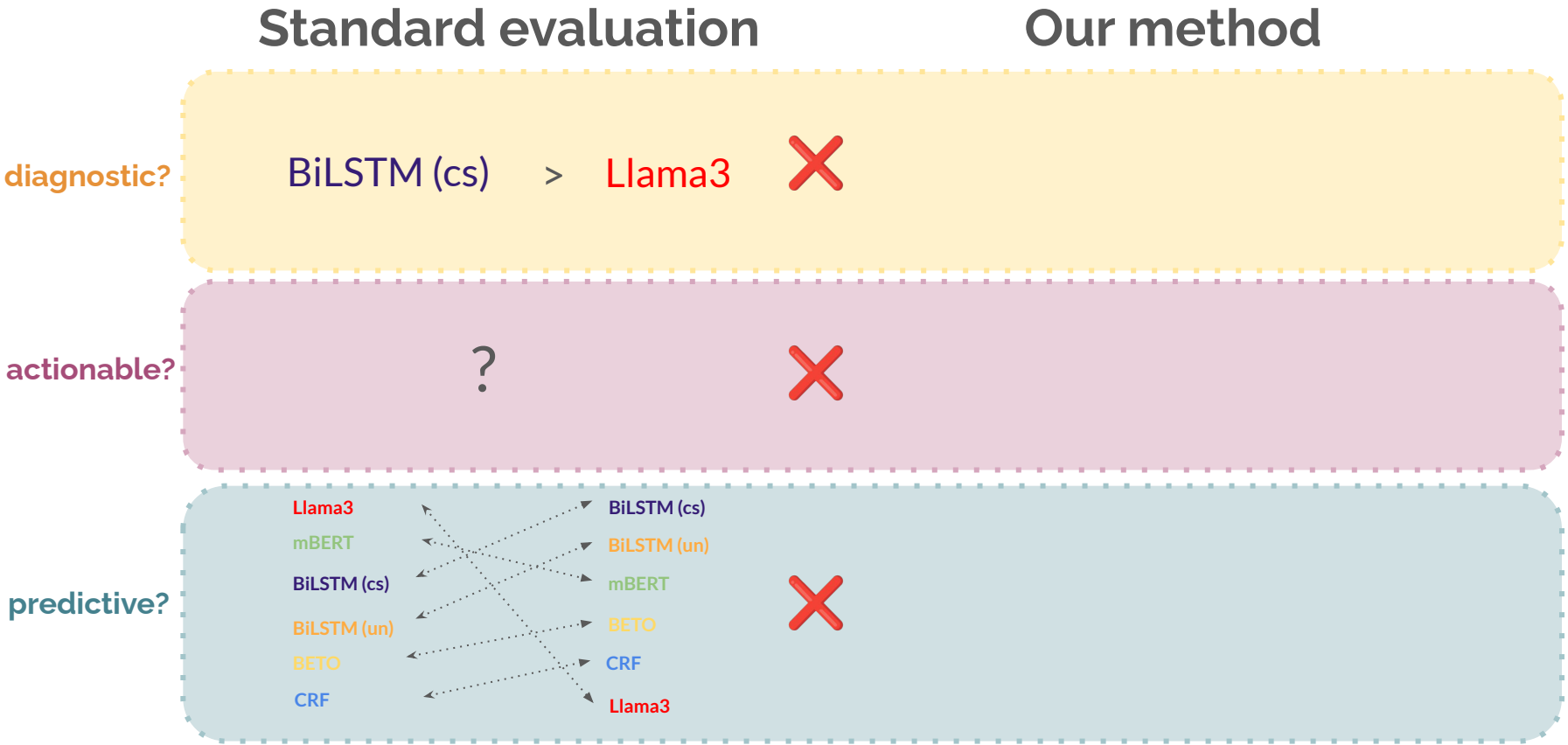


Median correlation over 5 external datasets = 0.85


Standard evaluation
Our method
diagnostic?
actionable?
predictive?


Standard evaluation
Our method
diagnostic?
BiLSTM (cs) > Llama3
actionable?
predictive?


Standard evaluation
Our method
diagnostic?
BiLSTM (cs)
>
Llama3
actionable?
?
predictive?


Standard evaluation
Our method
diagnostic?
BiLSTM (cs) > Llama3
actionable?
?
predictive?
Llama3
mBERT
BiLSTM (cs)
BiLSTM (un)
BETO
CRF
BiLSTM (cs)
BiLSTM (un)
mBERT
BETO
CRF
Llama3

Standard evaluation

Our method

diagnostic?

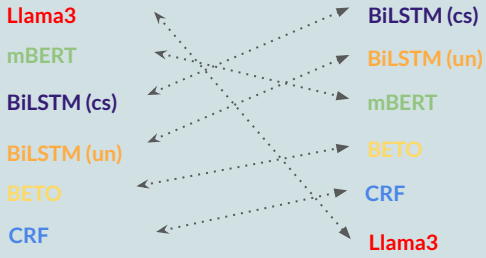
BiLSTM (cs) > Llama3 ✗

BiLSTM (cs): sentence initial spans
Llama3: spans w/o quotation marks ✓

actionable?

? ✗

predictive?



✗

Standard evaluation

Our method

diagnostic?

BiLSTM (cs) > Llama3 ✗

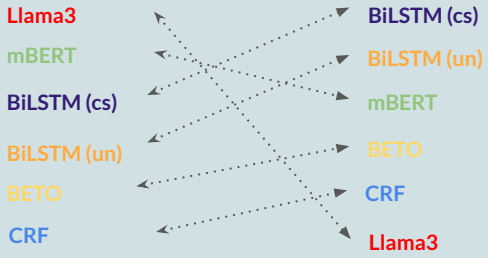
BiLSTM (cs): sentence initial spans
Llama3: spans w/o quotation marks ✓

actionable?

