Tesis doctoral
2025

Lexical borrowing detection as a sequence labeling task

Data, modeling and evaluation methods for anglicism retrieval in Spanish

Elena Álvarez Mellado

Dirigida por
Julio Gonzalo Arroyo
Constantine Lignos

Doctorado en Sistemas Inteligentes

Juzgo importante la conservación de la lengua de nuestros padres en su posible pureza, como un medio providencial de comunicación y un vínculo de fraternidad entre las varias naciones de origen español derramadas sobre los dos continentes. Pero no es un purismo supersticioso lo que me atrevo a recomendarles. El adelantamiento prodigioso de todas las ciencias y las artes, la difusión de la cultura intelectual y las revoluciones políticas piden cada día nuevos signos para expresar ideas nuevas, y la introducción de vocablos flamantes, tomados de las lenguas antiguas y extranjeras, ha dejado ya de ofendernos, cuando no es manifiestamente innecesaria, o cuando no descubre la afectación y mal gusto de los que piensan engalanar así lo que escriben.

Andrés Bello

Le temps altère toutes choses; il n'y a pas de raison pour que la langue échappe à cette loi universelle.

Ferdinand de Saussure

This is a lesson everyone who studies language eventually learns. You cannot stop language change. You may not like it; you may regret the arrival of new forms and the passing of old ones but there is not the slightest thing you can do about it. Language change is as natural as breathing. It is one of the linguistic facts of life.

David Crystal

ABSTRACT

Languages do not exist in isolation. Languages coexist; speakers of different languages interact. Contact between languages is one of the major inducers of language change. Speakers may take words that belong to one language and incorporate them into another language, a process that is known as *lexical borrowing*.

As lexical borrowings are a source of new words, the automatic identification of borrowings has been shown to be useful both for Linguistics (to assist data-driven research in contact linguistics, historical linguistics and lexicography) and in Natural Language Processing (as a pre-processing step for NLP downstream tasks, or when working on low-resource settings). However, previous work on lexical borrowing within Linguistics has mainly relied on manual inspection of corpora, an approach that is insufficient to account for an on-going and sparse phenomenon like borrowing incorporation. On the other hand, research within NLP has usually framed lexical borrowing identification as a word classification problem, in which isolated words deprived from context are classified as being a borrowing or not. This approach ignores the fact that context is key when deciding whether a given word is a borrowing or not, and is insufficient to account for multiword borrowings, adjacency or contextual ambiguity, which are all prevalent phenomena in real-world data.

This dissertation aims to provide computer-assisted methods for automatically retrieving lexical borrowings from text, with a focus on the identification of English lexical borrowings (or *anglicisms*) in Spanish. Contrary to previous work, we propose to frame lexical borrowing identification as a sequence labeling task, in which relevant spans of text are retrieved from sentences, in a similar fashion to how named entity recognition and multiword expressions are handled.

In the first half of this dissertation, we propose methods for collecting and annotating datasets of lexical borrowings in context, and apply them to produce COALAS, a novel corpus of Spanish journalistic texts annotated with anglicisms using BIO encoding. We use COALAS to analyze the performance and errors of several supervised sequence labeling models (CRF, BiLSTM-CRF, and Transformer-based models) and a large language model on a few-shot approach (8B Llama3). Our results show that a BiLSTM-CRF model fed with word and subword embeddings outperforms all other models (including Transformer-based models and the LLM, which performs poorly), and that all models are sensitive in varying degrees to capitalization, sentence position and shape.

In the second half of this dissertation, we explore the limitations that standard evaluation based on aggregated metrics produces when evaluating sequence labeling models in general and lexical borrowing identification systems in particular. We propose an evaluation methodology based on formal dimensions to characterize spans in sequence labeling tasks and apply it to build BLAS, a linguistically-motivated benchmark for anglicism identification in Spanish that can predict how a system will generalize to outside data.

Finally, we present *Observatorio Lázaro*, an observatory of anglicism usage in the Spanish press that implements the techniques and resources introduced in the course of this dissertation. The observatory monitors the daily usage of English lexical borrowings in the Spanish press and detects novel anglicisms that appear in Spanish newspapers. As of February 2025,

it has registered more than 1,400,000 occurrences of anglicisms since it was launched (with an average of 1,050 anglicisms retrieved daily; 450 of them unique, 45 of those previously unattested), which, to the best of our knowledge, makes *Observatorio Lázaro* the largest continuously-growing self-populating context-based database of anglicism usage ever compiled.

The results of this dissertation show that automatic identification of borrowings can indeed be framed as a sequence labeling task and that this approach is better suited than previous methods for capturing the linguistic nuances of borrowing usage in the wild and can successfully be applied to real-world scenarios.

RESUMEN Y CONCLUSIONES

Las lenguas no viven aisladas. Las lenguas conviven. El contacto entre lenguas es uno de los principales motores del cambio lingüístico. Los hablantes de un idioma usan en ocasiones palabras provenientes de otras lenguas, un proceso que se conoce como *préstamo léxico*.

El préstamo léxico es una fuente prolífica de palabras nuevas y, en consecuencia, la identificación automática de préstamos léxicos es relevante tanto en Lingüística (pues facilita la recogida de datos en disciplinas como la lingüística histórica o la lexicografía) como en Procesamiento de Lenguaje Natural (como pre-procesamiento para otras tareas de PLN o a la hora de trabajar con lenguas con pocos recursos). Sin embargo, el trabajo de identificación de préstamos léxicos en Lingüística se ha hecho tradicionalmente a mano, lo que resulta muy limitante para estudiar un fenómeno continuo y disperso como es la incorporación de préstamos léxicos. Por otro lado, trabajos previos en PLN han abordado la detección de préstamos léxicos como un problema de clasificación binaria, en el que la tarea consiste en clasificar palabras sueltas desprovistas de contexto. Esta aproximación pasa por alto el hecho de que es fundamental considerar el contexto en el que aparece un término para decidir si es un préstamo, y no permite dar cuenta adecuadamente de fenómenos como los préstamos multipalabra, los préstamos adyacentes o la ambigüedad contextual, fenómenos prevalentes todos ellos en la lengua real.

El objetivo de esta tesis es proponer métodos computacionales para la recuperación automática de préstamos léxicos en textos, con especial interés en la identificación de préstamos léxicos del inglés (también llamados *anglicismos*) en español. A diferencia de otros trabajos previos, proponemos abordar la identificación de préstamos léxicos como un problema de etiquetado de secuencias, en el que el objetivo es recuperar *spans* relevantes de longitud variable a partir de una frase, de un modo semejante a como se aborda el reconocimiento de entidades nombradas o la identificación de expresiones multipalabra.

En la primera mitad de esta tesis proponemos métodos para la recolección y anotación de conjuntos de datos que contengan préstamos en contexto y los aplicamos para producir COALAS, un corpus de prensa española anotado con anglicismos usando codificación BIO. A continuación, usamos COALAS para analizar el rendimiento y los errores de varios modelos supervisados de etiquetado de secuencias (un CRF, modelos basados en Transformers y varios BiLSTM-CRF con distintos tipos de *embeddings*), así como de un gran modelo de lenguaje en modo *few shot* (8B Llama3). Nuestros experimentos muestran que un BiLSTM-CRF alimentado con *embeddings* de palabra y subpalabra obtiene mejores resultados que todos los demás modelos (incluyendo los modelos basados en Transformers y el LLM, que obtiene resultados decepcionantes), y que todos los modelos son sensibles en mayor o menor medida a la posición que ocupa el anglicismo dentro de la oración, la presencia de mayúsculas o la combinación de caracteres que lo conforman.

En la segunda mitad de la tesis ahondamos en los problemas que acarrea usar métricas agregadas a la hora de evaluar tareas de etiquetado de secuencias como la nuestra. Proponemos una nueva metodología basada en dimensiones formales para caracterizar lingüísticamente los *spans* que aparecen en conjuntos de datos de etiquetado de secuencias y la aplicamos en la creación de BLAS, un *benchmark* para la identificación de anglicismos en

español con capacidad predictiva: los resultados que los modelos obtienen sobre el *benchmark* BLAS permiten anticipar el rendimiento que esos mismos modelos obtendrán en colecciones de datos externas.

Por último, presentamos *Observatorio Lázaro*, un observatorio del anglicismo en la prensa española que implementa lás técnicas y recursos producidos a lo largo de esta tesis. El observatorio monitoriza el uso diario de anglicismos en la prensa española e identifica nuevos préstamos que aparecen en los periódicos. Desde su lanzamiento en abril de 2020 hasta febrero de 2025, el observatorio ha registrado más de 1.400.000 apariciones de anglicismos (con una media de 1.050 anglicismos diarios detectados; 450 de ellos únicos, de los cuales 45 serán nuevos), lo que convierte a *Observatorio Lázaro* en la mayor base de datos léxica pública de este tipo automáticamente generada y en constante crecimiento.

Los resultados de esta tesis muestran que la identificación automática de préstamos léxicos puede efectivamente ser formulada como un problema de etiquetado de secuencias, y que esta aproximación permita dar cuenta mejor de la variabilidad lingüística que caracteriza el uso de préstamos léxicos en la lengua real.

Agradecimientos

Esa pregunta se la han hecho un poco todas las madres a lo largo de la historia. Cómo lo voy a hacer, se han dicho millones a sí mismas, igual que me decía yo en silencio, mientras aparentaba tenerlo todo bajo control.

"La historia de los vertebrados", Mar García Puig

El autobús de la línea 83 recorre el trayecto que va desde la Ciudad Universitaria de Madrid hasta el Barrio del Pilar. Un día de 2009 yo viajaba en el 83 camino a casa desde la Facultad de Filología de la Complutense. Hacía unos meses que había abandonado la carrera de Medicina y aún buscaba mi sitio como nueva estudiante de Lingüística. Por las mañanas estaba aprendiendo a programar en Prolog. Por las tardes iba a clase de Historia de la lengua. Acababa de descubrir que programar me parecía lo más divertido que había hecho en mi vida, cosa que chocaba frontalmente con la posición de rechazo casi identitaria que había mantenido hasta entonces hacia todo lo que tuviera que ver con los ordenadores y la tecnología.

Aquel día, mientras viajaba en autobús, se me ocurrió una idea: convertir las reglas que estaba estudiando en clase de Historia de la lengua en reglas de Prolog. Si todas las palabras en castellano terminan en *r*, *s*, *l*, *n*, *d*, *z* o vocal, entonces podía hacer un programa que identificara como imposibles en castellano todas aquellas palabras que no cumplieran esa restricción. De aquella idea feliz que tuve montada en el 83 nació AZRAEL, un programita informático *amateur* e ineficiente a más no poder que reconocía palabras que podían ser castellano a partir de su estructura silábica.

Unos años después, en 2014, yo trabajaba en la Fundéu, el servicio de consultas y recomendaciones lingüísticas dirigido a periodistas y hablantes menesterosos que por aquel entonces estaba ubicado en la Agencia EFE. Todas las mañanas mis compañeros se reunían para comentar las noticias del día susceptibles de generar dudas a los hablantes. ¿Cómo se escribe esta nueva palabra? ¿Qué hacemos con este anglicismo? ¿Habéis oído esta expresión? Yo no me dedicaba a la parte de resolver dudas, pero asistía embelesada a aquellas reuniones matutinas en las que en la planta 14 de un edificio de la Avenida de Burgos un grupo de lingüistas se juntaba para analizar la lengua a través del cristal de las noticias. De mi paso por la Fundéu me quedó la costumbre de andar siempre ojo avizor a la caza de las novedades lingüísticas que se abren paso en los titulares de prensa y también la convicción de que el trabajo de rastreo e identificación de novedades léxicas que había visto hacer manualmente a mis compañeros podía hasta cierto punto automatizarse.

En el otoño de 2019 yo estudiaba en la Universidad de Brandeis. Ese semestre se había

incorporado al departamento un profesor nuevo que había trabajado en un tema del que yo no había oído hablar: el *codeswitching* o la alternancia de código, ese fenómeno por el que los hablantes multilingües saltan de una lengua a otra, a veces dentro de una misma oración. Su trabajo consistía en identificar automáticamente en un texto los puntos de *codeswitching*. Yo no sabía nada de *codeswitching*, pero sus clases sobre etiquetado de secuencias me gustaban mucho, así que me puse a pensar posibles temas de investigación que pudieran hacer coincidir sus intereses y los míos. Y me acordé de AZRAEL, y me acordé de la Fundéu, y se me ocurrió que la detección de préstamos léxicos (que es un fenómeno similar aunque distinto al *codeswitching*) podía abordarse tal vez como un problema de etiquetado de secuencias y quizá así podría hacer el observatorio lingüístico que había imaginado en las reuniones de la Fundéu y automatizar el trabajo de amanuense que había visto hacer a mis compañeros. *Lo llamaré Lázaro*, pensé entonces, acordándome de Carreter y sus dardos. Aquel profesor nuevo era Constantine Lignos, que hoy codirige esta tesis.

Si me remonto en el tiempo no es por nostalgia, sino porque el origen de esta investigación no se circunscribe al trabajo de estos últimos cuatro años, sino que viene de mucho más atrás. A veces siento que llevo quince años dándole vueltas a lo mismo. Los agradecimientos, por tanto, deben remontarse también otros tantos años atrás.

Quiero dar las gracias a mis directores de tesis. A Constantine Lignos por acompañarme y guiarme desde el comienzo de este proyecto y por enseñarme cuanto sé del etiquetado de secuencias. A Julio Gonzalo por su fe ciega en mi criterio y por dejarme plena libertad para investigar en un tema absolutamente de nicho por el único motivo de que a mí me interesaba.

Gracias a mis compañeros del departamento y de doctorado por la solidaridad y la camaradería, muy especialmente a los doctorandos *julianos* (Guillermo, Eva, Pablo), y a los veteranos Agustín y Andrés. Algunas de las conversaciones más interesantes de estos años las he tenido con ellos. Me siento particularmente en deuda con Andrés Duque: su presencia en el papel de hermano mayor (ya fuera resolviendo cuestiones mundanas sobre plazos y burocracias, respondiendo preguntas de investigación o dando apoyo moral en general) ha sido para mí fundamental a lo largo de este proceso y le estoy muy agradecida. Gracias también a Víctor Fresno por acompañar esta tesis en sus comienzos y por su entusiasmo.

Quiero dar las gracias también a Luis Espinosa y Jordi Porta, con quienes organizamos la primera edición de ADoBo en 2021. Fue Jordi quien me animó a proponer la tarea, lo que me llevó a confeccionar el corpus para la competición que acabaría siendo COALAS y que en último término propició que siguiera adelante con este tema de investigación.

Por motivos ahora evidentes, esta tesis es deudora de mi paso por la Fundéu entre los años 2014 y 2015. Me acuerdo del equipo con el que coincidí en aquel tiempo, y especialmente de Javier Lascuráin, con quien compartía mesa entonces. Él fue una de las primeras personas a las que le enseñé el observatorio cuando estaba aún en estado embrionario.

El núcleo duro de Archiletras (Arsenio, Pilar, Zanya) ha seguido con interés los avances del proyecto desde sus orígenes y me han dado un espacio donde mis ideas disparatadas sobre divulgación lingüística son escuchadas y tienen cabida.

Álex Grijelmo cubrió el proyecto con gran generosidad cuando apenas era un esbozo (¿qué estudiante de doctorado espera ver su investigación apenas echada a andar reseñada en *El País*?) y ha seguido mi investigación desde entonces con mucho cariño.

Muchas gracias a todas las personas que a lo largo de estos años han utilizado (ya fuera de manera más informal o como parte de su trabajo) el observatorio, los *datasets* o los modelos

creados. Sus sugerencias y comentarios han enriquecido esta tesis y saber que había alguien al otro lado usando los datos me ha resultado tremendamente gratificante.

Gracias a los lenguaraces por las conversaciones estimulantes y porque siempre hay alguien que conoce la referencia bibliográfica adecuada o que está dispuesto a prestarse como conejillo de indias para experimentos.

Gracias a mis médicas (Belén, Cris, Dori, Gise, Laura, Leila, Leire, Palomas, Paula, Raquel, Rebe, Sara) por ser mis cuarteles de invierno en todas mis zozobras.

Gracias a Anita Robla Vicario por cuidarme, por escucharme y por sacarme de concierto.

Gracias a Carlos Gil Bellosta, la persona con la que puedo hablar de modelos de lenguaje y de crianza en una misma conversación; las comidas periódicas con él me han mantenido cuerda este tiempo.

Gracias a Paula Ruiz-Castillo por su vitalidad incombustible y por su lealtad.

Gracias a Javi Gómez-Serrano (que ya estaba allí cuando AZRAEL despertó) por su mirada clara y por el apoyo, el cariño y la alegría durante los coletazos finales de esta tesis.

Me siento eternamente en deuda con Elena Bordons, que fue mi profesora de Lengua en bachillerato, por introducirme a la Lingüística y descubrirme así una de las mayores fuentes de felicidad de mi vida. Sus clases de sintaxis resuenan todavía en mi cabeza.

La elaboración de una tesis es una tarea que se extiende a lo largo de varios años, y que por lo tanto no es ajena a los acontecimientos vitales que atraviesa la persona que la escribe. Mi recuerdo de muchos de los desarrollos de esta tesis están inevitablemente ligados a las circunstacias en las que se produjeron. El conjunto de datos de COALAS lo recopilé en las tardes oscuras de un invierno gélido y aún pandémico en Massachusetts. El artículo que dio pie a lo que hoy es el capítulo 4 lo escribí en el sillón del acompañante de la planta post-UCI del Hospital La Paz. La primera versión estable de la librería `pylazaro` la terminé unos días antes de parir, con unas columnas de humo amenzantes alzándose en el horizonte mientras los helicópteros de emergencias nos sobrevolaban intentado controlar uno de los muchos incendios que asolaron España en el verano de 2022, y yo sentía (en parte por la estampa apocalíptica, en parte por la particular cuenta atrás hacia lo desconocido en la que estaba inmersa) que se aproximaba el fin del mundo.

Pero sin duda el acontecimiento que ha marcado profundamente esta tesis ha sido el nacimiento de mi hijo Blas. La frontera que separa las partes II y III de este documento no es puramente formal o metodológica, es también vital. La tesis y la crianza comparten algunas similitudes sorprendentes: ambas son tareas maratonianas que conllevan mucha dedicación, mucha soledad y algunos momentos de felicidad incomparable. Buena parte del trabajo de estos años se ha llevado a cabo en las condiciones poco favorables que caracterizan la maternidad reciente: con escasas horas de sueño en el cuerpo, arañando ratos de trabajo en las siestas o entre tomas, yendo de congreso con un sacaleches en el bolso, capeando virus imprevistos, y, en general, con un agotamiento infinito.

Dicen que hace falta a toda la tribu para criar un niño. Diría que la frase es especialmente cierta cuando la madre está haciendo una tesis. En primatología se llama *alomadres* a los individuos de un grupo que cuidan crías que no son suyas. Quiero dar las gracias a todas las alomadres que han cuidado de Blas y que nos han brindado su ayuda de una manera u otra a lo largo de estos años: abuelos Isa y Amador, tía Ana, tía Marigel, abuelos Javier y María, tía Elena, tío Miguel, Andrea, Dara, Ariadna. Sobre su tiempo y su apoyo también se construye esta tesis.

Por último, hay cuatro personas que han sido centrales en este proceso y a las que quiero dar las gracias:

A Andrés: por ser un compañero leal, generoso y paciente, virtudes todas ellas fundamentales tanto a la hora de compartir la crianza como para acompañarme en esta tesis. Algunas de las ideas que aquí se exponen las discutí con él antes incluso que con mis directores.

A Paloma, mi pilar de carga, mi muro de las lamentaciones, mi dique de contención: por acudir en mi auxilio cuando lo he necesitado y por hacerme sentir incondicionalmente comprendida y apoyada siempre.

A mi madre Isabel: en estos años he sentido que había personas con las que podía compartir mis desvelos por la tesis y personas con las que podía compartir los desgastes propios de la crianza. Mi madre, que conocía bien ambas facetas, es la persona que mejor ha sabido ver y comprender los sacrificios y renuncias que esa ambivalencia me ha conllevado.

A mi hijo Blas: aprendiste a decir la palabra "tesis" inusitadamente pronto y con dos años jugabas a hacer la tesis simulando que tecleabas sobre la mesa. Tengo la certeza de que si no hubiera sido madre esta tesis hubiera sido mejor de lo que es. Pero también sé que sin tu presencia luminosa me hubiera perdido toda esta felicidad.

A todos ellos, muchas gracias.

En el Barrio de la Concepción, Madrid, marzo de 2025.

Table of Contents

Abstract vi

Resumen y conclusiones viii

Agradecimientos xii

Table of Contents xiii

List of Tables xix

List of Figures xxv

I The problem: Lexical borrowing 1

1 Introduction 3
- 1.1 Definition: What is lexical borrowing? . . . 5
  - 1.1.1 Scope of the phenomenon . . . 5
  - 1.1.2 Anglicisms in Spanish . . . 6
  - 1.1.3 Borrowing vs codeswitching . . . 7
- 1.2 Motivation: Why automatizing lexical borrowing retrieval? . . . 9
  - 1.2.1 Applications in Linguistics . . . 9
  - 1.2.2 Applications in Natural Language Processing . . . 11
- 1.3 What this thesis is about . . . 11
- 1.4 What this thesis is not about . . . 12
- 1.5 Structure of this thesis . . . 13

2 Literature review 15
- 2.1 Previous work on borrowing detection . . . 16
  - 2.1.1 Computer-assisted methods to borrowing detection . . . 16
  - 2.1.2 A typology of borrowing detection systems . . . 20
- 2.2 Methodological limitations from previous work . . . 24
  - 2.2.1 Limitations of pattern matching . . . 24
  - 2.2.2 Limitations of character n-gram probability . . . 25
  - 2.2.3 Limitations of lexicon lookup . . . 25
  - 2.2.4 Limitations of word classification approaches . . . 26

2.2.5 On the importance of multiword borrowing detection 28
2.3 Differences with previous work 31
2.3.1 Differences with work within historical linguistics 31
2.3.2 Differences with work within code-switching identification 33
2.3.3 Differences with previous work on anglicism detection in Spanish . . . 33
2.4 Summary 36

II Data & models 39

3 COALAS: Corpus of Anglicisms in the Spanish Press 41
3.1 Limitations of previous work and motivation 42
3.2 Corpus desiderata 43
3.3 Data curation: sources and filtering 44
3.4 Annotation process 46
3.5 Corpus description 46
3.6 Discussion 47
3.7 Summary 50

4 Model experimentation 53
4.1 Supervised models for lexical borrowing detection 53
4.1.1 Conditional random field model 54
4.1.2 Transformer-based models: BETO and mBERT 56
4.1.3 BiLSTM-CRF with unadapted embeddings 57
4.1.4 BiLSTM-CRF with codeswitch embeddings 60
4.1.5 Results 62
4.1.6 Error analysis 63
4.2 A large language model's attempt at lexical borrowing detection 70
4.2.1 Model and prompting methodology 71
4.2.2 Results 73
4.2.3 Error analysis 74
4.3 Discussion 76
4.4 Summary 78

5 ADoBo: A shared task on Automatic Detection of Borrowings 81
5.1 Task description 81
5.1.1 Data 82
5.1.2 Evaluation 82
5.1.3 Competition timeline 83
5.2 Participants' submissions 83
5.2.1 A CRF model with data augmentation 84
5.2.2 Transformer-based models using STILTs 85
5.3 Results 87
5.3.1 Systems' performance 87
5.3.2 Experiments with outputs 88

5.4 Discussion . . . 90
5.5 Summary . . . 91

6 Annotating lexical borrowing vs codeswitching 93
6.1 The elephant in the room: the codeswitching vs borrowing distinction . . . 93
6.2 Challenges in annotating codeswitching and borrowing . . . 94
6.3 Dataset development . . . 97
6.3.1 Annotation guidelines . . . 97
6.3.2 Data selection . . . 100
6.3.3 Dataset description . . . 103
6.4 Model experimentation . . . 104
6.5 Results . . . 106
6.6 Discussion . . . 107
6.7 Summary . . . 108

III Evaluation 111

7 Evaluation approaches 113
7.1 Evaluation scenarios and desiderata . . . 114
7.2 Extraction task evaluation: approaches and limitations . . . 117
7.2.1 Problems with span-based evaluation . . . 118
7.2.2 Problems with token-based evaluation . . . 122
7.2.3 A silver bullet: the token and separator model . . . 125
7.3 Experiments: Revisiting COALAS outputs . . . 126
7.4 Discussion . . . 129
7.5 Summary . . . 130

8 Dimensions in evaluation 133
8.1 Possible explanations for the paradoxical results from chapter 7 . . . 134
8.2 Rationale . . . 138
8.3 Span characterization: precedents from NER . . . 140
8.4 Dimensions definition . . . 142
8.5 Experiments across dimensions . . . 144
8.5.1 Evaluation of the dataset . . . 144
8.5.2 Evaluation of the models . . . 146
8.6 Discussion . . . 148
8.6.1 Benefits of the dimensions . . . 149
8.6.2 Differences with prior work . . . 150
8.6.3 Applicability to other span identification tasks . . . 152
8.6.4 Limitations of the notion of *unseenness* . . . 153
8.7 Summary . . . 158

9 BLAS: A Benchmark for Loanwords and Anglicisms in Spanish 159
9.1 Motivation . . . 160
9.1.1 What if...? . . . 160
9.1.2 Limitations of the dimension-based evaluation from chapter 8 . . . 161
9.1.3 Towards diagnostic, actionable and predictable benchmarking in NLP 162
9.2 Benchmark creation . . . 166
9.2.1 Benchmark desiderata . . . 166
9.2.2 Span attribute selection . . . 167
9.2.3 Data curation . . . 168
9.2.4 Punctuation transformations . . . 170
9.2.5 Benchmark description . . . 171
9.2.6 Evaluation metrics . . . 174
9.3 Experiments & results . . . 178
9.3.1 Experiments across models . . . 178
9.3.2 Predictive ability of BLAS . . . 183
9.3.3 Statistical significance of the dimensions . . . 184
9.4 Discussion . . . 189
9.4.1 Limitations . . . 191
9.4.2 Relation to previous work . . . 192
9.4.3 The cycle of benchmark creation . . . 193
9.5 Summary . . . 195

IV Results 197

10 Lázaro: an observatory of anglicism usage in the Spanish press 199
10.1 Motivation . . . 199
10.2 Implementation . . . 201
10.3 Public outreach . . . 204
10.4 Discussion . . . 206
10.5 Summary . . . 209

11 Conclusions 211
11.1 Contributions . . . 211
11.2 Resources . . . 214
11.3 Limitations . . . 216
11.4 Future Work . . . 218
11.5 Publications . . . 219

V Appendices 221

A Annotation guidelines 223

B Data statement 241

C A practical comparison between evaluation approaches **245**

D Guidelines for the creation of examples in BLAS **253**

E Results over BLAS benchmark **259**

F Predictive ability of BLAS **297**

References **305**

List of Tables

2.1 Borrowing detection systems classified by purpose, method, type of input, approach and languages. 23

3.1 Number of tokens and spans in COALAS. 45
3.2 Most frequent spans in the COALAS training set. 47
3.3 Most frequent spans in the COALAS test set 48
3.4 Comparison between the dataset from Álvarez Mellado (2020a) and COALAS. 49
3.5 Data sources per split included in COALAS. 52

4.1 Results of the CRF model on the COALAS development set and test set (results from a single run). 56
4.2 Results of BETO and mBERT on the COALAS development set and test set (average scores and standard deviation reported over 10 runs with different random seeds for initialization). 56
4.3 Results of a BiLSTM-CRF model on the COALAS development set with different word and subword embedding combinations (results on `ALL` labels from a single run). 59
4.4 Results of a BiLSTM-CRF model on COALAS development set and test set using different combinations of unadapted embeddings: BETO and BERT embeddings, BPE embeddings, and optionally character embeddings (average scores and standard deviation reported over 10 runs with different random seeds for initialization). 60
4.5 Results of a BiLSTM-CRF model on the COALAS development set and test set using different combinations of codeswitch embeddings and subword embeddings (average scores and standard deviation reported over 10 runs with different random seeds for initialization). 61
4.6 Summary of results obtained by all supervised models (for `ALL` labels) on the COALAS development set and test set (for the CRF, results from a single run are reported; for all other models, average scores over 10 runs are reported). 63
4.7 Most frequent errors made by each model on the COALAS test set. FN = false negative; FP = false positive. 66
4.8 Most frequent errors made by Llama3 on COALAS test set. FN = false negative; FP = false positive. 74
4.9 Final ranking of models based on the performance obtained on the COALAS test set. 76

5.1 Participants' results at ADoBo on the standard version of COALAS test set. 84
5.2 Participants' results at ADoBo on the lowercase version of COALAS test set. 85
5.3 Participants' results at ADoBo on the unquoted version of COALAS test set. 86
5.4 Participants' results at ADoBo on the lowercase unquoted version of COALAS test set. . . . 87

6.1 Token counts by label on the LM2013 dataset. . . . 103
6.2 Most frequent tokens in the LM2013 dataset for each label: Spanish, English, and borrowing . . . 104
6.3 Results obtained by baseline models on the reannotated version of the LM2013 dataset (results from a single run). . . . 106

7.1 Comparison between different encoding schemes for token-based annotation of borrowings: BIO, BILOU, IO, IOB and TS. . . . 127
7.2 Ranking of models evaluated on the token & separator evaluation model on the COALAS test set. . . . 128

8.1 Number of error types made per model according to MUC error typology. . . 135
8.2 Number of types of overlapping errors per model: *Missing* refers to overlapping errors where at least a token was missed. *Spurious* refers to overlapping errors where at least a spurious token was added. *Both* refers to overlapping errors where there were both a missing token and a spurious token. . . . 136
8.3 Recall obtained by each model on different span dimensions. . . . 157
8.4 Precision obtained by each model on different span dimensions. . . . 157

9.1 Number of sentences in BLAS per span type. . . . 170
9.2 Casing and quotation mark perturbations considered in BLAS. . . . 172
9.3 Extended version of the MUC error typology considered for BLAS. . . . 177
9.4 Span recall per model on BLAS and COALAS. . . . 179
9.5 Overall span recall on BLAS per model on different casing and quotation mark configurations. . . . 179
9.6 Span recall obtained by each model over BLAS for different types of spans and contexts (on standard casing, with and without quotation marks). . . . 181
9.7 Correlation between the predicted and true recall scores and ranking positions over five datasets. . . . 183
9.8 Comparison in recall results per model across different dimensions in BLAS (asterisk indicates statistical significance with $p < 0.05$). . . . 188

10.1 List of media sources tracked by Lázaro Observatory . . . 202
10.2 Average numbers of URLs, tokens and anglicisms processed by Observatorio Lázaro per day. . . . 203

C.1 Ranking of outputs on strict span level evaluation (blocks of color represent predicted spans). . . . 247
C.2 Ranking of outputs on lenient span level evaluation (blocks of color represent predicted spans). . . . 248

C.3 Ranking of outputs on token level evaluation with BIO encoding. 249
C.4 Ranking of outputs on token level evaluation with IO encoding. 249
C.5 Ranking of outputs on the token & separator evaluation. 250
C.6 Spearman correlation coefficients between pairs of ranking orders produced by different evaluation approaches. 251

E.1 Results obtained by CRF over BLAS with standard casing, with quotation marks. 260
E.2 Results obtained by CRF over BLAS with standard casing, without quotation marks. 260
E.3 Results obtained by CRF over BLAS with lowercase text, with quotation marks. 261
E.4 Results obtained by CRF over BLAS with titlecase text, with quotation marks. 261
E.5 Results obtained by CRF over BLAS with uppercase text, with quotation marks. 262
E.6 Results obtained by CRF over BLAS with titlecase spans, with quotation marks. 262
E.7 Results obtained by CRF over BLAS with uppercase spans, with quotation marks. 263
E.8 Results obtained by CRF over BLAS with lowercase text, without quotation marks. 263
E.9 Results obtained by CRF over BLAS with titlecase text, without quotation marks. 264
E.10 Results obtained by CRF over BLAS with uppercase text, without quotation marks. 264
E.11 Results obtained by CRF over BLAS with titlecase spans, without quotation marks. 265
E.12 Results obtained by CRF over BLAS with uppercase spans, without quotation marks. 265
E.13 Results obtained by BETO over BLAS with standard casing, with quotation marks. 266
E.14 Results obtained by BETO over BLAS with standard casing, without quotation marks. 266
E.15 Results obtained by BETO over BLAS with lowercase text, with quotation marks. 267
E.16 Results obtained by BETO over BLAS with titlecase text, with quotation marks. 267
E.17 Results obtained by BETO over BLAS with uppercase text, with quotation marks. 268
E.18 Results obtained by BETO over BLAS with titlecase spans, with quotation marks. 268
E.19 Results obtained by BETO over BLAS with uppercase spans, with quotation marks. 269
E.20 Results obtained by BETO over BLAS with lowercase text, without quotation marks. 269
E.21 Results obtained by BETO over BLAS with titlecase text, without quotation marks. 270

E.22 Results obtained by BETO over BLAS with uppercase text, without quotation marks. . . . 270
E.23 Results obtained by BETO over BLAS with titlecase spans, without quotation marks. . . . 271
E.24 Results obtained by BETO over BLAS with uppercase spans, without quotation marks. . . . 271
E.25 Results obtained by mBERT over BLAS with standard casing, with quotation marks. . . . 272
E.26 Results obtained by mBERT over BLAS with standard casing, without quotation marks. . . . 272
E.27 Results obtained by mBERT over BLAS with lowercase text, with quotation marks. . . . 273
E.28 Results obtained by mBERT over BLAS with titlecase text, with quotation marks. . . . 273
E.29 Results obtained by mBERT over BLAS with uppercase text, with quotation marks. . . . 274
E.30 Results obtained by mBERT over BLAS with titlecase spans, with quotation marks. . . . 274
E.31 Results obtained by mBERT over BLAS with uppercase spans, with quotation marks. . . . 275
E.32 Results obtained by mBERT over BLAS with lowercase text, without quotation marks. . . . 275
E.33 Results obtained by mBERT over BLAS with titlecase text, without quotation marks. . . . 276
E.34 Results obtained by mBERT over BLAS with uppercase text, without quotation marks. . . . 276
E.35 Results obtained by mBERT over BLAS with titlecase spans, without quotation marks. . . . 277
E.36 Results obtained by mBERT over BLAS with uppercase spans, without quotation marks. . . . 277
E.37 Results obtained by BiLSTM with unadapted embeddings over BLAS with standard casing, with quotation marks. . . . 278
E.38 Results obtained by BiLSTM with unadapted embeddings over BLAS with standard casing, without quotation marks. . . . 278
E.39 Results obtained by BiLSTM with unadapted embeddings over BLAS with lowercase text, with quotation marks. . . . 279
E.40 Results obtained by BiLSTM with unadapted embeddings over BLAS with titlecase text, with quotation marks. . . . 279
E.41 Results obtained by BiLSTM with unadapted embeddings over BLAS with uppercase text, with quotation marks. . . . 280
E.42 Results obtained by BiLSTM with unadapted embeddings over BLAS with titlecase spans, with quotation marks. . . . 280
E.43 Results obtained by BiLSTM with unadapted embeddings over BLAS with uppercase spans, with quotation marks. . . . 281

E.44 Results obtained by BiLSTM with unadapted embeddings over BLAS with lowercase text, without quotation marks. 281
E.45 Results obtained by BiLSTM with unadapted embeddings over BLAS with titlecase text, without quotation marks. 282
E.46 Results obtained by BiLSTM with unadapted embeddings over BLAS with uppercase text, without quotation marks. 282
E.47 Results obtained by BiLSTM with unadapted embeddings over BLAS with titlecase spans, without quotation marks. 283
E.48 Results obtained by BiLSTM with unadapted embeddings over BLAS with uppercase spans, without quotation marks. 283
E.49 Results obtained by BiLSTM with codeswitch embeddings over BLAS with standard casing, with quotation marks. 284
E.50 Results obtained by BiLSTM with codeswitch embeddings over BLAS with standard casing, without quotation marks. 284
E.51 Results obtained by BiLSTM with codeswitch embeddings over BLAS with lowercase text, with quotation marks. 285
E.52 Results obtained by BiLSTM with codeswitch embeddings over BLAS with titlecase text, with quotation marks. 285
E.53 Results obtained by BiLSTM with codeswitch embeddings over BLAS with uppercase text, with quotation marks. 286
E.54 Results obtained by BiLSTM with codeswitch embeddings over BLAS with titlecase spans, with quotation marks. 286
E.55 Results obtained by BiLSTM with codeswitch embeddings over BLAS with uppercase spans, with quotation marks. 287
E.56 Results obtained by BiLSTM with codeswitch embeddings over BLAS with lowercase text, without quotation marks. 287
E.57 Results obtained by BiLSTM with codeswitch embeddings over BLAS with titlecase text, without quotation marks. 288
E.58 Results obtained by BiLSTM with codeswitch embeddings over BLAS with uppercase text, without quotation marks. 288
E.59 Results obtained by BiLSTM with codeswitch embeddings over BLAS with titlecase spans, without quotation marks. 289
E.60 Results obtained by BiLSTM with codeswitch embeddings over BLAS with uppercase spans, without quotation marks. 289
E.61 Results obtained by Llama3 over BLAS with standard casing, with quotation marks. 290
E.62 Results obtained by Llama3 over BLAS with standard casing, without quotation marks. 290
E.63 Results obtained by Llama3 over BLAS with lowercase text, with quotation marks. 291
E.64 Results obtained by Llama3 over BLAS with titlecase text, with quotation marks. 291
E.65 Results obtained by Llama3 over BLAS with uppercase text, with quotation marks. 292

E.66 Results obtained by Llama3 over BLAS with titlecase spans, with quotation marks. . . . 292
E.67 Results obtained by Llama3 over BLAS with uppercase spans, with quotation marks. . . . 293
E.68 Results obtained by Llama3 over BLAS with lowercase text, without quotation marks. . . . 293
E.69 Results obtained by Llama3 over BLAS with titlecase text, without quotation marks. . . . 294
E.70 Results obtained by Llama3 over BLAS with uppercase text, without quotation marks. . . . 294
E.71 Results obtained by Llama3 over BLAS with titlecase spans, without quotation marks. . . . 295
E.72 Results obtained by Llama3 over BLAS with uppercase spans, without quotation marks. . . . 295

F.1 Number of spans per type in COALAS test set. . . . 298
F.2 Number of spans per type in COALAS development set. . . . 299
F.3 Number of spans per type in the 2020a development set. . . . 300
F.4 Number of spans per type in the 2020a test set A. . . . 301
F.5 Number of spans per type in the 2020a test set B. . . . 302
F.6 Pearson and Spearman correlation coefficients over true recall scores, F1 scores and ranking obtained by six models over different datasets and the predicted recall scores and rankings obtained by extrapolating BLAS scores. . . . 303
F.7 Pearson and Spearman correlation coefficients obtained from comparing recall score distributions and ranking between the true results obtained on a dataset, the predicted results obtained by extrapolating BLAS scores and BLAS results. 303

List of Figures

3.1 Data selection flowchart for COALAS. 51

4.1 Prompt provided to Llama3 inspired by Ashok and Lipton's template: Definition of the task with examples, question and task, response by Llama3. . . 72

7.1 Comparison between results obtained on the strict span evaluation and the token and separator evaluation. 129

8.1 Ratio of spans per dimension in the COALAS test set. 155
8.2 Ratio of spans per dimension in the COALAS training set. 156

9.1 Ratio of spans per dimension in BLAS. 173
9.2 Creation process behind each of the splits in BLAS. 174
9.3 The GEM cycle of benchmark creation. Image from Gehrmann et al. (2021) 194

10.1 Observatorio Lázaro: Pipeline and output. 201
10.2 Observatorio Lázaro website. 204
10.3 Anglicism sheet for *reality show*. 205
10.4 Search section from Observatorio Lázaro's website 206
10.5 A tweet by `@lazarobot` with the word *binge-watching*. 207

A.1 Decision steps to follow during the annotation process to decide whether to annotate a word as an anglicism. 227

Part I

The problem: Lexical borrowing

Chapter 1

Introduction

> *It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you'll be swept off to.*
>
> JRR Tolkien, writer

No language is an island. Languages coexist; speakers of different languages interact. Multilingualism and contact between languages have been the norm throughout history, not the exception. As Thomason (2003) states:

> "All languages are mixed in a weak sense: there is no natural human language in which foreign material is wholly lacking."

Language mixing is therefore a linguistic behavior that humans display universally and that lies at the heart of communication. Understanding how speakers mix languages provides us with relevant information about how languages interact and about the nature of language as a human faculty in general.

Contact between languages is in fact one of the major inducers of language change. Speakers may take patterns that belong to one language and incorporate them into another language, a process that is known as *linguistic borrowing*. When the process in question concerns the lexicon (that is, when the foreign elements incorporated are *words*), we talk about lexical borrowing.

The study of lexical borrowing provides valuable insight both into the social and cultural contexts that shape languages, as well as into the internal workings of a language as a system. There are many reasons why speakers of a language may decide to borrow a word from another language:

Sometimes a new artifact or technique may be imported from one community to another, and the name of the new invention will be imported along with the object, thus revealing a cultural exchange between communities.

Other times, a word may be imported because the foreign word is perceived as more prestigious than the native one, thus reflecting a social hierarchy that impacts language use.

In that regard, borrowings are a valuable piece of evidence that can attest contact between languages and illustrate the social dynamics between cultures and communities throughout time (Zenner et al., 2012).

When a lexical borrowing becomes established, it will go through a process of morphological and phonological adaptations to make it comply with the rules of the recipient language, a process that is called *assimilation*. The borrowability of certain words (that is, the fact that some words are more likely to be borrowed than others), the linguistic status that the borrowed word receives in the recipient language and the process of assimilation that the borrowed word may undergo provide an interesting insight on the grammatical patterns and speakers' expectations that operate on the recipient language and on the general constraints on language change (Van Hout and Muysken, 1994; Haspelmath, 2009). The linguistic expectations that speakers of the recipient language have about their own language become outstandingly distinct at the time of borrowing. After all, a language is nothing more (and nothing less) than a set of symbols (signs, morphemes, syntactic patterns) that have the ability to convey meaning within a system. The process of borrowing makes explicit what symbols that convey meaning within the donor system do not have the ability to mean in the recipient language, and what are the formal restrictions that a foreign symbol is required to have in order for it to work within the recipient system.

1.1 Definition: What is lexical borrowing?

1.1.1 Scope of the phenomenon

Linguistic borrowing is the process of reproducing in one language elements and patterns that come from another language (Haugen, 1950). Linguistic borrowing therefore involves the exchange between two languages and has been widely studied within the field of contact linguistics (Weinreich, 1963). Various typologies have been proposed to classify linguistic borrowings according to different criteria, such as typological characteristics, levels of language involved, integration of the borrowed element within the recipient language, etc. (Thomason and Kaufman, 1992; Matras and Sakel, 2007; Haspelmath and Tadmor, 2009).

Lexical borrowing in particular involves the incorporation of single lexical units from one language into another language and is usually accompanied by morphological and phonological modification to conform with the patterns of the recipient language (Poplack et al., 1988; Onysko, 2007). For example, all borrowed verbs in Spanish must undergo the process of adding the suffix *-ar* or *-ear* in order to comply with the morphological paradigm of Spanish verbs (*tuitear*, from *to tweet*). Similarly, unadapted borrowings that do not comply with phonological restrictions of Spanish will also be transformed. That phonological transformation might even end up showing on its spelling: the unadapted borrowing *spoiler* coexists with the adapted form *espóiler*, that better reflects the Spanish pronunciation and graphotactics.

All these phonological, orthographic and morphological transformations can lead to a borrowing eventually becoming fully assimilated into the recipient language lexicon, which may lead to native speakers losing the perception of the term being "foreign" (Lipski, 2005). Some authors establish the need of a borrowing being recognized as foreign by native speakers in order to be considered as such (Zenner et al., 2012). For instance, a word like *bar* was originally borrowed from English into Spanish, but it has been so assimilated that is now perceived as a native word by monolingual speakers of Spanish, and its English nature is

only seen as etymological. On the other hand, a word like *whisky* (that has also been used in Spanish for some time) has never been fully assimilated and is perceived as a foreign word.

In this thesis, we will be dealing with unassimilated lexical borrowings in Spanish, that is, words from other languages that have been incorporated into Spanish but that have not undergone the process of assimilation. Our research will focus on lexical borrowings from English (also known as *anglicisms*), as they represent the vast majority of current borrowings in Spanish.

1.1.2 Anglicisms in Spanish

The study of the influence of English on the European languages has been a hot topic in linguistics for decades (Furiassi et al., 2012), particularly concerning anglicism usage in Spanish (Núñez Nogueroles, 2017b).

Previous work estimated that a reader of French newspapers encounters a new lexical borrowing every 1,000 words (Chesley and Baayen, 2010), English borrowings outnumbering all other borrowings combined (Chesley, 2010). In Chilean newspapers, lexical borrowings account for approximately 30% of neologisms, 80% of those corresponding to anglicisms (Gerding et al., 2014). In European Spanish, it was estimated that anglicisms could account for 2% of the vocabulary used in Spanish newspaper *El País* in 1991 (Rodríguez González, 2002), a number that is likely to be higher today.

Concerning anglicism classification and analysis of anglicism usage in Spanish, seminal work on was performed by Pratt (1980), where any word that entered Spanish via the English language was considered an anglicism, regardless of its previous history or the original language the word was coined in. Unlike prior approaches to this issue, Pratt's work is purely descriptivist and is in fact openly critical towards the purist view that prior authors had on borrowing, an attitude that he considers to be unscientific and even xenophobic (Pratt, 1980, pg. 17). Many of the forecasts outlined in 1980 by Pratt concerning the influence of English on the Spanish language are in fact common practice in today's Spanish, like the use of the

suffix *-s* in English-derived plural forms (*fans, tuits*) or the proliferation of words derived with *-ismo* or *-ización* (Pratt, 1980, pg. 239).

Later works on anglicisms have proposed different classifications (Lorenzo, 1996; Medina López, 1998; Rodríguez González, 1999; Núñez Nogueroles, 2018a) and have focused on various aspects of anglicism usage, such as orthographic integration (Rodríguez González, 2018), diachronic shifts (Gimeno Menéndez and Gimeno Menéndez, 2003), typological characteristics (Gómez Capuz, 1997), syntactic changes (Rodríguez Medina, 2002), sociocultural aspects (Gómez Capuz, 2004) or lexicographical coverage (Balteiro, 2011a).

Research on anglicism usage in Spanish has mainly been informed by corpus linguistics. As we will see in section 1.2.1, this approach suffers from an undeniable limitation: the fact that the current state of corpus linguistics resources in Spanish (where corpus lookup is restricted to manual inspection of examples) is insufficient to account for an on-going process like borrowing usage.

1.1.3 Borrowing vs codeswitching

The concept of *linguistic borrowing* covers a wide range of linguistic phenomena, but is generally understood as the process of introducing words, elements or patterns of one language (the donor language) into another language (the recipient language) (Haugen, 1950; Weinreich, 1963). The process of lexical borrowing requires a language from which a certain word is taken (the donor language) and a second language into which that word in inserted (the recipient language). In that sense, lexical borrowing is somewhat similar to linguistic codeswitching, the mixing between two or more languages in discourse that is typical of bilingual communities (Poplack, 1980). Unlike borrowings, codeswitches are by definition not integrated into the recipient language (Poplack, 2012), and they do not produce a change in the lexicon of the recipient language. As Haspelmath puts in a nutshell:

> "Code-switching is not a kind of contact-induced language change, but rather a kind of contact-induced speech behavior."