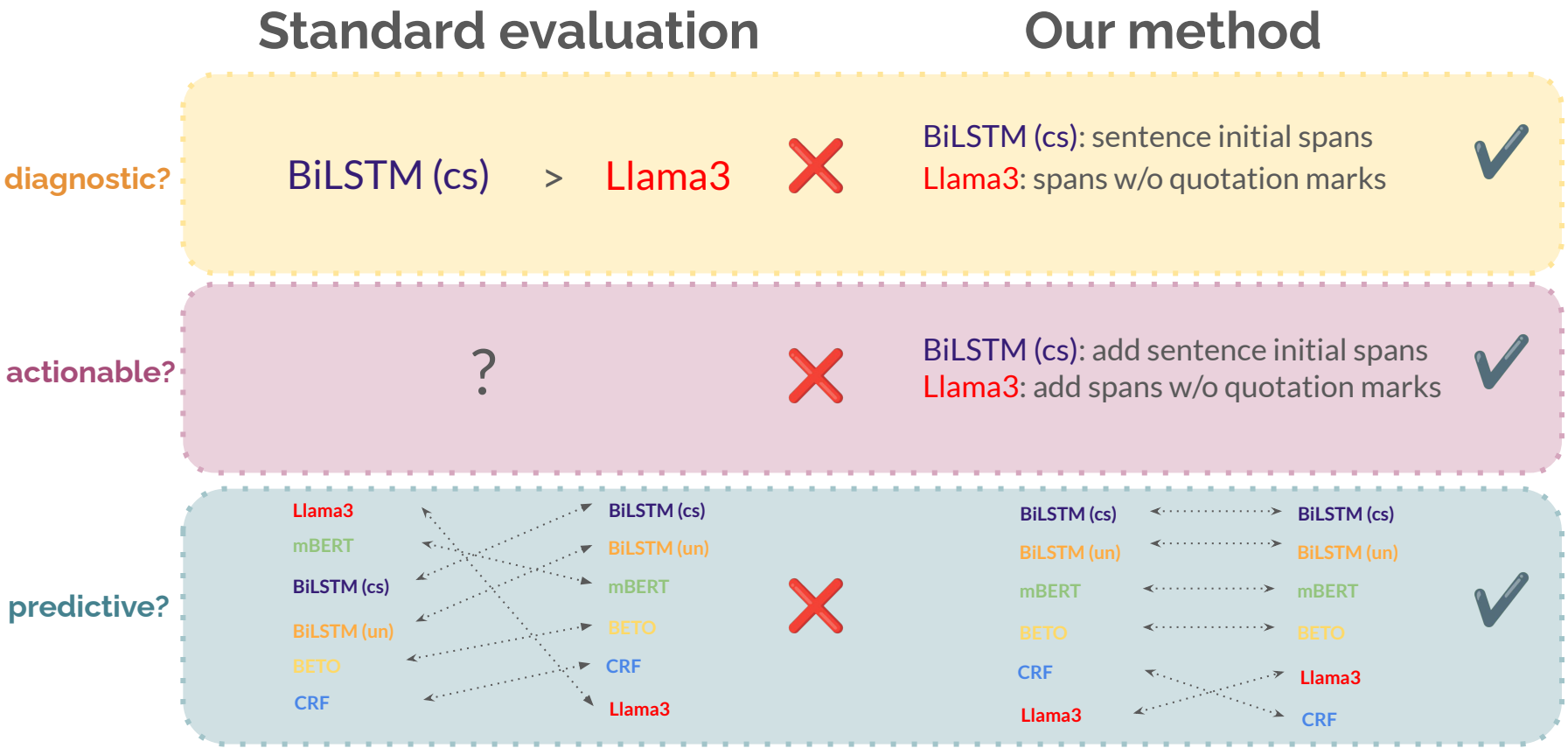
? ✗

BiLSTM (cs): add sentence initial spans
Llama3: add spans w/o quotation marks ✓

predictive?



✗


Standard evaluation
Our method
diagnostic?
BiLSTM (cs) > Llama3
BiLSTM (cs): sentence initial spans
Llama3: spans w/o quotation marks
actionable?
?
BiLSTM (cs): add sentence initial spans
Llama3: add spans w/o quotation marks
predictive?
Llama3
mBERT
BiLSTM (cs)
BiLSTM (un)
BETO
CRF
BiLSTM (cs)
BiLSTM (un)
mBERT
BETO
CRF
Llama3
BiLSTM (cs)
BiLSTM (un)
mBERT
BETO
CRF
Llama3
BiLSTM (cs)
BiLSTM (un)
mBERT
BETO
Llama3
CRF