However, a clear-cut distinction between borrowing and codeswitching remains elusive, particularly when it comes to non-integrated foreign language items that appear in contexts that are otherwise monolingual (sometimes referred as "lone other-language items" or LO-LIs). In fact, borrowing and codeswitching have sometimes been described as a continuum with a fuzzy frontier between the two. According to that view, borrowings would originate as codeswitches from a bilingual community that eventually become frequent and widespread enough to end up being gradually incorporated into the recipient language and assimilated by monolingual speakers as part of their lexicon. Supporters of this view therefore see intrasentential codeswitches (especially LO-LIs) and lexical borrowing as indistinguishable, or at least as two manifestations of the same phenomenon (Clyne et al., 2003).

Other authors, however, disagree with this approach and support that borrowing and codeswitching are two clearly distinct phenomena. Poplack and Dion (2012) argue that borrowings do not originate as codeswitches that eventually become integrated. On the contrary, Poplack and Dion hold that speakers decide whether the foreign item they are going to use is going to be borrowed or codeswitched at the time of accessing the item, regardless of the borrowing being frequent and well-established or nonce. Their results suggest that the linguistic mechanisms involved in borrowing and codeswitching are therefore independent from one another and the process of integration is abrupt, not gradual (Poplack and Sankoff, 1984; Poplack et al., 1988, 2020).

The frontier between both phenomena is therefore controversial (Lipski, 2005) and disagreement on what a borrowing is (and is not) exists (Gómez Capuz, 1997). Consequently, multiple criteria have been proposed to distinguish between lexical borrowing and codeswitching, among them frequency (Stammers and Deuchar, 2012), level of integration (Poplack and Sankoff, 1984), length (Calude et al., 2020; Deuchar, 2020), bilingual competence of the speaker (Pfaff, 1979) or listedness (Treffers-Daller, 2023). As Treffers-Daller states:

"The dilemma of distinguishing between lexical borrowing and code-switching can

> be seen as a specific instantiation of the problems involved in delimiting what belongs in the lexicon (fixed, arbitrary patterns) and what is computed online (productive rules), and should therefore be considered as part of the grammar."

As today, the borrowing-codeswitching distinction debate is far from settled (Poplack, 2012; Deuchar and Stammers, 2012) and, as Poplack and Dion state:

> "Distinguishing codeswitching and borrowing is the thorniest issue in the field of contact linguistics today."

As a result, some authors opt for an agnostic terminology and talk about donor-language items (Deuchar and Stammers, 2016) or code-mixing in general (Muysken, 2000).

For the purpose of this research, we will follow the approach taken by Poplack and Dion and consider codeswitching and borrowing as two separate phenomena. We will consider that codeswitches are fluent multiword interferences that normally comply with grammatical restrictions in both languages and that are produced by bilingual speakers in bilingual discourses (usually in spoken language), while we will define lexical borrowings as words of foreign origin that are used by monolingual individuals without knowledge of the donor language.

In this thesis we will be concerned mainly with the retrieval of lexical borrowings from monolingual Spanish text, although we will tackle the codeswitching-borrowing distinction in chapter 6.

1.2 Motivation: Why automatizing lexical borrowing retrieval?

The automatic identification of lexical borrowings from text has been shown to be useful in various areas of research. We describe here its applications in the areas of Linguistics and Natural Language Processing.

1.2.1 Applications in Linguistics

Borrowings are a source of new words. Consequently, the empirical study of borrowings is key in the areas of neology, language change and lexicography (Gerding Salas et al., 2018;

Freixa Aymerich, 2022).

Consequently, data driven research in the field of corpus linguistics has extensively studied the phenomenon of anglicism usage in Spanish (Rodríguez Medina, 2002; Vélez Barreiro, 2003; Gimeno Menéndez and Gimeno Menéndez, 2003; Diéguez, 2004; Balteiro, 2011b; Patzelt, 2011; De la Cruz Cabanillas and Tejedor Martínez, 2012; Oncíns Martínez, 2012; Núñez Nogueroles, 2016, 2017a, 2018b).

These works, however, have mainly relied on manual inspection of previously compiled general corpora such as reference corpus CREA[1] (*Corpus del Español Actual*) and CORPES[2] (*Corpus del Español del Siglo XXI*), or on the creation of new tailor-made corpora designed to analyze specific genres, varieties or phenomena.

The main limitation of these corpus-assisted projects is that the retrieval of anglicisms and their concordances relies on the manual lookup of the corpus. That implies that either the whole corpus is processed and annotated by a human (in order to identify all anglicisms it contains), either the research is limited to a list of predefined terms of interest to be queried. This approach seems insufficient to account for an on-going phenomenon like anglicism incorporation, because borrowings are notorious for being volatile (Poplack, 2017). In addition, as lexical borrowings are a sparse phenomenon, large corpora are required in order to collect exhaustive repertoires of anglicisms in use, something that is simply not feasible without efficient automatization.

Some corpus interfaces do allow searching for borrowings in general, such as CORPES. Their results, however, leave a lot to be desired. At the time of writing this dissertation, the result of concordances obtained by CORPES when querying for borrowings include true borrowings, along with foreign proper nouns (such as *Spice Girls*, *Modern Family* or *Madama Butterfly*), unusual Spanish words (such as *chundachunda* or *fu*, in the idiom *ni fu ni fa*), ill-tokenized words or complete Latin or English phrases (which would correspond more to the phenomenon of codeswitching than borrowing).

[1] `http://corpus.rae.es/creanet.html`
[2] `https://www.rae.es/corpes/`

Ultimately, both the dissatisfaction with the CORPES anglicism results and the obvious limitations of having to manually process the corpus point in the same direction: the need for robust computer-assisted methods in corpus linguistics in general and in borrowing studies in particular that can facilitate data-driven research on language contact, language change and the lexicon.

This thesis aims to produce such methods.

1.2.2 Applications in Natural Language Processing

From the point of view of Natural Language Processing, borrowings and neologisms contribute to out-of-vocabulary words, that is, words that are unknown to models because they were not seen during training, which can hinder performance (Zheng et al., 2024).

The automatic identification of lexical borrowings from text has therefore proven to be relevant for NLP downstream tasks, such as parsing (Alex, 2008) or text-to-speech synthesis (Leidig et al., 2014).

In addition, the identification of borrowings has been used as bootstrapping technique in machine translation to enlarge the available vocabulary when working on very low-resourced languages that heavily borrow from high-resourced languages (Tsvetkov et al., 2015; Tsvetkov and Dyer, 2016).

1.3 What this thesis is about

This thesis is concerned with the task of retrieving unassimilated English lexical borrowings (or *anglicisms*) from Spanish text, i.e. words from English origin that are introduced into Spanish without any morphological or orthographic adaptation (Pratt, 1980; Gómez Capuz, 1997; Núñez Nogueroles, 2018a).

The task of extracting unassimilated anglicisms from Spanish text is a more challenging undertaking than it might appear to be at first. To begin with, lexical borrowings can be single or multiword expressions (e.g., *prime time*, *tie break* or *machine learning*).

Second, linguistic assimilation is a diachronic process and, as a result, what constitutes an unassimilated borrowing is not clear-cut. For example, words like *bar* or *club* were unassimilated lexical borrowings in Spanish at some point in the past, but have become so widespread and frequent that the process of phonological and morphological adaptation is now complete and they cannot be considered unassimilated borrowings anymore.

In addition, not all English words that appear in Spanish texts will be anglicisms: proper names, literal quotations and other language-mixing phenomena can naturally occur in Spanish text, without any of them being true examples of lexical borrowing. To make things worse, there are non-related words that exist both as English and Spanish words (*quince*, *come*, *primer*, *pie*). These words may or may not be a borrowing depending on the context they appear in.

All these subtleties make the automatic retrieval of lexical borrowings non-trivial. Consequently, in prior work on anglicism extraction from Spanish text, plain dictionary lookup produced very limited results with F1 scores of 47% (Serigos, 2017a) and 26% (Álvarez Mellado, 2020a). This justifies the need for a more NLP-heavy approach to the task that can succeed where manual inspection or plain dictionary lookup cannot.

Unlike prior approaches, we defend that lexical borrowing detection should not be approached on a word-by-word basis in which units are assumed to be single words and in which context is considered irrelevant. On the contrary, we defend that lexical borrowing detection must be framed as a sequence labeling task, in which relevant spans of texts (which may be single-word or multiword) are extracted from sentences, in a similar fashion to how NER or MWE are approached.

1.4 What this thesis is not about

This thesis is concerned with the methodological issues that arise when creating models for automatizing the task of retrieving lexical borrowings (particularly anglicisms) from Spanish text.

This thesis does not propose any psycholinguistic model or usage-based theory on the process of borrowing or on anglicism usage, but aims to produce methods that may facilitate getting relevant data that can eventually inform such models and hypotheses.

Our research is concerned with the *hows* (how can we frame this problem, how can we evaluate it), rather than the *whys* or the *whats*. The reader of this thesis will not find in this document data or figures portraying how many borrowings are used in Spanish or whether anglicism usage in the Spanish press is on the rise. That analysis should be left to corpus linguistics researchers, who, hopefully, will find some of the results and application proposals of this thesis useful for their own research.

Finally, our research neither proposes novel NLP modeling techniques, but it simply repurposes already-existing NLP approaches and analyzes the mismatches and issues that may arise (in terms of annotation, modeling and evaluation) when applying them to the task of retrieving lexical borrowings from text.

That pre-Linguistics post-NLP space is the realm this thesis inhabits.

1.5 Structure of this thesis

This thesis is organized as follows:

We first review previous research on automatic retrieval of lexical borrowings and identify the limitations of existing methods that our research aims to address (chapter 2).

In part II we cover the the fundamentals of compiling data and training models for the task of automatically retrieving lexical borrowings from text: we introduce a new dataset for the task of retrieving anglicisms and other lexical borrowings from Spanish journalistic text (chapter 3); we evaluate several supervised sequence labeling models and a large language model for the task of retrieving lexical borrowings from Spanish text (chapter 4); we present the results of ADoBo, a shared task on automatic detection of borrowings held at IberLEF 2021 (chapter 5); and finally, we propose annotation guidelines and models for the task of distinguishing between codeswitches and lexical borrowings on social media text (chapter 6).

In part III we investigate the methodological issues that arise when evaluating span identification tasks, such as lexical borrowing detection: we first review existing approaches to sequence labeling evaluation and identify their blind spots (chapter 7); we propose a set of formal dimensions that can assist when documenting and evaluating span identification tasks, and use them to identify systematic weaknesses in our experimental results from part II (chapter 8); finally, informed by our findings from chapters 7 and 8, we introduce BLAS, a benchmark for the task of automatically retrieving anglicisms from Spanish text that facilitates a more rigorous and informative evaluation of borrowing detection systems (chapter 9).

To conclude, in part IV we document Observatorio Lázaro, a practical application that showcases the results of our research (chapter 10) and summarize the main findings of this thesis, along with the limitations that future work should address (chapter 11).

Chapter 2

Literature review

¡Qué gran tónico sería para el novel observador el que su maestro, en vez de asombrarlo y desalentarlo con la sublimidad de las grandes empresas acabadas, le expusiera la génesis de cada invención científica, la serie de errores y titubeos que la precedieron, constitutivos, desde el punto de vista humano, de la verdadera explicación de cada descubrimiento! Tan hábil táctica pedagógica nos traería la convicción de que el descubridor, con ser un ingenio esclarecido y una poderosa voluntad, fue, al fin y al cabo, un hombre como todos.

Santiago Ramón y Cajal, histologist

In this chapter we present the existing work on the field of lexical borrowing detection, particularly on the identification of anglicisms. We classify prior work according to different criteria (approach, method, aim, languages involved). We then compare our proposal with the theoretical and methodological assumptions that prior and related works have made, point out their limitations and describe how our work deviates from those previous works.

2.1 Previous work on borrowing detection

2.1.1 Computer-assisted methods to borrowing detection

The automatic detection of borrowings has not been a major topic in NLP. Most of the work on automatic and semiautomatic approaches to borrowing detection has come from the realm of corpus linguistics, where the manual retrieval of borrowings of interest has traditionally been the norm. More recently some works have proposed different techniques to automatize the retrieval of borrowing candidates, initially through applying a series of heuristics to semiautomatically filter anglicism candidates based on pattern matching and character probability (Furiassi and Hofland, 2007; Andersen, 2012; Serigos, 2017a), and more recently based on machine learning techniques (Losnegaard and Lyse, 2012).

In the area of NLP, lately some automatic approaches to borrowing detection have been proposed, normally aimed towards improving how natural language processing systems deal with out-of-vocabulary words and low-resourced languages (Alex, 2008; Mansikkaniemi and Kurimo, 2012; Leidig et al., 2014; Tsvetkov and Dyer, 2016).

Several projects have approached the task of extracting lexical borrowings in various European languages, such as German (Alex, 2008; Garley and Hockenmaier, 2012; Leidig et al., 2014), Italian (Furiassi and Hofland, 2007), French (Alex, 2008; Chesley, 2010), Finnish (Mansikkaniemi and Kurimo, 2012), and Norwegian (Andersen, 2012; Losnegaard and Lyse, 2012), with a particular focus on anglicism extraction.

Among examples of heuristic-based methods, Furiassi and Hofland (2007) explored corpora lookup and character n-grams to extract pseudoanglicisms from a corpus of Italian newspapers.

Alex (2008) combined lexicon lookup and a search engine module that used the web as a corpus to detect English inclusions in a corpus of German and French texts.

Also in French, Chesley (2010) extracted anglicisms from newswire using pattern-matching search.

Andersen (2012) used dictionary lookup, regular expressions and lexicon-derived frequencies of character n-grams to detect anglicism candidates in the Norwegian Newspaper Corpus (NNC) (Hofland, 2000).

Nevěřilová (2016) built a pipeline for the automatic identification of language mixing in Czech texts (which included proper names, expressions and foreign words) based on dictionary and corpora lookup.

Moreno Fernández and Moreno Sandoval (2018) applied a series of pattern restrictions along with dictionary lookup and manual inspection to a corpus of 89,000 tweets in order to identify 500 anglicism candidates from a corpus of tweets in US Spanish.

Mikušová (2020) combined pattern matching and corpus lookup to identify anglicism candidates in a corpus of Finnish, and proposed a system to rank them based on their likelihood of actually being an anglicism.

Mæhlum and Ivanova (2023) proposed a loanword detection system for Sakha based on phonotactic rules to leverage POS-tagging and parsing in low-resourced settings.

Other projects, on the other hand, have taken a more machine-learning heavy approach. Goldberg and Elhadad (2008) proposed a loosely supervised method for detecting transliterated foreign names and borrowings in Hebrew. To do so, they gathered words from a corpus of old Hebrew texts as data for native forms (under the assumption that borrowings would be non-existent in old texts), and automatically generated pseudoborrowings using a set of transliteration rules.

Losnegaard and Lyse (2012) explored a machine learning approach to anglicism detection in Norwegian by using TiMBL (Tilburg Memory-Based Learner, an implementation of a k-nearest neighbor classifier) with character trigrams as features.

Mansikkaniemi and Kurimo (2012) proposed a method based on n-gram perplexity to detect foreign names in morphologically-rich languages like Finnish.

Garley and Hockenmaier (2012) trained a maxent classifier with character n-gram and morphological features to identify anglicisms in German online communities.

Also in German, Leidig et al. (2014) used language model perplexity and lexicon and corpora lookup to build an English inclusion classifier based on decision trees, SVM and a majority voting system.

Koo (2015) presented an unsupervised approach based on the expectation-maximization algorithm to detect loanwords in Korean.

Tsvetkov et al. (2015) and Tsvetkov and Dyer (2016) proposed a finite-state pipeline of morpho-phonological transducers based on Optimality Theory (OT) to identify plausible donor-loan words pairs. They devised this model as a way of leveraging OOV words for machine translation in low-resourced scenarios and applied them to Arabic-Swahili, French-Romanian and Italian-Maltese pairs.

Spektor (2021) built a finite-state pipeline based in linguistic properties and trained an SVM and a feed-forward neural net to identify loanwords in Russian.

Also in Russian, Lukichev et al. (2023) trained Transformer-based models (ruBERT, ruRoberta, XLM-RoBERTa) to identify anglicisms in Russian within a pipeline aimed at substituting anglicisms in Russian texts.

Zaitsev and Minchenko (2022) trained a logistic regression classifier based on character probability to detect borrowings in Andic languages.

Nath et al. (2022) explored several language-agnostic models (logistic regression, SVM, random forest and a deep neural network) for borrowing detection based on edit distance, semantic similarity measures and phonetic alignment, and trained them on data harvested from Wikipedia.

Colakovac and Borucinsky (2024) built a model based on pattern matching, character n-grams and an SVM classifier to identify borrowed words in a corpus of Croatian.

Dinu et al. (2024a) tested various ensembles of machine learning classifiers (SVM, Naive Bayes) and Tranformer-based models to identify the direction of borrowing in word pairs from various Romance languages and to distinguish between borrowings and cognates (Dinu et al., 2024b) using the RoBoCoP dataset (Dinu et al., 2023).

Lately, Uighur has received considerable attention in terms of borrowing identification systems. Uighur is a Turkic language that receives heavy influence from its surrounding neighbours. It is a low-resourced language that consistently borrows from a Arabic, Russian, Chinese, Turkish and Farsi, which makes the study of borrowing in Uighur particularly attractive. As a result, several borrowing detection systems have been proposed for Uighur to leverage OOV words in machine translation: Mi et al. (2018a,b) proposed a cross-lingual word embedding model to identify loanword candidates based on semantic similarity. In Mi (2023b), they proposed a BiLSTM-CRF model enriched with linguistic features trained on augmented data from Wikipedia. And in Mi (2023a) they proposed a BiLSTM-CRF model for loanword identification in social media texts in Uighur where loanword identification was approached as multitask learning along with POS, NER and spelling error correction. The identified loanwords were later processed in order to identify potential donor words based on edit (phonetic) distance, and semantic distance.

Concretely for the task of retrieving anglicisms from Spanish text, Serigos (2017a,b) built an extractor of anglicisms from a corpus of Argentinian newspapers and TV and film subtitles. This anglicism identifier calculated the probability of a word being English or Spanish based on character n-grams; when the difference between both probabilities was smaller than a given threshold, a lexicon lookup module adjudicated whether the word was an anglicism or not. In order to avoid mistaking named entities with anglicisms, a capitalization module was set to ignore capitalized words. This system achieved an F1 score of 74.50% on unseen data.

Also in Spanish, Stewart et al. (2021) proposed a pipeline for collecting various forms of loanwords using dictionaries and manual filtration to analyze the integration of anglicisms in verb constructions in social media texts.

As lexical borrowings can be single-word (e.g. *app*, *online*, *smartphone*) or multiword spans (e.g. *machine learning*), they are a natural fit for applying chunking-style approaches. In Álvarez Mellado (2020a) we first framed the task of anglicism detection in Spanish as an

extraction task and introduced chunking-based annotation with BIO encoding along with sequence labeling models for borrowing detection in Spanish media, producing an F1 score of 86.41%.

Chiruzzo et al. (2023) also applied chunking-based approaches to the IberLEF 2023 shared task on distinguishing between codeswitches and unassimilated loanwords from Spanish into Guarani on a codemixed dataset of newswire and tweets. The best-performing system yielded an F1 score of 38.36%, which was obtained by a multilingual BERT model that was first fine-tuned on monolingual Guarani data (Pugh and Tyers, 2023), followed by a multitask learning approach based on a NLLB model (a model trained for machine translation aimed at low-resourced languages) with a CRF as a decoder (F1=19.39) (Muñoz Ortiz and Vilares, 2023).

2.1.2 A typology of borrowing detection systems

The works we have just presented comprise very different types of projects, both in terms of methods applied and final aim. For systematization purposes, we propose to classify existing borrowing detection projects according to the following four aspects:

1. Purpose: The motivation and final aim of the project will guide some of the decision-making behind the development. While some borrowing detection systems were developed within humanities disciplines and are more geared towards the study of language itself, other systems were designed as part of an NLP pipeline and therefore are more engineering-driven. We may distinguish three main types of purposes:
   (a) As an NLP component: the borrowing detection system is designed as a component within an NLP pipeline, with the aim of improving downstream results or as a way to leverage existing resources to account for low-resourced languages. In that regard, borrowing detection systems have been designed to be incorporated into parsing (Alex, 2008), machine translation (Tsvetkov et al., 2015; Tsvetkov and Dyer, 2016; Mi et al., 2020), automatic speech recognition (Mansikkaniemi and Kurimo, 2012;

Pritzen et al., 2021) or text-to-speech synthesis (Leidig et al., 2014).

   (b) As a tool for corpus analysis: the borrowing detection system is designed as a term extractor to retrieve borrowings from corpora and thus facilitate the analysis of language synchronically. That is the case of Garley and Hockenmaier (2012); Andersen (2012); Serigos (2017a,b); Bogunović (2023); Colakovac and Borucinsky (2024).

   (c) As a tool for historical linguistics: the borrowing detection system is devised as a way to establish phylogenetic relationships between languages, attest language contacts and reconstruct ancient protolanguages. That is the case of works by List (2019); Miller et al. (2020b); Ciobanu and Dinu (2020); Cristea et al. (2021); Dinu et al. (2024a,b)

2. Method: some systems will rely on heuristics to identify borrowing candidates, such as lexicon and corpora lookup (Alex, 2008; Nevěřilová, 2016; Moreno Fernández and Moreno Sandoval, 2018), pattern matching (Chesley, 2010) or character combination probabilities (Mansikkaniemi and Kurimo, 2012), or a combination of these (Andersen, 2012; Serigos, 2017a,b). Others will train machine learning models, either classical supervised machine learning models (Losnegaard and Lyse, 2012; Leidig et al., 2014), deep learning models (Spektor, 2021) or transformer-based models (Nath et al., 2022; Pugh and Tyers, 2023; Lukichev et al., 2023; Dinu et al., 2024a).

3. Approach: some systems will take full sentences as input and the task will consist in identifying borrowing candidates within context (Serigos, 2017a,b), whereas in other works the input will be isolated words from word lists, with no surrounding context whatsoever (Miller et al., 2020b). This implies that in some projects the context will be taken into account during the tagging process (Mi, 2023a; Pugh and Tyers, 2023), while for others there will be no context whatsoever (Spektor, 2021; Miller et al., 2020b). As a result, different projects will frame the task of borrowing detection in different ways:

   (a) Binary classification task: a word is given and the system must classify it as bor-

rowing or not.

(b) Word pair classification: two words are given and the system has to determine whether one has been borrowed from the other.

(c) Span identification: relevant spans of text of variable length are retrieved from text.

4. Recipient and donor languages: borrowings are foreign words that are used in a text that is otherwise monolingual. In that regard, some projects have taken a fully language-agnostic approach to the problem, and aim to identify borrowings from any donor language into a collection of recipient languages (Nath et al., 2022), while others have taken a more restricted approach and limited the identification of borrowings to a given donor language (English, for Alex (2008)), to a given recipient language (Uyghur, for Mi et al. (2020); French for Zurbuchen and Voigt (2024)) or both (English borrowings in Spanish for Serigos (2017a,b); Spanish borrowings in Guarani for Chiruzzo et al. (2023); Pugh and Tyers (2023)).

The aspects we have just enumerated are not completely independent from one another: systems designed with a linguistic purpose tend to orbitate towards heuristic methods, while NLP developments are more geared towards machine learning solutions. Systems that take word lists as input will necessarily approach the problem as a word classification task and there will be no context to consider. Although the opposite is not necessarily true: systems that take sentences as input can also approach the problem as a classification problem on a word-by-word basis, and ignore context altogether. Some projects approach the problem on a word-by-word basis, but then fuse adjacent words into spans applying heuristics, transforming it into a span retrieval problem (Serigos, 2017a,b).

Table 2.1 classifies existing works on borrowing detection according to the aspects we have just enumerated. Although these aspects may not be exhaustive, this classification may help us get a better understanding of how previous work have framed the problem of borrowing detection, how systems are different to one another and where our research lies.

Work	Purpose	Method		Approach			Languages	
		Heuristics/Features	Model	Input	Context	Framing	Recipient	Donor
Furiassi and Hofland (2007)	Corpus analysis	Lexicon lookup, char n-gram		Sentences	No	Binary classification	it	en
Alex (2008)	NLP component (parsing)	Lexicon/corpus lookup		Sentences	No	Binary classification	fr, de	en
Goldberg and Elhadad (2008)	NLP component	Lexicon lookup	NB	Words	No	Binary classification	he	multi
Chesley (2010)	Corpus analysis	Pattern matching		Sentences	No	Binary classification	fr	en
Andersen (2012)	Corpus analysis	Pattern matching, lexicon lookup, char n-gram probability		Sentences	No	Binary classification	no	en
Losnegaard and Lyse (2012)	Corpus analysis	Character n-gram probability	k-NN	Sentences	No	Binary classification	no	en
Mansikkaniemi and Kurimo (2012)	NLP component (ASR)	Character n-gram probability		Sentences	No	Binary classification	fi	multi
Garley and Hockenmaier (2012)	Corpus analysis	Lexicon lookup, char n-gram probability	Maxent	Sentences	No	Binary classification	de	en
Leidig et al. (2014)	NLP component (text-to-speech)	Lexicon lookup, char n-gram probability	Decision tree, SVM	Words	No	Binary classification	de	en
Koo (2015)		Character n-gram probability	EM		No	Binary classification	ko	en
Tsvetkov et al. (2015),2016	NLP component (MT)	Morpho-phonological features (OT)		Words	No	Word pair classification	multi	multi
Nevěřilová (2016)	Corpus analysis	Lexicon lookup		Sentences		Span identification	cs	multi
Serigos (2017b)	Corpus analysis	Pattern matching, lexicon lookup, char n-gram probability		Sentences	No	Binary classification	es	en
Moreno Fernández and Moreno Sandoval (2018)	Corpus analysis	Pattern matching, lexicon lookup		Sentences	No	Binary classification	es	en
Mi et al. (2018b)	NLP component (MT)	Edit (phonetic) distance, semantic distance, char n-gram probability, POS tagging	Log-linear model, BiLSTM	Words	No	Binary classification	ug	ru, ar, tr, zh
Mi et al. (2018a)	NLP component (MT)	Edit (phonetic) distance, char embeddings, POS tagging	BiLSTM-CRF	Sentences	Yes	Binary classification	ug	ru, ar, tr, zh
Álvarez Mellado (2020a)	Corpus linguistics	Character n-grams	CRF, BiLSTM-CRF	Sentences	Yes	Span identification	es	multi (en)
Ciobanu and Dinu (2020)	Historical linguistics	Character n-grams	Decision tree, SVM, NB	Words	No	Word pair classification	ro	it, pt, es, tr
Miller et al. (2020b)	Historical linguistics	Bag of words (sounds), sound n-gram probability	SVM	Words	No	Binary classification	multi	multi
Mikušová (2020)	Corpus analysis	Pattern matching, lexicon lookup		Sentences	No	Binary classification	fi	en
Miller et al. (2021)	Historical linguistics	IPA transcription	Transformer-based classifier, recurrent network	Words	No	Binary classification	multi	multi (es)
Spektor (2021)	Corpus analysis	Character n-gram probability	Logistic regression, SVM, FFNN	Sentences	No	Binary classification	ru	en
Cristea et al. (2021)	Historical linguistics	Character n-gram, edit distance, diachronic rules of sound evolution	Random Forests, Gradient Boosting, Multi-layered Perceptron, XGBoost, Bi-LSTM and Support Vector Machines (SVM)	Words	No	Binary classification	ro, it, fr, es, pt, ca	la
Nath et al. (2022)	NLP component	Edit (phonetic) distance, semantic (cosine) distance, phonetic alignment	Logistic regression, SVM, random forest, deep neural network	Words	No	Binary classification	multi	multi
Zaitsev and Minchenko (2022)	NLP component	Char n-gram, phoneme n-gram	Logistic regression	Words	No	Binary classification	akv, ani, bph, cji, kpt, kva, gdo, tin	en
Mi (2023a)	NLP component (MT)	POS, NER, SEC (multitask learning)	BiLSTM-CRF	Sentences	Yes	Binary classification, word pair classification	ug	ru, ar, tr, zh, fa
Lukichev et al. (2023)	NLP component (anglicism substitution)		ruBert, ruRoBERTa, XLM-RoBERTa	Sentences	No	Binary classification	ru	en
Pugh and Tyers (2023)	NLP component		mBERT	Sentences	Yes	Span identification	gn	es
Mæhlum and Ivanova (2023)	NLP component	Phonotactic rules		Sentences	No	Binary classification	sah	ru
Zurbuchen and Voigt (2024)	Corpus analysis	Lexicon lookup, MUSE embeddings	k-NN	Sentences	No	Binary classification	fr	multi
Colakovac and Borucinsky (2024)	Corpus analysis	Pattern matching, lexicon lookup, char n-gram	SVM	Sentences	No	Binary classification	hr	en
Dinu et al. (2024a,b)	Historial linguistics	Char n-gram, phonemic n-gram	SVM, Naive Bayes, Transformer	Words	No	Word pair classification	it, r, pt, es , ro	it, r, pt, es , ro

Table 2.1: Borrowing detection systems classified by purpose, method, type of input, approach and languages.

2.2 Methodological limitations from previous work

In the previous section we introduced and systematized the existing work on borrowing detection systems. We now describe some of the general limitations from previous work and lay out why some of the approaches we have just presented may be insufficient or unsuited for the use case we envision. Our critique will focus on the limitations of the three most mainstream techniques (pattern matching, character n-gram probability, lexicon lookup) and the most popular methodological framing (binary classification). We will also justify the importance of considering multiword borrowings when building borrowing detection systems.

2.2.1 Limitations of pattern matching

Pattern matching consists in defining a set of patterns (substrings, character combinations, etc.) that are unlikely to appear in native words from the recipient language but are reliable indicators of a word being a borrowing. For instance, a native Spanish word is not expected to contain the digraph *wh*, which is very frequent in English. We could therefore hypothesize that any word in Spanish text that contains the pattern *wh* is likely to be borrowed.

Pattern matching was for instance the approach followed by Chesley (2010), who manually searched for certain character combinations in a corpus to retrieve borrowing candidates in French newswire.

The main issue with pattern matching is that it will only retrieve borrowings that match the given patterns. For instance, in Spanish we may be tempted to search for words that end in *-ing*, with the hope that it will yield English borrowings. And we will certainly be able to identify a good amount of anglicisms that match that pattern (*streaming*, *warning*, *topping*, *restyling*), but there will be many other borrowings that will necessarily go unnoticed because they do not match the defined pattern (or any other unusual pattern whatsoever).

Moreover, pattern matching is not free from the curse of ambiguity: the digraph *sh* may look unambiguously foreign in Spanish and may seem as a good indicator of a word being

borrowed (as in *show)*, but it can also appear in prefixed Spanish native words such as *deshacer*.

2.2.2 Limitations of character n-gram probability

A more sophisticated approach than pattern matching that also incorporates the shape of the word into the decision-making is to calculate the probability of a word given its character combination, that is, by using a character n-gram model. For instance, that was the approach followed by Mansikkaniemi and Kurimo (2012) to retrieve foreign words from Finnish.

The problem with this approach is that, even when some character combinations will be more likely than others, it ignores the fact that some combinations and even some words can exist in both languages: words like *pie*, *social*, *come*, *sea* or *red* exists both in English and in Spanish. Whatever the result of the character n-gram model is, a system based on n-gram probability only will be by definition incapable of distinguishing between the English and non-English use of the word *pie*.

2.2.3 Limitations of lexicon lookup

Lexicon lookup systems have also been a popular approach to borrowing detection, as in Furiassi and Hofland (2007); Alex (2008); Serigos (2017a,b). In this approach, a lexicon of the donor language (that is, a dictionary or some sort of curated word list) is chosen as a lookup reference. Let's say that we take a dictionary of English words and check whether the words in a Spanish corpus are included in that English dictionary. If a word is present, then it is considered an anglicism. Other works make use of already-compiled lists of anglicisms, such as Stewart et al. (2021).

An alternative approach is to use an exclusion word list, that is, a list of words that will not be considered a borrowing. In our previous example, we could use a dictionary of Spanish words and decide that we will label as a borrowing candidate any word that is not found in the Spanish dictionary. That was precisely the approach taken by Moreno Fernández

and Moreno Sandoval (2018): in order to retrieve anglicism candidates from a corpus, they applied a filter to discard all the words that were present in an inflected lexicon of Spanish.

Although this approach may be convenient (because it enables reusing previously compiled resources such as lexicons, automatically-populated pages from Wiktionary, etymological dictionaries, etc.), lexicon lookup suffers from several limitations. First, the fact that a word is listed or not in a reference word list may not indicate that is a borrowing: again, there are words that coincidentally exist as words in both languages. Automatically considering those as borrowings (or discarding them as candidate borrowings) will be misguided.

In addition, a strategy that simply labels as a borrowing anything that is present (or not present) in a given lexicon will necessarily be ill-equipped to deal with text where foreign non-borrowed words appear. That is the case of journalistic text, where foreign proper nouns, literal quotations and other language inclusions abound.

Finally, works that build on previously compiled anglicism lists assume that there is an already compiled list to begin with. In other words, this approach makes use of prior lexicographic work to facilitate borrowing identification. But who facilitates lexicographic work? How do these already compiled borrowings enter a dictionary to begin with? The purpose of our work is to devise computer-assisted methods that can assist lexicographers identify borrowings that appear in real-world data but are not registered in mainstream dictionaries or word lists yet. An approach that works by matching borrowings from a lexicon will be useless precisely for the task of assisting at automatically compiling those very same lists.

2.2.4 Limitations of word classification approaches

The limitations from the three techniques that we listed above (pattern matching, n-gram probability, lexicon lookup) can be subsumed into the following more general point: the fact that context is key when determining if something is a borrowing or not.

The huge majority of approaches to borrowing detection systems we have seen framed

the task as a binary classification problem, in which isolated words with no surrounding context whatsoever are analyzed to determine if it is a borrowing or not. Even in works where the word in question appeared in context (that is, works in which the input was a full sentence), the problem was tackled on a word-by-word basis and the surrounding context of the word was considered irrelevant to determine whether the word was a borrowing or not, as in Furiassi and Hofland (2007); Andersen (2012); Serigos (2017a,b).

The problem with this approach is that binary classification of isolated word assumes that words can be classified as a borrowing or not regardless of the context they appear in. Consider the following real sentence, extracted from the satyrical news site *El Mundo Today*:

Example 1. *Cada vez más personas optan por el van life y se van a vivir a una autocaravana o a una furgoneta camperizada.*[1][2]

The word *van* will be part of a multiword borrowning in *van life*, but will be a native word when it appears as a conjugated form of the verb *ir*, as in *se van a vivir*.

This problem extends to many other borrowings that also exists as homograph native words. A word like *pie* will be a non-borrowing in *caer de pie*[3], but it will be a borrowing in *receta de pie de limón*[4]. Similar issues arise when we deal with proper names: the word *cash* will be a borrowing when it appears in a sentence such as *Estoy sin 'cash'*[5], but it will be part of a proper name (thus, not a borrowing) in the sequence *Johnny Cash*. The key issue is that in order to correctly decide whether a given word is a borrowing (or part of a borrowing) we need to condition on the context. A system that ignores context when identifying borrowings will be incapable of distinguishing between borrowings and other phenomena, such as native homograph words or other foreign language inclusions like proper nouns, literal quotations, etc., phenomena all of them absolutely prevalent in standard corpus

[1] "Everyday, more and more people opt for a 'van life' and decide to move into an RV or a motorhome."

[2] `https://www.elmundotoday.com/2023/10/como-convertir-tu-piso-en-una-autocaravana-para-tener-una-autocaravana-grande-en-vez-de-un-piso-pequeno/`

[3] "To land on your feet"

[4] "Lemon pie recipe"

[5] "I'm broke"

material such as newspapers.

These issues are no different to the type of ambiguities that other contextual task such as named entity recognition (NER) or multiword expression identification (MWE) encounter: depending on the context, *Jobs* can be a proper name (a named entity) or a noun (Nadeau et al., 2006); the sequence *to make a face* can be an idiom, but also a literal phrase (Fazly et al., 2009). And, in all of these cases, there is no way to determine which is the right category deprived from context, in other words, in isolation. In NER and MWE, we frame the problem as a sequence labeling task in which the labeling of each tag is done contextually and where spans of interest are retrieved from text. To put it another way, we frame these tasks as extraction tasks. We defend that borrowing detection should be treated similarly.

Binary classification approaches (where words are classified in isolation) may still be useful in historical linguistics, where we want to trace the etymological origin of words or establish contact relations between languages based on vocabulary lists. But if we are interested in retrieving in-context borrowings from real-world sentences (as corpus linguist and lexicographers normally do), the word-by-word approach is essentially flawed.

2.2.5 On the importance of multiword borrowing detection

Multiword borrowings are borrowings that consist of two or more words, such as *machine learning*, *pole position*, *all male panel* or *fake news*. Multiword borrowings are by no means infrequent. However, they have only seldom been considered by borrowing detection systems.

Borrowing detection systems that take isolated words as input disregard the phenomenon of multiword borrowing altogether. Given that these systems approach the problem as a classification task on isolated words and there is no context from which the borrowings need to be retrieved, there is no point in considering the possibility of borrowings of variable length.

To the best of our knowledge, only a handful of works have explicitly considered multiword borrowings when building their systems:

1. The lexicon lookup system by Nevěřilová (2016) considered foreign multiword expressions as part of their processing system, but their extraction was conflated with other language mixing multiword phenomena, such as proper names and quotes.

2. The system proposed by Serigos (2017a,b) does consider the existence of multiword borrowings (or phrase borrowings). However, the approach followed by Serigos assigns a binary tag on a word-by-word basis and simply assumes that any two adjacent words in a sentence that have been tagged as a borrowing by their system will necessarily be a multiword borrowing. This approach prevents the system from distinguishing between true multiword borrowings, where the full span was borrowed and works as a lexical unit (as in *machine learning*), from two separate borrowings that happen to be collocated in a sentence (as in *burger gourmet*, where *burger* and *gourmet* were borrowed independently but may appear next to one another).

3. Mi et al. (2018a) applied chunking-based annotation (BIOES) at character level for borrowing detection in Uighur sentences using sequence labeling models, although their output layer transformed the final label into a binary classification task and they did not report their treatment of multiword borrowings. In addition, their approach also considered foreign proper names and other named entities as borrowings and their identification system was limited to nouns.

4. The annotation on the task on borrowing versus codeswitching proposed by Chiruzzo et al. (2023) at IberLEF 2023 was implemented following chunking-based BIO encoding, and therefore the systems that were submitted to the task allowed for the correct identification of multiword borrowing from Spanish into Guarani (Pugh and Tyers, 2023; Muñoz Ortiz and Vilares, 2023).

5. The system proposed by Colakovac and Borucinsky (2024) for anglicism identification in Croatian was originally aimed at also retrieving multiword borrowings, but their implementation was shown to be insufficient for the task and multiword borrowings had eventually to be incorporated via manual corpus search.

In our approach, we consider multiword borrowing a crucial aspect of in-context borrowing detection that should not be overlooked. We lay out here the theoretical motivation behind our criterion:

Multiword borrowing is a manifestation of a more general aspect of the structure of the language: the fact that the lexicon is not organized based on tokenized strings, but in multiword units, and that the productivity of the lexicon is governed by the principle of recursion, which accounts for the existence of multiword units. In fact, some estimations argue that multiword units in general account for between 30%-50% of the words in a language (Conklin and Schmitt, 2012).

Moreover, the existence of multiword borrowing has been pointed out as a central issue in the definition of lexical borrowing itself. As Treffers-Daller (2023) states:

> "Because the distinction between borrowing and code-switching is essentially a specific instantiation in bilingualism of the basic distinction between words and rules, studying how single words and MWUs from one language are used in another language can also contribute to theory building on the distribution of labour between vocabulary and grammar."

In other words, it is not just a matter of how frequent multiword borrowings are (and whether their frequency justifies ignoring them or dealing with them straightforwardly), but a matter of how accurate our depiction of lexical borrowing as a linguistic phenomenon is. Consequently, a system that aims to model lexical borrowing as a phenomenon should be able to adequately account for multiword borrowings.

Given the key implications that multiword borrowings have, we propose to move from the conventional approach that frames borrowing detection as a binary classification problem into an extraction problem, where spans of variable length are retrieved from context (thus enabling both single-item and multiword borrowing retrieval). The purpose of this dissertation is precisely to explore the consequences in terms of annotation, modeling and evaluation that reframing entails.

2.3 Differences with previous work

We now summarize the ways in which our approach to in-context lexical borrowing identification differs from other adjacent research areas (such as codeswitch identification or loanword identification within historical linguistics) and identify the two existing works that are the closest in nature to our research, and how they compare to our approach.

2.3.1 Differences with work within historical linguistics

There is a growing interest in the realm of historical linguistics in applying computer-assisted techniques to the comparative method (List, 2019). The automatic identification of loanwords has seen a surge in historical linguistics (Minett and Wang, 2003; Miller et al., 2020b; Köllner, 2021; Miller et al., 2021; Miller and List, 2023; Miller, 2024), mainly as a way of establishing phylogenetic relations between languages, attesting language contact and reconstructing ancient protolanguages or proto-words through cognate identification (Ciobanu and Dinu, 2020).

Computational approaches to language contact within historical linguistics tend to fall in either of the following two approaches (List, 2019): (a) identifying certain traits that cannot be explained within a phylogenetic tree (and that could be explained via language contact) and (b) identifying similarities across languages on non-cognate words.

The purpose of these projects is to identify words in the lexicon of a language whose shape exhibits a conflict in regard to the shape that the word should have according to its expected phonological evolution and that can be explained by hypothesizing a possible borrowing process. For example, Miller et al. cite the English word *mountain* as an example of phylogenetic conflict, as its similarity to the words meaning 'mountain' in the Romance languages (Italian *monte*, Spanish *montaña*) suggests that it must have been borrowed from a Romance language, instead of being originated natively following the expected shape that Germanic languages have (German *Berg*, Dutch *berg*).

Recently, there have been several projects that have developed computer-assisted resources for the task of establishing the etymological origin of words and identifying loanwords in the lexicon of the language, such as `LingPy`[6] (List and Moran, 2013; List and Forkel, 2021), `PyBor`[7] (Miller et al., 2020a) and `LoanPy`[8] (Martinović, 2023).

Although related to our work (as loanword identification is our common interest), the scope, methodology and purpose of our research is quite different to the work developed within historical linguistics. We list here the main differences:

1. The subject of study in computer-assisted historical linguistics is vocabulary lists (such as WOLD (Haspelmath et al., 2014)), while our work focuses on retrieving previously unlisted lexical borrowings from context sentences.
2. The loanword identification task is therefore framed as a word classification problem, in which an isolated word is classified as a loanword or not. Our work approaches the problem as a sequence labeling task in which spans of interest are retrieved from context.
3. Given that their aim is to unmask established words in a language that may have originated in another language, these projects focus on identifying *assimilated* loanwords, while our purpose is to identify novel, *unassimilated* borrowings that may have not even be registered in dictionaries today.
4. The purpose of their approach is to identify the etymology of words as a way to establish phylogenetic relations or possible contacts between languages in a diachronic fashion. Our aim is to provide corpus linguists and lexicographers with a term-extraction technology that may assist them at identifying candidate borrowings in a corpus (for dictionary addition, for instance) and for monitoring changes in the lexicon synchronically.

Although these projects are closely related to our own work, we consider this research a different endeavour as it is steered by different methods and purposes, which makes our efforts complementary but non-comparable.

[6] `https://lingpy.org/`
[7] `https://github.com/lingpy/pybor`
[8] `https://github.com/LoanDB/loanpy`

2.3.2 Differences with work within code-switching identification

While borrowing detection is quite niche as an NLP topic, codeswitching identification on the other hand has seen a lot of attention from the NLP community, essentially aimed at processing social media messages from multilingual speakers. There is a thriving community that works on NLP applied to codeswitched text, as well as established shared tasks on the topic (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018; Chen et al., 2022) and benchmarks (Aguilar et al., 2020).

As mentioned in section 1.1.3, the scope and nature of lexical borrowing is, however, somewhat different to that of code-switching. In fact, applying code-switching models to lexical borrowing detection has previously proved to be unsuccessful, as they tend to overestimate the number of anglicisms (Serigos, 2017a).

Moreover, in terms of approaching the problem, automatic code-switching identification has been framed as a sequence modeling problem where every word receives a language ID label (as in a part-of-speech tagging task).

In this thesis we argue that in-context retrieval of lexical borrowings is a good fit for sequence-labeling approaches but, unlike codeswitching identification, the problem should be framed as a span identification task (as in named entity recognition or chunking), where strings of words of variable length will be retrieved from sentences.

2.3.3 Differences with previous work on anglicism detection in Spanish

There are two prior works that specifically tackle the identification of in-context anglicisms in Spanish using computer-assisted methods: Serigos (2017a,b), and our own work from Álvarez Mellado (2020a).

Serigos's work was the first to show that computational methods could successfully be applied to perform automatic detection of anglicisms in Spanish newswire (Serigos, 2017a), and also the first to apply distributional techniques to analyze the semantic specificity of those extracted anglicisms (Serigos, 2017b). Serigos, however, makes some assumptions that

we deviate from. Here are some of the limitations of the anglicism identifier presented by Serigos (2017a):

1. Although the input to their system are sentences and therefore anglicisms appear in context, the system analyzes each word in isolation, i.e. the previous word is considered irrelevant to determine whether the current word is an anglicism.
2. The way the capitalization module works (discarding any word in titlecase as a named entity, including at the beginning of the sentence) assumes that anglicisms can never appear on the first position of the sentence. Serigos argues that nouns in the first position of a sentence are extremely rare in Spanish. Although this is true in many contexts, sentence initial anglicisms are very normal in newspaper headlines, which is precisely the type of sentence that has been shown to be particularly challenging for other sequence labeling tasks, such as part-of-speech tagging (Gorman and Bedrick, 2019). Moreover, capitalized anglicisms (such as *Big Data*) are in fact not rare, even in mid sentence positions.
3. Although the system includes multiword anglicisms (i.e. anglicisms that are formed by more than one word, such as *prime time*), given that the system considers every word in isolation (without considering the previous word), it assumes that all anglicisms occurring sequentially will be a single borrowing. This is definitely the most frequent case, but this assumption prevents the system from considering any other possibilities, such as two anglicisms being adjacent.
4. Only two language tags were considered: Spanish and English. No tags for any other languages were included and borrowings from other languages were labeled as Spanish during the annotation process. As Serigos points out on the results discussion, this decision hurt the final metrics, as the anglicism identifier tended to identify non-English borrowings as anglicisms.
5. The system was evaluated on token level exclusively, and no span level evaluation was given for multiword borrowings. Token level takes into account each tag individually

and any tag that is correct will be counted as a true positive, regardless of whether the full span that the individual token belongs to was correctly labeled or not. For example, in token level evaluation, if only *night* in *late night* was detected, *night* would still count as a true positive. In span level evaluation, only full matches over the entire span count, i.e. the entire span *late night* would have to be correctly labeled in order to count as a true positive. Span level evaluation may seem too harsh, because no credit is given to partial matches. However, token level evaluation can overstate results; after all, a model that would only detect English function words could get away with just detecting *and* in *rock and roll* or *by* in *stand by* and still get a generous result.

Concerning our own previous work, in Álvarez Mellado (2020a) we already proposed an annotated dataset and two models for the task of automatically retrieving lexical borrowings (mainly anglicisms) from Spanish text, in which the task was framed as an extraction task. At that time, we released a dataset of 325,000 tokens of journalistic texts written in European Spanish, annotated with anglicisms and other lexical borrowings following BIO encoded annotation. We explored two sequence labeling models for the task of automatically retrieving borrowings from text: a CRF model fed with handcrafted features (along with `word2vec` embeddings) and a BiLSTM-CRF model with word and character embeddings. The best result was obtained by the CRF model, which produced and F1 score of 86.41% on the test set.