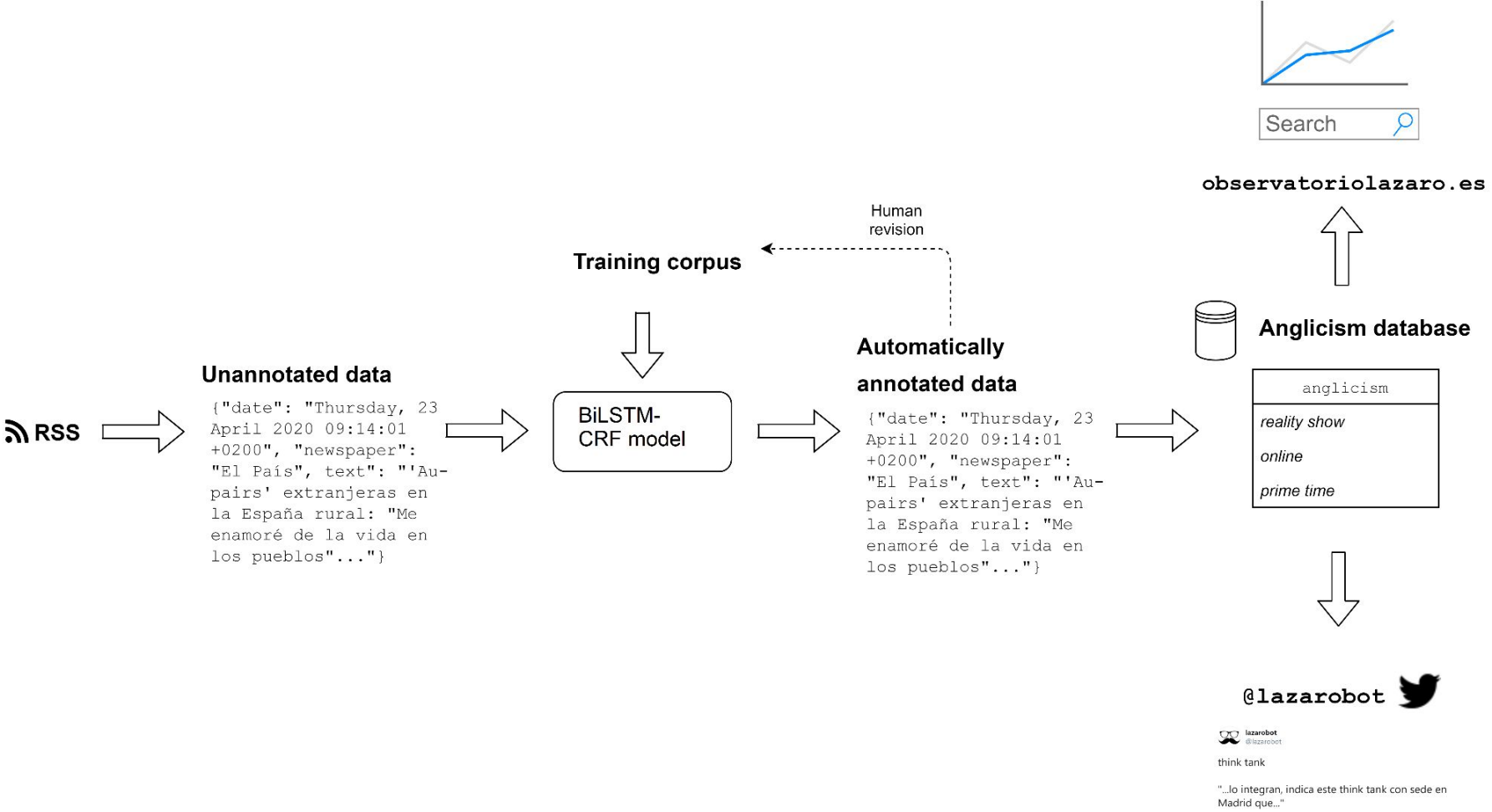
Contributions and final remarks

Observatorio Lázaro

Observatorio del anglicismo en la prensa española

22	759 255 030	1 511 233	60 359
Medios observados	Palabras analizadas	Anglicismos totales	Anglicismos únicos

Observatorio Lázaro analiza y extrae automáticamente los anglicismos aparecidos en las noticias del día de 22 medios españoles de prensa escrita, entre ellos elDiario.es, El País, El Mundo, ABC, La Vanguardia, El Confidencial, 20minutos, Agencia EFE, La Marea, El Economista, Marca, Fotogramas, Rolling Stone, Elle o El Mundo Today. Más sobre Observatorio Lázaro.


RSS
Unannotated data
{"date": "Thursday, 23 April 2020 09:14:01 +0200", "newspaper": "El País", text": "'Au-pairs' extranjeras en la España rural: "Me enamoré de la vida en los pueblos"..."}
BiLSTM-CRF model
Training corpus
Human revision
Automatically annotated data
{"date": "Thursday, 23 April 2020 09:14:01 +0200", "newspaper": "El País", text": "'Au-pairs' extranjeras en la España rural: "Me enamoré de la vida en los pueblos"..."}
Anglicism database
anglicism
reality show
online
prime time
Search
observatoriolazaro.es
@lazarobot
lazarobot
@lazarobot
think tank
"...lo integran, indica este think tank con sede en Madrid que..."

rider

Anglicismo: sí

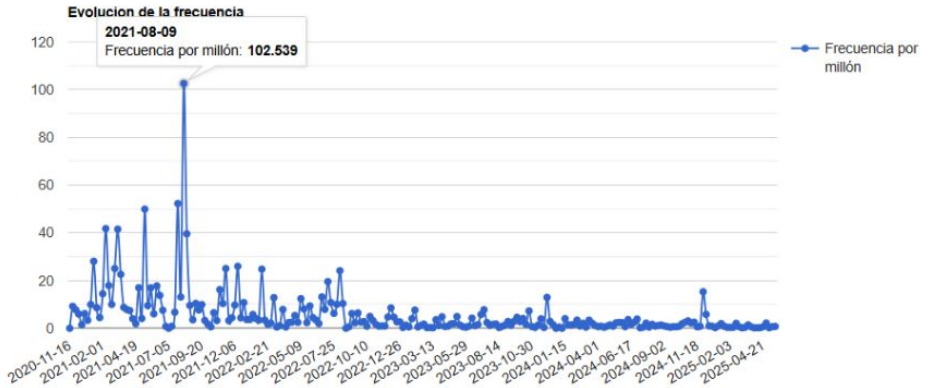
Formas: rider riders

Frecuencia media de aparición*: 4.148

Frecuencia de aparición en el último mes*: 0.796

Secciones habituales: Economía Portada Tecnología Deporte España

* Frecuencia por cada millón de palabras medida desde agosto de 2020.