We regard that preliminary work as a successful proof of concept that shows that borrowing identification in Spanish can be framed as an extraction task in which relevant spans of texts are retrieved from sentences. However, that previous work suffered from several limitations in terms of data, modeling and evaluation:

1. The dataset was limited in terms of sources considered and phenomena included.
2. The best results were obtained by a CRF model, and more sophisticated approaches were only timidly (and unsuccesfully) explored.
3. The highest results obtained were well below 90%, so in terms of performance there is

ample room for improvement for the task, particularly in terms of recall (R=80), which is critical in lexicography-oriented technology and was consistently lower than precision across all the evaluations performed.

4. The generalization capabilities of the model were limited: when the model was confronted with a different test set (with texts from the same genre, but on topics that were out of distribution), its performance dropped to a modest 70% in F1 score. In addition, 80% of the false negatives were previously unseen borrowings, and only 17% of the previously unseen borrowings in the test set were correctly identified. This means that although the model was not simply memorizing (as it could retrieve some borrowings never seen during training), its generalization capabilities left much to be desired.
5. The evaluation was exclusively performed at span level, and other metrics were disregarded.

In this work we propose to continue the approach we took on that first attempt from 2020a. We further explore the methodological issues and implications that arise when framing borrowing detection as an extraction task and critically revise the decisions and assumptions we made on that previous work in terms of data and modeling (part II), as well as evaluation (part III).

2.4 Summary

In this chapter we have reviewed previous computer-assisted methods to lexical borrowing detection. We have summarized the existing research on the subject; we have proposed a typology of systems based on end purpose, methods used, approach to the problem and languages considered; and we have identified the limitations that these existing methods have.

We have identified three relevant issues that past research failed to tackle and that are key when performing in-context retrieval of lexical borrowings that seeks to assist linguistic work or serve as a pre-processing step in NLP pipelines:

1. The modeling approach should be able to consider and capture multiword borrowings.
2. The modeling should be able to account for adjacent borrowings and distinguish them adequately from multiword borrowings (*machine learning* versus *burger gourmet*).
3. Word classification approaches (based on pattern matching, lexicon lookup, or plain character n-gram probability) are not enough to fully represent the problem because borrowings in isolation can be ambiguous (*cash*, *pie*). It is the *context* a given word or phrase appears in that informs whether something is a borrowing or not.

Finally, we have outlined the ways in which our approach to in-context lexical borrowing identification differs from other adjacent research areas or similar projects, such as codeswitch identification or loanword identification within historical linguistics.

Part II

Data & models

Chapter 3

COALAS 🐨: Corpus of Anglicisms in the Spanish Press

Without data, you're just another person with an opinion.

Edwards Deming, statistician

Suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world (...) and from that maybe they'll come up with some generalizations or insights. Well, sciences don't do this.

Noam Chomsky, linguist

In Álvarez Mellado (2020a) we already proposed a dataset of Spanish text annotated with anglicisms and other unassimilated lexical borrowings. We could rightfully ask whether it makes sense to compile more data for the task of detecting lexical borrowings in Spanish. Was there not enough data in the previous dataset? Was the annotation defective?

In this chapter we enumerate the limitations that the previous dataset had and that motivate the need for a new dataset of anglicism usage. We introduce the methodology followed in order to compile a brand new dataset of anglicism usage from scratch. We then

describe the data selection process and the annotation process.

The resulting dataset is COALAS[1], a corpus of anglicisms in the Spanish press annotated with lexical borrowings. This corpus is composed of 370,000 tokens and comprises 3,161 lexical borrowings (of which 1,600 are unique). COALAS is larger, more borrowing-dense, OOV-rich, and topic-varied than previous corpora available for the task of lexical borrowing detection. Finally, we list some of the limitations of COALAS that users of this resource should bear in mind.

3.1 Limitations of previous work and motivation

In Álvarez Mellado (2020a) we introduced a corpus of emerging anglicism annotated for the task of lexical borrowing detection. That dataset consisted of a collection of Spanish journalistic texts made of 325,665 tokens annotated with unassimilated lexical borrowings following chunking-based BIO encoding. That previous dataset was a first attempt at compiling a corpus aimed at lexical borrowing detection and it suffered from significant limitations, which we enumerate here:

1. The dataset focused exclusively on a single source of news, the Spanish newspaper *elDiario.es*.
2. The dataset consisted only of headlines.
3. The number and variety of borrowings were limited.
4. The dataset was annotated by one annotator only, and no inter-annotator agreement evaluation was provided.
5. There was a significant overlap in borrowings between the training set and the test set: 46% of the borrowings in the test set were unseen (with only 26% of the borrowings in the test being truly unseen, and an additional 20% were seen in training but with a label that was either unseen or ambiguous).

[1] `http://github.com/lirondos/coalas`

Although all these limitations are important, limitation #5 is critical. Lexical borrowings are a source of new words and senses. Their interest for linguists resides precisely in the fact that they are emergent elements in a language. Moreover, our aim is to build technology that can assist lexicographers during dictionary creation; consequently, these systems are expected to be able to detect novel borrowings. In other words, the type of borrowing that a lexicographer will be most interested in retrieving are previously unseen borrowings.

However, in the test set from Álvarez Mellado (2020a), 3 out of 4 borrowings in the test set appeared in the training set, and 1 out of 2 were seen with exactly the same unambiguous tag it had in the training set. This prevents us from truly assessing how good the model is at retrieving borrowings because, as Bernier-Colborne and Langlais (2020) state, in sequence labeling tasks like ours if we removed all unseen or ambiguous tokens from the test set it would suffice to assign each token the most frequent tag seen in training to get a perfect score. Consequently, the dataset from Álvarez Mellado (2020a) is essentially flawed and cannot serve our purposes: we need a better dataset of anglicism usage in Spanish that is rich in previously unseen borrowings so that we can truly assess the generalization capabilities of the models.

3.2 Corpus desiderata

With the limitations from section 3.1 in mind, we can now list the requirements that an ideal corpus for lexical borrowing detection should satisfy:

1. Our work is mainly concerned with detecting lexical borrowings from journalistic text. In that regard, our corpus should be restricted to newspaper articles, blog posts and other written sources from the journalistic domain.
2. Ideally, we aim to collect texts from sources that publish under flexible licensing that allow for resharing.
3. As we stated in section 3.1, what we envision is a system that is trained on existing wide-spread borrowings and can successfully retrieve novel previously unseen borrow-

ings. Assuming that, like all lexical phenomena, lexical borrowings follow a zipfian distribution (with a few borrowings that are very frequent and a very long tail of infrequent borrowings), what we seek is a system that can correctly identify the infrequent borrowings in the long tail after having learned what the frequent borrowings look like.

The downside to requirement #3 is that lexical borrowing is a sparse phenomenon and borrowed words are typically rare in monolingual discourse (Poplack et al., 1988). Previous work has estimated that anglicisms amount to 1%-2% of the language (Poplack et al., 1988; Rodríguez González, 2002; Chesley, 2010; Serigos, 2017a). This means that most sentences in a text will not contain borrowings. Even those that do, it is very likely that the contained borrowing will be one of the top frequent ones. We need to collect a corpus that is rich in borrowings, and compile a test set that contains rare ones so that we can assess models' generalization abilities. This means that we are hoping to build a corpus that is rich in a phenomenon that is not that frequent (because, as prevalent as borrowings may be, most sentences in texts will contain no borrowings whatsoever) and that incorporates a decent amount of rare examples of that infrequent phenomenon in the test set, so that we can assess generalization. These two requirements would call for a gigantic corpus, but that would require a massive annotation process that goes beyond the scope of this dissertation. Therefore, we need to find a way to satisfy these two requirements while keeping the annotation effort within the limitations of time and resources of this PhD project.

3.3 Data curation: sources and filtering

In order to satisfy the requirements from section 3.2, data was collected separately for the training, development, and test sets. This was done to ensure minimal overlap in borrowings, topics, and time periods.

- The training set consists of a collection of *articles* appearing between August and December 2020 in *elDiario.es*, a progressive online newspaper based in Spain.
- The development set contains *sentences* from the same source, but extracted from arti-

Set	Tokens	ENG	OTHER	Unique
Training	231,126	1,493	28	380
Development	82,578	306	49	316
Test	58,997	1,239	46	987
Total	372,701	3,038	123	1,683

Table 3.1: Number of tokens and spans in COALAS.

cles that were published in January 2021.

- The data in the test set consisted of annotated sentences extracted in February and March 2021 from a diverse collection of online Spanish media that covers specialized topics rich in lexical borrowings and usually not covered by *elDiario.es*, such as sports or videogames (see table 3.5).

To focus annotation efforts on articles that were anglicism-rich and sentences that contained good borrowing examples, we built a filtering pipeline that served us a bootstrapping system to retrieve texts that were good candidates for annotation. For that filtering pipeline we used the CRF model from Álvarez Mellado (2020a) along with a dictionary look-up pipeline that included English and Spanish word lists and a database of anglicisms used in Spanish press (see chapter 10). The following data selection process was followed (see figure 3.1 for a flowchart of the process):

- For the training set, articles from *elDiario.es* that contained more than 5 borrowing candidates (flagged by the CRF model) were selected for annotation.
- For the development and test sets, the main goal of the data selection was to create borrowing-dense, OOV-rich datasets, allowing for better assessment of generalization. To that end, the annotation was based on sentences instead of full articles. If a sentence contained a word either flagged as a borrowing by the CRF model, flagged by the wordlist pipeline or simply not present in the training set, it was selected for annotation. This data selection approach ensured a high number of borrowings and OOV words, both borrowings and non-borrowings.

3.4 Annotation process

The corpus resulting from the process described in section 3.3 was annotated with BIO encoding adapted from Ramshaw and Marcus (1995) to denote the boundaries of each span. The corpus was annotated by a native speaker of Spanish with a background in linguistic annotation (see appendix B) using Doccano (Nakayama et al., 2018). The annotation guidelines (provided in appendix A) were based on those from Álvarez Mellado (2020a) but were expanded to account for a wider diversity of topics. Following Serigos's observations and our previous work from 2020a, English lexical borrowings were labeled `ENG`, other borrowings were labeled `OTHER`. Here is an example from the training set:

Example 2. ***Benching*** *[ENG], estar en el banquillo de tu* ***crush*** *[ENG] mientras otro juega de titular.*[2]

In order to assess the quality of the guidelines and the annotation, a sample of 9,110 tokens from 450 sentences (60% from the test set, 20% from training, 20% from development) was divided among a group of 9 linguists for double annotation. The mean inter-annotation agreement computed by Cohen's kappa was 0.91, which is above the 0.8 threshold of reliable annotation (Artstein and Poesio, 2008).

3.5 Corpus description

The result of the process we have just described is COALAS (COrpus of AngLicisms in the SpAnish presS). Our dataset consists of Spanish newswire annotated for unassimilated lexical borrowings. The dataset is made of 372,701 tokens, and contains 3,038 spans labeled as `ENG` and 123 labeled as `OTHER` (see table 3.1).

Table 3.2 displays the most frequent spans in the training set, table 3.3 displays the most frequent spans in the test set. All of the sources used are European Spanish online

[2] "Benching: being on your crush's bench while someone else plays in the starting lineup."
`https://www.eldiario.es/red/que-es/benching-banquillo-crush-juega-titular_1_6484819.html`

Count	Type	Span	Count	Type	Span
197	ENG	online	15	ENG	streaming
196	ENG	app	14	ENG	ranking
58	ENG	apps	14	ENG	fake news
41	ENG	marketing	14	ENG	call center
37	ENG	lobby	12	ENG	black
29	ENG	software	12	ENG	podcasts
26	ENG	bluetooth	11	ENG	boom
26	ENG	influencers	10	ENG	startups
24	ENG	podcast	10	ENG	startup
20	ENG	bots	10	ENG	email
19	ENG	thriller	9	ENG	lobbies
18	ENG	parking	9	ENG	loot boxes
16	ENG	riders	9	ENG	trafficker

Table 3.2: Most frequent spans in the COALAS training set.

publications (newspapers, blogs, and news sites) written in European Spanish from Spain and published under open licensing (see table 3.5).

Our annotation is licensed with a permissive CC BY-NC-SA-4.0 license. With one exception, the sources included in our dataset release their content under Creative Commons licenses that allow for reusing and redistributing the material for non commercial purposes. This was a major point when deciding which news sites would be included in the dataset. The exception is the source *Microsiervos*, whose content was used with explicit permission from the copyright holder. Table 3.5 lists the URL and license type for each source.

3.6 Discussion

COALAS is similar in some ways to our previous dataset from Álvarez Mellado (2020a): both datasets are composed of Spanish journalistic texts, borrowings (which can be single-word items or multiword items) are annotated in context and spans are labeled using BIO encoding.

But in other ways, both datasets are clearly distinct. Table 3.4 summarizes the differences between both datasets. While in terms of number of tokens the difference between the

Count	Type	Span	Count	Type	Span
18	ENG	streaming	5	ENG	cropped top
17	ENG	look	5	ENG	rover
15	ENG	top	4	ENG	street style
8	ENG	online	4	ENG	cool
8	ENG	post	4	ENG	deepfakes
8	ENG	apps	4	ENG	onboarding
8	ENG	looks	4	ENG	gamers
6	ENG	software	4	ENG	gaming
6	ENG	oversize	4	ENG	spa
5	ENG	total look	4	ENG	hardware
5	ENG	podcast	4	ENG	midi

Table 3.3: Most frequent spans in the COALAS test set

previous corpus and ours is not much (less than 50,000 tokens), there is a substantial difference in the the amount of borrowings and the number of OOV spans between both datasets. In addition, the test set of COALAS is very borrowing-dense: while the training set contains 6 borrowings per 1,000 tokens, the test set contains 20 borrowings per 1,000 tokens. The test set is also very rich in out-of-vocabulary spans (or previously unseen spans): 92% of the borrowings in the test set did not appear in training or appeared with an ambiguous tag (83% were truly unseen, 9% were seen but with different tagging). In comparison, 46% of the borrowings in the previous test set were unseen (with only 26% being truly unseen). As we will see in chapter 4, these differences (subtle in terms of quantity, but significant in terms quality) will prove to be relevant when assessing for generalization abilities of models.

Like all resources, COALAS has significant limitations that its users should be aware of. To begin with, the corpus consists exclusively of news published in Spain and written in European Spanish. This fact by no means implies the assumption that European Spanish represents the whole of the Spanish language. On the contrary, the notion of assimilation is usage-based and community-dependent, and thus the dataset we present and the annotation guidelines that were followed were designed to capture a very specific phenomena at a given time and in a given place: unassimilated borrowings in the Spanish press.

In order to establish whether a given word has been assimilated or not, the annotation guidelines rely on lexicographic sources such as the prescriptivist *Diccionario de la Lengua Española* (Real Academia Española, 2020) by the Royal Spanish Academy, a dictionary that aims to cover world-wide Spanish but whose Spain-centric criteria has been previously pointed out (Blanch, 1995; Fernández Gordillo, 2014). In addition, prior work has suggested that different Spanish dialects and Spanish-speaking regions may borrow anglicisms at different rates (McClelland, 2021). Consequently, we limit the scope of the dataset to European Spanish not because we consider that this variety represents the whole of the Spanish-speaking community, but because we consider that the approach we have taken here may not account adequately for the whole diversity in borrowing assimilation within the Spanish-speaking world.

Finally, the corpus and annotation guidelines presented in this chapter were designed exclusively for identifying borrowings in Spanish journalistic monolingual text. Consequently, our guidelines do not consider the distinction between borrowing and codeswitching, as codeswitches are not expected to be found. In chapter 6 we will expand the guidelines we have presented here and include directions to account for the distinction between borrowing and codeswitching in conversational genres such as social media texts, where both codeswitches and borrowings abound.

	Álvarez Mellado (2020a)	**Ours**
# journalistic sources	1	21
# tokens	325,665	372,701
# `ENG` spans	1,304	3,038
# `OTHER` spans	102	123
% unseen spans (in test)	46%	92%

Table 3.4: Comparison between the dataset from Álvarez Mellado (2020a) and COALAS.

3.7 Summary

In this chapter we have presented COALAS, a novel dataset of Spanish newswire annotated with unassimilated lexical borrowings. We have analyzed the limitations that previous resources available for the task had, which justifies the need for this novel corpus. We have listed the desired properties that an ideal dataset for borrowing detection should have and presented the pipeline we built to compile such a corpus from Spanish journalistic texts. Finally, we have described the annotation process we followed and the inter-annotator agreement obtained.

The result of this process is COALAS, a sequence labeling dataset annotated for lexical borrowing following chunk-based BIO encoding. COALAS is composed of 372,701 tokens extracted from various journalistic sources written in European Spanish. It contains 3,161 spans labeled as `ENG` or `OTHER` and is more varied and rich in borrowings than previous available resources.

We expect that this dataset should be more challenging for models than previous datasets available for the task. In chapter 4 we will assess if already-existing models can successfully perform on COALAS or if more sophisticated techniques are required.

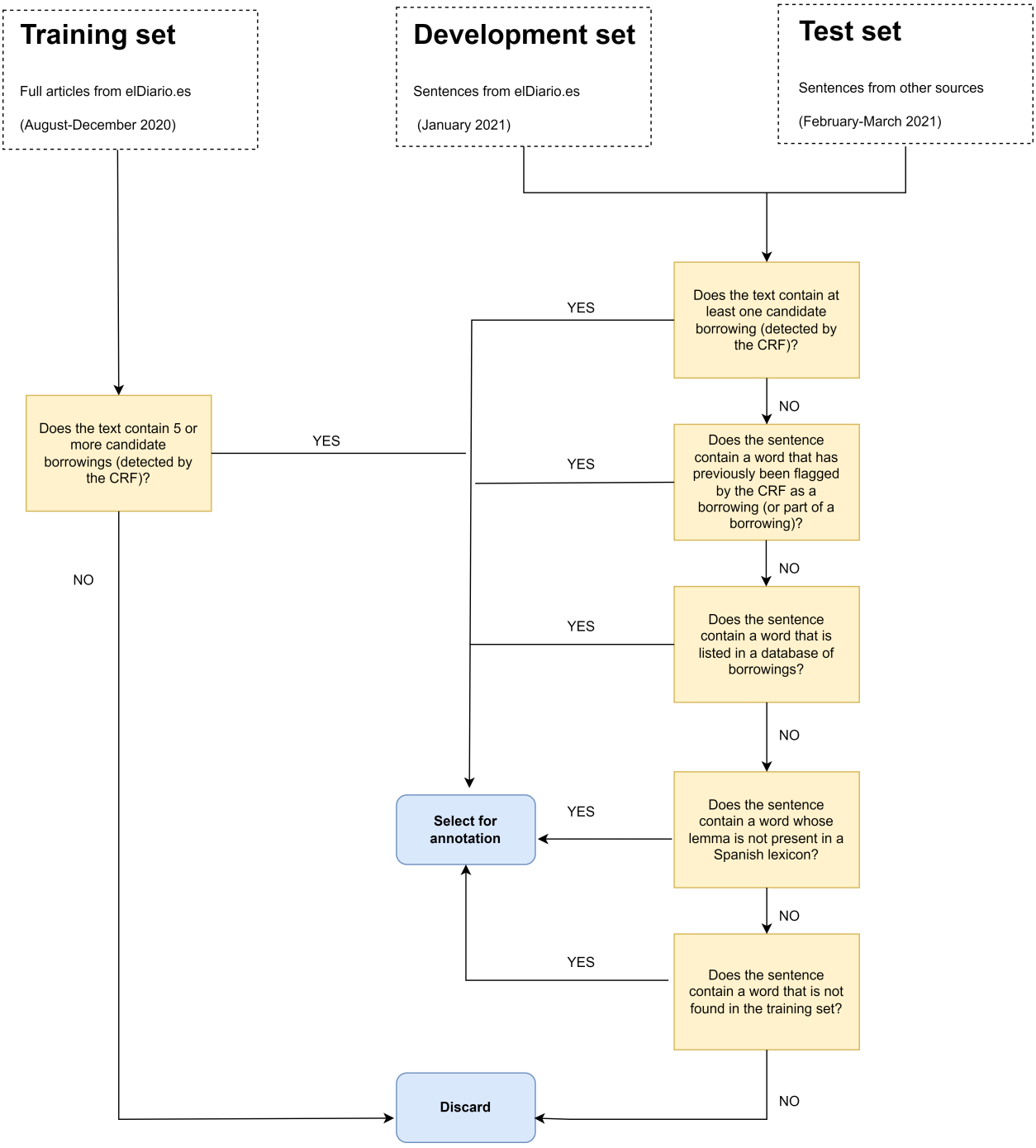

Figure 3.1: Data selection flowchart for COALAS.

Media	Topics	URL	License	Set(s)
ElDiario.es	General newspaper	https://www.eldiario.es/	CC BY-NC 4.0	Train, Dev.
El orden mundial	Politics	https://elordenmundial.com/	CC BY-NC-ND	Test
Cuarto poder	Politics	https://www.cuartopoder.es/	CC BY-NC 3.0	Test
El salto	Politics	https://www.elsaltodiario.com/	CC BY-SA 3.0	Test
La Marea	Politics	https://www.lamarea.com/	CC BY-SA 3.0	Test
Píkara Magazine	Feminism	https://www.pikaramagazine.com/	CC BY-NC-ND	Test
El blog salmón	Economy	https://www.elblogsalmon.com/	CC BY-NC 3.0	Test
Pop rosa	Gossip	https://www.poprosa.com/	CC BY-NC 3.0	Test
Vida extra	Videogames	https://www.vidaextra.com/	CC BY-NC 3.0	Test
Espinof	Cinema & TV	https://www.espinof.com/	CC BY-NC 3.0	Test
Xataka	Technology	https://www.xataka.com/	CC BY-NC 3.0	Test
Xataka Ciencia	Technology	https://www.xatakaciencia.com/	CC BY-NC 3.0	Test
Xataka Android	Technology	https://www.xatakandroid.com/	CC BY-NC 3.0	Test
Genbeta	Technology	https://www.genbeta.com/	CC BY-NC 3.0	Test
Microsiervos	Technology	https://www.microsiervos.com/	Used with written permission	Test
Agencia Sinc	Science	https://www.agenciasinc.es/	CC BY 4.0	Test
Diario del viajero	Travel	https://www.diariodelviajero.com/	CC BY-NC 3.0	Test
Bebe y más	Parenthood	https://www.bebesymas.com/	CC BY-NC 3.0	Test
Vitónica	Lifestyle & sports	https://www.vitonica.com/	CC BY-NC 3.0	Test
Foro atletismo	Sports	https://www.foroatletismo.com/	CC BY-NC 2.1	Test
Motor pasión	Automobiles	https://www.motorpasion.com/	CC BY-NC 3.0	Test

Table 3.5: Data sources per split included in COALAS.

Chapter 4

Model experimentation

All models are wrong, but some are useful.

George Box, statistician

Ce qui est simple est toujours faux.

Ce qui ne l'est pas est inutilisable.

Paul Valéry, poet

In chapter 3 we introduced COALAS, a new dataset of Spanish press annotated with unassimilated lexical borrowings following chunking-based BIO encoding. In this chapter, we will use COALAS to train and evaluate several models for the task of retrieving borrowings from Spanish text. We will first experiment with a collection of different supervised models and analyze their performance on the COALAS test set (section 4.1). We will then tentatively explore whether a large language model can be used for the task of retrieving lexical borrowings in a few-shot fashion (section 4.2).

4.1 Supervised models for lexical borrowing detection

In this section we describe the process of training and testing a collection of supervised models for the task of identifying lexical borrowings in Spanish using the COALAS test set.

We will first replicate the CRF model we introduced in Álvarez Mellado (2020a) (which at the time was the available SOTA model for borrowing identification in Spanish) and assess how well it performs on the COALAS test set (section 4.1.1).

We will then train and evaluate more sophisticated models: two Transformer-based models (section 4.1.2), and two BiLSTM-CRF models with different types of word embeddings and subword embeddings combinations (sections 4.1.3 and 4.1.4).

All the scores reported on this chapter are calculated at span level and require extracted spans to match the annotation exactly in terms of boundaries and type to be correct. Evaluation was performed with `SeqScore`[1] (Palen-Michel et al., 2021), using `conlleval`-style repair for invalid label sequences.

4.1.1 Conditional random field model

As we saw in section 2.3.3, in previous work we had already introduced a sequence-labeling supervised model for the task of automatically retrieving lexical borrowings from Spanish: a CRF model fed with `word2vec` embeddings and linguistic handcrafted features (Álvarez Mellado, 2020a).

As baseline model for COALAS dataset, we replicated the CRF model with handcrafted features from Álvarez Mellado (2020a). Given that COALAS was designed to be richer, more borrowing-dense and more varied than the previous available dataset, we expected that the CRF model (which yielded an overall F1 score of 87% on the test set from the 2020a dataset) would be revealed as insufficient when confronted with the COALAS test set.

The CRF model was built using `pycrfsuite`[2] (Peng and Korobov, 2014), a Python wrapper for `crfsuite` (Okazaki, 2007) that implements CRF for labeling sequential data. The model also uses the `Token` and `Span` utilities from `spaCy`[3] (Honnibal and Montani, 2017).

The same handcrafted binary features from Álvarez Mellado (2020a) were used for the

[1] `https://github.com/bltlab/seqscore`
[2] `https://github.com/scrapinghub/python-crfsuite`
[3] `https://spacy.io/api`

model, namely:

- Bias: active on all tokens to set per-class bias.
- Token: the string of the token.
- Uppercase: active if the token is all uppercase.
- Titlecase: active if only the first character of the token is capitalized.
- Character trigram: an active feature for every trigram contained in the token.
- Quotation: active if the token is any type of quotation mark (‘ ’ ” “ ” ≪ ≫).
- Suffix: last three characters of the token.
- POS tag: part-of-speech tag of the token provided by `spaCy` utilities.
- Word shape: shape representation of the token provided by `spaCy` utilities.
- Word embedding: provided by Spanish `word2vec` 300-dimensional embeddings by Cardellino (2019), one feature per dimension.
- URL: active if the token could be validated as a URL according to `spaCy` utilities.
- Email: active if the token could be validated as an email address by `spaCy` utilities.
- Twitter: active if the token could be validated as a possible Twitter special token: `#hashtag` or `@username`.

A window of two tokens in each direction was used for feature extraction. Optimization was performed using L-BFGS, with the following hyperparameter values chosen following the best results from Álvarez Mellado (2020a): c1 =0.05, c2 =0.01.

As shown in table 4.1, the CRF produced an overall F1 score of 66.15% on the development set (P=74.13, R=59.72) and an overall F1 of 55.44% (P=77.89, R=43.04) on the test set. The CRF results on COALAS are far below the F1 of 86.41% reported on Álvarez Mellado (2020a), showing the impact that a topically-diverse, OOV-rich dataset can have, especially on test set recall: the model's recall was already modest (just below 60%) when tested on the development set, which was similar in topic coverage to the training set. But when tested on the test set (which included topics not seen during training), the

Set	Precision	Recall	F1
Development			
ALL	74.13	59.72	66.15
ENG	74.20	68.63	71.31
OTHER	66.67	4.08	7.69
Test			
ALL	77.89	43.04	55.44
ENG	78.09	44.31	56.54
OTHER	57.14	8.70	15.09

Table 4.1: Results of the CRF model on the COALAS development set and test set (results from a single run).

	Development			**Test**		
	Precision	Recall	F1	Precision	Recall	F1
BETO						
ALL	73.36 ± 3.6	73.46 ± 1.6	73.35 ± 1.5	86.76 ± 1.3	75.50 ± 2.8	80.71 ± 1.5
ENG	74.30 ± 1.8	84.05 ± 1.6	78.81 ± 1.6	87.33 ± 2.9	77.99 ± 1.5	82.36 ± 1.5
OTHER	47.24 ± 22.2	7.34 ± 3.7	11.93 ± 4.9	36.12 ± 10.1	8.48 ± 3.5	13.23 ± 3.6
mBERT						
ALL	**79.96** ± 1.9	**73.86** ± 2.7	**76.76** ± 2.0	**88.89** ± 1.0	**76.16** ± 1.6	**82.02** ± 1.0
ENG	80.25 ± 2.6	84.31 ± 1.9	82.21 ± 1.9	89.25 ± 1.6	78.85 ± 1.0	83.64 ± 1.0
OTHER	66.18 ± 16.5	8.6 ± 6.8	14.41 ± 10.0	45.30 ± 11.3	7.61 ± 1.5	12.84 ± 2.3

Table 4.2: Results of BETO and mBERT on the COALAS development set and test set (average scores and standard deviation reported over 10 runs with different random seeds for initialization).

model's recall plummeted to the forties. These results demonstrate that we have created a more difficult task than the one we proposed in 2020a and motivate using more sophisticated models.

4.1.2 Transformer-based models: BETO and mBERT

In light of the unsatisfying results obtained by the CRF model in section 4.1.1, we decided to explore more sophisticated models for the task of retrieving lexical borrowings from COALAS. We evaluated two Transformer-based models:

– BETO base cased model: a monolingual BERT model trained for Spanish (Cañete et al., 2020)

– mBERT: multilingual BERT, trained on Wikipedia in 104 languages (Devlin et al., 2019)

Both models were run using the `Transformers`[4] library by `HuggingFace` (Wolf et al., 2020). The same default hyperparameters were used for both models: 3 epochs, batch size 16, and maximum sequence length 256. Except where otherwise specified, we report average results for 10 runs that use different random seeds for initialization. We perform statistical significance testing using the Wilcoxon rank-sum test (Dror et al., 2018).

As shown in table 4.2, the mBERT model (F1=82.02) performed better than BETO (F1=80.71), and the difference was statistically significant ($p = 0.027$). The limited number of `OTHER` examples explains the high deviations in the results for this label. Both models performed better on the test set than on the development set, despite the difference in topics between them, suggesting good generalization. This is a remarkable difference with the CRF results, where the CRF performed substantially worse on the test set than on the development set.

4.1.3 BiLSTM-CRF with unadapted embeddings

We also explored several possibilities for a BiLSTM-CRF model fed with different types of Transformer-based unadapted embeddings and subword embeddings. By *unadapted* embeddings, we mean embeddings that have not been fine-tuned for the task of anglicism detection or any other related task (e.g. codeswitching). The purpose was to assess whether the combination of different embeddings that encode different linguistic information could outperform the Transformer-based models from section 4.1.2. All of our BiLSTM-CRF models were built using `Flair`[5] (Akbik et al., 2018) with default hyperparameters (hidden size=256, learning rate=0.1, mini batch size=32, max number of epochs=150) and embeddings provided by `Flair`.

[4] `https://github.com/huggingface/transformers`
[5] `https://github.com/flairNLP/flair`

Exploratory embedding combinations

We first ran exploratory experiments on the development set of COALAS with different types of embeddings using `Flair` tuning functionalities. We explored the following embeddings: Transformer embeddings (mBERT and BETO), fastText embeddings (Bojanowski et al., 2017), one-hot embeddings, byte pair embeddings (Heinzerling and Strube, 2018), and character embeddings (Lample et al., 2016).

The results obtained on the development set during these exploratory experiments ranged from 39% to 74% in F1 score. The best results were obtained by a combination of mBERT embeddings and character embeddings (F1=74.00), followed by a combination of BETO embeddings and character embeddings (F1=72.09). These results show that using contextualized embeddings unsurprisingly outperforms non-contextualized embeddings for this task, and that subword representation is important for the task of extracting borrowings that have not been adapted orthographically. The fact that subword representation contributes to performance is consistent with previous work: feature ablation experiments for borrowing detection have shown that character trigram features contributed the most to the results obtained by a CRF model (Álvarez Mellado, 2020a).

The worst result (F1=39.21) was produced by a model fed with one-hot vectors, and the second-worst result was produced by a model fed exclusively with character embeddings. While it performed poorly (F1=41.65), this fully unlexicalized model based exclusively on character representation outperformed one-hot embeddings, reinforcing the importance of subword information for the task of unassimilated borrowing extraction.

Optimal embedding combinations

In light of the exploratory embedding experiments and our earlier experiments with Transformer-based models, we fed our BiLSTM-CRF model with different combinations of contextualized word embeddings (including English BERT embeddings from Devlin et al.), byte-pair embeddings (BPE) and character embeddings. Table 4.3 shows development set

Word embedding	BPE embedding	Char embedding	Precision	Recall	F1
mBERT	-	-	82.27	69.30	75.23
mBERT	-	✓	79.45	72.96	76.06
mBERT	multi	✓	81.37	73.80	77.40
mBERT	es, en	-	83.07	74.65	78.64
mBERT	es, en	✓	80.83	77.18	78.96
BETO, BERT	-	✓	81.44	76.62	78.96
BETO, BERT	-	-	81.87	76.34	79.01
BETO, BERT	es, en	✓	85.94	**77.46**	81.48
BETO, BERT	es, en	-	**86.98**	77.18	**81.79**

Table 4.3: Results of a BiLSTM-CRF model on the COALAS development set with different word and subword embedding combinations (results on ALL labels from a single run).

results from different combinations of embeddings. The best overall F1 score on the development set was obtained by the combination of BETO embeddings, BERT embeddings and byte-pair embeddings. The model fed with BETO embeddings, BERT embeddings, byte-pair embeddings and character embeddings ranked second.

Several things stand out from the results in table 4.3. The BETO+BERT embedding combination consistently works better than mBERT embeddings, and BPE embeddings contribute to better results. Character embeddings, however, seem to produce little effect at first glance. Given the same model, adding character embeddings produced little changes in F1 or even slightly hurt the results. Although character embeddings seem to make little difference in overall F1, recall was consistently higher in models that included character embeddings, and in fact, the model with BETO+BERT embeddings, BPE embeddings and character embeddings produced the highest recall overall (77.46%). This is an interesting finding, as our results from sections 4.1.1 and 4.1.2 as well as prior work (Álvarez Mellado, 2020a) identified recall as weak for borrowing detection models.

The two best-performing embedding combinations from table 4.3 (BETO+BERT embeddings, BPE embeddings and optionally character embeddings) were evaluated on the test set. Table 4.4 gives results per type on the development and test sets for these two models. For both models, results on the test set were better (F1=82.92, F1=83.63) than on the devel-

Embeddings	Development			Test		
	Precision	Recall	F1	Precision	Recall	F1
BETO+BERT and BPE						
ALL	**85.84** ± 1.2	77.07 ± 0.8	**81.21** ± 0.5	**90.00** ± 0.8	76.89 ± 1.8	82.92 ± 1.1
ENG	86.15 ± 1.1	88.00 ± 0.6	87.05 ± 0.6	90.20 ± 1.9	79.36 ± 1.1	84.42 ± 1.1
OTHER	72.81 ± 13.7	8.8 ± 0.9	15.60 ± 1.6	62.68 ± 8.0	10.43 ± 0.9	17.83 ± 1.3
BETO+BERT, BPE, and char						
ALL	84.29 ± 1.0	**78.06** ± 0.9	81.05 ± 0.5	89.71 ± 0.9	**78.34** ± 1.1	**83.63** ± 0.7
ENG	84.54 ± 1.0	89.05 ± 0.5	86.73 ± 0.5	89.90 ± 1.1	80.88 ± 0.7	85.14 ± 0.7
OTHER	73.50 ± 11.4	9.38 ± 3.0	16.44 ± 4.8	61.14 ± 7.9	9.78 ± 1.8	16.81 ± 2.9

Table 4.4: Results of a BiLSTM-CRF model on COALAS development set and test set using different combinations of unadapted embeddings: BETO and BERT embeddings, BPE embeddings, and optionally character embeddings (average scores and standard deviation reported over 10 runs with different random seeds for initialization).

opment set (F1=81.21, F1=81.05), which highlights the generalization capabilities of deep learning models in contrast to the CRF. Although the best F1 score on the development set was obtained with no character embeddings, when run on the test set the model with character embeddings obtained the best score; however, these differences did not show to be statistically significant. Recall, on the other hand, was higher when run with character embeddings (R=78.34) than when run without them (R=76.89), and the difference was statistically significant ($p = 0.019$). This finding again corroborates the positive impact that character information may have in recall when dealing with previously unseen borrowings.

4.1.4 BiLSTM-CRF with codeswitch embeddings

As final experiment for the supervised approach, we decided to explore whether the task of detecting unassimilated lexical borrowings could be framed as transfer learning from language identification in codeswitching. As before, we ran a BiLSTM-CRF model using `Flair`, but instead of using the unadapted Transformer embeddings from section 4.1.3, we used codeswitch embeddings (Sarker, 2020), that is, fine-tuned Transformer-based embeddings that have been pretrained for language identification on the Spanish-English section of the LinCE dataset (Aguilar et al., 2020). Once more, we experimented by stacking different types of subword embeddings along with the codeswitch embeddings.

Embeddings	Development			Test		
	Precision	Recall	F1	Precision	Recall	F1
Codeswitch						
ALL	80.21 ± 1.7	74.42 ± 1.5	77.18 ± 0.5	90.05 ± 1.0	76.76 ± 3.0	82.83 ± 1.6
ENG	80.19 ± 1.6	85.59 ± 0.5	82.78 ± 0.5	90.05 ± 3.1	79.37 ± 1.6	84.33 ± 1.6
OTHER	85.83 ± 19.2	4.70 ± 2.3	8.78 ± 4.2	90.00 ± 12.9	6.52 ± 0.0	12.14 ± 0.1
Codeswitch + char						
ALL	81.02 ± 2.5	74.56 ± 1.5	77.62 ± 0.9	89.92 ± 0.7	77.34 ± 2.2	83.13 ± 1.2
ENG	81.00 ± 1.6	85.91 ± 0.9	83.34 ± 0.9	89.95 ± 2.3	80.00 ± 1.2	84.67 ± 1.2
OTHER	73.00 ± 41.6	3.67 ± 2.7	6.91 ± 4.9	68.50 ± 40.5	5.43 ± 2.9	9.97 ± 5.3
Codeswitch + BPE						
ALL	**83.62** ± 1.6	**75.91** ± 0.7	**79.57** ± 0.7	90.43 ± 0.7	78.55 ± 1.3	84.06 ± 0.7
ENG	83.54 ± 0.7	86.86 ± 0.8	85.16 ± 0.8	90.57 ± 1.3	81.14 ± 0.7	85.59 ± 0.7
OTHER	94.28 ± 12.0	7.55 ± 2.1	13.84 ± 3.6	67.17 ± 8.3	8.70 ± 1.4	15.30 ± 2.2
Codeswitch + BPE + char						
ALL	82.88 ± 1.8	75.70 ± 1.3	79.10 ± 0.9	**90.60** ± 0.7	**78.72** ± 2.5	**84.22** ± 1.6
ENG	82.90 ± 1.4	86.57 ± 1.0	84.66 ± 1.0	90.76 ± 2.6	81.32 ± 1.6	85.76 ± 1.6
OTHER	87.23 ± 14.5	7.75 ± 2.8	14.03 ± 4.7	66.50 ± 17.5	8.70 ± 2.3	15.13 ± 3.5

Table 4.5: Results of a BiLSTM-CRF model on the COALAS development set and test set using different combinations of codeswitch embeddings and subword embeddings (average scores and standard deviation reported over 10 runs with different random seeds for initialization).

Table 4.5 gives results for these models. The two best-performing models were the BiLSTM-CRF with codeswitch and BPE embeddings (F1=84.06) and the BiLSTM-CRF model with codeswitch, BPE and character embeddings (F1=84.22). The differences between these two models did not show to be statistically significant, but the difference with the best-performing model with unadapted embeddings from section 4.1.3 (F1=83.63) was statistically significant ($p = 0.018$).

Adding BPE embeddings showed to improve F1 score by around 1 point compared to either feeding the model with only codeswitch embeddings (F1=82.83) or only codeswitch and character embeddings (F1=83.13), and the differences were statistically significant in both cases ($p = 0.024$, $p = 0.018$). None of the models with codeswitch embeddings outperformed the best development set results obtained by the two best performing BiLSTM-CRF models with unadapted embeddings from section 4.1.3 (F1=81.21, F1=81.05), but they did outperform them when run on the test set.

The fact that borrowing detection can be productively framed as a transfer-learning task

from codeswitching identification may be unsurprising from the NLP perspective, as the codeswitch-based BiLSTM-CRF models benefit from the labeled data seen during pretraining for the language-identification task and thus the models are indirectly being exposed to more multilingual data than just the training data from our initial corpus. From the linguistics perspective, however, the fact that we can successfully approach borrowing detection as a transfer-learning task from codeswitching is interesting, because as we saw in section 1.1.3, previous work has made the case that when speakers access a code-mixed item they make the instantaneous decision of whether to treat it as a borrowing or a codeswitch and that therefore borrowing and codeswitching should be regarded as two independent phenomena (Poplack and Dion, 2012).

4.1.5 Results

Table 4.6 summarizes the results obtained by all supervised models over the development set and the test set of COALAS.

As we have seen, the diversity of topics and the presence of OOV words in the dataset can have a remarkable impact on results. The CRF model—which in previous work had reported an F1 score of 86%—saw its performance drop to a 55% when dealing with our dataset, despite the fact that both datasets (the current one and the previously available dataset from 2020a) consisted of journalistic European Spanish texts. This fact showcases the importance that data selection can have during the process of corpus creation.

On the other hand, neural models (Transformer-based models and BiLSTM-CRF models) performed better. All of them performed better on the test set than on the development set, which suggests good generalization ability. The BiLSTM-CRF model fed with different combinations of Transformer-based word embeddings and subword embeddings (either unadapted or codeswitch) outperformed multilingual BERT and Spanish monolingual BETO. The model fed with codeswitch, BPE, and character embeddings ranked first and was significantly better than the result obtained by the model fed with BETO+BERT, BPE, and

Model	Word emb	BPE emb	Char emb	Development			Test		
				Precision	Recall	F1	Precision	Recall	F1
CRF	w2v (es)	-	-	74.13	59.72	66.15	77.89	43.04	55.44
BETO	-	-	-	73.36	73.46	73.35	86.76	75.50	80.71
mBERT	-	-	-	79.96	73.86	76.76	88.89	76.16	82.02
BiLSTM-CRF	BETO+BERT	en, es	-	**85.84**	77.07	**81.21**	90.00	76.89	82.92
BiLSTM-CRF	BETO+BERT	en, es	✓	84.29	**78.06**	81.05	89.71	78.34	83.63
BiLSTM-CRF	Codeswitch	-	-	80.21	74.42	77.18	90.05	76.76	82.83
BiLSTM-CRF	Codeswitch	-	✓	81.02	74.56	77.62	89.92	77.34	83.13
BiLSTM-CRF	Codeswitch	en, es	-	83.62	75.91	79.57	90.43	78.55	84.06
BiLSTM-CRF	Codeswitch	en, es	✓	82.88	75.70	79.10	**90.60**	**78.72**	**84.22**

Table 4.6: Summary of results obtained by all supervised models (for ALL labels) on the COALAS development set and test set (for the CRF, results from a single run are reported; for all other models, average scores over 10 runs are reported).

character embeddings.

Recall was a weak point across all models: it was consistently lower than precision. In fact, the best recall score over the test set was 78.72%, which leaves room for improvement. Adding character embeddings showed to produce statistically significant improvements in recall, which opens the door to future work.

4.1.6 Error analysis

We divide the error analysis into two sections. We first analyze errors and successes made by each model (one model per model type), with the aim of gaining insight on what are the unique characteristics of each model in comparison with other models. We then identify which errors on the test set were common to the three best performing models, with the aim of identifying which instances of the dataset are more challenging for models.

Error analysis per model

In this section we analyze the errors made by each model. Table 4.7 displays the most frequent errors made per model as provided by SeqScore (Palen-Michel et al., 2021). With the aim of understanding what data points were handled uniquely well or badly by each model, we also report some of the unique answers (both mistakes and correct answers) made

by each model.

CRF The CRF model was the worst performing of all supervised models. The most prevalent error made by the CRF was that it consistently identified the prefix *post* as an anglicism when it was not (as in *post-guerra*, "post-war").

The CRF recall was particularly poor (R=43.04), and indeed it failed at retrieving even orthographically obvious borrowings such as *know how*, *speedrunner* or *smoothie*. The CRF also tended to mark as borrowings some fully Spanish (not borrowed) phrases that did not bear any apparent resemblance with a borrowing, such as *pre-atasco* or *pastilla abortiva*.

BETO BETO was surprisingly good at distinguishing between *post* as a borrowing and as a prefix: all occurrences of *post* in the COALAS test set were correctly labeled by BETO. This is remarkable, given that the ambiguity of the word *post* was a common source of errors for the other models.

On the other hand, the partial retrieval of multiword borrowings was an issue. This was particularly true for multiword borrowings where one of the words exists as a native word in Spanish (such as *red* in *red carpet* or *total* in *total look*).

mBERT Just like the CRF model, mBERT also had a tendency to over-identify *post* as an anglicism, even when it was used as the non-English prefix. As a result, false positives caused by the retrieval of the prefix *post* were the most frequent source of errors for mBERT.

In addition, there were 46 tokens that were incorrectly labeled as borrowings by the mBERT model, and that were not a source of errors for other models. These include foreign words used in reported speech or acronym expansion (21), proper names (11) and already assimilated borrowings (7).

On the other hand, there were 27 tokens that were correctly labeled only by the mBERT model. The mBERT model was particularly good at detecting the full span of multiword borrowings as in *no knead bread*, *total white*, *wide leg* or *kettlebell swings* (which were only

partially detected by other models) and at detecting borrowings that could pass for Spanish words (such as *fashionista*, *samples*, *vocoder*). The mBERT model also correctly labeled as `O` 12 tokens that the other two models mistook as borrowings, including morphologically adapted anglicisms, such as *craftear* (Spanish infinitive of the verb *to craft*), *crackear* (from *to crack*) or *lookazo* (augmentative of the noun *look*).

BiLSTM-CRF with unadapted embeddings There were 23 tokens that were incorrectly labeled as borrowings solely by this model, the most common types being assimilated borrowings (such as *fan*, *clon*) and Spanish words (*fiestones*) (9 each).

32 tokens were correctly labeled as borrowings only by this model. These include borrowings that could pass for Spanish words (*camel*, *canvas*). In addition, this model also correctly labeled as `O` 6 tokens that the other two mistook as borrowings, including borrowings that we can consider today as assimilated (such as *film* or *sake*) or the usage of *post* as a prefix of Latin origin (as in *post-producción*), which other models mistook with the English word *post*.

BiLSTM-CRF with codeswitch embeddings The codeswitch-based system incorrectly labeled 18 tokens as borrowings, including proper nouns (7), such as *Baby Spice*, and fully asimilated borrowings (5), such as *jersey*, *relax* or *tutorial*.

This model correctly labeled 27 tokens that were mistakenly ignored by other models, including multiword borrowings (*dark and gritty*, *red carpet*) and other borrowings that were non-compliant with Spanish orthographic rules but that were however ignored by other models (*messy*, *athleisure*, *multitouch*, *workaholic*).

The codeswitch-based model also correctly labeled as `O` 16 tokens that the other two models labeled as borrowings, including acronym expansions, lower-cased proper names and orthographically unorthodox Spanish words, such as the ideophone *tiki-taka* or *shavales* (a non-standard writing form of the word *chavales*, "guys").

Count	Error	Type	Span
11	FP	ENG	*post*
10	FP	ENG	*top*
6	FP	ENG	*tops*
6	FN	ENG	*oversize*
5	FP	ENG	*look*
5	FN	ENG	*total look*
5	FN	ENG	*cropped top*
5	FN	ENG	*rover*
4	FN	ENG	*cool*
4	FN	ENG	*midi*
4	FN	ENG	*blazer*
3	FP	ENG	*sports*

(a) CRF

Count	Error	Type	Span
5	FN	ENG	*total look*
5	FN	ENG	*top*
4	FP	ENG	*look*
4	FP	ENG	*top*
4	FN	ENG	*midi*
3	FP	ENG	*carpet*
3	FP	ENG	*tops*
3	FP	ENG	*top 3*
3	FN	ENG	*denim*
3	FN	ENG	*sexy*
3	FN	ENG	*red carpet*
3	FN	ENG	*spa*

(b) BETO

Count	Error	Type	Span
9	FP	ENG	*post*
6	FN	ENG	*top*
5	FN	ENG	*total look*
5	FN	ENG	*rover*
4	FP	ENG	*look*
4	FP	ENG	*top*
4	FN	ENG	*spa*
4	FN	ENG	*midi*
3	FP	ENG	*tops*
3	FN	ENG	*sitcom*
3	FN	ENG	*look*
3	FN	ENG	*curly*

(c) mBERT

Count	Error	Type	Span
5	FN	ENG	*total look*
4	FP	ENG	*look*
4	FP	ENG	*top*
4	FN	ENG	*spa*
4	FN	ENG	*midi*
3	FP	ENG	*carpet*
3	FP	ENG	*tops*
3	FN	ENG	*sexy*
3	FN	ENG	*red carpet*
3	FN	ENG	*rover*
3	FN	ENG	*top*
2	FP	ENG	*bell hooks*

(d) BiLSTM-CRF unadapted

Count	Error	Type	Span
5	FP	ENG	*post*
5	FP	ENG	*top*
5	FN	ENG	*total look*
5	FN	ENG	*rover*
4	FP	ENG	*look*
4	FN	ENG	*spa*
4	FN	ENG	*midi*
3	FN	ENG	*look*
3	FN	ENG	*sexy*
2	FP	ENG	*glamour*
2	FP	ENG	*bell hooks*
2	FP	ENG	*guay*

(e) BiLSTM-CRF codeswitch

Count	Error	Type	Span
16	FN	ENG	*look*
15	FN	ENG	*top*
8	FN	ENG	*apps*
8	FN	ENG	*looks*
7	FN	ENG	*post*
6	FP	ENG	*Android*
6	FN	ENG	*software*
5	FP	ENG	*Netflix*
5	FP	ENG	*Instagram*
5	FP	ENG	*YouTube*
5	FN	ENG	*total look*
5	FN	ENG	*cropped top*

(f) Llama3

Table 4.7: Most frequent errors made by each model on the COALAS test set. FN = false negative; FP = false positive.

Errors shared by the top-performing models

We now identify which instances in the test set were a source of error common to the three top-performing models: (1) the mBERT model, (2) the BiLSTM-CRF with BERT+BETO, BPE and character embeddings and (3) the BiLSTM-CRF model with codeswitch, BPE and character embeddings. The purpose is to identify which instances of the dataset are more challenging for models.

Borrowings labeled as `O` There were 137 tokens in the test set that were incorrectly labeled as `O` by all three models. 103 of these were of type `ENG`, 34 were of type `OTHER`. These errors can be classified as follows:

- Borrowings with capitalized initial (12), which tend to be mistaken by models with proper nouns:

  Example 3. *La mejilla en flor con el* ***Blush*** *[ENG] de la temporada.*[6]

  Example 4. *Una cerveza tipo* ***Ale*** *[ENG] o* ***Lager*** *[ENG].*[7]

  Example 5. *Análisis de empresa basados en* ***Big Data*** *[ENG].*[8]

- Borrowings in sentence-initial position (9), which were also capitalized and therefore consistently mislabeled as `O`:

  Example 6. ***Youtuber*** *[ENG], mujer y afroamericana: Candace Owen podría ser la alternativa a Trump.*[9]

  Example 7. ***Streamer*** *[ENG], locutor de deportes electrónicos en español y creador de contenido en Twitch.*[10]

  Example 8. ***Prowler push*** *[ENG], el ejercicio de fuerza con el que Elsa Pataky se prepara para su próxima película.*[11]

[6] "The blooming cheek with the blush of the season."
[7] "An ale or lager beer."
[8] "Business analytics based on Big Data."
[9] "Youtuber, woman and African-American: Candace Owen could be the alternative to Trump."
[10] "Streamer, e-sports commentator and content creator on Twitch."
[11] "Prowler push, the resistance training exercise that Elsa Pataky follows to prepare for her next film."

Sentence-initial borrowings are particularly tricky, as models tend to confuse these with foreign named entities. In fact, prior work on anglicism detection based on dictionary lookup (Serigos, 2017a) stated that borrowings in sentence-initial position were rare in Spanish and consequently chose to ignore all foreign words in sentence-initial position under the assumption that they could be considered named entities. However, these examples (and the difficulty they pose for models) prove that sentence-initial borrowings are not rare and therefore should not be overlooked.

- Borrowings that also happen to be words in Spanish (8), such as the word *primer*, that is a borrowing found in makeup articles (*un primer hidratante*, "a hydrating primer") but also happens to be a fully Spanish adjective meaning "first" (*primer premio*, "first prize"). Borrowings like these are still treated as fully unassimilated borrowings by speakers (and pronounced according to the English pronunciation, /ˈpɹaɪ.meɹ/), even when the form is exactly the same as an already-existing Spanish word. These were a common source of mislabeling, and also produced partial mismatches in multiword borrowings: *red* (which exists in Spanish meaning "net") in *red carpet*, *tractor* in *tractor pulling* or *total* in *total look*.

- Borrowings that could pass as Spanish words (58): most of the misslabeled borrowings were words that do not exist in Spanish but that could orthographically pass for a Spanish word. That is the case of words like *burpees* (hypothetically, a conjugated form of the non-existing verb *burpear*), *gimbal*, *mules*, *bromance* or *nude*.

- Other borrowings (50): a high number of mislabeled borrowings were borrowings that were orthographically implausible in Spanish, such as *trenchs*, *multipads*, *hypes*, *riff*, *scrunchie* or *mint*. The fact that none of our models were able to correctly classify these orthographically implausible examples leaves the door open to further exploration

of character-based models, investigating character-level perplexity as a source of information and a better assessment of the impact on retrieval that certain contexts may have.

Non-borrowings labeled as borrowings 29 tokens were incorrectly labeled as borrowings by all three models. These errors can be classified in the following groups:

- Metalinguistic usage and reported speech: a foreign word or sentence that appears in the text to refer to something someone said or wrote.

  Example 9. *Escribir "**icon pack**" [ENG] en el buscador.*[12]

- Lower-cased proper nouns: such as websites.

  Example 10. *Acceder a la página **flywithkarolg** [ENG]*[13]

  Example 11. *Prolífica escritora y activista feminista y académica **bell hooks** [ENG].*[14]

- Computer commands: the test set included blog posts about technology, which mentioned computer commands (such as *sudo apt-get update*) that were consistently mistaken by our models as borrowings. These may seem like an extreme case—after all, computer commands do contain English words—but they are a good example of the real data that a borrowing-detection system may encounter.

  Example 12. ***sudo apt - get update** [ENG]*

- Foreign words within proper nouns: lower-cased foreign words that were part of multi-word proper nouns.

  Example 13. *La serie "10.000 **ships** [ENG]" cuenta la odisea de la princesa Nymeria.*[15]

- Acronyms and acronym expansions:

  Example 14. *El entrenamiento HITT (**high intensity interval training** [ENG]).*[16]

[12] "Type 'icon pack' on the search box."
[13] "Access the website flywithkarolg."
[14] "Prolific writer and feminist activist bell hooks."
[15] "The series '10,000 ships' tells the story of princess Nymeria."
[16] "HITT training (High-intensity interval training)."

- Assimilated borrowings: certain borrowings that are already considered by RAE's dictionary as fully assimilated were labeled by all models as anglicisms.

  Example 15. *Labios rojos, a juego con el* ***top*** *[ENG].*[17]

Type confusion Three tokens of type `OTHER` were marked by all models as `ENG`. There were no `ENG` borrowings that were labeled as `OTHER` by all three models.

Example 16. *Había* ***buffet*** *[ENG] libre.*[18]

4.2 A large language model's attempt at lexical borrowing detection

In section 4.1, we exclusively focused on supervised approaches to lexical borrowing detection. In other words, our approach to lexical borrowing detection assumed the existence of annotated data that can be used to train and evaluate machine learning models (or the possibility to produce such data). This assumption may be realistic in some situations and for certain languages. But in many scenarios it is simply not feasible to collect and annotate the amount of data that machine learning models require. This is particularly true for projects where the motivation of the work is the study of language itself and resources may be limited (as in lexicographic work), and where the linguistic phenomenon we seek to study is particularly sparse (as in lexical borrowing).

To account for data-scarce scenarios, pure lexicon lookup systems based on ready available word lists (which require no annotated data) have been tested. However, as we exposed in chapter 2, approaches exclusively based on lexicon lookup disregard context, which prevents these systems from distinguishing between anglicisms and foreign proper names (*cash*) or homograph native words (*social media*, *primer*). As convenient as they may be, prior research has shown that lexicon lookup systems produce disappointing results: in Álvarez Mellado (2020a) we experimented with various lexicon lookup methods (such as checking whether the

[17] "Red lips, matching top."
[18] "There was a free buffet."

token is found in a lexicon of English words or whether its lemma is absent from a Spanish dictionary), with F1 scores ranging from 1.02% to 25.40%. Consequently, we still lack an approach to lexical borrowing detection that is less data-intensive than the supervised models we explored in previous chapters and that can serve even as a bootstrapping method to build systems when data is scarce.

To that end, in this section we propose to test whether generative large language models can work for the task of retrieving lexical borrowings from Spanish text. We do not intend to fine-tune any of the existing LLMs using the training material we already have, although we are aware that that recipe would most certainly yield better results. On the contrary, we want to limit ourselves to the few-shot approach, which mimics the situation that lexicographers and other end-users in low-resource scenarios could see themselves in.

Our objective is to assess if these recent LLMs can successfully be used for lexical borrowing detection when annotated data is not at our disposal and see how their results compare to the supervised paradigm we explored in section 4.1. This is a tentative experiment, and it should not be seen as an exhaustive exploration of the application of LLMs to lexical borrowing, but rather as a first proof of concept that may come in handy when data is scarce and that hopefully can open the door to future research.

4.2.1 Model and prompting methodology

Large language models have previously been applied to information extraction tasks such as NER, following various methods and approaches (Xu et al., 2024). Our purpose in this experiment is to provide an LLM with a set of sentences that contain anglicisms and see how well the model performs at identifying them with minimal prompting, in a few-shot or almost zero-shot fashion.

Our prompting methodology was inspired by the work of Ashok and Lipton (2023), which offers prompted-based heuristics to apply to generative LLMs for the task of NER. Their experiments show that in addition to the few-shot examples, providing the model with a

```
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"
```

Figure 4.1: Prompt provided to Llama3 inspired by Ashok and Lipton's template: Definition of the task with examples, question and task, response by Llama3.

modular definition of the relevant entity types in the prompt produces better results than pure few-shot prompting.

To that end, we prompted the model with a definition of what an anglicism is and a description of the task the model was expected to complete. The definition included a list of four examples of decontextualized anglicisms (*online*, *machine learning*, *podcast*, *blockchain*) and a description of some of the characteristics and contexts an anglicism may appear in (single-word, multiword, adjacent, quoted or unquoted) illustrated by a minimal collection of the following three examples:

Example 17. *El evento se retransmitirá por streaming.*[19]

Example 18. *La marca ha sacado a la venta un nuevo concept car.* [20]

Example 19. *Compré mi smartphone online.*[21]

The definition explicitly mentioned that proper names such as entity names should not be considered anglicisms.

Figure 4.1 displays the prompt we used, which draws from the style of prompting for NER proposed by Ashok and Lipton.

The model of choice for our experiments was Llama3[22] (Grattafiori et al., 2024; AI@Meta, 2024). The reasoning behind this choice was that Llama3 was at the time of conducting our experiments among the best performing models while being open source, which ensures that our conclusions could be tested, replicated and easily transferred to other languages and scenarios. The model we used was the small Llama3 (8B parameter), as it is less resource intensive than larger versions. We queried it through `ollama` library[23], with default context window size of 2048 and temperature set to 0.0.

4.2.2 Results

The results obtained by Llama3 on COALAS test set were very disappointing: it produced a precision of 49.17%, a recall of 32.14% and an overall F1 score of 38.87%. These results are well below the worst performing model from chapter 4 (P=77.89, R=43.03, F1=55.44), which were obtained by the CRF.

As we did when reporting scores in section 4.1, results were calculated using `SeqScore` (Palen-Michel et al., 2021), so that the results would be comparable to the scores obtained by the supervised models. This means that the output produced by Llama3 had to be converted

[19] "The event will be broadcast live via streaming".
[20] "The company has released a new concept car".
[21] "I bought my smartphone online".
[22] `https://github.com/meta-llama/llama-models/blob/main/models/llama3/MODEL_CARD.md`
[23] `https://ollama.com/`

Count	Error	Tokens	Count	Error	Tokens
16	FN	*look*	5	FN	*cropped top*
15	FN	*top*	5	FN	*rover*
8	FN	*apps*	4	FP	*Xbox*
8	FN	*looks*	4	FP	*VPN*
7	FN	*post*	4	FP	*Amazon*
6	FP	*Android*	4	FP	*Google*
6	FN	*software*	4	FN	*gamers*
5	FP	*Netflix*	4	FN	*hardware*
5	FP	*Instagram*	4	FN	*midi*
5	FP	*YouTube*	4	FN	*oversize*
5	FN	*total look*	4	FN	*blazer*

Table 4.8: Most frequent errors made by Llama3 on COALAS test set. FN = false negative; FP = false positive.

from spans (as seen in the example from figure 4.1) to a CoNLL-formatted file that could be taken as input by `SeqScore`. However the conversion was not straightforward, as the spans returned by Llama3 did not always match exactly the span in the sentence: casing in the response gave by Llama3 was sometimes changed, the spelling was different (*cryptos*, instead of *criptos*), quotation marks were frequently returned as part of the response, etc. Consequently, the output required some manual post-processing in order to produce a valid output. We opted for a lenient approach and fixed those replies manually so that they would produce a valid CoNLL output. It should be noted that up to 7 times the span returned by Llama3 was fully hallucinated and not present at all in the given sentence. These 7 spurious spans were simply ignored when converting the output to CoNLL format (as there was no easy way of including them) and are therefore not computed to calculate scores. This means that the precision was in fact *worse* than the scores we report, as there are 7 hallucinated spans that were ignored but that could have been counted as false positives.

4.2.3 Error analysis

Table 4.8 displays the most frequent errors in the output produced by Llama3. The same table is also displayed on table 4.7 to facilitate comparison with the errors produced by the

supervised models.

Among the most frequent errors, some surprises stand out: *look* is the most frequent missed span, which was missed 16 times. This means that Llama3 failed to retrieved almost all the occurrences of *look* in COALAS test set, which appeared 17 times in the test set. This is remarkable, given that even a simple system based on dictionary lookup, pattern-matching or character probability would easily succeed at retrieving *look*. It should be noted that although *look* can also be found among the top recurrent errors in other models (see the CRF, BETO and the BiLSMT-CRF with unadapted embeddings in table 4.7), for the supervised models *look* was always a false positive, which means that the error was caused by the imperfect retrieval of a multiword borrowing such as *total look*. In the case of Llama3, *look* was a false negative, which means that the error was caused because it was fully missed, not because it was partially retrieved.

Llama3 also failed to retrieve all occurrences of the span *top* (15 occurences), *apps* (8 occurrences) and *looks* (8 occurrences).

Llama3 performance was also somewhat erratic, with the very same anglicism being identified or missed depending on the context it appeared in (*online*).

False positives were also pervasive: the most frequent cases were foreign proper names (such as *Android*, *Netflix*, *YouTube*, *Instagram*), even when the prompt we provided Llama3 explicitly mentioned that proper names should not be considered a borrowing.

Llama3 tended to label as a borrowing any span written between quotation marks. While this strategy earned Llama3 some true positives (*balance of performance*, *script doctors*, *feel-good movie*) it was also the cause behind some of the false positives (*cero emisiones*[24], *póntelo-pónselo*[25]). This tendency of Llama3 to identify as a borrowing spans of text between quotation marks is surprising, as none of the examples we prompted Llama3 with was written with quotation marks.

It is worth mentioning that the prompt gave no instructions whatsoever on how to proceed

[24] "Zero emissions"

[25] "Put it on, put it on him", a famous slogan from the 90s to promote condom use.

Model	Precision	Recall	F1
BiLSTM-CRF (codeswitch)	**90.60**	**78.72**	**84.22**
BiLSTM-CRF (unadapted)	89.71	78.34	83.63
mBERT	88.89	76.16	82.02
BETO	86.76	75.50	80.71
CRF	77.89	43.04	55.44
Llama3	49.17	32.14	38.87

Table 4.9: Final ranking of models based on the performance obtained on the COALAS test set.

with non-English borrowings (which were labeled as `OTHER` in the dataset). As a result, non-English borrowings were consistently ignored by Llama3 (or labeled as `ENG`). However, given the low number of `OTHER` spans in the COALAS test set, the results computed only on `ENG` spans are virtually the same as the scores reported overall (P=49.17, R=33.33, F1=39.73).

4.3 Discussion

Table 4.9 provides the ranking of the models explored on this chapter according to the results obtained on the COALAS test set over `ALL` labels. For both BiLSTM-CRFs, results reported in table 4.9 are those with BPE embeddings and character embeddings, which yielded the best scores (see tables 4.4 and 4.5).

Overall, two supervised models produced the most promising results: the BiLSTM-CRF fed with codeswitch embeddings, along with BPE and character embeddings (F1=84.22) and the BiLSTM-CRF model fed with unadapted embeddings (BETO and BERT embeddings), along with BPE and character embeddings (F1=83.63).

Transformer-based models (mBERT, BETO), while competitive, produced worse results than the aforementioned BiLSTM-CRF models (and the differences showed statistical significance).

The big disappointment in our results are the scores obtained by Llama3 (P=49.17, R=32.14, F1=38.87). Not only did Llama3 rank last both across precision and recall, but also precision results were remarkably poor, with a difference of almost 30 points from the

precision obtained by the CRF, which ranked second to last. Still, recall was lower than precision also for Llama3, and ten points below the CRF. Llama3's poor results were mostly caused by a tendency to over-identify non-borrowed words as borrowings (especially foreign proper nouns and spans between quotation marks) combined with an inconsistent recall performance, with the same anglicisms being retrieved or not depending on the context. Future work should assess if larger models or different styles of prompting (with more examples and detailed definitions) can produce better results.

Recall was a weak point for all models we examined, as it was consistently lower than precision for all models. This finding suggests the need to continue working towards better recall-oriented models, not only as a way of improving the F1 score in general, but also because users of term extraction systems as these may be more interested in minimizing false negative results (i.e. getting as many candidates as possible, even if some are wrong), even if it is at the cost of getting some false positives (which can always be manually filtered out).

Among false negatives, capitalized borrowings (such as *Big Data*) and borrowings in sentence-initial position were frequent, as they tend to be mistaken with named entities. This finding suggests that borrowings with capitalized initial should not be overlooked. Similarly, words that exist both in English and Spanish (like *total* or *red*) are not rare and were also a common source of error, especially for multiword borrowings. In terms of false positives, the most frequent source of mistakes were non-borrowed foreign words, such as proper names, metalinguistic usages or reported speech.

The supervised models introduced in section 4.1 have been publicly released and are available through `HuggingFace`'s modelhub[26], as well as via Python package `pylazaro`[27] (see chapter 11.2).

[26] `https://huggingface.co/models?other=arxiv%3A2203.16169`
[27] `https://pylazaro.readthedocs.io/en/latest/`

4.4 Summary

In this chapter we have explored various models for the task of extracting lexical borrowings from Spanish text. We have used the COALAS corpus we presented in chapter 3 to train and test several sequence labeling models.

The previous SOTA model (a CRF model that produced an F1 score of 86 on the previous dataset from Álvarez Mellado (2020a)) produced an F1 score of only 55 when evaluated on COALAS test set, which showcases the importance that data selection can have during the process of corpus creation. These results prove that COALAS is a more challenging dataset that allows for a more rigorous evaluation of borrowing detection models than prior resources.

We first experimented with different supervised sequence labeling models (CRF, BiLSTM-CRF, and Transformer-based models) and explored various embedding combinations (Transformer-based embeddings, character embeddings, BPE embeddings, fine-tuned embeddings). Results show that a BiLSTM-CRF model fed with Transformer-based embeddings pretrained on codeswitched data along subword embeddings produced the best results (F1 = 84.22, 84.06), followed by a combination of contextualized word embeddings and subword embeddings (F1=83.63). Our error analysis identified that certain common traits (like being capitalized, being homograph of a native Spanish word or appearing in sentence initial position) hurt performance even for the best-performing models.

Second, we tentatively explored if LLMs could be used for the task of lexical borrowing detection in data-scarce scenarios where the supervised approach is not feasible. We queried Llama3 on the test set of COALAS following a few-shot approach and analyzed its results. The scores obtained by Llama3 on the COALAS test set were surprisingly low (F1=38.87), and they were even outperformed by the worst-performing supervised model.

In light of our experiments, we can conclude that supervised models clearly outperform Llama3 for the task of identifying lexical borrowing in Spanish, at least on COALAS test set.

In part III we will go back to the results we have presented in this chapter: we will critically re-examine the evaluation approach we have followed and investigate whether the performance of these models can be predicted based on the formal properties of text.

Chapter 5

ADoBo: A shared task on Automatic Detection of Borrowings

Whosoever is delighted in solitude, is either a wild beast or a god.

Aristotle, philosopher

The COALAS dataset we presented in chapter 3 was used to organize ADoBo[1], a shared task on automatic detection of borrowings that took place within IberLEF 2021 (Iberian Languages Evaluation Forum). In this chapter we describe the purpose and scope of the shared task, introduce the systems that participated in it, and share the results obtained during the competition.

5.1 Task description

The ADoBo shared task focused on the identification of unassimilated lexical borrowings, words from another language that are used in Spanish without orthographic modification and that have not (yet) been integrated into the recipient language—for example (*running*, *smartwatch*, *influencer*, *holding*, *look*, *hype*, *prime time* and *lawfare*).

[1] `https://adobo-task.github.io/2021.html`

5.1.1 Data

The COALAS corpus introduced in chapter 3 was the chosen dataset for the shared task. The annotated versions of the training set and the development set were distributed among participants. Participants were expected to develop borrowing identification systems based on that data. A few weeks later, an unannotated version of the test set was distributed among participants, who had to run their systems on the unannotated test set and submit the annotation produced by their system.

The data was distributed in CoNLL format. An additional collection of documents that was not evaluated (the background set) was released as a part of the test set to ensure that participating teams were not be able to easily perform manual examination of the evaluated part of the test set.

The following two limitations were established for participants during training:

1. No additional human annotation was allowed for training. Given that the main purpose of the shared task was to evaluate how different models perform for the task of borrowing detection, using external data annotated with borrowings would prevent a fair evaluation of different model approaches.
2. Although the usage of regular lexicons and linguistic resources was accepted, no automatically compiled lexicons of borrowings (such as those produced by already-existing models that perform borrowing extraction) were allowed. The reason for this limitation was that we were interested in evaluating how different approaches to borrowing detection performed when dealing with previously unseen borrowings, and models that piggyback on already-existing systems's output would prevent that.

5.1.2 Evaluation

The predicted outputs of the test set produced by the participating systems were evaluated and ranked according to their performance.

The evaluation metrics used for the task were the standard precision, recall and F1 score over spans. F1 score was used as the official evaluation metric to rank the participating teams. Evaluation was done exclusively at the span level. This means that only exact matches were considered, and no credit was given to partial matches. For example, given the multiword borrowing *late night*, the entire span would have to be correctly labeled in order to count as a true positive.

Scores were computed using `SeqScore`[2] (Palen-Michel et al., 2021), a Python package for evaluating sequence labeling tasks, configured to emulate the `conlleval` evaluation script (the same package we used to evaluate models from chapter 4). Along with precision, recall and F1 score, we report the predicted number of borrowings per system and the number of correct predictions.

5.1.3 Competition timeline

The following calendar was followed during the competition:

- February, 15 2021: Sample set, evaluation script and annotation guidelines released.
- March, 15 2021: Training set released.
- April, 5 2021: Development set released.
- May, 10 2021: Unannotated test set released.
- May, 17 2021: Systems output submissions.
- May, 20 2021: Results posted.
- September, 21 2021: ADoBo overview session at IberLEF 2021.

5.2 Participants' submissions

We received nine submissions from four different teams. However, only two teams submitted system description papers. As a result, we have no description whatsoever for two of the

[2] `https://github.com/bltlab/seqscore`

Team	System	Type	P	R	F1	Pred.	Corr.
Marrouviere	(1)	ALL	88.81	81.56	85.03	1,180	1,048
		ENG	90.70	82.65	86.49	1,129	1,024
		OTHER	47.06	52.17	49.48	51	24
Versae	(2)	ALL	88.77	81.17	84.80	1,175	1,043
		ENG	90.31	82.73	86.35	1,135	1,025
		OTHER	45.00	39.13	41.86	40	18
Marrouviere	(3)	ALL	89.40	66.30	76.14	953	852
		ENG	90.98	67.55	77.54	920	837
		OTHER	45.45	32.61	37.97	33	15
Marrouviere	(4)	ALL	92.28	61.40	73.74	855	789
		ENG	93.43	63.12	75.34	837	782
		OTHER	38.89	15.22	21.88	18	7
Versae	(5)	ALL	62.76	46.30	53.29	948	595
		ENG	62.97	47.62	54.23	937	590
		OTHER	45.45	10.87	17.54	11	5
Mgrafu	(6)	ALL	65.15	37.82	47.86	746	486
		ENG	65.31	38.90	48.76	738	482
		OTHER	50.0	8.69	14.81	8	4
BERT4EVER	(7)	ALL	75.27	27.47	40.25	469	353
		ENG	75.43	28.25	41.10	464	350
		OTHER	60.00	6.52	11.76	5	3
BERT4EVER	(8)	ALL	76.29	25.29	37.99	426	325
		ENG	76.48	25.99	38.80	421	322
		OTHER	60.00	6.52	11.76	5	3
BERT4EVER	(9)	ALL	76.44	24.75	37.39	416	318
		ENG	76.64	25.42	38.18	411	315
		OTHER	60.00	6.52	11.76	5	3

Table 5.1: Participants' results at ADoBo on the standard version of COALAS test set.

participating systems, including the one that obtained the best results. We provide a brief summary of the two participating systems for which we received a submission, and refer the reader to their respective task description papers for further details.

5.2.1 A CRF model with data augmentation

The BERT4EVER team (Jiang et al., 2021) submitted a system based on combining several CRF models trained on different portions of the task's training data. The models were used to label a freely-available open corpus in Spanish, and individual models were then re-trained on the output. Results suggest that this strategy improves two F1 points on the

Team	System	Type	P	R	F1	Pred.	Corr.
Marrouviere	(1)	ALL	73.66	82.49	77.83	1,439	1,060
		ENG	76.31	83.45	79.72	1,355	1,034
		OTHER	30.95	56.52	40.00	84	26
Marrouviere	(4)	ALL	81.49	63.04	71.08	994	810
		ENG	82.70	64.81	72.67	971	803
		OTHER	30.43	15.22	20.29	23	7
Marrouviere	(3)	ALL	72.66	67.63	70.05	1,196	869
		ENG	75.49	68.85	72.01	1,130	853
		OTHER	24.24	34.78	28.57	66	16
Versae	(2)	ALL	59.57	82.33	69.13	1,776	1,058
		ENG	61.34	84.02	70.91	1,697	1041
		OTHER	21.52	36.96	27.20	79	17
Versae	(5)	ALL	42.27	48.48	45.16	1,474	623
		ENG	42.37	49.72	45.75	1,454	616
		OTHER	35.00	15.22	21.21	20	7
Mgrafu	(6)	ALL	52.17	39.22	44.78	966	504
		ENG	52.30	40.36	45.56	956	500
		OTHER	40.00	8.69	14.29	10	4
BERT4EVER	(7)	ALL	70.29	28.72	40.77	525	369
		ENG	70.38	29.54	41.61	520	366
		OTHER	60.00	6.52	11.76	5	3
BERT4EVER	(8)	ALL	69.92	26.23	38.14	482	337
		ENG	70.02	26.96	38.93	477	334
		OTHER	60.00	6.52	11.76	5	3
BERT4EVER	(9)	ALL	70.49	25.84	37.81	471	332
		ENG	70.60	26.55	38.59	466	329
		OTHER	60.00	6.52	11.76	5	3

Table 5.2: Participants' results at ADoBo on the lowercase version of COALAS test set.

test set when compared to a trained-on-task-data-only baseline. The paper combines two well-known items in the machine learning toolbox, namely CRFs and data augmentation, and shows that bootstrapping an additional dataset is indeed useful.

5.2.2 Transformer-based models using STILTs

The Versae team (De la Rosa, 2021) submitted a system that experimented with using STILTs—supplementary training on intermediate labeled-data tasks (Phang et al., 2019). They experimented with training using part of speech, named entity recognition, code-switching, and language identification datasets, but found that models trained this way

Team	System	Type	P	R	F1	Pred.	Corr.
Marrouviere	(1)	ALL	90.35	82.33	86.16	1,171	1,058
		ENG	91.18	83.45	87.15	1,134	1,034
		OTHER	64.86	52.17	57.83	37	24
Versae	(2)	ALL	88.71	80.08	84.17	1,160	1,029
		ENG	90.19	81.60	85.68	1,121	1,011
		OTHER	46.15	39.13	42.35	39	18
Marrouviere	(3)	ALL	90.84	66.38	76.71	939	853
		ENG	91.09	67.64	77.63	920	838
		OTHER	78.95	32.61	46.15	19	15
Marrouviere	(4)	ALL	91.39	60.31	72.67	848	775
		ENG	92.75	61.99	74.31	828	768
		OTHER	35.00	15.22	21.21	20	7
Versae	(5)	ALL	62.76	46.30	53.29	948	595
		ENG	62.97	47.62	54.23	937	590
		OTHER	45.45	10.87	17.54	11	5
Mgrafu	(6)	ALL	66.81	36.50	47.21	702	469
		ENG	67.00	37.53	48.11	694	465
		OTHER	50.0	8.69	14.81	8	4
BERT4EVER	(7)	ALL	78.37	25.37	38.33	416	326
		ENG	78.40	26.07	39.13	412	323
		OTHER	75.00	6.52	12.00	4	3
BERT4EVER	(8)	ALL	79.03	22.88	35.49	372	294
		ENG	79.08	23.49	36.22	368	291
		OTHER	75.00	6.52	12.00	4	3
BERT4EVER	(9)	ALL	79.34	22.41	34.95	363	288
		ENG	79.39	23.00	35.67	359	285
		OTHER	75.00	6.52	12.00	4	3

Table 5.3: Participants' results at ADoBo on the unquoted version of COALAS test set.

consistently perform worse than fine-tuning multilingual language models. The Versae team also explored other multilingual Transformer-based models (without STILTs), namely multilingual BERT, RoBERTa, and models trained on small sets of languages.

Their best result was obtained by fine-tuning and running XLM-RoBERTa, which reportedly scored 84.97% on F1 score, followed by multilingual BERT (84.78%) (De la Rosa, 2021).

Team	System	Type	P	R	F1	Pred.	Corr.
Marrouviere	(1)	ALL	78.04	82.96	80.42	1,366	1066
		ENG	78.67	83.94	81.22	1,322	1040
		OTHER	59.09	56.52	57.78	44	26
Marrouviere	(3)	ALL	77.96	67.70	72.47	1,116	870
		ENG	78.28	68.93	73.30	1,091	854
		OTHER	64.00	34.78	45.07	25	16
Marrouviere	(4)	ALL	81.14	61.95	70.26	981	796
		ENG	82.36	63.68	71.83	958	789
		OTHER	30.43	15.22	20.29	23	7
Versae	(2)	ALL	60.07	81.48	69.15	1,743	1,047
		ENG	61.76	83.05	70.84	1,666	1,029
		OTHER	23.38	39.13	29.27	77	18
Versae	(5)	ALL	42.41	48.48	45.24	1,469	623
		ENG	42.48	49.72	45.82	1,450	616
		OTHER	36.84	15.22	21.54	19	7
Mgrafu	(6)	ALL	54.56	37.74	44.62	889	485
		ENG	54.60	38.82	45.38	881	481
		OTHER	50.0	8.69	14.81	8	4
BERT4EVER	(7)	ALL	72.96	26.46	38.83	466	340
		ENG	72.79	27.20	39.60	463	337
		OTHER	100	6.52	12.24	3	3
BERT4EVER	(8)	ALL	72.75	23.89	35.97	422	307
		ENG	72.55	24.54	36.67	419	304
		OTHER	100	6.52	12.24	3	3
BERT4EVER	(9)	ALL	73.17	23.35	35.40	410	300
		ENG	72.97	23.97	36.09	407	297
		OTHER	100	6.52	12.24	3	3

Table 5.4: Participants' results at ADoBo on the lowercase unquoted version of COALAS test set.

5.3 Results

In this section, we report the results obtained by participants' systems on the test set of ADoBo shared task. We first focus on the overall performance, we then report the results obtained on some experiments we performed on the systems' output.

5.3.1 Systems' performance

Scores are summarized in table 5.1. For each team, precision, recall and F1 score are provided, along with the predicted number of borrowings and the number of correct predictions.

F1 scores ranged from 37.29% to 85.03%, with the Marrouviere team scoring highest (F1=85.03, P=88.81 and R=81.56), close to the next-highest scores from the Versae team (F1=84.80, P=88.77 and R=81.17). These results indicate that, although the Transformer-based approaches followed by Versae team were among the top performing systems (positions #2 and #5 in the ranking), the solutions proposed by Marrouviere team outperformed them and produced an F1 score of 85.03%. Unfortunately, Marrouviere team did not submit a system description paper and consequently we cannot know what their solution was.

All the CRF models proposed by BERT4EVER ranked last (positions 7-9 in the ranking), well behind the Mgrafu team, which ranked #6.

5.3.2 Experiments with outputs

To get a better understanding of the systems that took part in the shared task, we performed some experiments on the output that was submitted by participants.

Combining outputs

In order to assess the complementarity of the submitted systems, we combined the different outputs. The combination consisted of the union of all detected terms. Since the number of systems is not very high, all combinations of systems were explored. In terms of F1 score, the best performing combination was (1), (2), and (4), with F1= 87.83, P= 87.83, and R= 89.26, a result that outperforms the scores obtained separately by each individual system.

Stress tests on punctuation and casing

In order to assess the robustness to casing and punctuation perturbations of the submitted systems, we performed a series of stress tests.

Three variations of the test set were included in the background set (the additional collection of documents released along with the test set):

1. A lowercase version, where all uppercase letters in the original test set were transformed to lowercase.
2. A no-quotation-mark version, where all quotation marks in the original test set (“ ” ‘ ’ « ») were removed.
3. A lowercase no-quotation-mark version, where all uppercase letters where transformed to lowercase AND all quotation marks were removed.

None of these versions was used to rank the systems but to observe the systems difference in performance on different textual characteristics. The rationale for these experiments was to assess how well systems performed if certain orthotypographic cues that usually appear along with borrowings (such as quotation marks) were removed. After all, a borrowing is still a borrowing regardless of whether it is written with or without quotation marks and it would be of little use to have a model that systematically labeled anything between quotation marks as a borrowing, or that only detected borrowings if they are written between quotation marks.

Similarly, many of the foreign words that appear in newswire are usually proper names, where the uppercase can serve as cue to distinguish them from borrowings. Given that speakers are capable of distinguishing borrowings from proper names in spoken language — where no case distinction exists—and that these cues are not present in other textual genres (e.g. social media), we were interested in assessing how well the models performed when no case cue was available.

Results for these experiments are presented in tables 5.2, 5.3 and 5.4. Focusing on the best two performing systems, we observe a drop of global F1 due to a consistent drop on precision not compensated with a slight increase of recall for the lowercase versions of the test set. In general, the drop in system (2) is more pronounced than in system (1), which causes its repositioning in the corresponding rankings. For the unquoted version of the test set, system (1) increases its F1 and system (2) decrements it slightly.

System (1) seems to be more robust to the casing and quotating perturbations we have

introduced, but as we have no information on the system was built, we can not attribute any of the differences to any characteristics of the systems.

5.4 Discussion

Participating systems of the ADoBO shared task obtained results that ranged from 37% to 85% in terms of F1 score (see table 5.1). The top-performing results were obtained by system (1) and they are in line with our best results from chapter 4 (an average F1 score of 84.22% over 10 runs).

The results obtained from ADoBo 2021 shared task offer several interesting insights.

Although data augmentation techniques such as the one proposed by BERT4EVER team may be a workable solution to improve performance, Transformer-based models (like the ones that Versae experimented with: mBERT, XLM-RoBERTa,) will still produce better results with less data.

Supplementary training on intermediate labeled-data tasks can sometimes be useful, but it did not show to be effective for Versae's experimentation. Transfer learning techniques from codeswitch identification tasks had however produced good results when used in the form of fine-tuned embeddings in chapter 4.

The stress tests we performed based on casing and punctuation transformations produced mixed results, but showed that models are indeed sensitive to these perturbations.

Finally, for all nine systems recall scores were lower than precision scores. This finding corroborate our observations from chapter 4 that recall is the Achilles' heel of lexical borrowing detection models.

In terms of the shared task itself, our goal with ADoBo was to raise awareness about a topic that, although highly relevant in the linguistics literature, has been mostly neglected within NLP. Although there are some already well-established shared tasks on mixed-language settings, they have focused exclusively on code-switched data on social media text (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018), which is close to borrowing but

different in scope and nature (see section 1.1.3), and no specific venue existed on borrowing detection in NLP so far. To the best of our knowledge, ADoBo 2021 was the first shared task specifically devoted to linguistic borrowing.

Although the participation for this first edition was modest (nine systems submitted from four different teams), the response was positive and it seems to indicate that there exists a moderate population within the community that is interested in borrowing as an NLP task. In fact, a post-task survey distributed among registered participants showed that 85% of respondents were interested in seeing future editions around this phenomenon, particularly on languages other than Spanish and including both semantic and diachronic borrowings.

5.5 Summary

In this chapter we have presented the results of the ADoBo 2021, a shared task on extracting lexical borrowings from Spanish newswire. The challenge proposed to participants was to submit systems to retrieve lexical borrowings from the COALAS dataset that we introduced in chapter 3. Participants results ranged from F1 scores of 37% to 85%, recall scores being consistently lower than precision. The scores obtained by the top-performing systems are in-line with our best results from chapter 4. The results from this shared task show that lexical borrowing detection is not a trivial task and it is indeed an open problem that requires further research.

Chapter 6

Annotating lexical borrowing vs codeswitching

Codeswitching, rather than representing debasement of linguistic skill, is actually a sensitive indicator of bilingual ability.
Shana Poplack, linguist

6.1 The elephant in the room: the codeswitching vs borrowing distinction

In chapters 3, 4 and 5 we presented a dataset and a suite of models for the task of automatically retrieving lexical borrowings (mainly anglicisms) from Spanish text. The resources we introduced, along with the annotation guidelines we developed, were specifically designed for the identification of lexical borrowings in monolingual texts. This is a valid approach, given that our aim was to devise models that can retrieve lexical borrowing from standard canonical texts, such as journalistic texts, which we expect to be monolingual. This approach, however, dodges the long-standing debate over the borrowing-codeswitching continuum that has loomed over Linguistics and language contact studies for decades (Poplack and Sankoff, 1984; Deuchar and Stammers, 2012; Poplack, 2012; Poplack and Dion, 2012; Treffers-Daller,

2023) (see chapter 1.1.3).

In this chapter we draw our attention to the elephant in the room: the borrowing-codeswitching distinction that we conveniently ignored when we designed our annotation guidelines in chapter 3. We propose additional strategies and annotation guidelines to distinguish between the two phenomena in social media texts and apply them to an already existing dataset of English-Spanish bilingual tweets. We reannotate the dataset following these guidelines and explore whether Transformer-based models can successfully distinguish between borrowing and codeswitching based on our reannotation.

6.2 Challenges in annotating codeswitching and borrowing

As we lay out in chapter 2, the identification of lexical borrowings in context has seldom been explored within NLP. Codeswitching, on the other hand, has received substantial attention from the NLP community over the last decade and, as a result, various shared tasks exist (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018), as well as a steady benchmark (Aguilar et al., 2020). Due to the nature of codeswitching, these efforts have primarily focused on social media datasets.

Prior work on creating datasets of codeswitching has framed the task of language identification for codeswitched utterances as a sequence labeling task, where every token receives a language identification tag (Maharjan et al., 2015, among others), in a similar fashion to how part-of-speech tagging is approached. For example, in a collection of English-Spanish codeswitched tweets, tokens in Spanish will be labeled with a language identification tag for Spanish, and tokens in English will be labeled with a language identification tag for English. Additional labels have been proposed for use in codeswitching datasets, such as *ambiguous* for words whose language is difficult to determine even in context, *other* for tokens in languages other than the main languages under study, *mixed* for intralexical codeswitching (words that combine morphemes from different languages), *NE* for named entities, and *none* for punctuation marks, emoji, Twitter mentions, etc.

This repertoire of labels has become the usual tagset in codeswitching shared tasks (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018). However, this tagset does not take into account that not all other-language inclusions are necessarily a codeswitch, nor do they establish the status that lexical borrowings should receive. After all, language mixing can happen in cases that are not necessarily codeswitching. Given an utterance that is mostly monolingual, should we assume that any token from another language is a codeswitch, or could it be something else? How should borrowings be considered in these annotation projects? In the prior work we examined, we were not able to identify explicit, published definitions or guidelines on what should constitute a codeswitch—or perhaps more crucially, what is *not* a codeswitch.

As we saw in chapter 1.1.3, lexical borrowing is a process that resembles codeswitching in that it conveys the inclusion of tokens from another language into otherwise monolingual contexts, but it is different in nature (Poplack and Dion, 2012; Poplack, 2017). Codeswitches are usually produced by multilingual individuals, and thus they comply with the grammatical structure of both languages *at the same time*. By definition, codeswitches are not integrated into the recipient language, unlike established borrowings.

On the other hand, lexical borrowings can be produced by monolingual speakers, and they are often accompanied by morphological and phonological modification. Lexical borrowings comply with the grammatical patterns of the recipient language and can eventually become assimilated into it until the perception of being an inclusion of "foreign" origin disappears (Lipski, 2005).

Foreign items that appear in contexts that are otherwise monolingual are usually called *lone other-language inclusions* (or LO-LIs) and are subject to great debate within the contact linguistics literature, as it is not always clear whether they should be considered intrasentential codeswitches or borrowings. Let's consider the following examples:

Example 20. *Intentando comprar online uno de los nuevos discos duros que sacó Samsung,*

pero qué lata tener que rellenar tanto formulario.[1]

Example 21. *I got it, but prefiero usar mi Dell para cosas sencillas.*[2]

Although some cases may be debatable, it seems unlikely that in a sentence such as the one in example 20 the word *online* could be considered a codeswitch. Our analysis of existing codeswitching corpora suggests that tokens like *online* have usually been annotated as codeswitches, although in the absence of formal annotation guidelines, it is unclear whether that was intentional or not. On the other hand, a sentence such as the one in example 21 is clearly a codeswitched utterance.

Given that lexical borrowing and codeswitching are two distinct linguistic phenomena, we believe that the annotation guidelines we proposed in chapter 3 for identifying lexical borrowings in monolingual texts are not well-suited to account for domains that are rich in codeswitching and where the prevalence of lexical borrowings is also high, such as social media messages (Luján García, 2017; Sanou, 2018; Stewart et al., 2021). Consequently, in this chapter we provide specific guidelines on how to deal with the codeswitching-borrowing distinction when annotating a dataset for codeswitching. We believe that this distinction should not be overlooked by the NLP community working on codeswitched data: if codeswitching is the phenomenon of interest, then having a collection of tweets that are rich in other-language inclusions is not sufficient. As Poplack and Dion state:

> "Distinguishing codeswitching and borrowing is the thorniest issue in the field of contact linguistics today."

However, previous work on codeswitching dataset design has rarely attempted to establish an explicit difference between intrasentential codeswitching—codeswitching that happens inside one sentence—and lexical borrowing. Previous work in Hindi by Bali et al. (2014) proposed a statistical approach based on frequency and linguistic typology to capture the distinction between borrowing and mixing, while Patro et al. (2017) proposed different

[1] "Trying to buy one of the new hard disks released by Samsung online, but what a pain it is to have to fill in so many forms."

[2] "I got it, but I'd rather use my Dell for simple things."

metrics to compute the likelihood of a word being borrowed or being codeswitched. While many codeswitching datasets include Spanish data, the codeswitching/borrowing distinction has not been explicitly explored, and to the best of our knowledge, no annotated dataset has made this distinction.

In fact, a glance at the most frequent codeswitches in an English-Spanish codeswitched Twitter dataset (Maharjan et al., 2015) reveals that social media abbreviations and well-established internet terms (such as *lol*) are among the most frequent codeswitches annotated. It is debatable whether these words are in fact true codeswitches, as they can be considered examples of the internet dialect that monolinguals also use rather than evidence of bilinguals switching between languages, as Maharjan et al. point out:

> "In the case of abbreviations, some of them such as "lol" and "lmao" have become social media lingo rather than abbreviations of English words and thus cross language barriers."

6.3 Dataset development

With the previously mentioned challenges in mind, we developed a dataset consisting of tweets annotated to reflect codeswitching between Spanish and English as well as borrowing of English into Spanish. This section describes our annotation guidelines and the process of creating this dataset. In section 6.3.1, we propose a set of guidelines to distinguish between codeswitching and borrowing that can be followed when annotating a corpus of English-Spanish codeswitched tweets. In section 6.3.2, we describe how we reannotated an existing codeswitching dataset to match our guidelines. Section 6.3.3 presents the counts of the resulting dataset and the most frequent tokens per label in the annotation. The stand-off annotation is released under a Creative Commons CC BY 4.0 license[3].

6.3.1 Annotation guidelines

Our annotation uses the following labels:

[3] `https://github.com/lirondos/borrowing-or-codeswitching`

- SPA: for tokens in Spanish

- ENG: for tokens in English

- OTH: for tokens in languages other than ENG or SPA

- BOR: for recent borrowings (in English or other languages)

- ENT: for named entities

- N: for punctuation marks, Twitter symbols (such as hashtags and mentions), URLs, etc.

In general, tokens in Spanish were annotated as SPA, and tokens in English were annotated as ENG. Internet acronyms (such as *lol*) were considered part of the internet jargon and annotated as Spanish whenever they occurred in monolingual Spanish contexts. Non-Spanish words that appeared in contexts where no other codeswitching was occurring and where there was a strong indication that the word was being borrowed (rather than codeswitched) were annotated as BOR.

Our goal with the borrowing tag was to address relatively recent borrowings into the Spanish language that might be spoken aloud in a normal conversational context, unlike the internet jargon referenced above.

The following were identified as borrowings, differentiating them from short codeswitches embedded in an otherwise monolingual segment:

1. English words related to Twitter terminology: such as *Follow Friday*, *tweet*, *follower*, etc.

2. Technology words: *server*, *hosting*, *user*, *post*, *blog*, *internet*, *template*, *app*, *online*, *chat*.

3. English words that are already registered in *Diccionario de la lengua española*[4] by Real Academia Española (Real Academia Española, 2020), the general dictionary of standard Spanish compiled by the Royal Spanish Academy, the main prescriptive institution on Spanish language.