black

Anglicismo: sí

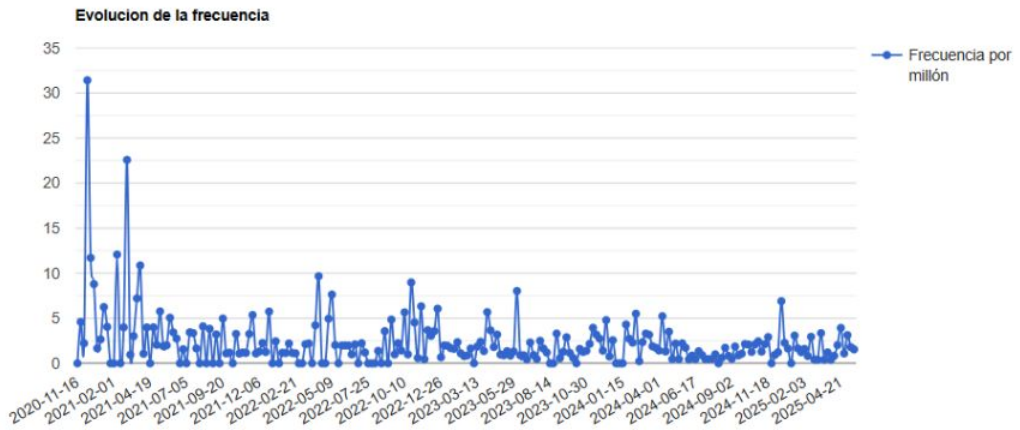
Formas: black blacks

Frecuencia media de aparición*: 2.382

Frecuencia de aparición en el último mes*: 2.655

Secciones habituales: Portada Economía Femenina Moda España

* Frecuencia por cada millón de palabras medida desde agosto de 2020.



prompt

Anglicismo: sí

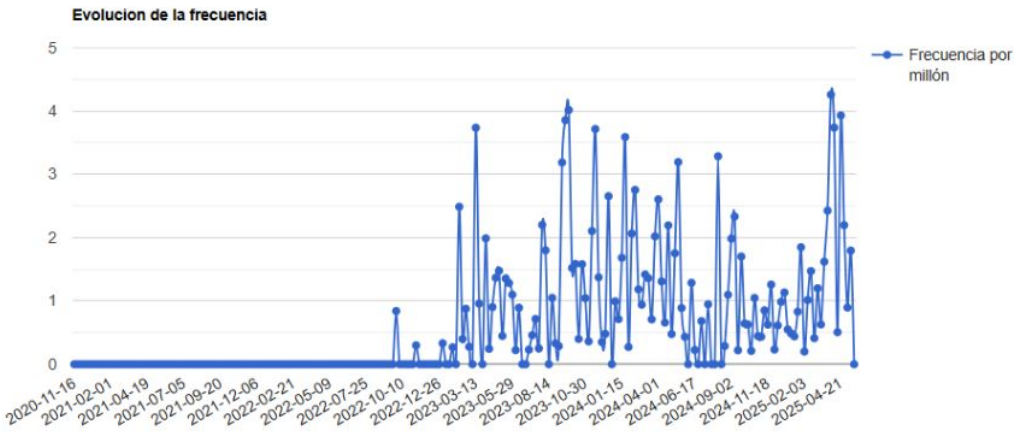
Formas: prompt prompts

Frecuencia media de aparición*: 0.789

Frecuencia de aparición en el último mes*: 2.23

Secciones habituales: Tecnología Economía Portada Cultura Ciencia

* Frecuencia por cada millón de palabras medida desde agosto de 2020.



chatbot

Anglicismo: sí

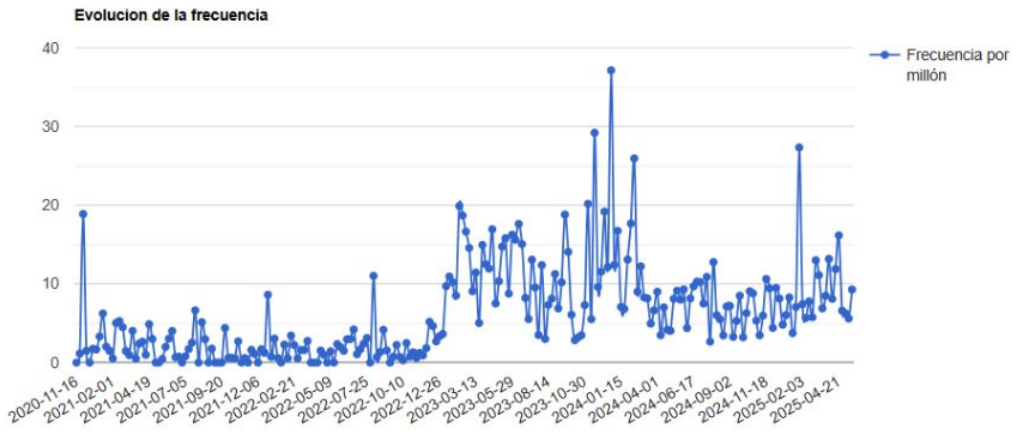
Formas: chatbot chatbots

Frecuencia media de aparición*: 7.074

Frecuencia de aparición en el último mes*: 8.549

Secciones habituales: Tecnología Economía Ciencia Portada Salud

* Frecuencia por cada millón de palabras medida desde agosto de 2020.



Daily numbers from Lázaro

Sources	22
Articles	850
Anglicisms	1050
Unique anglicisms	450
New anglicisms	45

Contributions

- Lexical borrowing identification as a sequence labeling task
- Data: collection and annotation methods for lexical borrowing
- Models: supervised models, LLM
- Evaluation: a methodology that is diagnostic, actionable and predictive

Resources

- 2 annotated datasets (COALAS, LM2013)
- Extensive annotation guidelines
- Suite of trained models (`HuggingFace`)
- 2 Python packages (`pylazaro`, `cascabel`)
- Benchmark for anglicism identification in Spanish (BLAS)
- A shared task (ADoBo 2021, 2025)
- A real-time pipeline that tracks anglicism usage in the Spanish press (Observatorio Lázaro)

Future work

- Real-world scenarios:
  - LLMs
  - Ensemble of models
  - Which metrics?

- How can we use dimensions?

- Diachronic dimension

Publications

- **Álvarez Mellado, E.**, Espinosa Anke, L., Gonzalo, J., Lignos, C., Porta Zamorano, J. (2021). Overview of ADoBo 2021: Automatic Detection of Unassimilated Borrowings in the Spanish Press. *Procesamiento del Lenguaje Natural*, *67*, 277-285. SJR: Q1 (Linguistics), Q2 (Computer Science). JCR: Q2 (Linguistics), Q4 (Computer Science).

CH. 5

- **Álvarez Mellado, E.**, Lignos, C. (2022). Detecting Unassimilated Borrowings in Spanish: An Annotated Corpus and Approaches to Modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3868-3888). GGS: A++.

CH. 3 + CH. 4

- **Álvarez Mellado, E.** (2022). Aproximaciones cuantitativas al estudio del anglicismo. In *Anglicismos en el español contemporáneo: Una visión panorámica*, ed. Rodríguez González, F. (pp. 87-113). Peter Lang.

CH. 2

- **Álvarez Mellado, E.**, Lignos, C. (2022). Borrowing or Codeswitching? Annotating for Finer-Grained Distinctions in Language Mixing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3195-3201). European Language Resources Association. GGS: B.

CH. 6

- **Álvarez Mellado, E.**, Gonzalo, J. (2024). Characterizing Spans for Sequence Labeling: A Case on Anglicism Detection. *Procesamiento del lenguaje natural*, *73*, 235-246. SJR: Q1 (Linguistics), Q2 (Computer Science). JCR: Q2 (Linguistics), Q4 (Computer Science).

CH. 8

Other publications

- Rueda, A., **Álvarez Mellado, E.**, Lignos, C. (2024). CoNLL#: Fine-grained Error Analysis and a Corrected Test Set for CoNLL-03 English. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 3718-3728). GGS: B.

- Amigó E., **Álvarez Mellado, E.**, Gonzalo J., Carrillo-de-Albornoz J. (forthcoming). Evaluating Sequence Labeling on the basis of Information Theory. Accepted to appear at *63th Annual Meeting of the Association for Computational Linguistics (ACL 2025) (Long Papers)*. GGS: A++.

- Ahmadi S., Hess M., **Álvarez Mellado E.**, Battisti E., Ding C., Göhring A., Gao Y., Jiang Z., Michail A., Morad P., Niklaus J., Panagiotopoulou M., Perrella S., Opitz J., Shaitarova A., Sennrich R. (forthcoming). ConLoan – A Contrastive Multilingual Dataset for Evaluating Loanwords. Accepted to appear at *63th Annual Meeting of the Association for Computational Linguistics (ACL 2025) (Long Papers)*. GGS: A++.

Impact: projects using Lázaro's data

- De Hoyos, J. C. (2023). Anglicismos en la lengua de la economía: entre el préstamo crudo y la adaptación léxica. *CLINA Revista Interdisciplinaria de Traducción Interpretación y Comunicación Intercultural*, *9*(1), 113-134.
- González, J. Á., Obrador, I. B., Herrero, Á. R., Sarvazyan, A. M., Chinea-Ríos, M., Basile, A., & Franco-Salvador, M. (2025). IberBench: LLM Evaluation on Iberian Languages. *arXiv preprint arXiv:2504.16921*.
- Lillo, A. (2022). Colloquial Anglicisms in Spanish toponymy. *Lebende Sprachen*, *67*(1), 133-167.
- Luján García, C. (2023). Anglicisms in Spanish gastronomy: new words for new eating habits. *Sintagma: revista de lingüística: 35*, *2023*, 51-69.
- Luján García, C. (2023). 'Drink for thought': Anglicismos en el campo de la bebida en la prensa digital española. *Borealis–An International Journal of Hispanic Linguistics*, *12*(2), 343-360.
- Luján García, C., & Nogueroles, E. E. N. (2024). On Political dream teams and Financial killers: Sports Anglicisms and Metaphorical Uses in Spanish Digital Press. *International Journal of English Studies*, *24*(1), 77-97.
- Núñez Nogueroles, E., & Luján García, C. I. (2022). Perceptions and reported use of information technology (it) anglicisms by spanish university students. *Miscelanea*.

Summary and conclusions

- Lexical borrowing identification as a sequence labeling task
- Data: collection and annotation methods for lexical borrowing
- Models: supervised models, LLM
- Evaluation: a methodology that is diagnostic, actionable and predictive