[4] `https://dle.rae.es/`

4. English words that are already registered in *Diccionario de Americanismos*[5] by Asociación de Academias de la Lengua Española (Asociación de Academias de la Lengua Española, 2010), a specialized dictionary that covers the vocabulary spoken in American Spanish and that has a rich representation of well-established lexical borrowings from English used in Latin America.

5. English words that are the headword of an entry in Spanish Wikipedia, such as music styles, genres and other cultural things (*hip hop*, *whisky*).

6. Words that have English origin but were used following Spanish grammatical structure, such as noun-adjective word order (*mensajes offline*, *rating online*).

The intuition behind guidelines #3–#5 was that if an English word either appears in some of the main prescriptive dictionaries of Spanish (which tend to be conservative in their approach to language change) or as the headword of a Wikipedia entry, then that would demonstrate that that English word is already well established among monolinguals in at least part of the Spanish-speaking world and should therefore be considered a borrowing, not a a codeswitch.

The motivation for guideline #6 is that codeswitches by definition normally comply with the grammatical restrictions of both languages involved. On the other hand, when an item is borrowed from another language—as opposed to codeswitching into the other language—the borrowing will be produced with all the morphosyntactic requisites required by the recipient-language (Poplack and Dion, 2012). Consequently, when we encountered an English word that was used in an otherwise Spanish context and followed all the expectations of Spanish syntax, we annotated it as a borrowing, not a codeswitch.

The annotation guidelines we propose build from the previous annotation guidelines we designed to assist annotators when identifying English borrowings in Spanish monolingual texts (see chapter 3 and appendix A). However, those previous guidelines were designed ex-

[5] `https://www.asale.org/recursos/diccionarios/damer`

clusively to identify borrowings in Spanish newspapers. As a result, they focused exclusively on Spanish monolingual text from Spain and used sources whose Spain-centric criteria has been previously pointed out (Blanch, 1995; Fernández Gordillo, 2014). This makes those guidelines insufficient to identify well-established borrowings in a more geographically and linguistically diverse collection of texts as our dataset, such as datasets of social media texts rich in codeswithes. Consequently, our guidelines were adapted so that they could account for a more diverse representation of Spanish dialectal varieties than the previous guidelines from chapter 3 by considering additional lexicographic sources, such as Wikipedia or *Diccionario de Americanismos*. Our guidelines ensured that English words that are well-established borrowings in many parts of the Spanish speaking Americas (such as *man*,[6] *nice*[7] or *party*[8]) would not be mistaken for codeswitches.

We tried to create guidelines for distinguishing borrowing and codeswitching that can be applied somewhat objectively and in the most reliable fashion possible. As we have stated previously, the distinction is challenging; no annotation guidelines will be perfect, and there is not one universal set of distinctions that all researchers or speakers of the languages involved will agree on. However, we believe these guidelines can at the very least produce reliable and replicable annotation, and that is an important step forward in this area.

6.3.2 Data selection

To reduce the effort required to create our corpus and place it within existing work on codeswitching, we decided to select an existing corpus already annotated for codeswitching which we could reannotate to match our guidelines. We selected the codeswitching-dense corpus LM2013 (Lignos and Marcus, 2013) for several reasons.

First, all tweets in the original LM2013 corpus had been annotated by two annotators and the original decisions made by each annotator were readily available, allowing us to

[6] `https://www.rae.es/damer/man`
[7] `https://www.rae.es/damer/nice`
[8] `https://www.rae.es/damer/party`

identify and revisit any disagreements between the two annotators.

Second, the LM2013 annotators had been instructed to annotate the data to attempt to differentiate borrowing and codeswitching. While the annotation does not have a borrowing tag, the annotators were instructed not to annotate borrowed words as codeswitches. In the LM2013 annotation process, this was referred to as the "monolingual grandparent" rule: if it is unclear whether a word is a borrowing or codeswitch, if one's monolingual grandparent would recognize it as part of their language, do not annotate it as a codeswitch. While this was intended to draw upon the life experiences of many bilinguals in that they are able to identify which words are understood by monolingual speakers, because it referred to an older relative it may have unintentionally had the effect of causing the annotators to annotate more recent borrowings as codeswitches.[9]

Third, unlike some other codeswitch datasets, the data used by LM2013 is primarily Spanish, not English, which gives more opportunities to annotate borrowings into Spanish, especially those from English. The tweets were selected from the Spanish subset of Burger et al. (2011), which was designed to be a general Spanish Twitter dataset created using automatic language identification without any intent to study language mixing. However, that dataset includes a significant amount of Spanish-English codeswitched tweets, which makes it a useful source for codeswitching data. The subset of the dataset selected by LM2013 was chosen to contain codeswitched tweets at a higher rate than the rest of the corpus.

The LM2013 dataset was reannotated following the guidelines from section 6.3.1 by a native speaker of Spanish that had a background in linguistics and had previous experience in linguistic annotation. Our reannotation differs from the original annotation of LM2013 in several ways. We completed adjudication of all 9,500 tweets annotated by LM2013. Due to limited adjudicator availability, they had only completed adjudication of 7,018 tweets, with the remainder having unresolved disagreements. The LM2013 dataset assigned language

[9]Perhaps future work would be better served by a "monolingual cousin" rule instead.

tags to named entities—names were annotated Spanish or English based on the surrounding language, not the source language of the name—but we found that this annotation was unreliable and we removed it, making our data more consistent with other codeswitching datasets that do not make this distinction.

LM2013's original annotation also distinguished between two types of entities: works of art (e.g., book titles) and other named entities. We removed this distinction as it was inconsistently annotated and most subsequent annotation either uses standard named entity types (PER, LOC, ORG, etc.) or a general "named entity" tag (Maharjan et al., 2015) like our annotation.

Finally, we explicitly annotated borrowings as such. The original LM2013 annotation did not do so and the LM2013 annotators were instructed to not treat borrowings as codeswitches, but the annotators followed this advice inconsistently. While the annotators were not consistent, the fact that they were instructed to make this distinction was useful, as it resulted in inconsistencies that we could easily identify in adjudication.

As discussed in section 6.2, a major difference between our annotation and previous codeswitching corpora that we are aware of (including the original LM2013 annotation) is that we do not count "internet-speak" as a codeswitch into English, reflecting the fact that monolingual communities may use internet shorthand adapted from English even in purely monolingual non-English contexts.

In order to identify borrowing candidates that could fall under the definition established in section 6.3.1, we manually looked for frequently used English borrowings and checked for inter-annotator disagreements, as in:

Example 22. *Me mandas tu <u>mail</u> y lo envío.*[10]

We also automatically flagged inconsistencies in how certain tokens were sometimes labeled as SPA and sometimes annotated as ENG (such as *man*, *server* or *tweet*) across the whole corpus, even if annotators agreed on the individual annotations.

[10] "Send me your email address and I will sent it to you."

Tag	Tokens
SPA	134,110
N	39,280
ENT	15,373
ENG	6,819
BOR	2,857
OTH	267
Total	198,706

Table 6.1: Token counts by label on the LM2013 dataset.

These techniques proved to be successful and revealed a high number of tokens that were actually borrowings and not true codeswitches (see section 6.3.3).

The result of our reannotation process is a corpus that makes the distinction between borrowing and codeswitching and has a narrower definition of codeswitching than previous work, allowing for investigation of a finer-grained linguistic phenomenon.

6.3.3 Dataset description

The corpus consists of 9,500 tweets that were annotated following the guidelines established in section 6.3.1. Out of those 9,500 tweets, 2,017 contained at least one borrowing, 2,495 contained at least one codeswitch, and 403 contained both a borrowing and a codeswitch. In total there are 198,706 tokens in the corpus, and counts for each label are given in table 6.1. The majority of tokens were labeled as SPA (134,110), 15,373 tokens were labeled with ENT, 6,819 tokens were labeled with ENG, 2,857 tokens were labeled with BOR, and 267 tokens were labeled with OTH. The largest sources of tokens labeled OTH belong to tweets with codeswitches into languages that are neither Spanish nor English, typically Portuguese and Catalan.

Table 6.2 summarizes the most frequent tokens per label. As there are essentially zero purely-English tweets in this dataset, the counts of English tokens reflect the count of tokens of codeswitches into English. The most frequent tokens show a noticeable difference with the most frequent tokens from some other codeswitched data counts. While Maharjan et al.

Spanish	Count	English	Count	Borrowing	Count
de	6,175	the	119	blog	231
que	4,298	I	105	post	130
y	3,281	you	84	web	107
el	3,257	to	81	internet	106
a	3,178	my	69	followers	55
la	3,175	a	69	online	41
en	3,120	it	62	blogs	41
no	2,410	is	60	software	33
me	2,086	in	56	Internet	33
es	1,764	on	49	timeline	32

Table 6.2: Most frequent tokens in the LM2013 dataset for each label: Spanish, English, and borrowing

(2015) reported internet terms and abbreviations such as *lol*, *lmao* or *idk* among the most frequent codeswitched tokens, the most frequent English tokens in our dataset are exclusively function words distributed in an approximately Zipfian fashion. This matches what we would expect in a fully monolingual English corpus and is similar to the distribution that tokens labeled as Spanish follow.

This suggests that what we are actually seeing labeled as English are true English sub-utterances—true codeswitches—and not the well-established borrowings that we find under the most frequent BOR column, or the "internet-speak" that we did not annotate as either a borrowing or codeswitch. However, it should be noted that the comparison of the amount and complexity of codeswitching in the data with other datasets is difficult because a quantitative codeswitching measure like that of Gambäck and Das (2016) cannot be meaningfully applied across corpora with different annotation guidelines.

6.4 Model experimentation

We wanted to evaluate how well Transformer-based models perform in predicting the annotated labels of our dataset and see if our annotation could serve to train a model to make the codeswitching-borrowing distinction. The data was divided into a standard training,

development, and test set split in 80/10/10 proportions.

We evaluated four Transformer-based models on the dataset:

- mBERT: multilingual BERT, trained on Wikipedia in 104 languages (Devlin et al., 2019)
- BETO base cased model: a BERT-based model trained on Spanish (Cañete et al., 2020)
- RoBERTa BNE: a RoBERTa based model trained on Spanish (Gutiérrez-Fandiño et al., 2021)
- RoBERTa Twitter: a RoBERTa based model trained on English Twitter data (Barbieri et al., 2020)

The aim behind this selection of models was to assess how Transformer-based models that had been trained on different types of text would perform on codeswitched data.

The only multilingual model that we included was mBERT.

BETO is a BERT-based model trained on a diverse set of international Spanish texts from different origins, such as OpenSubtitles, Global Voices, and the United Nations (Cañete, 2019).

RoBERTa BNE is a RoBERTa-based model that has been trained exclusively on data crawled from .es websites—those using the top-level domain for Spain—by the National Library of Spain. This means that while BETO training data is smaller (3 billion tokens vs BNE's 135 billion tokens), the model has probably been exposed to a more varied representation of the different varieties of Spanish spoken around the globe, which includes the area of Spanish where codeswitching may be more prevalent.

Finally, we added a RoBERTa-based model trained on Twitter data under the assumption that an English monolingual model trained on social media text could potentially do better than other models trained on other genres.

All models were run using the Transformers library by `HuggingFace` (Wolf et al., 2020) with the same default (untuned) hyperparameters: 3 epochs, batch size 32, and a maximum

Model	Accuracy	Precision	Recall	F1
mBERT	96.88	**96.69**	**96.61**	**96.65**
BETO	**96.91**	**96.69**	96.60	96.64
RoBERTa-BNE	93.73	93.19	93.23	93.21
RoBERTa Twitter	93.39	92.82	92.86	92.84

Table 6.3: Results obtained by baseline models on the reannotated version of the LM2013 dataset (results from a single run).

sequence length of 256.[11]

6.5 Results

Table 6.3 shows that both mBERT and BETO were the best performing models and achieved similar scores, with an F1 of 96.6%. While we have bolded the highest number in each column, we have no reason to believe that any differences between the performance of mBERT and BETO are meaningful. Interestingly, the RoBERTa-based model trained on the crawled data by the National Library of Spain performed worse, despite being a larger model. This seems to suggest the importance that a diverse set of training material may have when modeling certain linguistic phenomena that are impacted by geographical and dialectal variation, such as codeswitching and borrowing.

While we cannot directly compare our scores to other results as they are the first results on this dataset, these results are similar to those obtained on other datasets. For example, in the LinCE benchmark baseline (Aguilar et al., 2020)[12], for the language identification task on Spanish-English codeswitching data the F1 scores were 94.16% with a BiLSTM, 98.12% with ELMo, and 98.53% with mBERT. The mBERT baseline F1 for our dataset is slightly lower at 96.65%. One possible explanation is that our task is slightly more difficult for the model as it must make the borrowing-codeswitching distinction, but further experimentation and tuning would be required to support that claim.

[11] `https://github.com/huggingface/transformers/`
[12] `https://ritual.uh.edu/lince/`

The token-level accuracy of approximately 96.9% for mBERT and BETO baseline matches the accuracy of 96.9% reported by Lignos and Marcus (2013) on their data with a heuristic-based tagging approach. However, their evaluation excluded any tokens with named entity tags and their data was not annotated for borrowing, so it is likely their approach would get lower accuracy if evaluated on our reannotation.

6.6 Discussion

In this chapter we have showed that, although the codeswitching-lexical borrowing distinction may be disregarded when annotating monolingual data such as the one from chapter 3, the distinction must be tackled when dealing with social media texts. To that end, we have introduced additional annotation guidelines that explicitly cover the frontier between both phenomena and applied them to an already existing dataset of Spanish tweets, rich in English codeswitches and lexical borrowings.

Although at the time of publishing this work (Álvarez Mellado and Lignos, 2022a), ours was the sole dataset to implement such distinction, since then an additional work has also implemented the codeswitching-borrowing distinction on a Guarani-Spanish dataset (Chiruzzo et al., 2023). However, in that work no annotation guidelines on how the distinction was made were provided. Consequently, up to the best of our knowledge, ours still remain as the sole guidelines that explicitly account for the codeswitching-borrowing distinction.

Our guidelines and annotation, however, should not be taken as bullet-proof. While our guidelines seek to establish a division as clearly as possible, the distinction between codeswitching and borrowing is fuzzy and is far from being solved in the Linguistics literature. Many gray areas exist between social media jargon, nonce borrowing, and true codeswitching.

In addition, our goal for creating this corpus was to study a narrowly-scoped slice of multilingual language usage on Twitter—how Spanish speakers codeswitch into English, how they use borrowings and how these two phenomena intertwine—with the cleanest and most carefully annotated dataset possible. This corpus is not designed to be an evenly-balanced

English-Spanish codeswitching dataset. It is primarily a Spanish dataset, mainly composed of Spanish tweets that may have an English codeswitch or a borrowing.

Another limitation of the dataset is that as the original data selection process used generic Spanish language identification—Burger et al. (2011) do not provide any details—we do not know the geographic distribution of the speakers in the dataset and what dialects of Spanish and English they used. While we do believe this is a useful dataset for studying codeswitching and borrowing, caution should be exercised before drawing any global conclusions about the usage and mixing of either language.

In terms of the models' results, our scores in line or above some of the best-performing token-level results obtained in early approaches to language identification in English-Spanish codeswitched tweets, such as 86% F1 score with a CRF model (Maharjan et al., 2015), 91% F1 score obtained with a logistic regression classifier (Shirvani et al., 2016), 94% with a SVM classifier (Bar and Dershowitz, 2014),

It should be noted, however, that none of this systems were based on neural models or Transformers and yet obtain similar or higher results than some of the Transformer-based models that we ran.

6.7 Summary

As a final task in the data and modeling part of this thesis, we decided to tackle the challenge that the codeswitching-borrowing distinction poses when annotating language-mixed data. Our work from chapters 3-5 disregarded that distinction, as we focused exclusively on unassimilated borrowings in monolingual journalistic text, where codeswitching is not expected. As a result, our guidelines from chapter 3 were unsuited for annotating codeswitched data, such as bilingual tweets.

To address that limitation, in this chapter we have proposed novel annotation guidelines that build from the guidelines we designed in chapter 3 but considers bilingual scenarios, dialectal variation among Spanish speakers and distinguishes lexical borrowing from true

codeswitching. This distinction has been previously pointed out as crucial in the contact linguistics literature, but, to the best of our knowledge, it has not been made in any of the previous codeswitching datasets.

We have applied our annotation guidelines to an already existing dataset of Spanish tweets that was rich in codeswitched data. Finally, we have experimented with different Transformer-based models for the task of language identification and compared results in our dataset to previous work on other codeswitching datasets.

Part III

Evaluation

Chapter 7

Evaluation approaches

Wissenschaft: Es ist nicht ihr Ziel, der unendlichen Weisheit eine Tür zu öffnen, sondern eine Grenze zu setzen dem unendlichen Irrtum.

Bertold Brecht, playwright

In chapter 4 we experimented with different types of models for the task of extracting unassimilated anglicisms from Spanish text. Our experiments led to a ranking of models ordered by obtained results, where some models *seemed* to perform better than others. However, the notion of a model being *better* than another one is not absolute. On the contrary, it is dependent on two aspects:

1. The way the evaluation was performed.
2. The data that the models were evaluated on.

The very same models tested on a different dataset or evaluated using a different metric or approach may result in a much different podium of best-performing models.

In part III we look critically at some of the evaluation decisions that guided our developments from chapter 4.

In this chapter we will be focusing on ways to frame evaluation for information extrac-

tion tasks, such as lexical borrowing detection: we will discuss some of the limitations of evaluating lexical borrowing detection systems based on span match; we will review some alternative evaluation techniques that have been used in other extractions tasks and see how they compare to span-based evaluation for evaluating lexical borrowing detection systems; finally, we will re-evaluate the results from chapter 4 following an alternative approach and see if this alternative evaluation produces a different ranking of models.

Although we focus on lexical borrowing detection systems, our analysis and critiques are common to other information extraction tasks and hopefully our observations can inform future work in other sequence labeling developments.

7.1 Evaluation scenarios and desiderata

In chapter 4 we used F1 score to evaluate and rank models for the task of retrieving unassimilated lexical borrowings from text. F1 score calculates the harmonic mean between precision and recall (Van Rijsbergen, 1979). F1 score has been a steady metric in NLP evaluation for decades (Christen et al., 2023). As it is a standard metric in NLP, F1 score over spans has been widely used as an evaluation metric in information extraction tasks. In fact, F1 score over spans is the norm in many popular NER benchmarks (Tjong Kim Sang and De Meulder, 2003; Wang et al., 2019b) and evaluation scripts (Nakayama, 2018; Palen-Michel et al., 2021). As a result, when we performed our model evaluation in chapter 4 we uncritically used F1 score over spans, without further consideration.

This way of proceeding, although standard, is misguided: there is no evaluation metric that can serve all purposes. Quite on the contrary, evaluation metrics measure how good a model is *at something*. Consequently, the evaluation metric of choice should be determined (or at least informed) by the use case scenario where the model under evaluation will be deployed.

Here we enumerate a non-exhaustive list of real-world scenarios where a lexical borrowing detection system may be in place:

1. A lexicographer interested in identifying established anglicisms in the press that are good candidates to be registered in a dictionary.

2. A researcher in stylometry or corpus linguistics that wants to quantify all the lexical borrowings in a given text or corpus.

3. A linguist monitoring the frequency and occurrences of a concrete borrowing over time.

4. A language regulator institution that wants to track novel anglicisms to promote native alternatives or spelling adaptations before the foreign word gets incorporated.

5. A text-to-speech system that applies a preprocessing step to identify foreign words that do not comply with the standard rules of pronunciation of the recipient language.

6. A corpus management software that, given a concrete borrowing, identifies sentences in a corpus that can illustrate its meaning and serve as good dictionary examples.

7. A writing app or grammar checker that assists journalists, writers, etc. and makes editing suggestions to optimize the clarity and correctness of a text and flags non-standard language.

Some of these tasks are more oriented towards the discovery of previously unseen borrowings (scenarios #1 and #4), while in others is the correct in-context identification of a known borrowing that matters (scenarios #3 and #6)[1].

Some common traits arise from these scenarios:

a. In most of these tasks, it will be preferable to ensure that the system can exhaustively identify all possible candidates, even if it is at the cost of getting more false positives. The paradigmatic scenario is the case of lexicographic work, where it will be preferable to have some noise (that can always be filtered out manually) rather than miss relevant candidates.

[1]The distinction between discovery-based tasks and identification-based tasks follows the distinction established in Constant et al. (2017) for multiword expressions.

b. On a similar note, in the scenarios we have envisioned the partial retrieval of a multiword borrowing will still be preferable than retrieving no borrowing whatsoever. This follows a similar logic to the previous point: in linguistic work (and lexicographic work in particular), partially retrieving a multiword candidate is more desirable than fully missing a true borrowing, because partial matches and false positives can always be filtered out manually.

c. Given that the retrieval of lexical borrowings will happen in context and that we are interested in retrieving lexical items that may be multiword, our evaluation should be able to distinguish between having one multiword borrowing that is composed of several words and two adjacent spans that happen to be collocated one next to the other. This requisite may be optional for scenario #5 (because boundary identification of adjacent spans may be irrelevant for a text-to-speech system) but is vital for lexical identification and when word order matters, such as in scenarios #1, #2, #3, #4, #6 and #7.

These observations can be translated into the following desiderata. Our evaluation approach should:

i. Prioritize recall over precision. In fact, as the results from chapters 4 and 5 showed, recall is a weak point for all borrowing detection systems. It still may not be the sole aspect we want to measure: an evaluation exclusively based on recall could end up favoring systems that retrieve *almost anything* as a borrowing, thus making the system impracticable in real world scenarios. But along with F1 score, recall should be central when evaluating these systems.

ii. Account for overlapping matches of multiword borrowings so that they count as partial matches, something that our evaluation from chapter 4 fully disregarded.

iii. Distinguish between a multiword lexical borrowing that is composed of several items and two adjacent single-word borrowings that happen to be collocated next to one another. This means that the approach we take should consider not only whether a given token is part of a borrowing or not but also span boundaries.

In the following sections, we will investigate whether different evaluation approaches satisfy these conditions.

7.2 Extraction task evaluation: approaches and limitations

In our research we have framed lexical borrowing detection as an extraction task where a span of variable length of adjacent tokens is labeled as belonging to a given class, in this case, as being a borrowing. This approach is similar to other extraction tasks, such as NER, MWE or chunking.

This is a special type of sequence labeling task, because in addition to the tokens in the sequence being classified based on their surrounding context (as it occurs in classical sequence labeling tasks, such as part-of-speech tagging), adjacent tokens in the sequence may constitute larger units to be extracted, which we call spans. This type of sequence labeling task has sometimes been been called span identification tasks (Papay et al., 2020), segmentation tasks (Esuli and Sebastiani, 2010), sequential segmentation tasks (Suzuki et al., 2006), chunk identification (Kudo and Matsumoto, 2000) or more generally chunking-based approaches or phrase-extraction tasks.

Span-identification tasks have traditionally be framed in terms of modeling as a token classification problem, where the task consists of assigning the correct label to each token in a sequence adding a prefix to each tag to denote span boundaries. This is precisely how our annotation from chapter 3 and our supervised models from chapter 4 framed the problem. In terms of evaluation, this presents a dilemma: in span-identification tasks, should we evaluate outputs based on spans (was a given span correctly retrieved?) or should our evaluation be based on token classification (was a given token in the sequence correctly classified?)? The answer to this question will determine what gets rewarded, what gets penalized and the blind spots that may result from our evaluation strategy.

In the following subsections we are going to analyze the limitations that both span-based approaches and token-based approaches produce when evaluating extraction models.

7.2.1 Problems with span-based evaluation

A span is a segment of text that is characterized by a starting point (character or token onset), an end point (character or token offset) and a label. Let's consider the following real example:

Example 23. *Sticky buns: los primos poco vistos de los rollitos de canela.*[23]

In example 23, we may represent the lexical borrowing in it (*Sticky buns*) as a span that starts at token 0, ends at token 2 and is labeled as `ENG`. In span-based evaluation, we evaluate systems by comparing the spans in the output to the spans in the goldstandard. This comparison can be done strictly or more leniently, each of them creating problems of their own, that we describe in what follows:

Strict span evaluation

In strict span-based evaluation, a span in the output needs to match the boundaries and the label of the span in the goldstandard in order to be considered correct. F1 score is then calculated based on the true positives, false positives and false negatives yielded by this approach. For our example in 23, that would mean producing a span that matches the onset and the offset (0 and 2) as well as the label (`ENG`) for it to be correct.

This evaluation approach (that we will call strict span evaluation) is the common practice in most span identification tasks such as NER. Consequently, it was also the evaluation we performed in chapter 4.

However, using F1 score over strict matching spans is an evaluation approach that is particularly ill-suited for span identification tasks such lexical borrowing extraction. In a normal classification task scenario, the precision-recall allocation will develop as follow: a positive item will either be considered as positive (true positive) or negative (false negative) by a system, and a negative item will either be considered as positive (false positive) or

[2] "Sticky buns: the lesser known cousins of cinnamon rolls"

[3] `https://elcomidista.elpais.com/elcomidista/2023/02/14/receta/1676379241_008619.html`

negative (true negative) by a system. Precision and recall metrics will be then calculated based on these numbers and F1 score will then summarize both scores.

In span identification tasks like ours, the set of possible errors expands: a model may correctly detect a sequence of tokens as a span, but may assign it a wrong label; or it may correctly assign a label to a span, but incorrectly identify its boundaries; or it may do both (a partially matched span that gets a wrong label) (Nadeau and Sekine, 2007). Going back to our example in 23, a system may retrieve the partially correct span *Sticky*, the partially correct span *buns* or retrieve the span *Sticky buns* and assign it the incorrect label `OTHER`. All these errors lead to an unexpected behavior of the F1 score, as they will cause a double penalization: a predicted span that partially overlaps with the goldstandard span such as *buns* will count both as a false negative (because the expected gold span *Sticky buns* was not retrieved, thus hurting recall), and also as a false positive (because an unexpected span *buns* was detected instead, thus hurting precision). Consequently, under the F1 score over strict spans evaluation, a model that proposes no span whatsoever will get a better score than a system that proposes a partially correct span or a mislabeled span.

This inadequacy of F1 score over strict spans matching was pointed out by Manning (2006), who states:

> "The thing to notice is that for the first 4 event types [of errors], you are either correct or you get 1 demerit, assessed to either precision or recall. In the simple classification case, that's the end of the story and the F1 measure is sensible. But when doing precision and recall over subsequences, there are these other three event types. Each of them is assessed a minimum of 2 demerits, with both precision and recall being hit. Therefore, it is fairly clear that optimizing for F1 in this context will encourage a system to do the following: if I'm moderately uncertain of either the class label or the boundaries of the entity, because a mistake would cost me a minimum of 2 demerits, I'm better off proposing no entity, which will cost me only 1 demerit."[4]

Manning's observations refer to NER, but they can easily be extended to other span identification tasks, such as our topic, lexical borrowing extraction. In fact, the shortcomings of F1 score over strict spans are particularly relevant for our use case. Our aim is to create

[4] `https://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html`

a system that can assist lexicographers and linguists in the task of identifying new words and borrowings, especially candidate borrowings that are previously unseen for the model. However, because of the double penalization we have just presented, evaluation over strict spans rewards systems that show a conservative behavior. In other words, F1 score over strict spans privileges systems that will only propose a candidate when they are sufficiently certain about both its label and its boundaries. This means that the best performing models according to F1 score over strict spans are likely to inhibit from proposing borrowings whose label or boundaries are tricky, namely previously unseen borrowings or multiword borrowings, precisely the type of items that a lexicographer may be more interested in retrieving because it is in those scenarios that plain dictionary lookup is less likely to succeed. F1 score over strict spans may be the metric of choice for other use cases, such as scenarios where we want to prioritize retrieving correct full spans over partially correct spans. But for the use cases we defined in section 7.1 (that is, identifying borrowing candidates in a corpus mainly as a pre-processing step for linguistic and lexicographic work), a system that proposes a possible candidate (even without absolute certainty) or a partial match is probably a more helpful tool than a system that will only propose a candidate when it is sufficiently certain about both its label and boundaries.

These considerations cast doubt on F1 score over strict spans as a desirable evaluation strategy for our task. Moreover, we question whether the system that was ranked first in chapter 4 following this approach (the BiLSTM-CRF with codeswitch and subword embeddings) was indeed the best performing one for our purposes. Perhaps some of the other models were capable of retrieving more borrowing candidates (even if they were only partially correct) but they ranked lower because of the particular behavior of F1 score over strict spans.

Although alternative span-based metrics to F1 score have been proposed throughout the years (Nouvel et al., 2016), such as error per response (Chinchor and Sundheim, 1993) or slot error rate (SER) (Makhoul et al., 1999), they all assume the strict matching spans paradigm

and therefore have similar limitations to F1 score over strict spans.

Lenient span evaluation

As a response to the limitations that strict evaluation over spans bears, a more generous evaluation approach was proposed: lenient span evaluation. This approach encloses several possibilities:

- Exact match: a span is considered correct if boundaries match, regardless of the type. In example 23, that would mean producing the span *Sticky buns*, regardless of whether it was labeled as `OTHER` or `ENG`.

- Partial match: a span is considered correct if there is some boundary overlap, regardless of the type. In example 23, producing *Sticky buns*, *Sticky* or *buns*, either as `OTHER` or `ENG` would be considered correct.

- Type match: a span is considered correct as long as there is some boundary overlap and the type matches. In 23, producing *Sticky buns*, *Sticky* or *buns*, will be considered correct as long as it was labeled as `ENG`.

This evaluation schema was proposed by Chinchor and Sundheim (1993) for MUC-5 and has been replicated since, as in Segura-Bedmar et al. (2013). This approach guarantees that partially matching spans will be given credit and count towards the score. The problem with this proposal is that, although in some scenarios (such as ours) it may be desirable to take partial matches into account, this approach will make no distinction between partial matches and full matches. An ideal evaluation approach should allow to distinguish between a perfect system and all other systems. But with the lenient span-based approach, a system that gets a perfect F1 score of 100 could either be a perfect system that fully matches the goldstandard or an imperfect system that partially matches the goldstandard but consistently gets boundaries, types or both wrong.

Esuli and Sebastiani (2010) illustrate the unexpected effects that the lenient span-based evaluation approach may produce in certain extreme cases: under this approach, a system may get credit for simply labeling the whole document as one sole span, as this simple trick will earn the system one point per span in the document.

7.2.2 Problems with token-based evaluation

An alternative to the limitations that span-based evaluation causes is to frame the problem as a token classification task, in which each token in the sequence receives a tag and evaluation is computed based on the correctness of the assigned tag to each token. When framing span-identification tasks as a token classification problem, prefixes (such as BIO prefixes) will be used to mark the position of the token within the span and denote span boundaries (Ramshaw and Marcus, 1995).

In token level evaluation each tag is taken into account individually and any tag that is correct will be counted as a true positive, regardless of whether the full span that the individual token belongs to was correctly labeled or not. In other words, in token-based evaluation a problem like NER will be evaluated in the same fashion as we would approach part-of-speech tagging, where F1 score is calculated over the correctness of labels assigned at the token level. This is the approach followed in codeswitch identification tasks (Solorio et al., 2014; Molina et al., 2016), as well as the default approach when evaluating sequence labeling tasks in some popular Python libraries such as `scikit-learn` (Pedregosa et al., 2011).

Framing extraction tasks as a token classification problem solves some of the shortcomings that span-based evaluation displayed: token-level evaluation will give some credit to predicted spans that overlap with the goldstandard but, unlike the lenient span evaluation, the credit given to overlapping spans will be relative to the number of actually matching tokens. Similarly, false positive spans will amount more or less demerits depending on the length of the span, that is, depending on the number of tokens it comprises (unlike span-

based evaluation, where a wrong span would amount one false positive only, regardless of its length).

Token-level evaluation has however its own drawbacks. In token level evaluation we assume that span boundaries are encoded at the token level following an encoding scheme (BIO, BILOU, etc), where each token tag is marked by a prefix that indicates the position of the token within the span. For instance, in BIO encoding the first token in the span will be marked by B (begin), subsequent tokens in the span will be prefixed with I (inside) and tokens not belonging to any span will be labeled as O (outside). For example, in NER the `LOC` entity *Santiago de Compostela* will be encoded at token level following BIO encoding as `B-LOC I-LOC I-LOC`. In BILOU encoding (sometimes also called BIOES or BMES encoding), additional prefixes are considered: single-token spans will be prefixed with U and the last token in multiword spans will be labeled with L (last). Following the same example, the entity *Santiago de Compostela* will be encoded as `B-LOC I-LOC L-LOC` on BILOU encoding. Although it has been claimed that BILOU encoding outperforms BIO (Ratinov and Roth, 2009), the claim is disputed (Lignos, 2020) and both schemes remain well-established. Table 7.1 displays an example for lexical borrowing that illustrates the difference between various encoding formats.

The issue with calculating F1 score at token level is that the number of possible token level tags (and consequently the whole evaluation) will be dependent on the encoding scheme of choice. This means that two outputs that are identical on the span level may receive different scores if they are annotated at the token level with different encoding schemes, because in BIO encoding tokens will be classified into $2 \times n + 1$ categories (one for B and one for I per span type, plus O tokens, with n being the number of categories considered in the annotation), while under BILOU scheme tokens will be classified into $4 \times n + 1$ categories (one for B, one for I, one for L and one for U per span type, plus O tokens).

More importantly, token-based evaluation displays an imperfect treatment of overlapping results. Going back to our example in 23, under BIO encoding our borrowing *Sticky buns*

will be encoded in the goldstandard with BIO labels as `B-ENG I-ENG`. Let's say that we have a system's output that labeled the segment as `B-ENG O`, which means that it partially retrieved the span, as it only marked *Sticky* as a borrowing. This scenario would yield a true positive (for retrieving *Sticky* as part of the span) and a false negative (for ignoring *buns*). If we are interested in allowing for partial matching, this seems as a desirable outcome. But let's consider what happens if the partial matching occurred not for *Sticky*, but for *buns*. In that case, our system will produce the tag sequence `O B-ENG`, which will result in no credit whatsoever, because *buns* is labeled in the goldstandard as `I-ENG`, not `B-ENG`. This difference between these very similar cases (with one of the outputs receiving some credit, but not the other) seems asymmetrical and infelicitous. Moreover, this will only get worse if more fine-grained encodings are used, such as popular encoding BILOU.

To account for this limitation (the fact that the first and last token of a span will have a different prefix depending on the encoding, which makes truly accounting for partial matching difficult), we could rely on simpler encodings such as IO. In IO encoding, tokens within a span will simply be labeled with prefix I (inside), regardless of the position of the token within the span. In IO encoding, the span *Santiago de Compostela* would simply be encoded as `I-LOC I-LOC I-LOC`. Similarly, our span from example 23 would be annotated as `I-ENG I-ENG`. This simpler encoding would facilitate identifying partial matches on the token level, regardless of the position of the matching token. IO encoding in fact has been shown to produce the best results in model training for NER, outperforming other popular encodings such as BIO or BILOU (Alshammari and Alanazi, 2021) and has been used for some NER tasks and datasets (Ding et al., 2021).

There is however an undeniable downside to IO encoding: its inability to capture and adequately represent adjacent spans. An NER sequence such as `I-LOC I-LOC` in IO encoding could either represent a multiword entity (such as *Buenos Aires*), as much as it could represent two independent spans that happen to be collocated (such as *Cambridge Massachussets*). In lexical borrowing detection, this limitation would prevent us from distinguishing

multiword borrowings like *sticky buns* from collocated borrowings (like *marketing online*). This is not a minor aspect, because correctly identifying multiword borrowings versus adjacent borrowings is of vital importance for lexicographic work and terminology extraction and was in fact pointed out as one of our desideratum in section 7.1. As a result, and although IO encoding can indeed account for partial matching in token-level evaluation while other encodings cannot, we cannot consider IO encoding as a viable encoding for borrowing extraction.

There is an additional (yet unfrequent) encoding that seeks to keep the best of both BIO and IO encoding: IOB encoding. In IOB encoding, the I prefix is used by default to indicate that the token belongs to a span, and B is exclusively used to label the first token in a span when it happens to be preceded by another span (Ramshaw and Marcus, 1995). In other words, in IOB encoding, B will only be used to resolve the ambiguity that IO encoding has when annotating adjacent entities. Although IOB is a clever solution to the limitation that IO encoding has, it has a major downside: when using IOB encoding, the tag used to label a token may vary depending on the context in which it appears. Because of this inconsistency we will not be considering IOB tagging withing our research.

7.2.3 A silver bullet: the token and separator model

In order to mitigate the shortcomings and undesired effects displayed by strict span-based evaluation (too harsh; partial matching not taken into account; it may cause a double penalization), lenient span-based evaluation (too generous; not adequately accounting for a varying degree of overlap) and token-level evaluation (dependent on the encoding scheme of choice; imperfect treatment of partial matches), Esuli and Sebastiani (2010) proposed an alternative evaluation approach for span identification tasks: the token & separator model.

The token & separator model frames the evaluation of span identification tasks as a token classification problem, but it eliminates the tag prefixes that we would normally use to indicate the position of the token within the span (B, I) and instead adds separators

(including whitespaces) to the universe of taggable items. This means that under the token & separator model, whitespaces will also receive a tag. In fact, whitespace tagging will denote the boundaries of the span: whitespaces that do not belong to any span will be labeled as `O`, while whitespaces within spans will receive the tag of the span they belong to. For example, in NER the entity *Santiago de Compostela* (that would normally be tagged as `B-LOC I-LOC I-LOC` under BIO encoding and as `B-LOC I-LOC L-LOC` under BILOU encoding) will be labeled as `LOC LOC LOC LOC LOC` under the token and separator model, because the whitespaces between the tokens would count as taggable units. If we apply this to the annotation of borrowings, our multiword borrowing from example 23 *Sticky buns* would be annotated as `ENG ENG ENG`. Adjacent spans will be delimited by the whitespace between them, that will be tagged as `O` (see table 7.1).

The token & separator model succeeds where previous evaluation approaches failed: it gives credit to partial matches (more or less depending on the degree of overlap between the predicted span and the goldstandard), overlapping results are adequately computed regardless of the encoding scheme or the position of the token within the span, while it also evaluates the ability of the system to correctly identify the boundaries of a span.

Although the token & separator model is an elegant solution that solves the flaws that rise from standard span identification evaluation techniques, it has not been adopted in the field.

Appendix C provides a practical example on how the metrics we have reviewed penalize and reward outputs differently.

7.3 Experiments: Revisiting COALAS outputs

In the previous section we reviewed some of the alternative strategies to the strict span-based evaluation approach we followed in chapter 4.

Our observations inevitably cast a shadow of doubt on our results from chapter 4. In chapter 4 we trained and tested a collection of ten models for anglicism detection, and

Scheme	*Cursos*		*de*		*blockchain*	,	*machine*		*learning*		*y*		*marketing*		*online*	.[1]
BIO	O		O		B-ENG	O	B-ENG		I-ENG		O		B-ENG		B-ENG	O
BILOU	O		O		U-ENG	O	B-ENG		L-ENG		O		U-ENG		U-ENG	O
IO	O		O		I-ENG	O	I-ENG		I-ENG		O		I-ENG		I-ENG	O
IOB	O		O		I-ENG	O	I-ENG		I-ENG		O		I-ENG		B-ENG	O
TS	O	O	O	O	ENG	O	ENG	ENG	ENG	O	O	O	ENG	O	ENG	O

[1] *Courses on blockchain, machine learning and online marketing.*

Table 7.1: Comparison between different encoding schemes for token-based annotation of borrowings: BIO, BILOU, IO, IOB and TS.

we ranked them according to their performance based on F1 score over strict spans. We concluded then that a particular model (the BiLSTM-CRF model fed with codeswitched embeddings and subword embeddings) was the best performing one, as it obtained the highest F1 score of all ten models. But based on our observations from section 7.2, we could argue that perhaps the ranking would be different had we evaluated taking a different approach.

We therefore decided to re-evaluate the predicted outputs for the COALAS test set produced by our ten models from chapter 4, with the hope that, as different evaluation approaches penalize and reward outputs differently, evaluating our outputs on a different approach would perhaps lead to a ranking somewhat different to the original one and that this new ranking would hopefully shed some light into the models' performance.

We chose to re-evaluate our outputs from COALAS using the token & separator model. The reason behind this decision was that, contrary to the other evaluation paradigms, the token and separator model distinguishes between a perfect system and a system that partially retrieves spans (unlike the lenient span evaluation); it accounts for partial matches in a balanced way (unlike BIO-encoded evaluation); and it adequately represents the difference between adjacent spans and multiword spans (unlike IO-encoded token evaluation). In addition, the token and separator model showed a low correlation with strict span evaluation (Spearman coefficient=0.38) in our experiment from appendix C. Consequently, the token & separator model was chosen as the ideal approach for our re-evaluation.

In order to perform the re-evaluation on the token & separator model, we transformed the

Model	Word emb	BPE emb	Char emb	Precision	Recall	F1 score
BiLSTM-CRF	Codeswitch	✓	✓	94.30	**82.78**	**88.15**
BiLSTM-CRF	Codeswitch	✓	-	**94.67**	82.38	88.08
BiLSTM-CRF	BETO+BERT	✓	✓	92.77	82.68	87.42
BiLSTM-CRF	Codeswitch	-	✓	93.63	81.86	87.33
BiLSTM-CRF	Codeswitch	-	-	93.88	81.24	87.06
BiLSTM-CRF	BETO+BERT	✓	-	93.30	81.19	86.80
mBERT	-	-	-	91.90	81.69	86.49
BETO	-	-	-	91.02	81.56	86.00
CRF	-	-	-	81.06	44.42	57.39
Llama3	-	-	-	49.18	39.05	43.53

Table 7.2: Ranking of models evaluated on the token & separator evaluation model on the COALAS test set.

original outputs from chapter 4 into the format that the token & separator model expects. In other words, we took the original CoNLL-formatted outputs labeled with BIO encoding that our systems had produced, removed the B and I prefixes and instead explicitly labeled the whitespaces between tokens. Whitespaces within multiword spans received the same label of the span, while whitespaces not within spans were labeled `O` (for an example of the annotation format expected by the token & separator model, see table 7.1). We performed this transformation for all the outputs obtained by the models from chapter 4. For the neural supervised models, we transformed the outputs of all the runs made with different random seeds on the test set, and then calculated the average scores across runs.

Table 7.2 displays average precision, recall and F1 score obtained by each model evaluated on the token & separator model. The scores are slightly higher than when we evaluated following the strict span approach. This is expected, as the token and separator model will account for partial matches, while the strict span-based evaluation will not. What is unexpected is that the ranking of models obtained in table 7.2 is virtually the same to the ranking we obtained in chapter 4 when evaluating over strict spans.

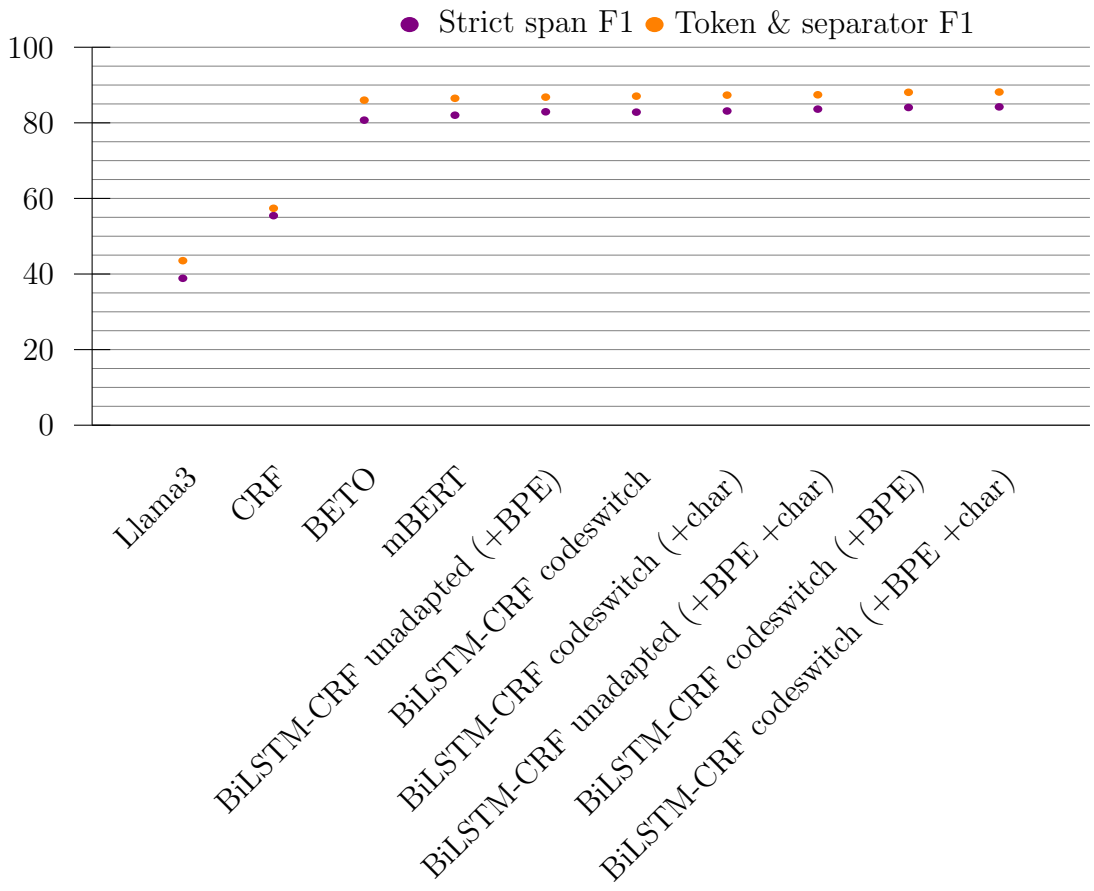

Figure 7.1: Comparison between results obtained on the strict span evaluation and the token and separator evaluation.

7.4 Discussion

In the previous sections we illustrated how different evaluation approaches will penalize and reward the very same outputs differently. Our analysis revealed some interesting findings.

The span-based strict evaluation is unable to account for partial matches at all. On the other hand, the span-based lenient evaluation is unable to distinguish between the perfect output and any imperfect output that partially overlaps with the goldstandard.

Token-based evaluation with BIO encoding displays an inconsistent behavior of overlapping matches, with some partial matches being computed while others not.

Token-based evaluation with IO encoding is simply incapable of representing span adjacency, deeming thus unfit for our task.

The token and separator model, on the other hand, gives credit to all partial matches in a balanced way, distinguishes between the perfect output and any imperfect output that partially overlaps with the goldstandard, and adequately accounts for span adjacency, which make it ideal for evaluating span-identification tasks such as ours.

Despite the substantial differences among evaluation approaches, our re-evaluation of the outputs produced in chapter 4 on the COALAS test set using the token & separator model produced the very same ranking to the one we obtained originally when we evaluated over strict spans (see table 7.2 and figure 7.1). More precisely, the ranking from table 7.2 and table 4.9 display a Spearman correlation coefficient of 0.98, they are almost a perfect copy of one another.

This is disappointing to say the least. Our observations seemed to prove that different approaches rank outputs in very different ways. We would have expected that two evaluation approaches as dissimilar as strict span-based evaluation and the token & separator model (which in our experiments from appendix C displayed a correlation of just 0.38) would yield a very different rankings. And yet, this did not happen: both evaluation approaches ranked our systems in almost the same way. How is this possible?

7.5 Summary

In this chapter we have looked critically at different evaluation approaches for span-identification tasks, such as lexical borrowing detection. We have analyzed the limitations that both span-based evaluation (strict and lenient) and token-based evaluation have, and we have identified an existing (yet underexploited) evaluation approach that overcomes those limitations: the token & separator model by Esuli and Sebastiani (2010). Our observations show that each evaluation approach ranks outputs in a different way.