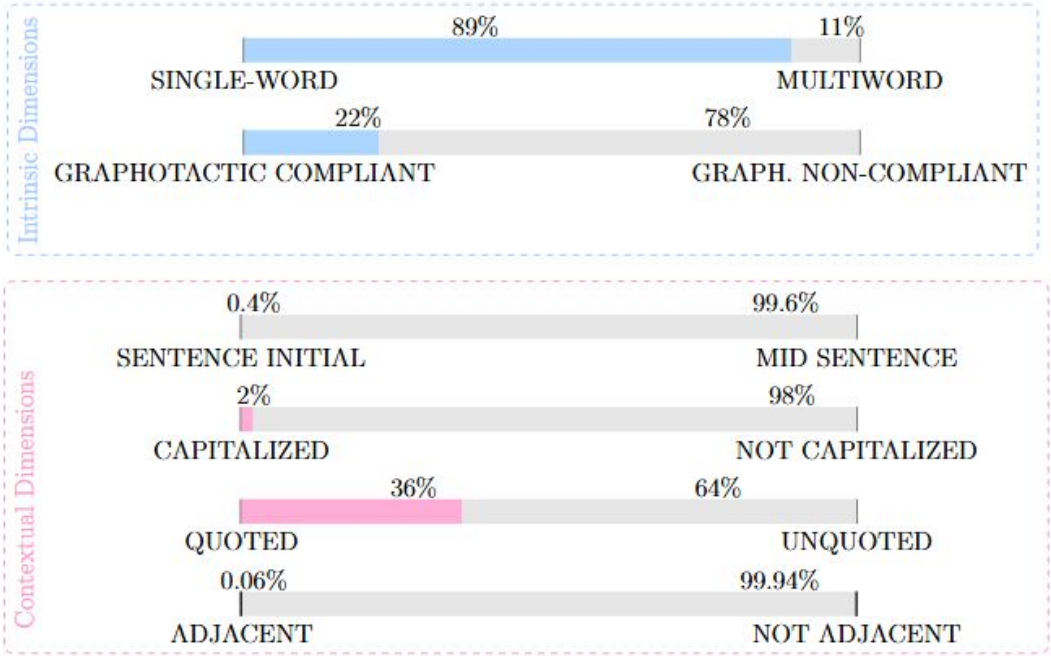
OMNIBVS MOBILIBVS MOBILIOR SCIENTIA

Rank	Model	Type	Model Size	↓ Lexical Analysis
9	microsoft/phi-4	Chat	14,145,704,960	31.98
4	utter-project/EuroLLM-9B-Instruct	Chat	9,152,319,488	25.87
12	utter-project/EuroLLM-9B	Pretrained	9,152,319,488	19.79
8	mistralai/Mistral-7B-Instruct-v0.3	Chat	7,113,805,824	17.84
3	Qwen/Qwen2.5-3B-Instruct	Chat	3,085,938,688	15.21
1	Qwen/Qwen2.5-7B-Instruct	Chat	7,070,619,136	13.46
2	IIC/RigoChat-7b-v2	Chat	7,615,616,512	13.26
14	google/gemma-2-2b-it	Chat	2,614,341,888	9.09
13	Qwen/Qwen2.5-1.5B-Instruct	Chat	1,543,714,304	6.24
6	microsoft/Phi-4-mini-instruct	Chat	3,836,021,760	3.39
5	meta-llama/Llama-3.1-8B-Instruct	Chat	7,504,924,672	3.23
11	meta-llama/Llama-3.2-3B-Instruct	Chat	3,212,749,824	2.8
10	HiTZ/Latxa-Llama-3.1-8B-Instruct	Chat	7,504,924,672	2.61

“[In sequence labeling tasks,] for tokens that are neither unseen nor ambiguous, predicting their label mainly involves memorizing their most frequent label in the training set”.

Bernier-Colborne, G., & Langlais, P. (2020). Hardeval: Focusing on challenging tokens to assess robustness of ner. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1704-1711).

Dimensions in COALAS training set



Recall across punctuation

Casing	CRF		BETO		mBERT		BiLSTM (un)		BiLSTM (cs)		Llama3	
	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Standard casing	0.22		**0.69**		0.61		0.66		0.60		0.66	
Text is lowercase	0.32		0.88		0.84		**0.90**		0.81		0.68	
Span is titlecase	0.00		5.20		0.05		0.06		0.06		**0.57**	
Text is titlecase	0.00		0.07		0.10		0.04		0.07		**0.33**	
Span is uppercase	0.00		0.00		0.00		0.00		0.01		**0.55**	
Text is uppercase	0.00		0.00		0.00		0.00		0.00		**0.22**	

*el perfil de las cuatro torres caracteriza el "**skyline**" madrileño*

*"**skyline**" de nueva york en la nueva película de woody allen*

Recall across punctuation

Casing	CRF		BETO		mBERT		BiLSTM (un)		BiLSTM (cs)		Llama3	
	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Standard casing	0.22		**0.69**		0.61		0.66		0.60		0.66	
Text is lowercase	0.32		0.88		0.84		**0.90**		0.81		0.68	
Span is titlecase	0.00		5.20		0.05		0.06		0.06		**0.57**	
Text is titlecase	0.00		0.07		0.10		0.04		0.07		**0.33**	
Span is uppercase	0.00		0.00		0.00		0.00		0.01		**0.55**	
Text is uppercase	0.00		0.00		0.00		0.00		0.00		**0.22**	

*el perfil de las cuatro torres caracteriza el "**skyline**" madrileño*

*"**skyline**" de nueva york en la nueva película de woody allen*

Recall across punctuation

Casing	CRF		BETO		mBERT		BiLSTM (un)		BiLSTM (cs)		Llama3	
	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Standard casing	0.22	0.07	**0.69**	0.45	0.61	**0.49**	0.66	0.45	0.60	0.46	0.66	0.23
Text is lowercase	0.32	0.16	0.88	0.57	0.84	0.65	**0.90**	0.61	0.81	**0.68**	0.68	0.28
Span is titlecase	0.00	0.00	5.20	5.20	0.05	0.06	0.06	5.78	0.06	0.06	**0.57**	**0.25**
Text is titlecase	0.00	0.00	0.07	0.06	0.10	0.05	0.04	0.03	0.07	0.07	**0.33**	**0.09**
Span is uppercase	0.00	0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.01	0.02	**0.55**	**0.45**
Text is uppercase	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.58	0.00	0.00	**0.22**	**0.06**