Finally, we revisited the outputs we produced for the COALAS test set in chapter 4 and re-evaluated them using the token & separator model. The ranking of models produced when evaluating under the token & separator model was virtually the same to the ranking

we obtained in chapter 4 when evaluating over strict spans. This result was unexpected, as strict span-based evaluation and the token & separator model have very different properties and both approaches had displayed very little correlation in our informal experiments from appendix C.

In chapter 8 we will hypothesize some of the possible causes behind this surprising result and propose a novel evaluation methodology that can help predict how different models will behave under different data scenarios.

Chapter 8

Dimensions in evaluation

To measure is to know. If you cannot measure it, you cannot improve it.
Lord Kelvin, physicist

In chapter 7 we explored some alternative approaches to evaluation over strict spans that can be used to evaluate span-identification tasks, such as lexical borrowing detection. We proved that different approaches will penalize and reward the same outputs in very different ways and, as a result, we anticipated that our ranking from chapter 4 would have looked very different had we followed an approach other than F1 score over strict spans. We tested our hypothesis by reevaluating the full collection of outputs produced by the models from chapter 4, but this time we used the token & separator evaluation model proposed by Esuli and Sebastiani (2010), which displayed very different properties to strict span match evaluation while keeping some of the properties our use case requires. Our suspicion was not confirmed. The ranking produced by Esuli and Sebastiani's approach was virtually the same as the one we had obtained when we evaluated on F1 score over strict spans.

In this chapter we look for possible explanations to the paradoxical results obtained in chapter 7. We will deep dive into two evaluation aspects that we overlooked in our previous enquires: the nature of the spans in our test set and the performance that different models display on them. We will propose a set of linguistic dimensions to characterize the spans

in a dataset and apply them to the COALAS dataset and to the models results obtained in chapter 4.

8.1 Possible explanations for the paradoxical results from chapter 7

There are three possible reasons that could explain why the re-evaluation of the outputs from COALAS using the token & separator model in section 7.3 produced basically the same ranking as the one produced by F1 score over strict spans in chapter 4, despite the difference in properties of both approaches:

Hypothesis 1. *Absence of multiword spans in the test set.*

The token & separator model is particularly good at giving credit to systems that partially retrieve spans, even if they are not fully correct, the scenario where span evaluation becomes the harshest. If our dataset included no examples of multiword spans then it would not be surprising that the resulting ranking would be the same under both evaluation paradigms. In other words, if our test set consisted mainly of single-word spans, then it would be expected that all evaluation paradigms (token evaluation, span evaluation, token & separator evaluation, etc.) produced similar rankings of systems, regardless of the approach. How expressive and adequate an evaluation approach is heavily depends on what the characteristics of the data under evaluation are and the purpose of the evaluation itself. The alleged advantages of the token & separator evaluation model over token evaluation or span evaluation will be useless if what we are evaluating is a dataset where no multiword spans ever happen, because there will hardly be any opportunities for the evaluation approach to discriminate across systems in the scenario where that approach is particularly expressive.

A simple count over the data contained in the COALAS test set rules out this hypothesis: out of the 1,285 spans labeled either as `ENG` or `OTHER` in the COALAS test set, 327 of them were multiword. However, these numbers do reveal an interesting point: the fact that only 1 in 4 spans in COALAS test set is multiword. While it is true that there are multiword

Model	Correct	Errors	Missing	Spurious	Overlap	Segmentation	Type
CRF	553	824	668	95	56	5	0
BETO	1045	335	160	103	37	17	18
mBERT	1021	334	201	82	25	18	8
BiLSTM (unadapted)	1030	303	200	57	22	16	8
BiLSTM (codeswitch)	1051	282	171	60	28	16	7

Table 8.1: Number of error types made per model according to MUC error typology.

spans in the test set, this imbalance in the numbers may very well be a contributing factor to the results we obtained in chapter 7.

Hypothesis 2. *Lack of diversity in the errors in the output.*

A second possible explanation for the unexpected results from chapter 7 is that all models behaved similarly when confronted with specific types of spans and only produced errors that were equally treated under all evaluation approaches. In other words, even if we are certain that the test set contained multiword borrowings, if none of the models ever produced a partially correct match (because the systems always either fully missed the span or perfectly retrieved it), that would explain why very different evaluation approaches produced the same ranking, because the type of error where both approaches differ never happened.

In order to verify this hypothesis, we applied to our outputs the MUC error typology proposed in Chinchor and Sundheim (1993) for NER. This typology classifies spans as:

- Spurious: a predicted span is not present in the goldstandard.

- Missed: a span in the goldstandard is not present in the prediction.

- Overlap: the predicted span partially overlaps with the goldstandard span.

- Type: the predicted span matches the boundaries of the goldstandard, but not the type.

We applied this typology to the spans in the outputs from chapter 4: if all the spans in those outputs were either spurious, missed or correct, then that would explain why the two evaluation paradigms (stric spans and token & separator) ranked the systems the same way.

Model	Missing	Spurious	Both	Total
CRF	49	7	0	56
BETO	29	8	1	37
mBERT	18	7	0	25
BiLSTM (unadapted)	19	3	1	22
BiLSTM (codeswitch)	25	3	1	28

Table 8.2: Number of types of overlapping errors per model: *Missing* refers to overlapping errors where at least a token was missed. *Spurious* refers to overlapping errors where at least a spurious token was added. *Both* refers to overlapping errors where there were both a missing token and a spurious token.

Table 8.1 displays the types of erroneous spans produced by each model. We added an additional error type to the original MUC error typology: segmentation. This was added to account for errors where the relevant tokens were correctly retrieved but the boundaries were wrong (such as fusing adjacent spans or splitting multiword spans). We made this decision because the original MUC guidelines did not explicitly specify how to classify these errors (they could be subsumed by the *overlap* error type, or under a more general class *incorrect*). Segmentation errors are particular relevant for our case, because along with overlapping errors, they are the type of phenomena where the two evaluation paradigms in question differ more drastically. We considered that it would be more informative to count them separate.

The numbers in table 8.1 rule out hypothesis 2. Although missing spans and spurious spans were the most frequent type of errors for all models, all outputs contained overlaping, segmentation and type errors. In fact, some interesting observations arise from these numbers: the model that missed the least amount of full spans was BETO (160). The high number of overlapping mistakes made by BETO probably contributed to the low number of missing spans, although this came at the price of retrieving the highest number of spurious spans (103) and getting the highest number of errors overall (335).

Overlapping errors can be caused by different issues: overlapping spans may have missed one or more tokens from the goldstandard span; they may contain one or more spurious

tokens that were not part of the goldstandard span; or they may do both (one or more tokens were missed in one end, and one or more tokens were added to the other end). In order to get a finer-grained analysis of the overlapping errors that the systems were making, we classified the overlapping errors according to those types. This distinction can provide us additional insight into whether the overlapping errors are related to token classification recall, token classification precision or both. Table 8.2 shows that overlapping errors caused by missing tokens clearly outnumber overlapping errors caused by spurious tokens. Overlapping errors caused by both missing tokens and spurious tokens are insignificant.

Hypothesis 3. *Aggregated average scores hinders performance analysis.*

A more lenient version of hypothesis 2 is that even if our models produced overlapping errors and segmentation errors, aggregated average evaluation (such as the ones we performed using F1 score) will inevitably hide these nuances and will prevent us from actually seeing the differences on how models perform on different types of spans. Ultimately, the chosen evaluation approach will be irrelevant if results are aggregated into an average.

After all, holistic scores over a dataset assume that all instances in a test set are equally difficult and that scores can be simply averaged. However, we know that certain spans will be more prone to errors than others. Multiword spans may cause overlapping errors; adjacent spans may produce segmentation errors. How well did our models perform on multiword borrowings? Did they manage to correctly identify the boundaries of adjacent spans? These questions become particularly relevant if the different types of spans are not equally represented in the test set, as we attested in hypothesis 1.

After ruling out the previous two hypotheses, we need to consider this as a possible explanation. In the next sections, we will analyze and formally characterize the type of spans contained in our dataset and assess the performance of our models on different types of spans.

8.2 Rationale

Not all anglicisms are created equal. Words such as *streaming*, *online* or *pie* may look equally unremarkable to a native speaker of English, but as anglicisms they all appear very different to the eye (and the ear) of a monolingual speaker of Spanish. Some borrowings display a combination of letters that looks unusual in the recipient language, while others could pass for a native word.

For instance, an anglicism like *streaming* violates the graphotactical expectations that Spanish speakers have, because words in Spanish are not supposed to begin with *str-* or end in *-ing*. *Streaming* could never pass for a word in Spanish.

On the other hand, a word like *online*, while still being an anglicism in Spanish, displays a combination of letters that does not look suspicious in Spanish. It is still a borrowing because it is used as such: the pronunciation of the word follows the English rules of pronunciation, not the Spanish ones. But even when we know that *online* is a borrowing in Spanish, there is nothing in its shape that makes it foreign or that prevents it from being a word in Spanish. A monolingual speaker of Spanish that encountered the word *online* for the first time in written text could easily think that it is a word in Spanish that is unknown to them, not necessarily assume that it must be a borrowing, unlike *streaming*.

Finally we have a word like *pie*, which is a fully Spanish word when it means "foot", but that may also be an anglicism when it refers to the dessert (pronounced as /paɪ/), as in "un *pie* de limón". When used as a borrowing, the word *pie* will be pronounced according to the English pronunciation rules, but there is nothing preventing the combination of letters p+i+e from being a word in Spanish, and, in fact, it is also a word in Spanish.

All these nuances can make the task of detecting anglicisms in text harder or easier for NLP models. A borrowing like *streaming* will be easier to spot, because just the combination of letters makes it a very good candidate as a foreign word. A model that incorporates the frequency of character n-grams (such as the rule-based model proposed in Serigos (2017a)

or our CRF model from Álvarez Mellado (2020a)) will probably be able to detect it.

A word like *online*, on the contrary, would probably be harder to spot based on character combination only, because there is nothing unusual about the combination of characters in the sequence. A model would need to have seen *online* labeled as a borrowing during training or rely on contextual information to correctly detect it as a borrowing.

Finally, the word *pie* will be the hardest case of all: not only could it easily escape any character-based model, but also the fact of having seen it as a Spanish word or as a borrowing during training does not guarantee that the model will be capable of making the right choice when confronted with *pie* in the test set.

Similarly, certain contexts will make the detection task harder or easier. According to the orthotypographic rules of Spanish, in standard text unassimilated borrowings are expected to be written surrounded by quotation marks (Real Academia Española, 2011). Consequently, anglicisms surrounded by quotation marks will stand a better chance of being retrieved, because the quotation marks may serve as a cue to the models.

Capitalization may also affect the detection process: as our error analysis from section 4.1.6 showed, anglicisms in upper case (such as *Big Data*) or in sentence-initial position have a higher chance of being mistaken with proper names, and are therefore are more likely to be missed.

What lies at the crux of the issue is that in any span identification task (such as borrowing detection), the *findability* of the span will be affected by certain aspects, such as whether the span was seen in training, its shape or the context it appears in. However, none of these issues were taken into account when we evaluated the performance of our models from chapter 4. How good are models at retrieving sentence initial position spans? Are they exclusively learning to detect certain formal cues, such as quotation marks? How well do these models perform when the cue is not present? Are graphotactical non-compliant spans (odd-looking borrowings) being privileged over compliant ones? It should not come as a surprise that a lexical borrowing detection model that obtained an excellent F1 score when

evaluated on a test set where all the spans contained contextual cues fails when faced with less obvious contexts.

From the results we obtained from chapter 4 and the re-evaluation we performed in chapter 7 we cannot really tell if our models are consistently relying on external cues, such as quotation marks, character combination or sentence position; neither can we know how well they perform on multiword borrowings or if they are adequately retrieving adjacent spans. In order to truly assess the generalization abilities of our models, we need to investigate how well our models perform in these different aspects. Differences in results across these traits will give us a better insight at how models perform and even help us anticipate how a model will perform on a different distribution.

8.3 Span characterization: precedents from NER

The idea of characterizing spans in span identification tasks is not new. After all, the issues we have pointed out for lexical borrowing detection are common to other task, such as NER.

Let's say that we have a NER model that is supposed to detect named entities (person, organization, location names) in Spanish text, and that we provide it with the following input sentence:

Example 24. *Se mudó de vuelta a Barcelona*[1].

Let's imagine that the NER model correctly marked *Barcelona* as a location entity. Was *Barcelona* correctly retrieved because the entity *Barcelona* was already seen during training? Or was it the result of the generalization abilities of the model? Would the output be any different had *Barcelona* not been in the training set? And what if the span had been written in lowercase?

Let's now consider the token *Podemos*, which can be a conjugated form of the Spanish verb *poder* (the present tense of the 1st person plural, "we can"), but is also the name of a

[1] "She moved back to Barcelona."

Spanish political left-wing party. How likely would it be for the NER system to produce a correct output for the `ORG` entity *Podemos* if it only saw *Podemos* used as a verb (labeled as `O`) during training?

Let's now imagine a third hypothetical scenario, where the entity in question is the Basque city *San Sebastián*, which is also the name of the Catholic martyr Saint Sebastian that the city is named after. Would our NER system be capable of recognizing *San Sebastián* as an entity of type `LOC` if it only saw the span *San Sebastián* marked as an entity of type `PER` during training?

As we have argued for lexical borrowing detection, certain characteristics of the spans or the context in which they appear will make NER easier or harder. As a result, prior research on NER has proposed different ways of classifying entity mentions with the aim of assessing how models perform on different types of entities and identifying challenging test instances.

Bernier-Colborne and Langlais (2020) proposed calculating the percentage of mislabeled tokens in the test set that were *unseen* (never seen during training) or *diff* (whose label was not the most frequent label in training) and found out that diff instances were more challenging and error-prone than unseen tokens. The authors decided to focus on these two subsets of tokens because, as Bernier-Colborne and Langlais point out, for all tokens that are neither unseen or diff, the task of assigning them a label is trivial, as it suffices to assign the most frequent tag seen in training.

Similarly, Tu and Lignos (2021) proposed evaluating NER models on what they call *tough mentions*, i.e. entities that were unseen (because they were not seen during training, or because they were seen but with a different label), or type-confusable (mentions that appeared in the test set with multiple labels). In other words, they provided a way of classifying named entities as more or less and challenging based on their presence in the training set and the stability of their label.

Fu et al. (2020) proposed an interpretable approach to NER evaluation by partitioning the test set into a series of buckets based on a set of attributes that entities may have, and

analyzed how different models performed on sets of entities that shared the same attributes. Their analysis considered local attributes defined on the basis of a concrete entity or sentence (such as entity length, sentence length, OOV density or entity density) and aggregated attributes that were calculated based on statistics across the whole training corpus (such as token frequency, entity frequency, token label consistency, span label consistency). Among other discoveries, they point out that label consistency is a key aspect, as it has more impact than frequency or model sophistication, which seems to corroborate Bernier-Colborne and Langlais's findings.

The proposal we present in section 8.4 heavily draws from all these precedents, combines them and apply the result to the field of lexical borrowing detection.

8.4 Dimensions definition

Following the reasoning from section 8.2, we now define a set of dimensions that can help us characterize the borrowing spans in our test set and assess in which way they may be difficult to detect for a model.

These dimensions may be grouped into three types:

1. Intrinsic dimensions: these are dimensions that affect the formal shape of the span itself.

   - Single-word vs multiword: the number of words that the span comprises (*online* vs *machine learning*).

   - Graphotactic-compliant vs graphotactic-noncompliant: whether the span complies or not with the graphotactic patterns of the recipient language. For example, an anglicism such as *healthy* could never be a word coined in Spanish, because we would never expect a Spanish word to have a *-th-* digraph or to end in *-thy*. On the other hand, an anglicism such as *prime time* fully complies with the graphotactics rules of Spanish.

2. Contextual dimensions: dimensions that reflect the contextual surroundings of the example in which the spans appears in.

- Sentence initial: whether the span appears in the first position of a sentence. Sentence-initial anglicisms can be mistaken with proper names and can be difficult to detect.

- Quoted vs unquoted: whether the span is surrounded by quotation marks or in italics. Quotations marks may function as a cue and make detection easier.

- Capitalization: whether the span is capitalized. Some borrowings may sometimes be written capitalized, even if they are not proper nouns or sentence-initial (for ex., *Big Data*), which may make detection harder, as systems can mistake them with proper nouns.

- Adjacency: whether the span is preceeded by another span. For example, the sequence *look total black* is comprised of two adjacent anglicisms (*look* and *total black*) that happen to be collocated. These nuanced combinations may lead to segmentation errors.

3. Novelty dimensions: these are dimensions that characterize the span in relation to its presence or absence from the training set.

   - Seen vs unseen: whether the span is contained in the training set.

   - Type confusable: the span was seen in training with at least two different tags. For example the word *post* may be seen during training both as the Spanish prefix of Latin orgin (as in *post guerra*, "post war"), and as the anglicism that refers to an online text (*un post de Facebook*, "a post on Facebook"). Similarly, *pie* will be a borrowing when talking about dessert (as in *un pie de limón*) but it will be a native Spanish word when it means "foot". The model will have to decide which of the labels seen in training is the appropriate one when labeling *pie* or *post* in the test set.

   - Unseen tagging: the span was seen in training, but not with the tag that it bears in the test set. For example, let's say that the Spanish word *red* (which means "net") was seen during training, and that the test set contains the previously unseen anglicism *red*

carpet. The model will have to apply a label to the word *red* that never saw applied to *red* during training.

The purpose of these dimensions is to assist evaluation by characterizing the different aspects that models may struggle with and identify error-prone instances from the test set. These dimensions are not mutually exclusive, and they may combine in various ways. In fact, these dimensions may help us analyze how different dimensions interact with one another. A given sentence may contain an anglicism that is multiword, unquoted, with a non-compliant shape and that was never seen during training.

8.5 Experiments across dimensions

We now propose to apply the dimensions introduced in section 8.4 to the dataset and modeling results from chapters 3 and 4. The aim is to investigate if by applying the dimensions to the COALAS dataset and to the models' results we can gain some insight into the composition of the dataset (how well represented are different types of spans in the corpus?) and into the models' performance (how good are the models at detecting different types of spans?).

8.5.1 Evaluation of the dataset

Figure 8.1 displays the distribution of spans in the COALAS test set per dimension. A look at the percentages can give us an idea of what the typical span in the test set looks like: the prototypical borrowing in the COALAS test set is a span of length 1 that does not comply with the graphotactical rules of Spanish spelling, that was not seen during training, that appears mid sentence, not collocated with other borrowings and that is written in lowercase, without quotation marks.

These numbers also show that, as OOV rich as the test set may be, certain phenomena are under-represented in the COALAS corpus. The test set is clearly biased towards grapho-

tactical non-compliant borrowings, which may very well explain why in the results obtained in chapter 4, models fed with subword embeddings produced better results.

The numbers of borrowings in sentence initial position, capitalized or adjacent to other borrowings are also minimal. This raises the question of how good models are at detecting borrowings in these challenging contexts, as in the hypothetical case that the models performed poorly on those contexts, there would be so few cases that the results would easily go unnoticed on the aggregated evaluation metrics.

Additionally, the low number of adjacent borrowings could had contributed to the paradoxical results from chapter 7, where both F1 score over exact spans and F1 score over tokens and separators produced the same ranking of models, despite the differences in their properties.

Although the novelty dimensions can only be applied to spans from the test set, we may also apply the intrinsic and contextual dimensions to the spans in the training set, and see how much the dimensions from the test set align with the dimensions from the training set. Figure 8.2 shows that the spans in the COALAS training set share many of the properties of the spans in the test set, but some of them are slightly different.

The main difference lies on the amount of quoted spans: 36% of the spans in the training set had quotation marks, but only 7% of the spans in the test set had them. This opens the door to the question of whether the models tended to perform better on the spans of the test set that had quotation marks or if they overgeneralized and inappropriately marked as a borrowing other quoted spans of non-borrowing text.

Other subtle differences between the dimensions from the test set and the training set are sentence initial spans: although spans in the first position of the sentence were infrequent in both splits, they were four times more frequent in the test set (2%) than in the training set (0.5%).

8.5.2 Evaluation of the models

With the percentages we have just provided, we can rightfully ask ourselves some questions: given that the majority of spans in the test set were single token, how well did the models from chapter 4 perform on multiword spans? How good was the performance on graphotactic compliant anglicisms? Was the unbalance in quotation marks between the test and the training set and issue for the models?

To answer these questions, we re-evaluated the models from chapter 4 controling for dimensions. We selected the following best performing model per family from chapter 4:

- The CRF with handcrafted features.

- The Spanish Transformer-based BETO.

- Multilingual BERT.

- The BiLSTM-CRF model fed with unadapted embeddings (and subword embeddings).

- The BiLSTM-CRF model fed with codeswitch embeddings (and subword embeddings).

- Llama3.

We performed the evaluation on the COALAS corpus but we splitted the test set into different subsets according to the different dimensions defined in section 8.4. For each dimension, results were calculated only on the spans concerned with the dimension in question. In other words, for the sentence initial dimension results, only the spans that satisfied the property of being in sentence initial position were taken into account. If other non-initial borrowings were present in the sentence, they were skipped. Tables 8.3 and 8.4 display span recall and precision per model on the different subsets of the test set defined by our dimensions. For each model, the reported results are those obtained by the best performing random seed. For both BiLSTM-CRF models, BPE embeddings and character embeddings were included.

The results obtained per dimension paint a very different picture to the results portrayed by aggregated precision and recall. The BiLSMT-CRF model with codeswitched embeddings (which obtained the highest F1 score overall) ranks poorly across type-confusable, sentence initial and quoted spans. In these dimensions, the BiLSTM-CRF model only outperforms the very modest CRF model and Llama3. It even ranks last on adjacent spans, with a recall of just 12.50. The BiLSTM-CRF model still outperforms all models across other dimensions, such as unseen spans, single token spans, capitalized spans, unquoted spans and graphotactic non-compliant spans, but the results across dimensions enable to detect some of its shortcomings.

The dimensions also allow to discover some strengths that had gone unnoticed in previous evaluations. The BETO model, which obtained a modest position in the original ranking from chapter 4 (just above the CRF model and Llama3) seems to obtain surprisingly good results in terms of recall across some of the dimensions where other models fail: BETO ranks first across type-confusable spans, sentence initial, quoted spans (with a spectacular R=97.98) and graphotactic compliant spans (the type of spans that could easily be missed by character-based models). BETO also ranks second in terms of recall across unseen spans, multiword spans, single token spans, unseen tagging, unquoted spans and graphotactic non-compliant spans. Multilingual BERT was better than other models at retrieving multiword spans and spans with unseen tagging, while the BiLSTM-CRF model fed with unadapted embeddings was the best at correctly identifying previously seen spans and adjacent spans.

The dimensions also show some of the weakness in the models: the CRF model got passable results at identifying previously seen or quoted spans, but its predictions are totally unreliable for sentence initial or capitalized spans.

Llama3's results per dimensions were as unsatisfying as the aggregated results, although it ranked first on sentence initial spans (tied with BETO and the BiLSTM-CRF with unadapted embeddings). Llama3 results over quoted spans were reasonably good (R=75.76), although this came at the price of producing a high number of spurious spans for that same dimension

(P=54.74). This is surprising, given that none of the examples we prompted the model with were quoted (see figure 4.1). Similarly, none of the examples in our prompt contained sentence initial borrowings, yet Llama3 seemed to perform on this dimension better than some of the supervised models. This suggests good generalization capabilities for Llama3 across dimensions where other models failed and for phenomena that we had not explicitly prompted the model with.

The results across dimensions also confirm our concerns about the unbalanced phenomena in the dataset: all models found quoted spans easier to retrieve than unquoted spans, as well as non-compliant spans were easier than compliant ones. Scores for sentence initial spans were catastrophic across all models, but if we had to choose a model for detecting sentence initial anglicisms, BETO, Llama3 and the BiLSTM-CRF with unadapted embeddings would probably be the safest bet, while the BiLSTM-CRF model with codeswitch embeddings would probably be the best call for capitalized spans.

8.6 Discussion

Our dimension-guided evaluation of COALAS revealed some interesting biases in the dataset: there were hardly any spans that were capitalized, adjacent or in sentence-initial position. This unbalance in COALAS can easily be explained by how the corpus was built to begin in: as we explained in chapter 3, we used a CRF model combined with a lexicon lookup pipepline to populate what ended up being our annotated corpus. This was necessary as lexical borrowings are very sparse and we wanted to ensure that our corpus was rich in borrowings, especially rare and previously unseen ones. Because sentence initial position borrowings were hardly found in the training set of COALAS, this translated into very poor recall results of all models over those types of contexts. The lack of sentence initial position borrowings should not come as a surprise: sentence-initial borrowings were the type of borrowings that were easily missed by the CRF model, which was precisely the type model that assisted us when building the corpus in chapter 3, to begin with. We discover ourselves

immersed in a vicious circle: we used a model to bootstrap the annotation process, thus producing a training corpus that is poor in the type of phenomena that the assisting model fails to capture, which ends up producing trained models that have hardly seen that type of borrowings and fail precisely at that particular type of borrowings. The solution to this vicious circle is not trivial: if we decided to use any of the other best performing supervised models to gather more data for more training, we would be facing a similar situation, given that they all found certain types of borrowings difficult to spot.

A possible solution to this issue could be working with an ensemble of models, where each model contributes to the annotation pile with examples that contain good candidates precisely of the type that that specific model is good at retrieving (see section Future Work in chapter 11).

8.6.1 Benefits of the dimensions

Our dimensions have shed new light into the nature of our dataset and the models' performance. On this matter, there are several areas where the proposed dimensions may be of use:

Model evaluation Evaluation is the *raison d'être* of these dimensions. Their main purpose is to help characterize spans, discover systematic blind spots in models' performance, and identify error-prone instances in the test set. In this regard, these dimensions satisfy the three requirements that Fu et al. (2020) defined for an ideal evaluation methodology: being fully automatic, allowing comparison across different datasets and allowing users to dig deeper into the strengths and weaknesses of the model.

Model selection for a given use case The results across dimensions may inform which model is the best-suited for a specific use case. For instance, in 7.1 we listed possible scenarios where a lexical borrowing detection system may be in place. Different scenarios will deem different dimensions more or less crucial: when selecting a model that assists lexicographers

identify new candidates for dictionary inclusion we may want to choose the model that obtains the best score on unseen spans or unseen tagging, regardless of the overall F1 score it obtains. Similarly, if our task is concerned with compiling a list of distinct borrowings in use, it may not be as important if the model excels at every context dimension: if a borrowing appears in a context that the model finds challenging (sentence-initial position, unquoted, collocated with another span) and the model fails to retrieve that particular occurrence, it may still be able to capture it in other contexts where the retrieval is easier (quoted, mid-sentence, not collocated with other spans).

Corpus documentation Recent research has emphasized the need to thoroughly document the type of data contained in NLP corpora (Bender and Friedman, 2018). Data documentation in sequence labeling usually focuses on reporting the number of spans, unseen tokens or provenance of the data, which does not tell us much about the characteristics of the data that the models will have to learn. Applying the dimensions to the COALAS corpus provided valuable information on the type of spans in the corpus and uncovered some of the limitations and under-represented phenomena that this dataset has. Reporting the number of spans per dimensions in sequence labeling datasets can help us get a better understanding of the type of phenomena contained in the test set and allows us to check how well-aligned the different splits are.

Data curation On a similar note, span dimensions may be of use during the process of creating a corpus, as it may help to decide the type of instances to allocate to the different splits. Dimensions may help identify that certain types of spans are ill-represented in a given split or, on the contrary, can be helpful when building adversarial splits to test generalization.

8.6.2 Differences with prior work

The dimensions we have just introduced draw from some of the previous work presented in section 8.3. Our unseen dimension, unseen tagging dimension and type-confusable dimension

build from the work in Bernier-Colborne and Langlais (2020), Fu et al. (2020) and Tu and Lignos (2021).

Our proposal however substantially diverges from prior work and complements some aspects. Bernier-Colborne and Langlais (2020) and Tu and Lignos (2021) exclusively focused on whether a span or its labeling was previously seen or not. Although it has been shown that this aspect heavily correlates with model performance (Bernier-Colborne and Langlais, 2020; Fu et al., 2020), the fact that a span was seen in training can hardly be the sole aspect that makes span detection challenging. In fact, we know that models are capable of correctly retrieving previously unseen borrowings. And we know this because models from chapter 4 obtained reasonably good results when tested on a test set that was extremely rich in previously unseen spans. But we also know that, while being able to retrieve *some* previously unseen borrowings, those models were not capable of retrieving *all* previously unseen borrowings, which implies that some previously unseen borrowings were more challenging to spot than others. A previously unseen span may or may not be challenging to retrieve because of other aspects, such as its shape or the context it appears in.

Fu et al. (2020) did consider some sentence attributes, such as sentence length or span density, but this stills overlooks the fact that certain properties of the span itself (such as casing) or of the context in which the span appears (such as position within the sentence or token surroundings) will affect the retrievability of the span. And that is the blind spot that our proposal seeks to address: spans in certain contexts will be easier to spot than others (thus, the sentence-initial dimension); the presence of typographic cues may help models flag certain spans (thus, the quotation marks dimensions); casing may contribute to some spans going unnoticed (thus the capitalization dimension). Being unseen may be an important attribute, but it is far from the only one and disregarding all other aspects will prevent us from understanding which models are good at what.

8.6.3 Applicability to other span identification tasks

We have defined these dimensions to assist on evaluating lexical borrowing detection systems, but this methodology can be applied to other span identification tasks, such as NER. As we saw in 8.3, polysemous spans are not rare in NER, and the retrievability of entities is also sensitive to context characteristics, such as capitalization and sentence position.

All the proposed dimensions are directly applicable to other tasks, except for the graphotactic compliance dimension. The capitalization dimension can inform how dependent an NER model is on casing and whether lowercase entities systematically go unnoticed. The quotation dimension can provide a more fine-grained evaluation of NER models retrieving works of arts (such as models trained on the OntoNotes ontology). Adjacent spans are equally prevalent when dealing with named entities, and a common source of error in standard NER benchmarks (Rueda et al., 2024).

In respect to the graphotactical dimension, this dimension points to a crucial issue that can also be applied to other tasks: the fact that, in any span-based task, there will be some spans that are more prototypical or salient than others. In borrowing detection, non-compliant borrowings will be more salient (and therefore, easier to spot), while compliant borrowings may go unnoticed. In NER, research has shown that models trained on data that is rich in English proper names will find it difficult to spot entities in other languages (Lin et al., 2021; Vajjala and Balasubramaniam, 2022).

At the end of the day, the graphotactical dimension in anglicism detection simply implements a more general aspect that affects other span-based tasks: the degree of distinctiveness in the spans. This aspect can be represented and measured in various ways. In the case of lexical borrowing, we chose to implement this feature as a binary dimension using formal constraints based on linguistic graphotactics, but continuous representations based on the Kullback–Leibler divergence have also been used in other span identification tasks (Papay et al., 2020).

To facilitate the application of our methodology to other spans identification tasks, we

released `cascabel`[2], a Python package that implements the dimension-based analysis from this chapter. Given a test set for a span identification task (such as the COALAS test set), `cascabel` provides information on how the spans in the given test set are distributed among intrinsic, contextual and novelty dimensions (see chapter 11).

8.6.4 Limitations of the notion of *unseenness*

The novelty dimensions we have presented in section 8.4 build on the notion of a span being unseen or not, that is, whether a token or span from the test set is also contained in the training set (and with the same label or not).

The notion of a span being *unseen* is a relevant one in span identification tasks: after all, when we train a model to detect spans, we normally want it to be capable of retrieving spans from the test set, even (or especially) if they were not seen during training.

In fact, the notion of a span being previously unseen is well-established in the field and has been used in the past to create challenging NER tasks, characterize datasets and analyze models' performance, as in Augenstein et al. (2017); Derczynski et al. (2017); Kim and Kang (2022); Ok et al. (2024).

However, the notion of unseenness can be challenged in at least three ways:

1. In languages with little morphological inflection, the decision of whether a span is seen or not may be clear. However, in morphologically rich languages, the frontier between the two becomes blurry: should morphological variations of the span (for example, with different case marking) be considered instances of the same span or not? Or should we only consider that a span was seen if it is contained in the training set with exactly the same shape it has in the test set?

2. A binary concept of unseenness may be suitable when working with models where single words are the smallest units in question. However, in more recent approaches (such as Tranformer-based architectures), tokenizers and embeddings (such as BPE embeddings)

[2] `https://github.com/lirondos/cascabel`

also build representations and assign a label to subword units. In that case, we cannot claim that the model saw or did not see a certain token labeled in a certain way, because it may very well be that what the model actually saw was a tokenized subword unit that was perhaps also seen (or not seen) at test time.

3. Finally, the notion of unseennes is perfectly clear in terms of recall: the ability of the model to retrieve a span that was not contained in the training set. But the notion becomes less obvious when we applied it to precision. For section 8.5.2 we still decided to provide precision scores as a way of contextualizing the obtained recall results (a perfect recall with a catastrophic precision tell us that the high ability to retrieve unseen spans is reached at the cost of over retrieving a lot of false positives), but as Tu and Lignos (2021) point out, it is unclear which false positives should be considered when computing precision over unseen spans.

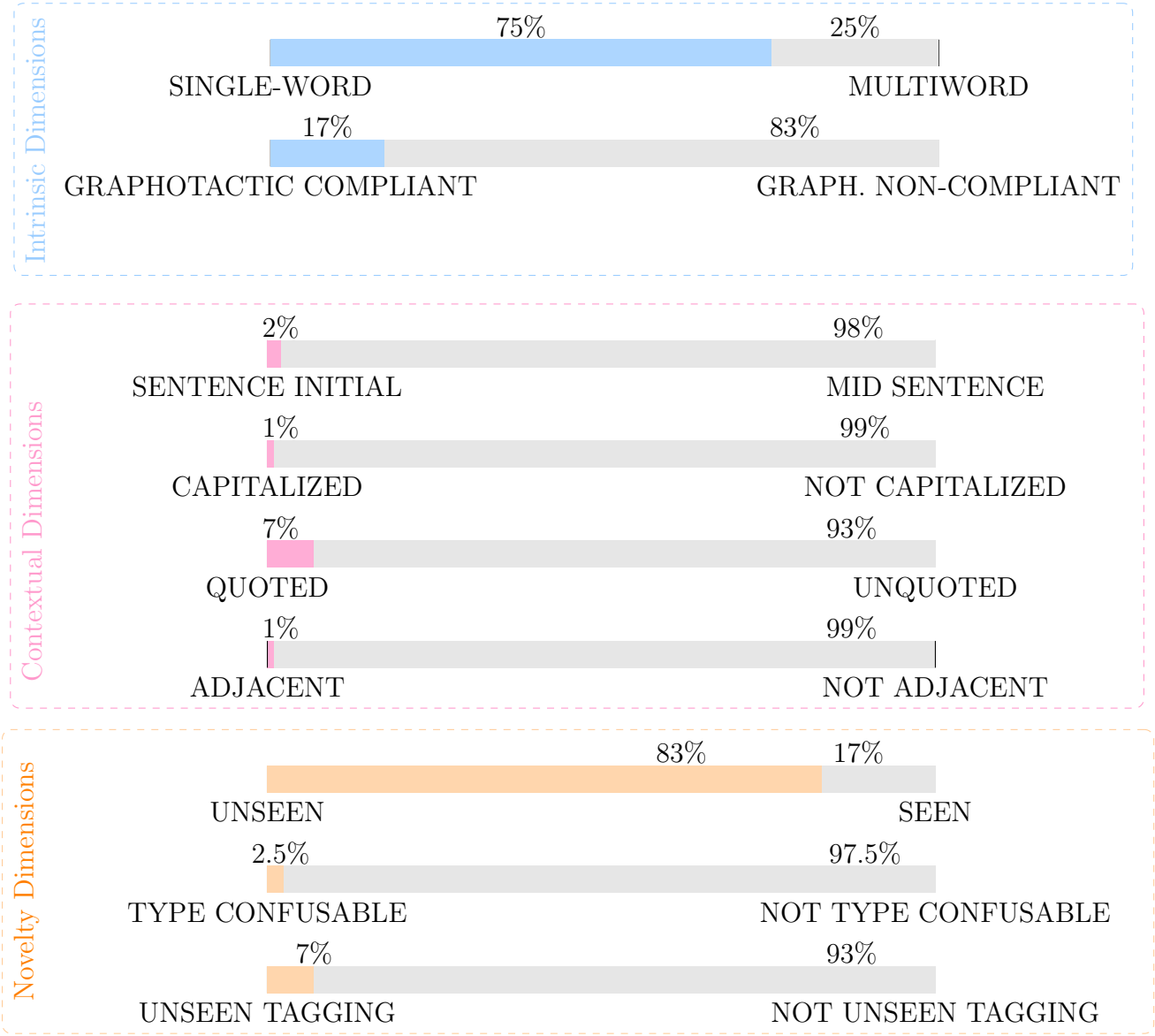

Figure 8.1: Ratio of spans per dimension in the COALAS test set.

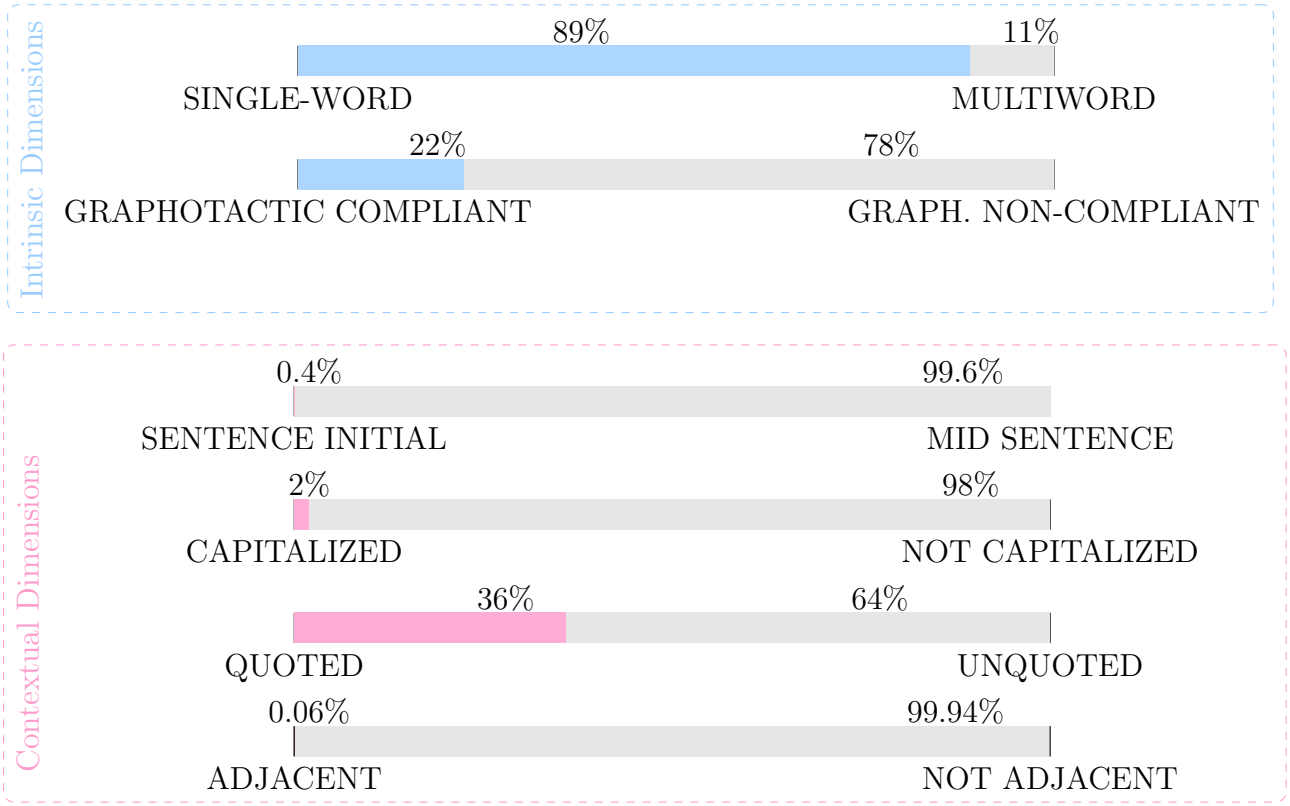

Figure 8.2: Ratio of spans per dimension in the COALAS training set.

Model	Overall recall	Unseen	Seen	Multi-word	Single-word	Type confusable	Unseen tagging	Sentence initial	Capital.	Quoted	Unquoted	Adjacent	Graph. n/ compl.	Graph. compl.
Llama3	32.14	32.11	32.30	44.34	27.97	35.48	38.71	41.67	31.25	75.76	28.50	18.75	33.30	25.76
CRF	43.04	34.75	81.86	36.39	45.30	51.61	24.73	8.33	12.50	74.75	40.39	25.00	46.09	26.26
BETO	81.32	79.51	89.82	83.79	80.48	93.55	66.67	41.67	25.00	97.98	79.93	18.75	83.72	68.18
mBERT	79.38	77.43	88.50	84.10	77.77	90.32	74.19	33.33	31.25	96.97	77.91	12.50	82.98	59.60
BiLSTM (unadapted)	80.16	77.71	91.59	82.87	79.23	74.19	62.37	41.67	18.75	95.96	78.84	31.25	83.26	63.13
BiLSTM (codeswitch)	81.79	79.79	91.15	83.79	81.11	70.97	63.44	33.33	50.00	93.94	80.78	12.50	85.37	62.12

Table 8.3: Recall obtained by each model on different span dimensions.

Model	Overall precision	Unseen	Seen	Multi-word	Single-word	Type confusable	Unseen tagging	Sentence initial	Capital.	Quoted	Unquoted	Adjacent	Graph. n/ compl.	Graph. compl.
Llama3	49.17	44.50	96.05	51.79	47.86	44.00	14.69	14.29	02.23	54.74	48.08	75.00	58.77	22.77
CRF	77.89	75.26	83.71	73.01	79.34	53.33	58.97	25.00	13.33	81.32	77.38	40.00	78.65	71.23
BETO	85.03	83.20	93.55	84.57	85.19	93.55	71.26	76.92	36.36	92.38	84.34	21.43	86.75	75.00
mBERT	88.08	87.98	88.11	86.75	88.48	62.22	75.82	61.54	55.56	92.31	87.58	18.18	89.75	76.62
BiLSTM (unadapted)	90.35	89.26	94.95	88.27	91.12	88.46	79.45	90.91	50.00	90.48	90.34	45.45	91.05	85.62
BiLSTM (codeswitch)	90.14	89.42	93.21	88.10	90.88	68.75	76.62	72.73	72.73	93.00	89.87	50.00	90.71	86.01

Table 8.4: Precision obtained by each model on different span dimensions.

8.7 Summary

In this chapter we have hypothesized three possible explanations for the unexpected results we obtained in chapter 7, when very different evaluation approaches (strict span evaluation and the token & separator model) produced virtually the same ranking of models on COALAS test set. After ruling out two of our hypotheses, we decided to further investigate the nature and type of spans contained in the COALAS test set and look into the models' performance on them.

In order to do so, we defined a set of formal dimensions that can help characterize the spans in a dataset (both in terms of shape, context and in relation to their presence in training set) and check how our models from chapter 4 performed across them.

The obtained results paint a very different picture to the F1 score ranking from chapter 4. The dimension-aided analysis uncovered limitations in the dataset and systematic errors in the models' performance that had gone unnoticed in the standard evaluation and error analysis we performed on chapter 4. The best-performing model on chapter 4 (the BiLSTM-CRF model with codeswitch embeddings) obtained mediocre results in some of the dimensions, while BETO and Llama3 (that had ranked last on the standard evaluation) produced competitive results on some types of spans, such as spans in sentence initial position or with quotation marks.

Chapter 9

BLAS: A Benchmark for Loanwords and Anglicisms in Spanish

Prediction is very difficult, especially about the future.

Niels Bohr, physicist

In this chapter we introduce BLAS, a test-only dataset to evaluate models in terms of recall on the task of identifying anglicisms in Spanish text. The design behind BLAS is linguistically-guided and informed by our findings from chapters 7 and 8.

As a test set, BLAS is rather atypical. Instead of being a large collection of real-world in-distribution scraped data, BLAS is a small set of linguistically-motivated examples made by hand that aim to exhaustively cover the different combinations of dimensions we introduced in chapter 8. Each sentence in BLAS is characterized by certain linguistic attributes, so that sentences with the same attributes can be grouped, thus providing a minimal but fine-grained evaluation of the performance of the model across different linguistic properties.

Our aim for BLAS is for it to serve as an evaluation resource that not only provides a score based on performance among competing systems, but that also assists researchers identify the strengths and weaknesses of a each model on certain linguistic phenomena.

In that regard, BLAS is an exhaustive, linguistically-motivated testbed that may help

practitioners identify which scenarios a model is good at (thus diagnostic), which model is the best choice for a given scenario (thus actionable) and how the model may perform on a different distribution (thus predictive).

9.1 Motivation

9.1.1 What if...?

In chapter 3 we described our motivation behind the creation of the COALAS corpus: we wanted to assess whether models were capable of retrieving previously unseen borrowings from a diverse set of journalistic sources, a feature that is key for lexicographic work but that was not ensured by the previously available dataset from Álvarez Mellado (2020a).

In chapter 8 we discovered that, as OOV-rich as the COALAS test set may be, there were certain linguistic phenomena and contexts that were under-represented in the corpus. Moreover, the dimension analysis from chapter 8 unmasked that models (even the best-performing ones) tended to perform poorly across some of these phenomena, but that these poor results went unnoticed when they were aggregated into holistic evaluation metrics.

This raises an interesting what-if question: let's suppose that any of our models from chapter 4 was confronted with a new dataset. This new imaginary dataset would be similar in terms of genre to the COALAS corpus the models were trained and evaluated on (i.e., Spanish journalistic texts rich in anglicisms). If the distribution of the dimensions in this new dataset was analogous to the distribution of dimensions in the COALAS test set (in terms of span length, graphotactical compliance, casing, sentence position, etc.), we could anticipate that the results obtained by our models on this new imaginary dataset should be in line with the results we obtained on the COALAS test set.

But what if the distribution of the dimensions in this new dataset was different? What if the new dataset contained nothing but anglicisms that were capitalized, collocated with other borrowings or in sentence-initial position? What if the vast majority of anglicisms in the new dataset were graphotactically compliant? If that were the case, we could safely

venture that the overall scores obtained by our model on the COALAS test set would not be a reliable indicator of its performance on the new dataset: with high probability, the scores obtained on this new imaginary dataset would be lower than the one obtained on COALAS, because as our analysis from chapter 8 revealed, those precise span attributes (being sentence initial, graphotactically compliant, capitalized or adjacent) tended to produce lower recall scores. The same logic applies if the case was the other way round: a dataset that consisted exclusively on previously seen, non-compliant, mid-position borrowings written with quotation marks would probably yield a higher score than the overall score obtained by our models on the COALAS test set.

This matter points to a larger lingering issue that plagues evaluation in NLP: the fact that, in general, the score obtained by a system on a given test set does not anticipate how that very same model will perform on a different sample. Excellent performance on a given test set does not necessarily indicate excellence performance on the task in general.

9.1.2 Limitations of the dimension-based evaluation from chapter 8

We could be tempted to claim that we already carried out an evaluation of our models based on linguistic dimensions in chapter 8. Results from chapter 8 indeed gave us an interesting insight into which linguistic attributes our models are good at and which aspects they fail to tackle. Our experiments across dimensions, however, had some limitations.

First, as our dimension analysis showed, some linguistic attributes were under-represented in the COALAS test set: adjacent borrowings, capitalized borrowings and sentence-initial borrowings were almost non-existent in our test set. If the number of spans with those characteristics is too low, the results obtained by our models on those particular type of phenomena may not be a reliable predictor of how the models perform on those phenomena.

But more importantly, the main downside to our analysis from chapter 8 is that whenever we were evaluating our models on a given dimension, we could not really be sure if the successes and errors of a given model on the instances under a given dimension were actually

being caused by other dimensions of the data, or if we would have obtained other results had any of the dimensions been any different. For example, if our model successfully retrieved an unseen borrowing, would that be proof that the system is good at retrieving previously unseen borrowings? Or could that have been caused or affected by the fact that the borrowing in question had quotation marks? Would it still have succeeded if the very same borrowing in the very same sentence was capitalized? Similarly, a mistake when retrieving a multiword borrowing could perhaps be caused by the fact that it appeared at the beginning of the sentence. That missed borrowing, could it perhaps had been retrieved if it appeared in mid-sentence position instead? Would the result be any different if it had quotation marks?

In a nutshell, the experiments across dimensions we ran in chapter 8 cannot fully inform us about the weaknesses and strengths of our models, because various dimensions may be in place at the same time and the interference across them will prevent us from actually knowing whether the model can retrieve certain types of borrowings or not and how different dimensions collide.

These questions lead us to the following conclusion: we lack a reliable resource for evaluating lexical borrowing detection systems in Spanish that can really inform us about what the capabilities of a model are and that can anticipate how the model will perform when confronted with a new dataset.

9.1.3 Towards diagnostic, actionable and predictable benchmarking in NLP

Dataset creation is a crucial aspect of NLP research: the availability of annotated data guides the progress in the field and the way datasets are constructed even shapes the way we frame and approach tasks (Paullada et al., 2021).

Various factors may influence the dataset creation process (language, annotation resources, etc.). More frequently than not, the availability of the data itself (the existence of previously compiled corpora or scrapable texts) is a major factor in dataset creation and determines its content and the linguistic phenomena that gets represented in it (Plank, 2016)

and, in consequence, the type of phenomena that models will be evaluated on.

When a dataset for a given task is released, systems begin to be trained and evaluated on it. Certain datasets become the standard test set against which models are evaluated, thus becoming a benchmark for the task, such as CoNLL03 (Tjong Kim Sang and De Meulder, 2003) or OntoNotes (Pradhan et al., 2013) for NER. Making progress on that given task now implies obtaining better scores on a particular test set. In other words, solving a general task may now be restricted to making systems capable of modeling the phenomena that appear in one specific test set. This means that the phenomena that our modeling efforts will consider will be heavily influenced by the linguistic attributes that our text sample (the dataset) contained to begin with.

After reaching a certain degree of progress, improvement over scores may become elusive: more work may be required to produce increasingly smaller improvements. Scores can eventually plateau, thus producing a glass ceiling situation (Stanislawek et al., 2019). Researchers and practitioners try to grasp what there is left to improve for that given task and look at the errors that their systems are making on the test set in order to identify the type of examples their models are failing at (Rueda et al., 2024). With the development of new techniques, significant improvement on the scores may finally be reached. If the scores are high enough so that there is little room for improvement, the dataset can be considered saturated and sometimes the task will even be claimed to be solved (Kiela et al., 2021).

Eventually, a mismatch between the optimistic scores obtained by models in lab conditions over benchmark data and the results obtained by those same models on real-world data will be identified (Lin et al., 2020). As a result, patches will be explored to account for the type of naturalistic data that models are struggling with (Heigold et al., 2018; Ribeiro et al., 2018; Namysl et al., 2021). Finally, a new and more challenging dataset will be released for the task, and the process starts again.

This Sisyphus-esque approach is the usual modus operandi in NLP development. This manner of working, however, poses certain issues.

First, under this paradigm models are trained and tested over different sets that, although assumed to be independent, are similarly sampled and identically distributed (Hupkes et al., 2023). Evaluation over an identically-distributed sample may be suitable in certain restricted or predictable scenarios, but it has been shown that its results can overestimate performance and lead to poor generalization (Gorman and Bedrick, 2019). Moreover, calling a task "solved" because outstanding scores were obtained on a split of a dataset where most of the data is alike and where the presence of worst-case scenarios or linguistic variations of the phenomenon in question is not ensured may be too optimistic. Unsurprisingly, minor perturbations that pose no challenge for humans (such as casing or out-of-domain topics) can catastrophically affect performance on allegedly solved tasks (Gururangan et al., 2018; Mayhew et al., 2019), which points to the lack of robustness of these models (Ma et al., 2023).

As a result, some voices have advocated for building adversarial test sets, that is, populating test sets with the most challenging cases possible (Søgaard et al., 2021). As useful as they may be, adversarial test sets are not a silver bullet. It is indeed important to know that a certain model succeeds on the worst case scenarios. But evaluating exclusively on these types of extreme cases will prevent us from actually learning what a model can do: if tested on challenging cases that fool most models, a model that is good but far from perfect may perform as poorly as a dummy baseline. An evaluation that aims to be useful and actionable should be informative about the models' capabilities, as well as granular enough that it can help us distinguish between a bad model, a mediocre model, a good model and a great model, something that exclusively evaluating on the most difficult cases or on the worst case scenarios cannot guarantee (Rodriguez et al., 2021).

Even if the test set is balanced in terms of the difficulty of the examples it contains, model comparison based on a score over the whole test set may be short-sighted too, because by doing so we are losing crucial information: the type of data that the model is succeeding at. Let's say that we have a task for which state-of-the-art models achieve a recall of 80%.

A model that only scores 30% will be deemed mediocre and probably irrelevant. But what if the data that the apparently mediocre model is succeeding at includes precisely the 20% of data that SOTA models are failing at? That would be a remarkable achievement that could illuminate the areas of research future work should concentrate on, but we have no way of knowing that because in score-based evaluation is the overall score that counts, not the section of data it was achieved on.

In order to better contextualize the information that is portrayed by numeric scores only, qualitative analysis are usually performed, generally through error analysis. Error analysis usually consists in identifying the type of examples where models are failing and consequently establishing areas that require further work. Although it may seem counterintuitive, focusing exclusively on errors as a way of deciding what areas of modeling require improvement may not be the best solution either: let's imagine that we have an NER system and that we are focusing strictly on the errors the system made. Let us suppose that 20 of the entities that the system failed to retrieve where of type `PER`, and only 10 were `ORG`. We could be tempted to say that the model is performing worse on `PER` entities, but we could only claim that if the amount of `PER` and `ORG` entities in the test set was balanced to begin with. But if the test set contained overall, let's say, only 12 `ORG` entities and 70 `PER` entities, then the picture is quite different. Error analysis can only be a productive approach towards the improvement of models if it is accompanied by a *success analysis* that contextualizes how prevalent certain phenomena are (Wu et al., 2019). Otherwise we risk the possibility of overstating the importance of some errors and minimizing others, which may lead to misguided decisions.

In a nutshell, it is a wide-known (yet unsolved) fact in NLP that standard evaluation based on providing a holistic score averaged over a test set (that, more often than not, will be identically distributed to the data contained in the training set) does not tell us much about what the model can and cannot do, which areas of the problem require further work, or how the model will generalize to a different distribution.

9.2 Benchmark creation

In this section we describe the creation of BLAS, a test-only dataset to evaluate models on anglicism detection in Spanish that accounts for the limitations we have just exposed. The methodology behind the creation of BLAS is quite different to the usual modus operandi in test set creation. The key is to create a test set, that, contrary to common practice, does not rely on gathering large amounts of real-world in-distribution scraped data, but consists in handcrafting a small set of linguistically motivated examples that exhaustively cover the range of span attributes (in terms of shape, length, casing, sentence position, etc.) a system may encounter in the wild. Although we focus our proposal on anglicism detection, our methodology could also be relevant for other span identification tasks.

9.2.1 Benchmark desiderata

We now list the desired properties that our test set should satisfy. Our test set should be:

- Linguistically driven: the test instances in BLAS will be selected and grouped based on the linguistic attributes that we aim to evaluate.

- Dimensionally balanced: the resulting collection of examples in BLAS should cover the diversity of dimension values a borrowing detection system may encounter in the wild (casing, quotations, etc).

- Diagnostic: scores on the benchmark should help us identify which areas need more work or which linguistic phenomena are beyond the current capabilities of the model.

- Actionable: scores on the benchmark should inform decision-making when choosing which model is the most appropriate one.

- Predictive: results over the benchmark should help us anticipate how the model will perform on new unseen data.

- Model-independent: our benchmark aims to serve as an evaluation resource for current and future models. Consequently, our benchmark should not assume that the model under evaluation was trained on a particular training set or presuppose a modeling approach (machine learning based, lexicon lookup based, LLM).

- Exhaustively evaluated: as we saw in chapter 7, different evaluation approaches are suited for different use case scenarios and no evaluation metric can paint the full picture. Consequently, evaluation on our benchmark should include different perspectives (span level, token level) so that results are informative and actionable.

- Recall-oriented: in section 7.1 we identified recall as a crucial metric for technology that seeks to assist lexicographic work, while also being the Achilles' heel of all borrowing detection systems. With BLAS, we want to evaluate the performance of models at retrieving different types of borrowings appearing in different types of contexts. Thus, and although not exclusively, performance over recall will be central.

- Permisive licensing: in order to guarantee reusability and facilitate resharing.

- Free of data contamination issues: the benchmark should be created in such a way that we can ensure that no model has previously been exposed to its data (Sainz et al., 2023).

9.2.2 Span attribute selection

The first step of our methodology for the data selection process is to define which span attributes should models be evaluated on (and thus be represented in our dataset). This way of proceeding radically diverges from the usual data-centric approach, in which the readiness of data (and not its linguistic characteristics) guides the data selection process.

Our aim is to propose an evaluation test set that will help us assess whether a lexical borrowing detection model could correctly identify borrowings, exhaustively exploring the different configurations in which a borrowing may appear (quoted, unquoted, graphotactically compliant or not, capitalized, mid-sentence, multiword, adjacent, ambiguous, etc.).

Informed from our dimension analysis from chapter 8, these are the main attributes that our test set will asses:

– Span length: being single-word or multiword (*streaming* vs *machine learning*).

– Span shape: complying with the graphotactic expectation of the recipient language or not (*pipeline* vs *healthy*).

– Sentence position: appearing mid sentence or in sentence initial position.

– Quotation marks: being surrounded by quotation marks or not.

– Casing: being lowercase, uppercase, titlecase, etc.

– Adjacency: being collocated to another span (as in *look total black*).

– Ambiguity: the same sequence can be found as a span or not, depending on the context (the word *pie* will be an anglicism when talking about dessert as in *un pie de limón*, but it will be a native Spanish word when it means "foot").

These will be the linguistic attributes that BLAS will evaluate.

9.2.3 Data curation

We populated our test set with sentences that exhaustively cover the different combinations of the span attributes we listed in section 9.2.2 (for example, non-ambiguous multiword, graphotactically-compliant uppercase anglicisms that appear in sentence initial position between quotation marks, etc.). Finding examples from real sources that covered all combination of attributes would have been an unmanageable endeavor, as anglicisms are a sparse phenomenon that amount to 1%-2% of the language (Rodríguez González, 2002).

Consequently, instead of populating our benchmark with real-world examples, we decided to have such examples written from scratch, inspired on real examples from journalistic sources (see appendix D for detailed guidelines on the writing process and anglicism

selection). This approach, although unconventional, offered several benefits. First, instead of being confined to evaluating systems on the data that is available on the web, we gained full control of the linguistic attributes that our dataset would cover. Creating our examples from scratch also avoided copyright issues that usually prevent datasets from being freely reshared. Finally, the fact that the sentences were new and not gathered from any other existing sources ensured that no model has previously been exposed to the data contained in the dataset, thus avoiding data contamination issues (Sainz et al., 2023).

The mechanics behind the creation of BLAS consistently followed the same procedure for all examples (see figure 9.2):

1. Identifying the span attributes we wanted to evaluate (being graphotactically compliant, being ambiguous, being multiword, etc.) (see section 9.2.2).

2. Curating a minimal but varied list of ten anglicisms that satisfied those attributes (see appendix D).

3. Writing sentence examples for each of the selected anglicisms in different sentence positions (see appendix D).

All examples in BLAS are original and were written from scratch, although some of them were inspired by real world examples and all of them try to mimic the journalistic style of the Spanish newspapers and media.

The result of this process is a collection of 153 annotated sentences that contain 173 spans labeled as anglicisms (111 of which are unique). Table 9.1 reflects number of sentences per linguistic attribute. In order to assess their quality and validity, the 153 sentences were given to a second annotator for double annotation. The inter-annotation agreement computed by Cohen's kappa at token level was 0.98 (0.99 of pairwise F1 score at span level), which indicates annotation reliability (Artstein and Poesio, 2008).

Type	Length	Position	# sentences
Compliant	Single-word	Initial	10
		Mid	10
	Multiword	Initial	10
		Mid	10
Non-compliant	Single-word	Initial	10
		Mid	10
	Multiword	Initial	10
		Mid	10
Mixed compliant	Multiword	Initial	10
		Mid	10
Ambiguous	Single-word	Mid	3
	Multiword	Mid	10
Mixed ambiguous	Multiword	Initial	10
		Mid	10
Adjacent	Single-word	Mid	10
	Multiword	Mid	10

Table 9.1: Number of sentences in BLAS per span type.

9.2.4 Punctuation transformations

The 153 sentences resulting from the writing process we have just described underwent a series of systematic perturbations. These perturbations revolved around two aspects: casing and quotation marks.

For casing, we considered the following perturbations:

- Standard casing (no perturbation added, text as is).

- all words are lowercase.

- ALL WORDS ARE UPPERCASE.

- All Words Are Capitalized.

- Only BORROWINGS are in uppercase.

- Only Borrowings are capitalized.

For quotation marks, the following perturbations were considered:

- All “borrowings” are written between quotation marks.

- No borrowing is written with quotation marks.

This amounts to a total of twelve different configurations (six casing transformations with quotation marks, six without them). Table 9.2 shows the twelve possible variations for every sentence in BLAS. Each sentence in the original collection of 153 sentences underwent these twelve transformations. This means that our dataset contains twelve variations of every sentence, where the sentence is the same but written with different casing and quotation marks configurations.

9.2.5 Benchmark description

The result of the process we have just described is BLAS (Benchmark of Loanwords and Anglicisms in Spanish), a collection of 1,836 annotated sentences in Spanish (37,344 tokens) that contains 2,076 spans labeled as anglicisms. Those 1,836 sentences are repetitions of 153 sentences, but with different quotation marks and casing configurations. Every sentence in BLAS contains one span labeled as anglicism, except for the sentences with adjacent spans, which contain two.

Figure 9.1 illustrates dimensions in BLAS. Because we are not assuming a given training set, we disregard novelty dimensions of BLAS. Instead, we consider an additional intrinsic dimension, *ambiguity*, that quantifies the amount of anglicisms in BLAS that are ambiguous with a native (not borrowed) word in Spanish (borrowings such as *pie*, which would amount to unseen tagging or type-confusable tagging under the novelty dimensions).

Sentences in BLAS were created ad hoc by hand by a linguist inspired on real examples from journalistic sources. Sentences are characterized by the linguistic attributes of the span

Quotation	Casing	Example
No	Standard	*Hacer burpees es un ejercicio efectivo para perder peso en poco tiempo.*
No	All words are lowercase	*hacer burpees es un ejercicio efectivo para perder peso en poco tiempo.*
No	All words are uppercase	*HACER BURPEES ES UN EJERCICIO EFECTIVO PARA PERDER PESO EN POCO TIEMPO.*
No	All words are capitalized	*Hacer Burpees Es Un Ejercicio Efectivo Para Perder Peso En Poco Tiempo.*
No	Borrowings are uppercase	*Hacer BURPEES es un ejercicio efectivo para perder peso en poco tiempo.*
No	Borrowings are capitalized	*Hacer Burpees es un ejercicio efectivo para perder peso en poco tiempo.*
Yes	Standard	*Hacer "burpees" es un ejercicio efectivo para perder peso en poco tiempo.*
Yes	All words are lowercase	*hacer "burpees" es un ejercicio efectivo para perder peso en poco tiempo.*
Yes	All words are uppercase	*HACER "BURPEES" ES UN EJERCICIO EFECTIVO PARA PERDER PESO EN POCO TIEMPO.*
Yes	All words are capitalized	*Hacer "Burpees" Es Un Ejercicio Efectivo Para Perder Peso En Poco Tiempo.*
Yes	Borrowings are uppercase	*Hacer "BURPEES" es un ejercicio efectivo para perder peso en poco tiempo.*
Yes	Borrowings are capitalized	*Hacer "Burpees" es un ejercicio efectivo para perder peso en poco tiempo.*

Table 9.2: Casing and quotation mark perturbations considered in BLAS.

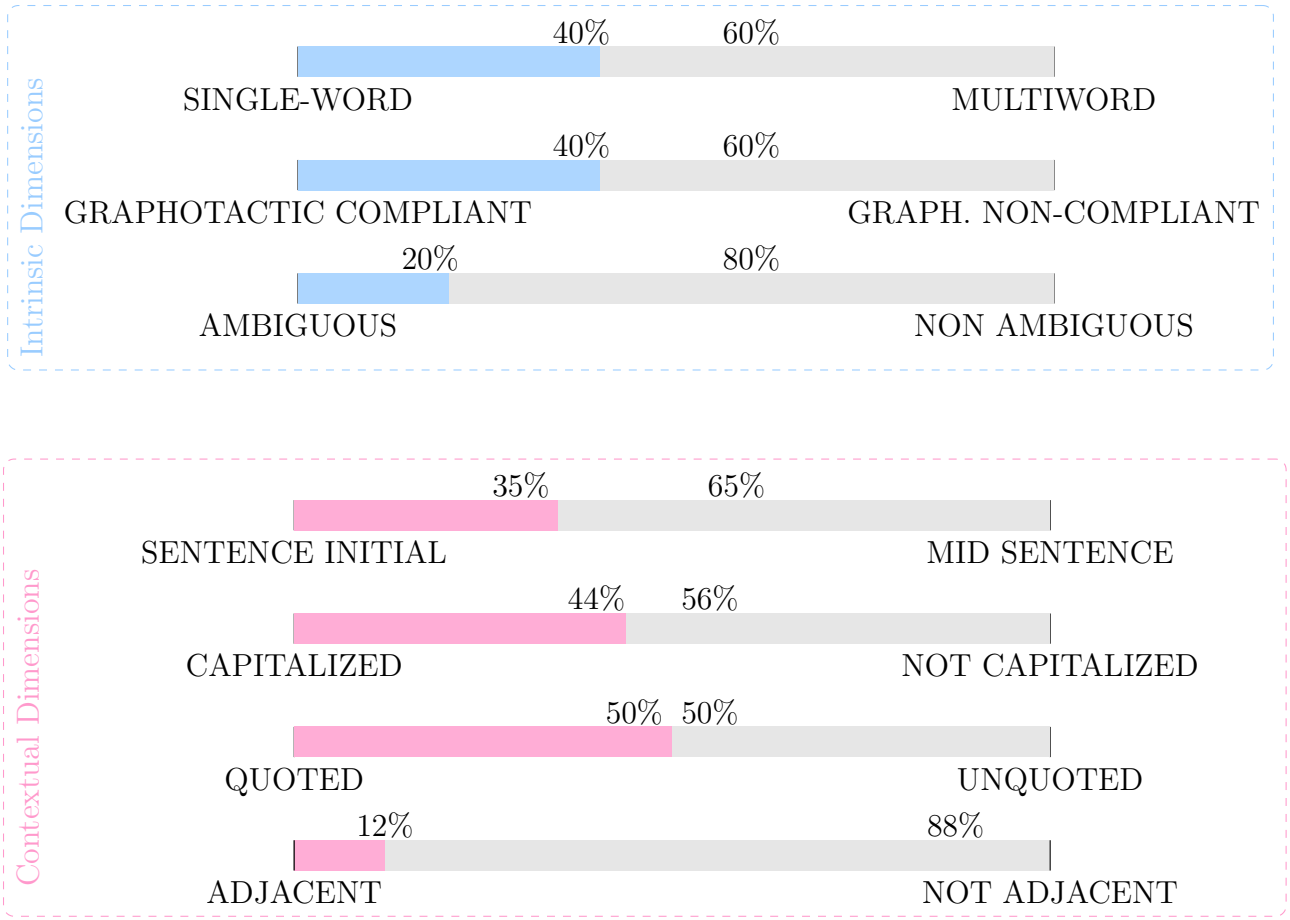

Figure 9.1: Ratio of spans per dimension in BLAS.

it contains: a group of sentences contains multiword anglicisms that are non-compliant and that appear mid sentence, another group of sentences contain the same multiword anglicisms but appearing in sentence initial position, etc. The sentences were created so that these groups with different attributes are balanced within the benchmark. This way, models can be evaluated on these sentences, and scores can be computed over sentences that share the same dimension attributes. This approach facilitates evaluating models on different dimension configurations separately.

Figuratively, this evaluation approach is inspired by the approach of unit test evaluation, in which a suite of tests seek to thoroughly evaluate a single aspect of the code. In a similar fashion, each group of sentences in BLAS seek to test how well models perform over different aspects of borrowing usage. A poor result over a given collection of sentences will give us

information on a concrete aspect of borrowing usage that the models find challenging and therefore are failing to capture.

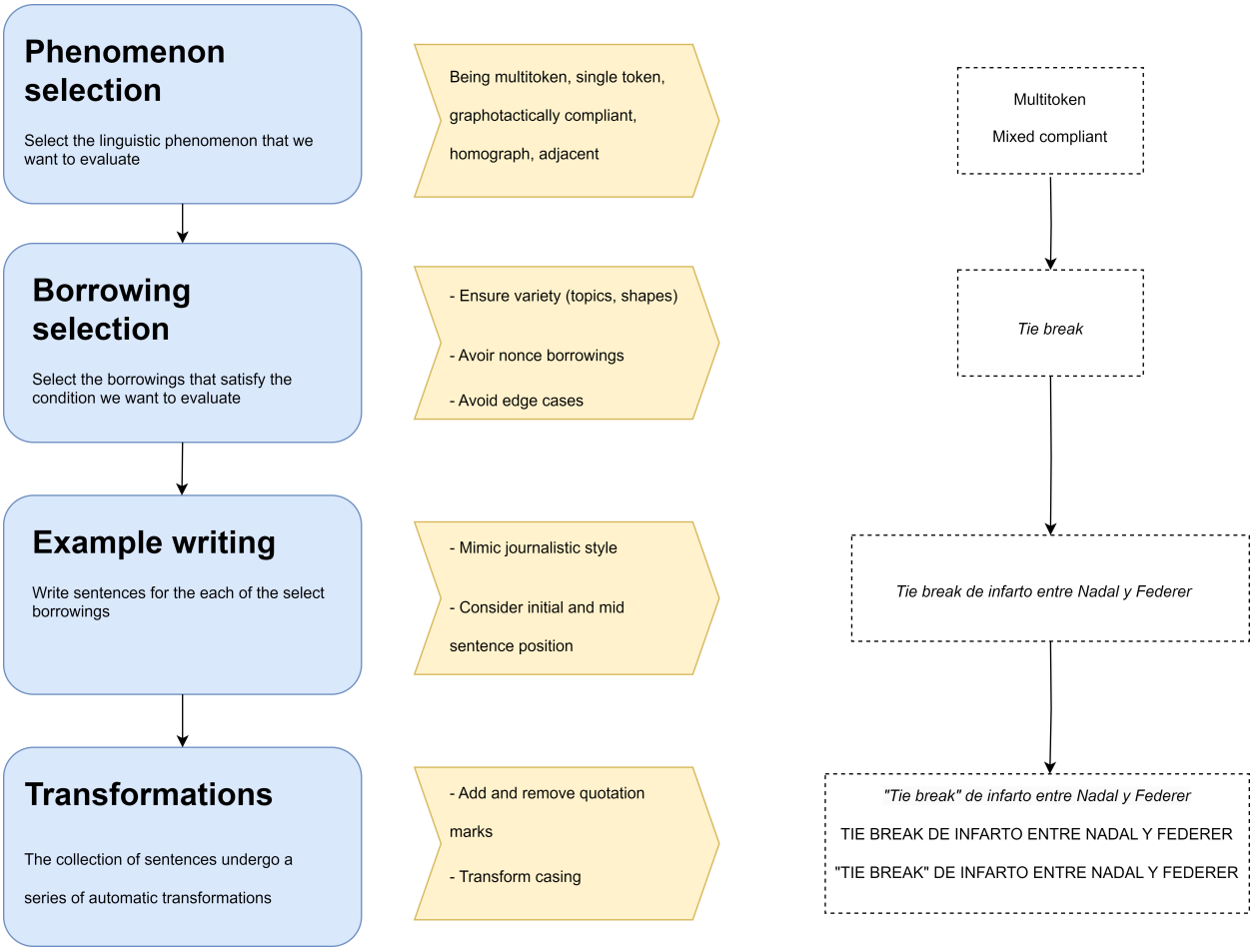

Figure 9.2: Creation process behind each of the splits in BLAS.

9.2.6 Evaluation metrics

In chapter 7 we explored some of the existing metrics for sequence labeling tasks. We analyzed them, identified some of their strengths and weaknesses and calculated the correlation among them. Based on our investigations, we distrust benchmarks that base all their evaluation on a single evaluation approach. Consequently, for BLAS we provide three evaluation paradigms:

1. Span based evaluation based on strict match (precision, recall, F1 score).

2. Token & separator evaluation model by Esuli and Sebastiani (2010) (precision, recall, F1 score).

3. An extended version of MUC error typology (see description below).

The rationale behind this choice was the following:

Span-based evaluation over strict matches is the most widely used metric in span identification tasks such as NER and MWE identification. In addition, span-based evaluation is a good choice for scenarios where strict matches (that is, where both the boundary and the type must be correct) are important. On the other hand, the token and separator model allows for partial matching (a feature that span-based evaluation lacks), while correctly accounting for mid-token retrieval or adjacent borrowings (unlike token evaluation over BIO tags and IO tags, respectively). We therefore conclude that providing both evaluation paradigms (six metrics in total) ensures a wide coverage as it provides both span-level perspective and token-level perspective.

In the next subsection we develop why we decided to also provide the error typology based on MUC and the extended version of it that we propose for BLAS.

An extended version of MUC error typology

As we saw in chapter 8, MUC error typology (Chinchor and Sundheim, 1993) is a classification that classifies erroneous spans in NER (or in any other span-identifications tasks) in the following types of errors: MISSED, SPURIOUS, OVERLAP and TYPE.

We find MUC typology to be a convenient addition to the traditional span evaluation and token evaluation, as it enables a better more informative error analysis. MUC error typology has the drawback of being less comparable among systems than standard metrics, but given that we are also providing evaluation metrics on span level and token-separator level, we consider that it can be a valuable addition. However, the error types defined in MUC are insufficient to account for certain mislabeling errors that we explicitly want to identify with BLAS, namely ill-segmented spans and differences across partially overlapping

spans. To that end, we propose to revisit the MUC error typology and extend it to cover some of the casuistic that was not included in the original typology by making the following additions:

1. Under MUC typology, all partial matches are considered as OVERLAP. However, we may distinguish between three types of overlapping spans: missing overlapping spans (where one or more gold standard tokens are missed), exceeding overlapping spans (where one or more tokens in the predicted spans are spurious) and a combination of both (the predicted span overlaps with the gold standard, but one end contains spurious tokens while the other end has missing tokens). These differences point to more general weaknesses in the models' performance: a predominance of missing overlapping spans probably indicates a more general trend of recall underperformance, while systems with a high number of exceeding overlapping spans will probably tend to perform worse across precision. Including all overlapping spans under the same category will prevent us from actually knowing what the main weaknesses of the models are and where we should focus our efforts. Consequently, we propose to subdivide the classic OVERLAP error from MUC into 3 types of errors: OVERLAP MISSING, OVERLAP SPURIOUS and OVERLAP MISSING+SPURIOUS.

2. A 2-token span that is retrieved by the system as two separate spans of length 1 each is nor missing, nor spurious, but poorly segmented. To account for these cases, we propose to add the categories SPLIT (a multiword span was retrieved as multiple shorter spans), FUSED (two adjacent spans where retrieved as a single longer span) and MISSEGMENTED (two adjacent spans were retrieved as two separate spans, but the boundary between them was incorrect and some tokens of one of the spans were wrongly assigned to the other span)

Table 9.3 displays the full list of errors inspired by MUC typology that we will be considering for BLAS.

Type	**Definition**	**Example**	**Is in MUC?**
Correct	The predicted span and the gold standard span match	*estampados tipo [patchwork]$_{ENG}$*	Yes
Missing	A span in the gold standard is not found in the prediction	*una noticia fake*	Yes
Spurious	A span in the prediction is not found in the gold standard	*la escritora [bell hook]$_{ENG}$*	Yes
Type	The span in the prediction has a different label than the span in the gold standard	*una [boutique]$_{ENG}$*	Yes
Overlap missing	The predicted span partially matches the gold standard span, but at least 1 token is missing	*estampado animal [print]$_{ENG}$*	As overlap
Overlap spurious	The predicted span partially matches the gold standard span, but at least 1 token is spurious	*con [escote halter]$_{ENG}$*	As overlap
Overlap spurious+missing	The predicted span partially matches the gold standard span, but at least 1 token is missing AND at least 1 token is spurious	*un baby [shower espléndido]$_{ENG}$*	As overlap
Split	One multiword span from the gold standard was retrieved as two adjacent shorter spans	*tecnología de [machine]$_{ENG}$ [learning]$_{ENG}$*	No
Fused	Two adjacent spans in the gold standard were retrieved as one long span	*llevaba un [look total black]$_{ENG}$*	No
Missegmented	Two adjacent spans in the gold standard were retrieved as two adjacent spans, but the boundary between them was wrong	*llevaba un [look total]$_{ENG}$ [black]$_{ENG}$*	No

Table 9.3: Extended version of the MUC error typology considered for BLAS.

9.3 Experiments & results

In this section we report results on the different experiments we conducted on BLAS. We first report results obtained by our models from chapter 4 on BLAS (section 9.3.1); we then evaluate the ability of BLAS to predict models performance on outside data (section 9.3.2); and finally we assess the statistical significance of our obtained results across dimensions (section 9.3.3).

Our analysis will focus on span level evaluation. Full report of metrics for all experiments (including recall, precision and F1 score at span and token and separator level) can be found in appendix E.

9.3.1 Experiments across models

The BLAS test set that resulted from section 9.2 was used to evaluate our six models from chapter 4: the CRF, Transformer-based BETO and mBERT, both BiLSTM-CRFs with different types of word embeddings (unadapted and codeswitch), and 8B-Llama3. Because BLAS is a test-only dataset, the models were not trained on any split of BLAS, but only evaluated on it.

As all sentences in BLAS are positive examples (they all contain at least one span), our analysis of results will focus on recall. As we will see in section 9.3.2, recall will also serve as a reliable proxy to F1 score performance.

Overall performance

Table 9.4 displays overall recall scores obtained by the six models on BLAS and compares them to the our previous scores on the COALAS test set (overall, and over the `ENG` label). Scores on BLAS range between 6% and 36%. These scores were obtained over the whole collection of 1,836 sentences in BLAS, which include all punctuation transformations.

Two facts stand out from these results. First, these results are far below our scores from

	BLAS		**COALAS** (ALL)		**COALAS** (ENG)	
Llama3	**36.37**	#1	32.14	#6	33.33	#6
mBERT	23.80	#2	76.16	#3	78.85	#3
BiLSTM-CRF (cs)	23.75	#3	**78.72**	#1	**81.32**	#1
BETO	23.55	#4	75.50	#4	77.99	#4
BiLSTM-CRF (un)	23.55	#5	78.34	#2	80.88	#2
CRF	6.50	#6	43.04	#5	44.31	#5

Table 9.4: Span recall per model on BLAS and COALAS.

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	21.97	7.51	**68.79**	45.09	61.27	**49.13**	66.47	45.09	60.12	46.24	66.47	23.12
Text is lowercase	31.79	16.18	87.86	56.65	83.82	65.32	**90.17**	61.27	80.92	**68.21**	67.63	27.75
Span is titlecase	0.00	0.00	5.20	5.20	4.62	5.78	6.36	5.78	5.78	6.36	**56.65**	**24.86**
Text is titlecase	0.00	0.00	7.51	5.78	10.40	5.20	4.05	2.89	6.94	6.94	**32.95**	**9.25**
Span is uppercase	0.00	0.58	0.00	0.58	0.00	0.00	0.00	0.00	1.16	2.31	**54.91**	**45.09**
Text is uppercase	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.58	0.00	0.00	**21.97**	**5.78**

Table 9.5: Overall span recall on BLAS per model on different casing and quotation mark configurations.

chapter 4 on COALAS test set, where the best performing model obtained an overall recall of 78% (and even reached a recall of 81% on ENG spans). This gap in recall proves that with BLAS we have created a test set that is more rigorous and challenging in terms of recall than the COALAS test set for the task of detecting anglicisms in Spanish. This increase in difficulty comes from carefully selecting the examples in the dataset based on linguistic criteria. If we take BLAS as an exhaustive collection of examples for evaluating the ability of models to retrieve anglicisms from a variety of linguistic shapes and contexts, then we can conclude that anglicism retrieval is far from being a solved task in Spanish, at least for the models we have explored.

Second, the best-performing model on BLAS is Llama3 (R=36.37), well above all the other models. This is remarkable, as Llama3 ranked last when tested on COALAS, even below the CRF. On the other hand, the BiLSTM-CRF with codeswitch embeddings (that ranked first when tested on COALAS), ranks third when tested on BLAS. In fact, results over BLAS and over COALAS show very low correlation, both in terms of recall scores and ranking: -0.04 for the recall scores, and -0.08 for the rankings.

Performance across punctuation

Holistic results such as the ones reported on table 9.4, however, do not tell us much about the type of examples models are failing to retrieve, or how to fix them. Table 9.5 displays the same results than table 9.4, but split into the different casing and quotation mark configurations of BLAS.

Two facts stand out from table 9.5. First, all models obtained better results on the versions of the dataset where the spans were written between quotation marks. The differences in recall between examples with quotation marks and without them are substantial, with scores dropping even around 20 points or more for some models. Our results reveal that Llama3 was by far the most vulnerable to quotation mark absence, with a loss of more than 40 points in recall when quotation marks were not present on the standard casing split. Note that these differences in scores are fully attributable to the absence of quotation marks, because both splits contain exactly the same sentences, but in one split the spans are surrounded by quotation marks, while in the other they are not.

Second, supervised models consistently obtained catastrophic results on the capitalized sections of BLAS, with models scoring zero across uppercase configurations. Llama3, on the other hand, while it still saw a drop in performance over the capitalized sections of BLAS (especially when the whole text was affected), produced *less* bad results than the supervised models, which explains its good position in the overall ranking from table 9.4.

The results from table 9.5 provide us with actionable conclusions that can inform decision-making when choosing which system to deploy for anglicism detection on different scenarios: if we are expecting well-edited text with standard casing and quotation marks, then BETO would probably be the best model. If quotation marks are not expected, then mBERT would be a better call.

The casing vulnerability points to an issue that future work should address: none of these models is reliable when applied to texts with non-standard capitalization. However, if we had to choose one model for such scenario, Llama3 would be our safest bet. On the other

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	40.00	**50.00**	20.00	40.00	20.00	40.00	30.00	30.00	**90.00**	20.00
		Mid	20.00	0.00	**100.00**	80.00	90.00	80.00	**100.00**	80.00	90.00	**90.00**	80.00	30.00
	Single	Ini	0.00	0.00	10.00	**20.00**	0.00	10.00	20.00	**20.00**	10.00	**20.00**	**50.00**	10.00
		Mid	10.00	0.00	70.00	50.00	80.00	80.00	90.00	80.00	**100.00**	**90.00**	40.00	10.00
Non-compliant	Multi	Ini	10.00	10.00	**90.00**	50.00	60.00	**70.00**	30.00	40.00	20.00	**70.00**	**90.00**	30.00
		Mid	60.00	10.00	**100.00**	80.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	90.00	40.00
	Single	Ini	0.00	0.00	20.00	20.00	20.00	30.00	20.00	10.00	10.00	**40.00**	**80.00**	**40.00**
		Mid	50.00	40.00	**100.00**	**100.00**	**100.00**	90.00	**100.00**	90.00	**100.00**	**100.00**	60.00	50.00
Mixed-compliant	Multi	Ini	0.00	0.00	30.00	**40.00**	20.00	30.00	20.00	30.00	20.00	20.00	**80.00**	20.00
		Mid	10.00	0.00	**100.00**	90.00	90.00	90.00	**100.00**	**100.00**	90.00	80.00	70.00	30.00
Ambiguous	Multi	Mid	0.00	0.00	50.00	30.00	50.00	**40.00**	**70.00**	20.00	50.00	10.00	**70.00**	10.00
	Single	Mid	0.00	0.00	**33.33**	**33.33**	**33.33**	0.00	**33.33**	0.00	0.00	0.00	**33.33**	0.00
Mixed-ambiguous	Multi	Ini	0.00	0.00	30.00	**40.00**	30.00	30.00	20.00	20.00	0.00	0.00	**90.00**	20.00
		Mid	10.00	0.00	70.00	50.00	70.00	50.00	**100.00**	**60.00**	60.00	50.00	80.00	40.00
Adjacent	Multi	Mid	45.00	10.00	**90.00**	20.00	65.00	**30.00**	80.00	5.00	80.00	15.00	55.00	10.00
	Single	Mid	60.00	25.00	95.00	15.00	95.00	25.00	95.00	**40.00**	**100.00**	35.00	30.00	15.00

Table 9.6: Span recall obtained by each model over BLAS for different types of spans and contexts (on standard casing, with and without quotation marks).

hand, if we are expecting lowercase text (such as social media text), then the BiLSTMs are the best choice.

Performance across different span types

We now analyze performance across different span types. We will focus our analysis on the standard casing configuration (where only proper names and sentence-initial words are capitalized), as standard casing is the most canonical and the results over the capitalized sections of BLAS were catastrophic across most models. We refer the reader to appendix E for the full results of the models across all other configurations.

Table 9.6 display recall results per model on the standard casing on different linguistic attributes, with and without quotation marks. The scores on this table display the degree of sensitivity of each model to anglicism type, span length, sentence position and quotation mark absence. Table 9.6 shows that certain span attributes produce better results than others, like mid-sentence spans. In fact, multiword non-compliant spans that appear in mid-sentence position with quotation marks can essentially be considered solved, with almost all models producing perfect or near perfect recall over those examples.

On the other hand, performance drops substantially when retrieving spans from sentence initial position. It is remarkable that the very same spans that were retrieved with a recall

score of 80%, 90% or even 100% when they appeared mid-sentence, produced a recall of 20%, 10% or even 0% when they appeared at the beginning of the sentence. Again, it should be noted these differences in recall between sentence-initial and mid-sentence spans can be fully ascribed to the differences in the contexts they appear in, as the spans were the same across both groups.

Similarly, the differences observed in performance when retrieving compliant and non-compliant spans can also be fully attributed to the shape of the spans themselves: BETO obtained a recall of 90% when retrieving non-compliant multiword spans from sentence initial position, but its performance dropped to 40% and 30% for compliant and mixed compliant spans on otherwise similar attributes. We can be sure that these differences are caused by the spans themselves and not the position within the sentence they appeared in, because in all three groups the sentence positions were the same and only the spans changed.

On the other hand, adjacency was not a challenge for supervised models: most of them produced scores over 80%, with quotation marks. The mediocre scores supervised models obtained on adjacent spans when quotation marks were absent (with models scoring 40% at most) are only apparent: full results on the token & separator evaluation in appendix E reveal that the retrieval at token level on unquoted adjacent spans were indeed good for most models (with some models even scoring around 90%), but in many cases the boundaries of the spans were incorrectly detected, which caused fusing errors (see tables E.14, E.26, E.38, E.50). Llama3, on the contrary, missed most adjacent spans and obtained poor results, both at span level and token level (see table E.62). This is the type of situation in which applying multiple evaluation approaches is revealed as useful, because otherwise the differences in performance between Llama3 and the supervised models on unquoted adjacent spans would have gone unnoticed.

Models	BLAS	COALAS (test set)				COALAS (dev set)				2020a corpus (test set A)				2020a corpus (test set B)				2020a corpus (dev set)			
	R	Pred. R		True R		Pred. R		True R		Pred. R		True R		Pred. R		True R		Pred. R		True R	
CRF	6.50	26.58	#6	44.31	#5	33.13	#6	68.63	#5	29.00	#6	65.57	#5	24.76	#6	53.17	#5	28.99	#6	64.84	#5
BETO	23.55	86.87	#3	77.99	#4	87.62	#3	84.05	#4	83.58	#4	83.49	#3	79.13	#4	81.75	#4	81.60	#4	86.30	#1
mBERT	23.80	86.86	#4	78.85	#3	87.10	#4	84.31	#3	83.86	#3	85.85	#1	83.17	#3	84.13	#3	82.73	#3	86.30	#1
BiLSTM (un)	23.55	87.86	#2	80.88	#2	88.73	#2	89.05	#1	84.85	#2	82.08	#4	84.04	#2	84.92	#2	83.06	#2	83.11	#4
BiLSTM (cs)	23.75	92.96	#1	81.32	#1	91.08	#1	86.57	#2	89.43	#1	83.96	#2	89.52	#1	85.71	#1	87.53	#1	85.84	#3
Llama3	36.37	42.93	#5	33.33	#6	55.46	#5	53.27	#6	48.53	#5	53.30	#6	43.17	#5	40.48	#6	48.49	#5	50.68	#6
Correlation				0.94	0.89			0.79	0.83			0.85	0.66			0.91	0.94			0.84	0.41

Table 9.7: Correlation between the predicted and true recall scores and ranking positions over five datasets.

9.3.2 Predictive ability of BLAS

The natural question that follows from our results from section 9.3.1 is: are the results per attribute obtained on BLAS representative of the ability of the models to perform on those linguistic attributes, not just over BLAS data, but in general? Could we use the scores obtained on BLAS to predict the overall recall that those same models would obtain on a different distribution? For instance, based on the results obtained on BLAS, we know that Llama3 produces a span recall of 90% over non-compliant multiword spans when they appear mid sentence and are written with quotation marks on standard casing. Could we extrapolate those results obtained by Llama3 on BLAS to other datasets and claim that that Llama3 will successfully retrieve 90% of the spans that share those same characteristics when tested on a different distribution?

In order to answer these questions, we ran the six models we ran on BLAS in section 9.3.1 on five additional datasets for anglicism identification in Spanish: the test and dev sets from COALAS from chapter 4 and the dev set and test sets A and B from our previous work from Álvarez Mellado (2020a) (see chapter 2.3.3).

The purpose of our experiment was to test if the scores the models obtained on BLAS across different span attributes could predict the results those same models would obtain on these external datasets. To do so, we counted the number of anglicism spans per type in each of these datasets (see appendix F) and calculated the expected number of spans that would be successfully retrieved by simply extrapolating from the recall results obtained per type by each model on BLAS. Going back to our example above: there are 20 spans in COALAS

test set that are non-compliant, multiword and that appear with standard casing, quotation marks and in mid sentence position. Given that the span recall of Llama3 obtained on BLAS on that particular type of spans is 90, we hypothesized that 90% of those 20 spans in COALAS would be successfully retrieved by Llama3, which amount to 18 true positives and 2 false negatives. We did the same for each type of span in all five datasets for every model, and calculated the expected overall span recall per model on each dataset based on BLAS scores.

Table 9.7 compares the true recall and ranking and the predicted recall and ranking derived from extrapolating the results produced on BLAS. Because BLAS contains only anglicisms (and no other borrowings), for the true recall scores and ranking on COALAS and on the 2020a datasets we report results obtained exclusively over the spans labeled as `ENG`. Results show that our method accurately predicts both recall values (with a median Pearson correlation of 0.85) and system ranking (median Spearman correlation of 0.83). The highest correlations were obtained on the COALAS test set and the 2020a test set B (0.94 and 0.91, respectively), precisely the two splits that were substantially different (in terms of out-of-distribution topics and vocabulary) to their training splits. Our predicted rankings over recall also display a strong correlation with F1 score rankings (median correlation 0.83; see appendix F for full results on prediction experiments and data).

It should be noted that overall scores on BLAS are not predictive themselves: it is the projection of the system performance across span attributes in BLAS that correlates with true scores (see tables 9.4 and F.7).

9.3.3 Statistical significance of the dimensions

In chapter 8 we defined a set of linguistic dimensions to characterize borrowings appearing in context. We hypothesized then that models would perform unevenly across these different dimensions, but that those differences would go unnoticed when averaged scores were computed. Our results across dimensions from chapter 8 seemed to corroborate our intuition

that dimensions do indeed affect performance. A reasonable question to ask is whether those differences we saw in performance can truly be explained because of the dimensions we defined, or whether the dimensions we defined were actually artificial and do not correlate with differences in performance. When we computed our results across dimensions in chapter 8, some dimensions had very few examples in order to establish a solid comparison. Moreover, because of the imbalance in COALAS test set, we could not really be sure if the differences in results between dimensions could actually be blamed on a single dimension or if there were more factors in place.

With BLAS, we have created a new test set that is balanced in terms of dimensions. In fact, we have twelve punctuation variations of the same dataset. This means that we can compare results across different variations of BLAS (with and without quotation marks, for instance) and be sure that the only difference in performance between the two will be explained by the punctuation differences in it.

In this section we propose to perform statistical significance tests over the results obtained by our models on BLAS. The idea is to split BLAS based on a given dimension (for instance, with and without quotation marks), compute recall results over the sentences in each of the two groups and then calculate whether the differences between both results is statistically significant.

This approach substantially diverges from the usual procedure of calculating significance tests in NLP. Statistical significance in NLP is usually computed across different models *on the very same dataset* (as we did in chapter 4 when we computed statistical significance across models initialized with different random seeds). What we propose now is slightly different: we propose to calculate significance on the results obtained by the very same model over two different datasets that are equal (or at least homogenenous) in all aspects except for one. If we obtain statistical significance over the distribution of results per sentence obtained by the very same model on two versions of the dataset that only differ in one single dimension (for example, quotation marks), then we can be moderately sure that that model in particular is

sensitive to that dimension in question.

Table 9.8 compares recall results obtained per model calculated over a single dimension value, indicating which differences are statistically significant, which suggests that the dimension in question is statistically significant for the model. Although the comparison across casing and quotation marks variations are the most solid ones (because in those comparisons we are certain that the dataset is exactly the same, except for a single aspect), we also include comparisons over other dimensions. For dimensions where only mid-sentence position examples existed (adjacent, ambiguous), the comparisons were controlled by sentence position. In other words, results across adjacent borrowings were compared across sentences without adjacent borrowings that were also mid-sentence. Similarly, dimensions that entailed multiword borrowings (mixed compliant, mixed ambiguous) were compared only with results over other sentences with multiword borrowings.

Statistical significance was calculated with Wilcoxon signed-rank test for casing and punctuation comparisons (where sample sizes are equal), and with Mann–Whitney U test for all the rest (where sample sizes where not equal) (Dror et al., 2018). Significance results over dimensions other than casing or punctuation, however, should be taken with a grain of salt: although BLAS is balanced in terms of dimensions (and the comparisons were performed controlling by sentence position and length) we cannot be really sure if the differences we observe are due to the dimension in place or if they are affected by other issues (borrowing selection, overall sentence context, etc.).