Recall across anglicism types

Type	Length	Position	CRF		BETO		mBERT		BiLSTM (unad.)		BiLSTM (codeswitch)		Llama3	
			W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Compliant	Multi	Ini		0.00		**0.50**		0.40		0.40		0.30		0.20
		Mid		0.00		0.80		0.80		0.80		**0.90**		0.30
	Single	Ini		0.00		**0.20**		0.10		**0.20**		**0.20**		0.10
		Mid		0.00		0.50		0.80		0.80		**0.90**		0.10
Non-compliant	Multi	Ini		0.10		0.50		**0.70**		0.40		**0.70**		0.30
		Mid		0.10		0.80		**1.00**		**1.00**		**1.00**		0.40
	Single	Ini		0.00		0.20		0.30		0.10		**0.40**		**0.40**
		Mid		0.40		**1.00**		0.90		0.90		**1.00**		0.50
Mixed-compliant	Multi	Ini		0.00		**0.40**		0.30		0.30		0.20		0.20
		Mid		0.00		0.90		0.90		**1.00**		0.80		0.30
Ambiguous	Multi	Mid		0.00		0.30		**0.40**		0.20		0.10		0.10
	Single	Mid		0.00		**0.33**		0.00		0.00		0.00		0.00
Mixed-ambiguous	Multi	Ini		0.00		**0.40**		0.30		0.20		0.00		0.20
		Mid		0.00		0.50		0.50		**0.60**		0.50		0.40
Adjacent	Multi	Mid		0.10		0.20		**0.30**		0.10		0.15		0.10
	Single	Mid		0.25		0.15		0.25		**0.40**		0.35		0.15

La actriz lució un look total black.

Recall across anglicism types

Type	Length	Position	CRF		BETO		mBERT		BiLSTM (unad.)		BiLSTM (codeswitch)		Llama3	
			W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Compliant	Multi	Ini	0.00	0.00	0.40	**0.50**	0.20	0.40	0.20	0.40	0.30	0.30	**0.90**	0.20
		Mid	0.20	0.00	**1.00**	0.80	0.90	0.80	**1.00**	0.80	0.90	**0.90**	0.80	0.30
	Single	Ini	0.00	0.00	0.10	**0.20**	0.00	0.10	0.20	**0.20**	0.10	**0.20**	**0.50**	0.10
		Mid	0.10	0.00	0.70	0.50	0.80	0.80	0.90	0.80	**1.00**	**0.90**	0.40	0.10
Non-compliant	Multi	Ini	0.10	0.10	**0.90**	0.50	0.60	**0.70**	0.30	0.40	0.20	**0.70**	**0.90**	0.30
		Mid	0.60	0.10	**1.00**	0.80	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.90	0.40
	Single	Ini	0.00	0.00	0.20	0.20	0.20	0.30	0.20	0.10	0.10	**0.40**	**0.80**	**0.40**
		Mid	0.50	0.40	**1.00**	**1.00**	**1.00**	0.90	**1.00**	0.90	**1.00**	**1.00**	0.60	0.50
Mixed-compliant	Multi	Ini	0.00	0.00	0.30	**0.40**	0.20	0.30	0.20	0.30	0.20	0.20	**0.80**	0.20
		Mid	0.10	0.00	**1.00**	0.90	0.90	0.90	**1.00**	**1.00**	0.90	0.80	0.70	0.30
Ambiguous	Multi	Mid	0.00	0.00	0.50	0.30	0.50	**0.40**	**0.70**	0.20	0.50	0.10	**0.70**	0.10
	Single	Mid	0.00	0.00	**0.33**	**0.33**	**0.33**	0.00	**0.33**	0.00	0.00	0.00	**0.33**	0.00
Mixed-ambiguous	Multi	Ini	0.00	0.00	0.30	**0.40**	0.30	0.30	0.20	0.20	0.00	0.00	**0.90**	0.20
		Mid	0.10	0.00	0.70	0.50	0.70	0.50	**1.00**	**0.60**	0.60	0.50	0.80	0.40
Adjacent	Multi	Mid	0.45	0.10	**0.90**	0.20	0.65	**0.30**	0.80	0.10	0.80	0.15	0.55	0.10
	Single	Mid	0.60	0.25	0.95	0.15	0.95	0.25	0.95	**0.40**	**1.00**	0.35	0.30	0.15

La actriz lució un "look" "total black".

Recall across punctuation

Casing	CRF		BETO		mBERT		BiLSTM (un)		BiLSTM (cs)		Llama3	
	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Standard casing		0.07		0.45		**0.49**		0.45		0.46		0.23
Text is lowercase		0.16		0.57		0.65		0.61		**0.68**		0.28
Span is titlecase		0.00		5.20		0.06		5.78		0.06		**0.25**
Text is titlecase		0.00		0.06		0.05		0.03		0.07		**0.09**
Span is uppercase		0.58		0.58		0.00		0.00		0.02		**0.45**
Text is uppercase		0.00		0.00		0.00		0.58		0.00		**0.06**

Recall across punctuation

Casing	CRF		BETO		mBERT		BiLSTM (un)		BiLSTM (cs)		Llama3	
	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Standard casing		0.07		0.45		**0.49**		0.45		0.46		0.23
Text is lowercase		0.16		0.57		0.65		0.61		**0.68**		0.28
Span is titlecase		0.00		5.20		0.06		5.78		0.06		**0.25**
Text is titlecase		0.00		0.06		0.05		0.03		0.07		**0.09**
Span is uppercase		0.58		0.58		0.00		0.00		0.02		**0.45**
Text is uppercase		0.00		0.00		0.00		0.58		0.00		**0.06**

Recall across punctuation

Casing	CRF		BETO		mBERT		BiLSTM (un)		BiLSTM (cs)		Llama3	
	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.	W. quot.	W/o quot.
Standard casing		0.07		0.45		**0.49**		0.45		0.46		0.23
Text is lowercase		0.16		0.57		0.65		0.61		**0.68**		0.28
Span is titlecase		0.00		5.20		0.06		5.78		0.06		**0.25**
Text is titlecase		0.00		0.06		0.05		0.03		0.07		**0.09**
Span is uppercase		0.58		0.58		0.00		0.00		0.02		**0.45**
Text is uppercase		0.00		0.00		0.00		0.58		0.00		**0.06**

streaming

Anglicismo: sí

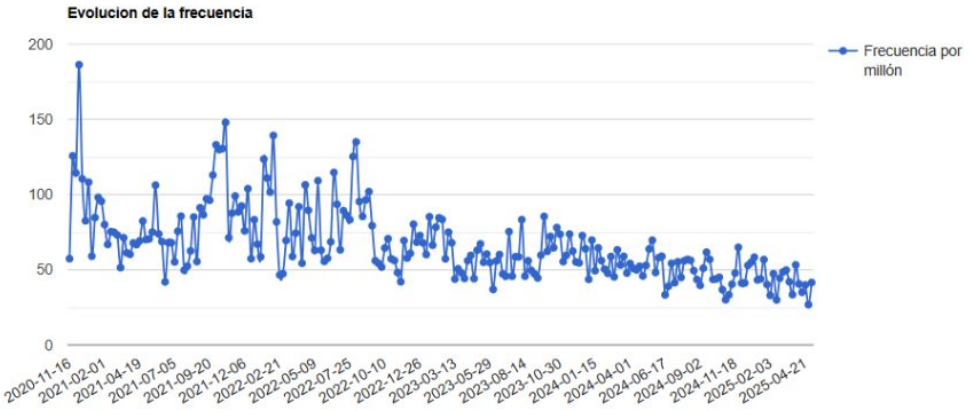
Formas: streaming streamings

Frecuencia media de aparición*: 61.429

Frecuencia de aparición en el último mes*: 38.422

Secciones habituales: Televisión Portada Cultura Tecnología Economía

* Frecuencia por cada millón de palabras medida desde agosto de 2020.



prime

Anglicismo: sí

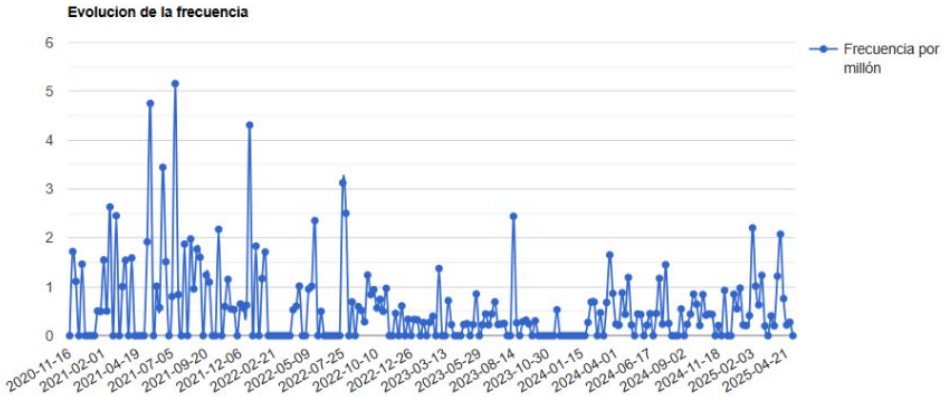
Formas: prime primes

Frecuencia media de aparición*: 0.521

Frecuencia de aparición en el último mes*: 0.276

Secciones habituales: Economía Deporte Portada Televisión Gente

* Frecuencia por cada millón de palabras medida desde agosto de 2020.



Show 10 entries

Search:

Anglicismo	Contexto	Medio	Fecha
prime	*Pero los alquileres no le andan a la zaga , con algunas rentas de propiedades ' prime ' que alcanzan cifras mareantes para el común de los mortales .*	elpais	01-05-2025
prime	*En esa línea , han aducido que prevén ofrecer esas oportunidades de forma " muy flexible " , tanto en inversión directa en activos ' prime ' Core y Core + , como a través de ' joint-ventures ' en " activos exclusivos " de mayor tamaño o a través de fondos en el caso de carteras o activos con mayor carga operativa en estrategias Value Add .*	eleconomista	23-04-2025
prime	*En ese mismo documento se explica que la ' yield ' o rentabilidad de este tipo de activos se sitúa en el 5 % en el caso de los activos ' prime ' de Madrid y Barcelona y del 6 % en propiedades ' prime ' de otras grandes ciudades .*	elpais	18-04-2025
prime	*Tras elogiar el paso de Carmen Alcayde por ' Supervivientes 2025 ' , Jorge Javier se centra en su crítica al artista , sumándose a las lanzadas por Laura Fa desde la propia Antena 3 : " Mientras Carmen está en su ' prime ' , Miguel Bosé no toca fondo " , introduce en su texto , señalando cómo " ha vuelto a nuestros vidas para anunciar , creo , una gira " .*	eldiario	16-04-2025
prime	*Está en su ' prime ' físico , ya se ha integrado plenamente en la dinámica de liderazgo y ha cubierto con éxito la bajada de la aportación de Facu .*	elconfidencial	11-04-2025
prime	*" A nivel artístico , digamos que estoy en mi ' prime ' .*	elmundo	10-04-2025
prime	*Omar Montes atraviesa ahora uno de los momentos más dulces de su carrera profesional , algo que él mismo afirmó : « A nivel artístico , digamos que estoy en mi ' prime ' .*	abc	10-04-2025
prime	*« A nivel artístico , digamos que estoy en mi ' prime ' .*	abc	10-04-2025
prime	*Ha comenzado el casting de ' OT 2025 ' con aquellos que aspiran a conseguir un pase ' prime ' cantándose una canción y colgando el momento en TikTok a través del hashtag # OTcover2025 .*	20minutos	07-04-2025
prime	*En cuanto al mercado de oficinas , " la tasa de ocupación se ha estabilizado en niveles próximos al 90 % en Barcelona y Madrid , e incluso en zonas ' prime ' se encuentra en niveles de plena ocupación , por encima del 95 %" .*	lavanguardia	07-04-2025

Showing 1 to 10 of 380 entries

Previous 1 2 3 4 5 ... 38 Next

Pulsa sobre la cabecera de las columnas para ordenar los resultados alfabéticamente, por fecha o por medio