Results from table 9.8 seem to corroborate some of our intuitions: although most models are sensitive to quotation marks, the BiLSTM-CRF with codeswitch embeddings is robust to quotation marks absence, as its differences in recall at span level with and without quotation marks were not significant. All models show strong statistical significance across differences in casing. Although this may have been obvious from simply skimming the results from section 9.3.1, we now have statistical corroboration of our hunch.

Interestingly, differences across mixed compliant and mixed ambiguous borrowings were

mostly not significant. We could be tempted to conclude then that those dimensions are irrelevant, as they cannot truly explain differences in performance across models (models find equally difficult detecting mixed compliant borrowings and mixed ambiguous borrowings). However, the sensitivity of a system to a given dimension is affected by the way the system works: it should not come as a surprise that systems that heavily rely on character representation (such as ours) disregard differences between compliant borrowings (borrowings that comply with the graphotactical rules of Spanish, even if the string does not exist in Spanish) and ambiguous borrowings (words that also happen to exist as Spanish words). A lexicon lookup system (that only identifies as a borrowing words that do not appear in a monolingual dictionary of Spanish) will systematically miss ambiguous borrowings but not non-ambiguous compliant ones, and its performance will very likely display significant differences between those two groups.

Llama3 seems to be less sensitive to some of the casing perturbations that were catastrophic for supervised models. In particular, the difference in performance between having the text in lowercase or having the spans in uppercase was not significant for Llama3, but that same comparison produced major differences in performance for the supervised models. On the other hand, the statistical tests corroborated the high sensitivity of Llama3 to the absence of quotation marks.

	CRF		BETO		mBERT		BiLSTM (unadapted)		BiLSTM (codeswitch)		Llama3	
	Span recall	TS recall	Span recall	TS recall	Span recall	TS recall	Span recall	TS recall	Span recall	TS recall	Span recall	TS recall
Unquoted, Quoted	**3.81, 7.84***	**6.59, 8.24***	**20.59, 27.61***	**26.12, 28.85***	**22.33, 26.53***	**25.68, 26.96***	**20.53, 27.45***	**24.7, 27.9***	23.09, 24.78	**27.02, 25.41***	**24.13, 52.45***	**26.0, 52.61***
Lowertext, Spans are uppercase	**22.55, 0.33***	**26.93, 0.33***	**74.84, 0.33***	**82.95, 0.33***	**77.45, 0.0***	**82.76, 0.0***	**78.59, 0.0***	**85.51, 0.0***	**76.8, 1.31***	**83.61, 1.14***	50.49, 52.29	51.58, 55.01
Lowertext, Spans are titlecase	**22.55, 0.0***	**26.93, 0.0***	**74.84, 5.23***	**82.95, 6.32***	**77.45, 5.56***	**82.76, 5.99***	**78.59, 5.72***	**85.51, 5.96***	**76.8, 6.05***	**83.61, 6.05***	**50.49, 42.65***	**51.58, 43.52***
Lowertext, Text is titlecase	**22.55, 0.0***	**26.93, 0.0***	**74.84, 7.03***	**82.95, 8.8***	**77.45, 8.17***	**82.76, 8.31***	**78.59, 3.43***	**85.51, 3.73***	**76.8, 6.86***	**83.61, 6.86***	**50.49, 22.55***	**51.58, 22.96***
Lowertext, Text is uppercase	**22.55, 0.0***	**26.93, 0.65***	**74.84, 0.0***	**82.95, 0.0***	**77.45, 0.0***	**82.76, 0.0***	**78.59, 0.33***	**85.51, 0.33***	**76.8, 0.0***	**83.61, 0.0***	**50.49, 14.71***	**51.58, 14.62***
Standard casing, Spans are uppercase	**12.09, 0.33***	**16.58, 0.33***	**57.19, 0.33***	**66.51, 0.33***	**55.39, 0.0***	**60.87, 0.0***	**55.88, 0.0***	**62.28, 0.0***	**52.61, 1.31***	**59.64, 1.14***	47.06, 52.29	**48.12, 55.01***
Standard casing, Spans are titlecase	**12.09, 0.0***	**16.58, 0.0***	**57.19, 5.23***	**66.51, 6.32***	**55.39, 5.56***	**60.87, 5.99***	**55.88, 5.72***	**62.28, 5.96***	**52.61, 6.05***	**59.64, 6.05***	**47.06, 42.65***	**48.12, 43.52***
Standard casing, Text is titlecase	**12.09, 0.0***	**16.58, 0.0***	**57.19, 7.03***	**66.51, 8.8***	**55.39, 8.17***	**60.87, 8.31***	**55.88, 3.43***	**62.28, 3.73***	**52.61, 6.86***	**59.64, 6.86***	**47.06, 22.55***	**48.12, 22.96***
Standard casing, Text is uppercase	**12.09, 0.0***	**16.58, 0.65***	**57.19, 0.0***	**66.51, 0.0***	**55.39, 0.0***	**60.87, 0.0***	**55.88, 0.33***	**62.28, 0.33***	**52.61, 0.0***	**59.64, 0.0***	**47.06, 14.71***	**48.12, 14.62***
Lowertext, Standard casing	**22.55, 12.09***	**26.93, 16.58***	**74.84, 57.19***	**82.95, 66.51***	**77.45, 55.39***	**82.76, 60.87***	**78.59, 55.88***	**85.51, 62.28***	**76.8, 52.61***	**83.61, 59.64***	50.49, 47.06	**51.58, 48.12***
Compliant, Non-compliant	**1.67, 12.08***	**1.67, 12.08***	**19.17, 32.08***	**19.17, 32.08***	**21.67, 36.67***	**21.67, 36.67***	25.83, 30.0	25.83, 30.0	**29.58, 35.83***	**29.58, 35.83***	**27.08, 49.58***	**27.08, 49.58***
Mixed compliant, Mixed ambiguous	2.5, 1.67	2.5, 3.06	**26.67, 18.75***	**28.06, 23.4***	23.75, 19.17	24.31, 21.25	**26.25, 19.58***	**26.81, 21.88***	**22.92, 14.17***	**24.03, 17.22***	41.25, 48.33	41.94, 48.47
Compliant, Ambiguous	1.67, 0.0	1.67, 5.56	20.0, 11.11	20.0, 11.11	**28.33, 5.56***	**28.33, 5.56***	**28.33, 5.56***	**28.33, 5.56***	**37.5, 0.0***	**37.5, 0.0***	**24.17, 8.33***	**24.17, 8.33***
Adjacent, Non adjacent	**14.17, 7.61***	**20.0, 8.33***	20.83, 26.81	**35.0, 26.81***	21.25, 31.88	32.92, 31.88	24.58, 28.99	34.58, 28.99	27.92, 34.78	37.92, 34.78	17.92, 31.16	25.42, 31.16
Mixed compliant, Non-compliant	**2.5, 11.25***	**2.5, 15.56***	**26.67, 31.67***	**28.06, 34.72***	**23.75, 33.75***	**24.31, 34.31***	**26.25, 29.58***	**26.81, 30.42***	**22.92, 29.58***	**24.03, 29.86***	**41.25, 53.75***	**41.94, 54.44***
Mixed compliant, Compliant	2.5, 3.75	2.5, 3.75	26.67, 28.33	28.06, 30.78	23.75, 24.58	24.31, 25.42	26.25, 23.75	26.81, 25.17	22.92, 23.75	24.03, 23.75	41.25, 37.5	41.94, 37.64
Sentence initial, Mid sentence	**5.0, 9.6***	**5.0, 11.87***	22.08, 25.0	**22.08, 29.29***	22.5, 28.66	**22.5, 32.2***	23.33, 27.65	**23.33, 30.68***	**25.42, 32.7***	**25.42, 35.73***	**42.08, 27.15***	**42.08, 29.42***
Multiword, single-word	**4.75, 7.86***	**6.43, 9.28***	**24.21, 23.9***	**27.97, 26.57***	**23.42, 26.34***	**25.15, 28.54***	**22.92, 26.02***	**25.45, 27.91***	**20.75, 29.95***	**23.24, 31.84***	41.21, 32.78	**42.01, 34.2***

Table 9.8: Comparison in recall results per model across different dimensions in BLAS (asterisk indicates statistical significance with $p < 0.05$).

9.4 Discussion

In section 9.2 we introduced BLAS, a linguistically-motivated benchmark for anglicism detection in Spanish. The methodology that guided the creation of BLAS is quite different to the standard way NLP test sets are created, where the examples are usually sampled from available data and identically distributed to the training set. On the contrary, BLAS is an artificially created test-only dataset that has been designed to exhaustively cover the different linguistic dimensions a system may encounter when processing data in the wild.

The results obtained by our models on BLAS are very different to the ones we obtained when evaluating on the test set of COALAS in chapter 4, both in terms of recall scores and ranking position.

In terms of overall recall, none of the models scored above 40% when evaluated on BLAS. This is remarkable, given that those same models had obtained recall scores over 75% when evaluated on COALAS test set. The rankings obtained on COALAS and on BLAS were also very dissimilar: both rankings showed a correlation of -0.08. Llama3, which had ranked last on COALAS, outperformed all supervised models on BLAS and ranked first. The good results obtained by Llama3 can partially be explained by the poor performance that all supervised models showed on the capitalized splits of BLAS. Although unusual casing had an undeniable impact on the models performance on BLAS, the differences in results between BLAS and COALAS cannot exclusively be attributed to differences in casing: the very same supervised models that scored above 75 in terms of recall when tested on COALAS scored in the sixties when confronted BLAS on the standard casing, and even below when quotation marks were removed (see table 9.5).

These results showcase the impact that test set design can have on the scores: even when COALAS was designed to be as difficult as possible as a test set (rich in OOV borrowings, topics in the test set not included in the training set, etc.), and both BLAS and COALAS are similar in nature and cover the same problem and genre, the representation and distribution

of linguistic phenomena in both datasets is very different, which leads to these very different results.

Results on BLAS show that we are far from having perfect performing models and that we cannot claim the task of identifying borrowings in Spanish to be solved. The fine-grained results from section 9.3.1 showed some of the systematic vulnerabilities of these models and point to issues that future work should address, such as the retrieval of graphotactically compliant anglicisms and anglicisms in sentence initial position.

Our statistical significance tests from section 9.3.3 corroborated that supervised models are specially vulnerable to non-standard capitalization, span length and span position within the sentence. Llama3 was especially sensitive to the absence of quotation marks, although it proved to be more robust to some capitalization perturbations.

In addition to serving as a test set to evaluate models and identify systematic weaknesses, our results from section 9.3.2 show that BLAS has predictive ability: the scores per attribute obtained by six models on BLAS were able to accurately predict the performance of those models on five external test sets with a median correlation of 0.85. This means that BLAS not only works as an exhaustive test bed against which models can evaluate their performance on a fine-grained selection of span types and contexts, but it also shows that scores obtained on BLAS can anticipate how a model will generalize to a different distribution.

Although we have built our case on anglicism detection, our methodology could in principle be applied to other span identification tasks, such as NER, where ambiguous and adjacent spans are ubiquitous and span retrievability is also sensitive to context, punctuation and sentence position. The graphotactic compliance distinction may be specific to anglicism identification, but this linguistic attribute is simply a concrete instantiation of a more general issue that is common to other sequence labeling tasks: the fact that in any span-based task, there will be some spans that will be more salient than others, a feature that has also been reported to be key in NER (Papay et al., 2020; Lin et al., 2021; Vajjala and Balasubramaniam, 2022).

The predictive ability displayed by BLAS fine-grained results suggests that linguistically-aided test set creation informed by error analysis may be a workable solution to the long quest for evaluation methodologies that can anticipate how a system will perform on new data, a feature that is considered key to measure the generalization abilities of NLP systems (Hupkes et al., 2023; Zhou et al., 2023).

9.4.1 Limitations

Our work has focused on evaluating the capability of models at retrieving true anglicisms of different types and shapes from different contexts. Consequently, our analysis and the creation of the dataset itself focused on recall. Prioritizing recall is suitable for tasks in which minimizing the number of false negatives is crucial (even if it is at the cost of having some false positives), as we established in chapter 7 .

Other tasks, however, may require taking precision into account and may wish to perform a similar evaluation over false positives. To evaluate precision in a similar way as we did with recall with BLAS, we would first need to investigate the type of linguistic phenomena that characterize false positive spans and then populate our test set with non-relevant spans that could pass for one.

It should be noted that predicting a precision score over an external dataset as we did with recall with BLAS would not be feasible with our methodology: in a span-based task, the sum of false negatives and true positives is known, as it is the number of total spans in the gold standard. Therefore, we can calculate the proportion of spans that are expected to be missed by a model, based on the recall scores it obtained on the benchmark. However, if we tried to do the same with precision, even if we had a precision-based benchmark, it would be unclear what the set of potential false positive spans of the external test set would be.

Finally, BLAS should not be taken as a definite benchmark for anglicism identification in Spanish, nor as the sole test set against a model should be measured. Quite on the contrary,

BLAS is intended as a very specific diagnosing testbed that may contribute documenting and assessing a model's capability. In that regard, BLAS follows the trend of other auxiliary validation sets (Ghaddar et al., 2021). How relevant the scores obtained by a model on BLAS are will be determined by the use case scenario where the model in question is expected to be deployed.

9.4.2 Relation to previous work

The ideas that have guided the creation of BLAS take inspiration from previous work on robustness evaluation in NLP, especially on NER models.

The gap in performance between the results obtained by NLP models when they are evaluated on benchmark data and when they are tested on real-world scenarios has previously been discussed (Kiela et al., 2021).

Lin et al. (2020) referred to this gap as the difference in performance between *regular NER* (NER performed in lab conditions on benchmark data) and *open NER* (NER performed on real world data) and pointed out three possible reasons behind this gap: strong name regularity in NER benchmark entities (i.e., most entities have a similar shape), overlap between the entities in the test set and the training set and test instances rich in context.

Additional research on NER has added two more issues as possible culprits for the lack of robustness of NER models: label inconsistency (Fu et al., 2020; Bernier-Colborne and Langlais, 2020; Tu and Lignos, 2021) and presence of adjacent spans (Rueda et al., 2024).

These phenomena that plague NER evaluation are no different to the type of issues that we have encountered in lexical borrowing detection evaluation. For BLAS, we have translated these issues (span regularity, overlap between training/test, obvious contexts, label inconsistency and adjacent spans) to the task of lexical borrowing detection through the analysis of dimensions and are well represented in BLAS so they can be evaluated.

The idea of applying systematic perturbations to inputs in order to assess models' robustness is not new and has previously been applied to different aspects of NER performance,

particularly in regard to casing (Mayhew et al., 2019), entity shape (Agarwal et al., 2021; Lin et al., 2021; Vajjala and Balasubramaniam, 2022), entity type or entity length (Wang et al., 2021). It also builds on previous work on stress test evaluation that assess models' robustness on specific linguistic phenomena (Naik et al., 2018; Wang et al., 2019a; Ribeiro et al., 2020).

Our motivation is similat to the work of Kovatchev and Lease (2024), but they opt for describing dataset items in terms of extrinsic features (how a set of systems perform on each of the instances), while our dimensions are intrinsic and linguistically motivated: they do not depend on the set of systems tested.

The idea of building a testbed by creating linguistically-motivated examples by hand draws from other expert-authored benchmarks from the natural language understanding literature (Bowman and Dahl, 2021), such as the FraCaS test suite (Cooper et al., 1996) or the the Winograd Schema Challenge (Levesque et al., 2012). It also follows the recommendations stated in Linzen (2020) of producing test-only benchmarks to avoid cuing the model with unintended statistical regularities from the training set, and those from Søgaard et al. (2021) that proposes evaluating on biased splits as evaluation protocol to determine data characteristics that affect performance. Similarly, the methodology followed for BLAS follows the design behind other benchmarks and evaluation campaigns such as Kiela et al. (2021) and Ettinger et al. (2017), which put the linguistic expertise at the center of NLP evaluation.

9.4.3 The cycle of benchmark creation

Finally, the benchmark BLAS we have presented in this chapter culminates the process we introduced in the course of chapters 3-9 and that proposes a series of resources for the task of automatically identifying anglicisms in Spanish text. The process we have followed and the results produced in the course of these chapters comply with the roadmap proposed by Gehrmann et al. (2021) in benchmark creation (see figure 9.3), namely:

i Improving data.

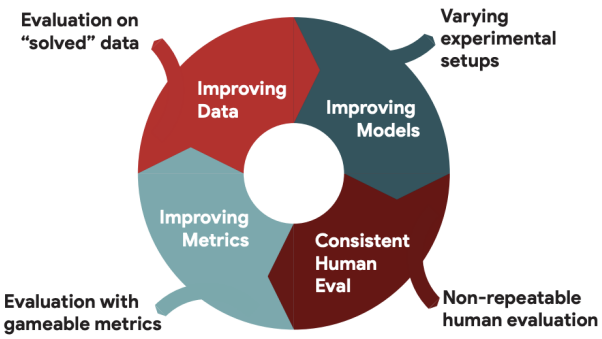

Figure 9.3: The GEM cycle of benchmark creation. Image from Gehrmann et al. (2021)

ii Improving models.

iii Performing consistent human evaluation.

iv Improving metrics.

We will now detail how the work we have presented here follows the Gehrmann et al.'s cycle of benchmark creation:

Previous SOTA results for lexical borrowing detection in Spanish scored 85% on F1 score (Álvarez Mellado, 2020a), which could perhaps lead to the false idea that lexical borrowing detection in Spanish could be considered a solved task (or at least a task with little room for improvement).

In chapter 3, we introduced COALAS, a new dataset for lexical borrowing detection in Spanish. Compared to the previous dataset for the task, COALAS includes more data, is more borrowing-dense, is richer in previously unseen borrowings and is more varied, as the data comes from more sources.

Our results from chapter 4 on the COALAS test set show that the existing SOTA model (the CRF) obtained mediocre results on this new test set, which proves that COALAS is a more challenging dataset than the previous one, and therefore allows for a more rigorous evaluation of models (i).

This led us to experiment with different models and embeddings for the task of lexical borrowing detection, which produced a suite of best-performing models that outperform

previous SOTA models (ii).

We performed a thorough analysis of our results (iii), both quantitatively (what scores are models getting on certain types of examples?) and qualitatively (what are the characteristics of the spans contained in the COALAS set?). Our analysis revealed two issues with our work:

- The discontents of strict span based evaluation to evaluate span identification tasks such as lexical borrowing detection (chapter 7).

- The skewness in the representation of certain linguistic phenomena in the COALAS dataset (chapter 8).

Consequently, we explored alternative evaluation metrics for sequence labeling tasks, and assessed how they behave on different scenarios (iv).

In this chapter we have introduced BLAS, which is an improvement over the evaluation data contained in COALAS, thus initiating a new iteration of the Gehrmann et al.'s cycle and setting the stage for a batch of future models that, hopefully, will outperform current ones.

9.5 Summary

In this chapter we have introduced BLAS, a new benchmark for the task of anglicism detection in Spanish. BLAS is linguistically exhaustive in terms of the phenomena contained in the dataset and is associated with several evaluation metrics in order to offer a bigger picture of what models can and cannot do.

BLAS is designed to serve as a benchmark that is *diagnostic* (because it can help identify weaknesses in the models' performance), *actionable* (because it can inform which model to choose for a given data scenario) and *predictive* (because the scores obtained on BLAS can anticipate how the model will perform on a new dataset).

Contrary to current practice in NLP, examples in BLAS were not scrapped from a real source, but designed by hand and written from scratch by a linguist, thus preventing data

contamination. The purpose of each group of examples in BLAS is to evaluate concrete aspects of systems' performance, such as their ability to retrieve anglicisms of a given length, shape or in different sentence positions.

Models results on BLAS paint a very different picture to the results we had obtained on the COALAS test set from chapter 4. Llama3, which had ranked last on COALAS, ranked first on BLAS, outperforming all supervised models on different aspects.

Our analysis showed that the linguistic dimensions that we defined on chapter 8 and that guided the design of BLAS were statistically significant and are indeed relevant in models' performance.

More importantly, we showed that the results obtained on BLAS could be used to predict the ranking of models on five external datasets for anglicism detection in Spanish, which proves that BLAS has predictive ability as it can anticipate how a model will perform on a new distribution.

Part IV

Results

Chapter 10

Lázaro: an observatory of anglicism usage in the Spanish press

En un diccionario no se puede dejar de trabajar. Constantemente estoy viendo en los periódicos o en las novelas expresiones que anoto para incluirlas. Ya tengo una gran colección de adiciones. Si no me muriera, seguiría siempre haciendo adiciones al diccionario.

María Moliner, lexicographer

10.1 Motivation

In the course of this dissertation we have introduced a series of techniques and resources for computer-assisted identification of unassimilated lexical borrowings (particularly anglicisms) in Spanish text. In this chapter we present Observatorio Lázaro (Lázaro Observatory), an observatory of anglicism usage in the Spanish press that applies the models and techniques we have introduced to a real-world scenario. Observatorio Lázaro monitors the daily usage of English lexical borrowings in the Spanish press and detects novel anglicisms that appear in Spanish newspapers.

The aim of the project is to facilitate the analysis of anglicism usage in real context, at large scale and in an systematic and data-driven fashion, which can assist lexicographic work and facilitate corpus linguistics research. The construction of a resource such as Observatorio Lázaro was the ultimate aim of the models and methods explored in previous chapters and its development has in fact guided the research we have conducted in the course of this thesis.

Observatorio Lázaro consists of an automatic pipeline that extracts anglicisms from Spanish newswire (such as *look*, *hype*, *prime time* or *lawfare*). Every day, the pipeline retrieves the news published in several Spanish media, processes them and extracts the anglicisms that appear on those articles. These extracted anglicisms are then stored and aggregated to produce visualizations and tables that are displayed in Observatorio Lázaro's website[1].

What lies at the core of Lázaro's pipeline are the supervised models from chapter 4 that were trained to automatically extract anglicisms from Spanish newswire. The observatory started running in April 2020 and it originally tracked eight Spanish newspapers and ran on the CRF model from Álvarez Mellado (2020a). In July 2022 the model was changed to the BiLSTM-CRF with codeswitch embeddings and subword embeddings we described in chapter 4, and the number of monitored media sites grew to 22 (see table 10.1).

Since it was launched, the observatory has registered more than 1,400,000 occurrences of anglicisms (50,000 of them unique). To the best of our knowledge, this the largest continuously-growing self-populating context-based database of anglicism usage ever compiled.

The name of the observatory, *Lázaro*, is an homage to Spanish linguist Lázaro Carreter, whose prescriptivist columns admonishing against the usage of anglicisms in the Spanish press became very popular in Spain during the decades of 1980s and 1990s.

[1] `https://observatoriolazaro.es`.
English version at `https://observatoriolazaro.es/en/`

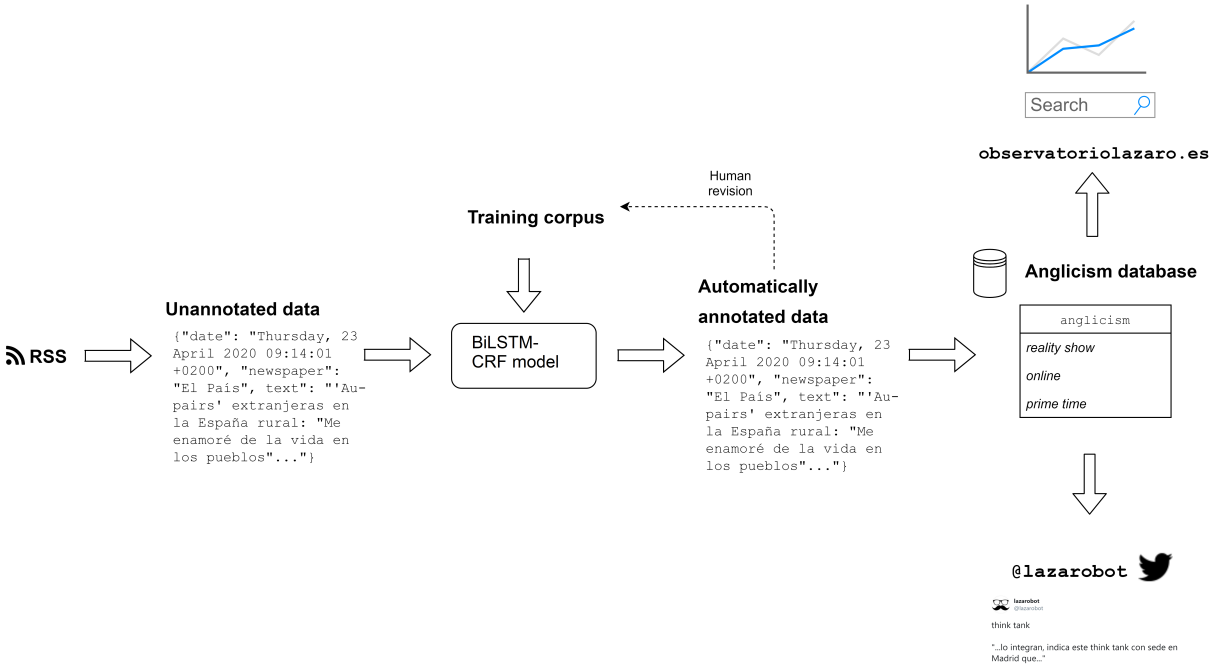

Figure 10.1: Observatorio Lázaro: Pipeline and output.

10.2 Implementation

The main component of Observatorio Lázaro is an automatic pipeline that extracts anglicisms from Spanish journalistic texts. This pipeline consists of 3 steps: article retrieval, anglicism extraction and anglicism storage. The stream of the process starting from the beginning of the pipeline to the output result is illustrated by figure 10.1.

Article retrieval and sources. The starting point of the pipeline is an automatic system that connects daily to the RSS feeds of several major news sources from Spain. The project originally started tracking 8 general newspapers: *El País*[2], *elDiario.es*[3], *El Mundo*[4], *ABC*[5],

[2]https://www.elpais.com/
[3]https://www.eldiario.es/
[4]https://www.elmundo.es/
[5]https://www.abc.es/

Media	URL	Topic
El País	https://www.elpais.com/	General news
elDiario.es	https://www.eldiario.es/	General news
ABC	https://www.abc.es/	General news
El Mundo	https://www.elmundo.es/	General news
La Vanguardia	https://www.lavanguardia.com/	General news
El Confidencial	https://www.elconfidencial.com/	General news
20 Minutos	https://www.20minutos.es/	General news
Agencia EFE	https://www.efe.com/	General news
Agencia Sinc	https://www.agenciasinc.es/	Science & Tech
Muy Interesante	https://www.muyinteresante.es/	Science & Tech
La Marea	https://www.lamarea.com/	Politics
El Salto	https://www.elsaltodiario.com/	Politics
El Economista	https://www.eleconomista.es/	Economy
Cinco Días	https://www.cincodias.elpais.com/	Economy
JotDown	https://www.jotdown.es/	Culture
El Mundo Today	https://www.elmundotoday.com/	Satirical news
Marca	https://www.marca.com/	Sports
Rolling Stones	https://es.rollingstone.com/esp/	Music
Fotogramas	https://www.fotogramas.es/	Cinema
Diez Minutos	https://www.diezminutos.es/	Gossip
Men's Health	https://www.menshealth.com/es/	Lifestyle
Elle	https://www.elle.com/es//	Lifestyle

Table 10.1: List of media sources tracked by Lázaro Observatory

La Vanguardia[6], *El Confidencial*[7], *20minutos*[8] y *Agencia EFE*[9]. In July 2022 the list of tracked sources was extended and currently 22 Spanish media sites are monitored by the observatory, including general newspapers and thematic journalistic sites (see table 10.1). Every day, the pipeline connects to these sources and extracts the texts of the articles published within the last 24 hours (see table 10.2). These articles are then preprocessed (for HTML tag removal, etc) and made ready for the step 2 of the system.

Anglicism extraction. The articles collected during step 1 are then sent to the extraction model that identifies anglicisms from Spanish newswire. The model processes the articles and extracts the anglicisms present in the texts. From April 2020 to July 2022, the model

[6] https://www.lavanguardia.com/
[7] https://www.elconfidencial.com/
[8] https://www.20minutos.es/
[9] https://www.efe.com/

	2020-2022	2022-2024
Sources	8	22
URLs	400	850
Tokens	145,000	520,000
Anglicisms (all)	430	1,050
Anglicisms (unique)	190	450
Anglicisms (new)	25	45

Table 10.2: Average numbers of URLs, tokens and anglicisms processed by Observatorio Lázaro per day.

behind the observatory was the CRF model from Álvarez Mellado (2020a). From July 2022 onwards and in light of the new results obtained in chapter 4, the CRF model was substituted by the BiLSTM-CRF model with codeswitch embeddings and subword embeddings trained on the COALAS corpus (see chapter 4).

The observatory therefore monitors the type of borrowings that the models we presented in the course of this thesis can extract: unassimilated lexical borrowings, with a special focus on unassimilated anglicisms. Consequently, other borrowings such as adapted borrowings (words whose spelling has been modified to comply with the morphophonological and orthographic patterns of the Spanish language, like *fútbol* from *football* or *tuitear* from *to tweet*) or assimilated borrowings (words that have already been registered in reference dictionaries without any italics or quotations, such as *bar* or *club*) are not expected to be extracted by the model and are therfore not tracked by the observatory. Similarly, other phenomena like semantic calques, syntactic anglicisms, acronyms and proper names are also considered beyond the scope of the project and are therefore not considered by the observatory.

Anglicism storage. The anglicisms extracted by the model in the previous step are collected and stored in a MySQL database. On an average day, a total of 1,050 anglicisms are registered (450 of which will be unique anglicisms) and added to the database. On average, the observatory identifies 45 novel anglicisms daily, that is, anglicisms that have not previously been registered by the observatory (see table 10.2).

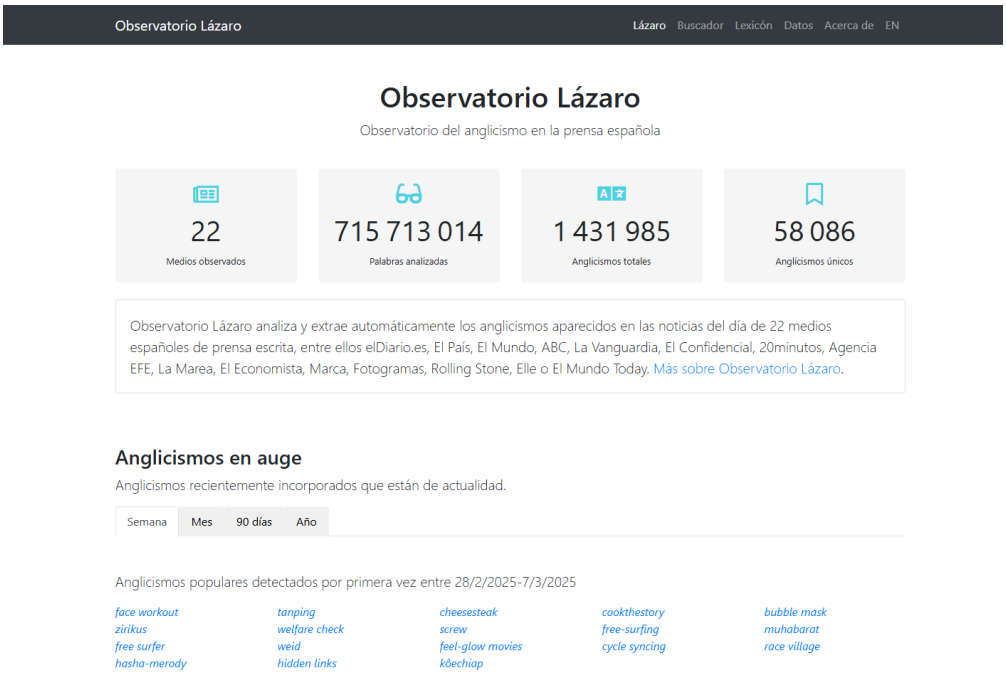

Figure 10.2: Observatorio Lázaro website.

For every anglicism, the date, context, newspaper and link to the article where the anglicism was found are stored. At the time of writing this document, the database stores more than 1,400,000 anglicisms (more than 50,000 distinct anglicisms), collected since April 2020.

10.3 Public outreach

The output of the pipeline is aggregated and used to produce various interactive graphs and tables. These visualizations are publicly available at the website `observatoriolazaro.es`[10] (see figure 10.2). The website offers a data-driven picture of the usage of anglicisms in the Spanish press that is constantly growing and is automatically updated daily.

The website also allows to retrieve all the information that the observatory has on a particular anglicism[11]. This information is aggregated and displayed in the form of a lexical

[10] English version at `https://www.observatoriolazaro.es/en/`
[11] `https://observatoriolazaro.es/en/lexicon.php`

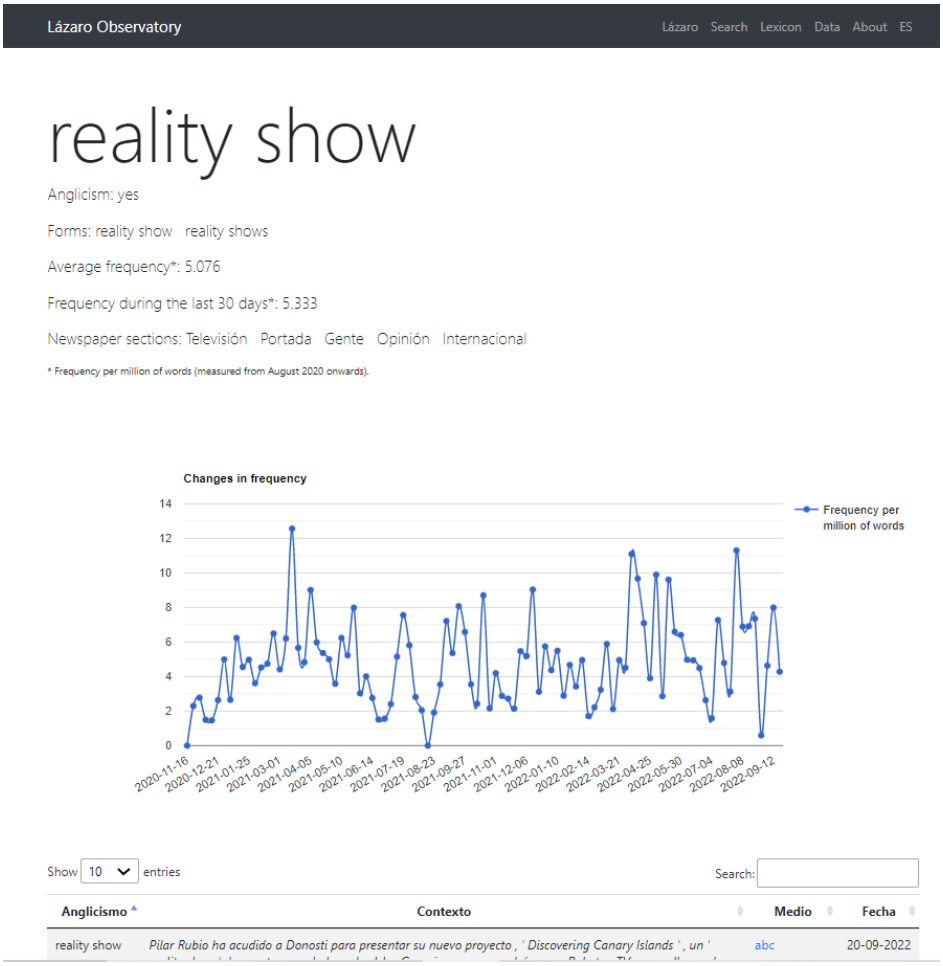

Figure 10.3: Anglicism sheet for *reality show*.

entry page[12] (see figure 10.3). For each anglicism, the lexical entry page includes normalized frequency, newspaper sections where the anglicism is usually found, frequency evolution graph, inflected forms associated with that anglicism and a list of all the concordances (appearances in context) tracked so far by the observatory.

The website also offers access to the 1,400,000 anglicism database, which can be queried from the *Search*[13] section of the website (see figure 10.4). The content of this database is also available as downloadable csv files under the section *Data*[14] of the website.

New anglicisms (ie, borrowings that are detected by the system and that have not been

[12]See for example: `https://observatoriolazaro.es/en/vocab.php?word=reality%20show`
[13]`https://www.observatoriolazaro.es/en/buscador.php`
[14]`https://www.observatoriolazaro.es/en/datos.html`

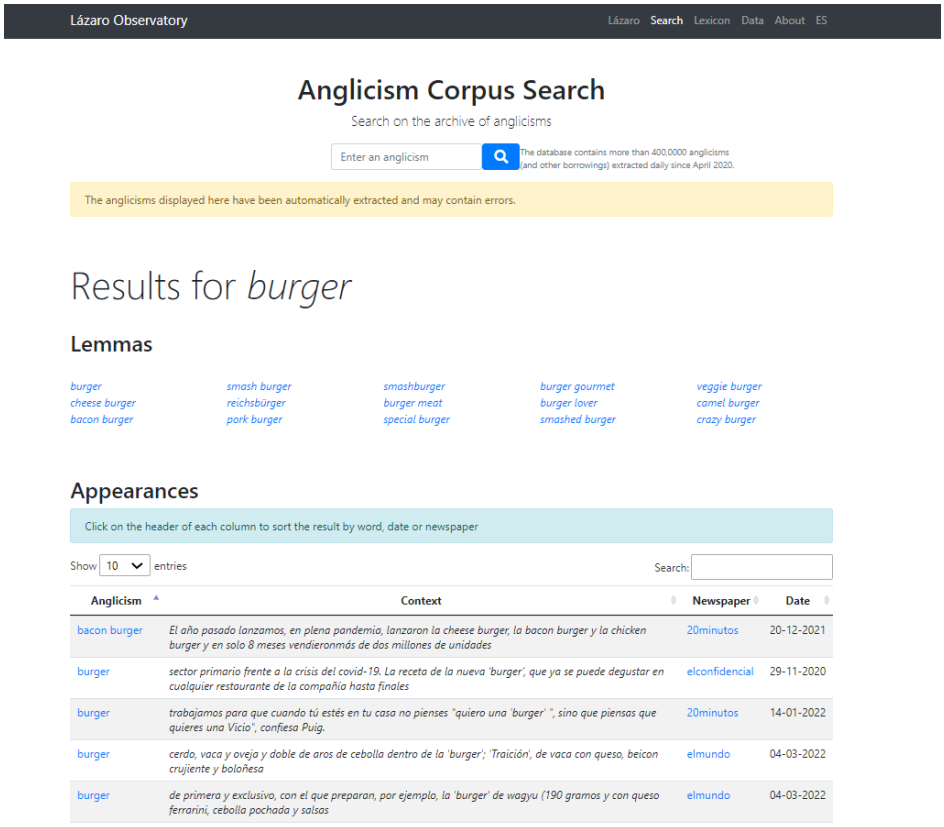

Figure 10.4: Search section from Observatorio Lázaro's website

registered before in the database) are automatically shared through social media accounts[15][16]. Each new anglicism found by the model is posted along with the context where the borrowing was found and the link to the newspaper article (see figure 10.5).

10.4 Discussion

Lázaro is the first publicly-available automatic system that monitors the usage of anglicisms in the Spanish press on large amounts of data and in real-time. This pipeline facilitates tracking anglicism frequency over time, and documents the incorporation of novel anglicisms along with their context, which can assist lexicographic work and corpus linguistics research. The model behind Lázaro's pipeline is the BiLSTM-CRF model with codeswitch embeddings

[15] `https://www.twitter.com/lazarobot`
[16] `https://bsky.app/profile/observatoriolazaro.es`

Figure 10.5: A tweet by `@lazarobot` with the word *binge-watching*.

and subword embeddings we described in chapter 4.

As a result, a database of more than 1,400,000 occurrences of anglicisms that is automatically populated on a daily basis has been released. To the best of our knowledge, this is the first database of its kind to be publicly available.

The visualizations, data and tables displayed at Observatorio Lázaro's website offer a data-driven report of the day-to-day usage of anglicisms in the Spanish press.

The content of the website and social media accounts provide a social dimension to the project geared towards non-specialized public. In this regard, Observatorio Lázaro seeks

to contribute to the public conversation about language change and anglicism usage from a descriptivist perspective. We hope that the development of Observatorio Lázaro and its satellite projects can contribute to raise awareness about the possibilities that machine learning has to offer to the study of language contact and corpus linguistics in general and the process of lexical borrowing in particular.

Finally, the purpose of Observatorio Lázaro is to analyze the usage of borrowings in the Spanish press. This project does not seek to promote or stigmatise the usage of borrowings, or those who use them. The motivation behind our research is not to defend an alleged linguistic purity, but to study the phenomenon of lexical borrowing from a descriptive and data-driven point of view.

The following third-party projects and publications have used and cited Observatorio Lázaro's data as part of their research so far:

- 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.

- Lillo, A. (2022). Anglicismos coloquiales en la toponimia española: Colloquial anglicisms in spanish toponymy. *Lebende Sprachen*, 67(1):133–167.

- Luján-García, C. I. (2021). Language contact in the media discourse: Anglicisms in section leisure in the newspaper la provincia. *Estudios de Linguistica Inglesa Aplicada (ELIA)*.

- Luján-García, C. (2021). Anglicisms in the area of food and drinks among spanish speakers. why are anglicisms used to refer to food and drinks-related terms in spanish? *Analele Universităţii din Craiova. Seria Ştiinţe Filologice. Lingvistică*, (1-2):180–190.

- Luján García, C. (2022). Are we concerned about eating healthily in spain? the increasing proliferation of anglicisms in the field of food and drink. In *45th AEDEAN Conference (Asociación Española de Estudios Anglo-Norteamericanos)*. Asociación Española de Estudios Anglo-Norteamericanos. AEDEAN.

- Luján-García, C. and Núñez-Nogueroles, E. E. (2022). Use of it English terms by Spanish university students: A case study. In *39º Congreso internacional de la Asociación Española de Lingüística Aplicada (AESLA)*.

- Luján-García, C. (2023). Adults only or pets welcome: Use of anglicisms in the tourist domain in spanish digital press. *Lengua y Habla*, 27.

- Luján-García, C. (2023a). Anglicisms in spanish gastronomy: new words for new eating habits. *Sintagma: revista de lingüística: 35, 2023*, pages 51–69.

- Luján-García, C. (2023b). '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. and Núñez-Nogueroles, E. E. (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. E. and Luján-García, C. (2022). Perceptions and reported use of information technology (it) anglicisms by Spanish university students. *Miscelánea: A Journal of English and American Studies*, 66:41–67.

10.5 Summary

In this chapter we have presented Observatorio Lázaro, an observatory of anglicism usage in the Spanish press. The observatory automatically monitors the presence of English lexical borrowings in a collection of Spanish media sites in real time.

The project is a practical application of the resources and methodologies explored in the previous chapters of this thesis and offers a public interface of the project that is informative and accessible for a non-specialized audience.

The observatory proves that computer-assisted methods for lexical borrowing detection (such as the ones developed in the course of this thesis) can successfully be used to build real-world applications that can inform linguistic work in a systematic and data-driven fashion.

Chapter 11

Conclusions

The narrowest hinge in my hand puts to scorn all machinery.

Walt Whitman, poet

11.1 Contributions

In the course of this thesis we have explored strategies concerning data annotation, modeling techniques and evaluation approaches for the task of automatically retrieving lexical borrowings from text, with a focus on anglicisms in Spanish.

The main novelty of this thesis is our proposal to frame automatic borrowing retrieval not as a classification task of isolated words as done in previous work, but as a sequence labeling task, in which spans of variable length are retrieved from text and where context is key when deciding whether something is a borrowing or not. Our approach frames the task in a similar fashion to other classic NLP tasks, such as NER or MWE.

Our results show that automatic borrowing identification can indeed be framed as span-identification task and that the problem can successfully be addressed applying sequence labeling methods. Our approach is better suited than previous methods for capturing the nuances of borrowing usage in the wild (multiword borrowings, span adjacency, in-context ambiguity), issues that are vital for linguistic-assisting technology. The span-identification

approach we have taken here is a major difference to all previous approaches to the task of automatic identifying borrowings and future research on the topic (either in Spanish or in other languages) may find this approach fruitful.

More concretely, in part II Data & models, we have proposed various data and modeling resources for the automatic identification of lexical borrowings in Spanish. We have first introduced COALAS, a novel dataset of monolingual Spanish journalistic text annotated with lexical borrowings (chapter 3), along with detailed annotation guidelines for the task (see appendix A). Our proposed dataset is larger, more diverse and borrowing dense than previous resources for the task and, as a result, previous SOTA models for the task obtained mediocre results when confronted with it.

We have then explored a collection of supervised sequence labeling models (traditional machine learning, deep learning models and Transformer-based models) for the task of automatically retrieving lexical borrowings from Spanish text, as well as a large language model (8B-Llama3) prompted in a few-shot fashion (chapter 4). Our results show that a BiLSTM-CRF model fed with different types of Transformer-based word embeddings (unadapted or fine-tuned for the task of language identification on codeswitched data) along with subword embeddings performs best on the task, and clearly outperforms 8B-Llama3.

We have reported the results obtained of organizing ADoBo, the first shared task on automatic detection borrowings held at IberLEF 2021, where participants were invited to train and submit their systems for lexical borrowing identification in Spanish using our COALAS dataset (chapter 5).

Finally, we have also proposed an extension of our guidelines to account for the distinction between lexical borrowing and codeswitching in multilingual settings, such as social media texts (chapter 6).

In part III Evaluation we have turned our eyes to the evaluation issues that plague span-identification tasks such as NER, MWE and ours, in-context lexical borrowing detection. We have identified the mismatches and dissatisfaction that both token-based and span-

based evaluation produce when evaluating span-identification tasks, and we have identified an alternative existing evaluation framework that, although hardly exploited within NLP, solves the shortcomings that other evaluation frameworks produce: the token & separator model (chapter 7).

We have then described a new methodology that proposes to evaluate span-identification tasks based on linguistic dimensions, in which models are evaluated on subsplits of the test set that share a given linguistic characteristic, such as sentence position, capitalization or relation to the training set. Our methodology was able to reveal systematic weaknesses in performance that had gone unnoticed when evaluating over the full test set and can be easily transferred to other span-identification tasks (chapter 8).

Finally, we have presented BLAS, a novel test-only dataset for the identification of anglicisms in Spanish. Contrary to the common practice in NLP, BLAS is a synthetic test set of hand-made sentences that serves as a testbed, where each group of examples seek to assess a particular linguistic aspect of model performance and where multiple scores are provided. BLAS is a linguistically-motivated benchmark that is diagnostic, actionable and predictable: our experiments showed that results across different linguistic dimensions in BLAS are statistically significant and that it has predictive ability, as the scores obtained by models on BLAS can anticipate how those same models will perform when confronted with a different distribution. We consider that the main contribution behind BLAS is not the benchmark itself but the methodology that guided its creation and that is transferable to other span identification tasks (chapter 9).

To conclude, we have introduced Observatorio Lázaro, an observatory of anglicism usage in the Spanish press that implements the developments and findings reported in this dissertation (chapter 10).

11.2 Resources

As a result of this thesis, we have produced two annotated datasets, two extensive annotation guidelines, a suite of trained models, two Python packages, a benchmark, a shared task and a website with a real-time pipeline that showcases the research carried out. We now enumerate the list of resources and deliverables that have been derived from this dissertation:

COALAS. The COALAS dataset of Spanish journalistic text annotated with lexical borrowings is available on GitHub under a permissive Creative Commons license that allows for reuse[1].

Annotation guidelines. Detailed annotation guidelines for the annotation of in-context lexical borrowings in Spanish monolingual newswire are provided in appendix A. An extension of those guidelines to account for the borrowing versus codeswitching distinction in multilingual social media text are provided in chapter 6.

Models. The neural supervised models fine-tuned for the task of automatically retrieving lexical borrowings from Spanish journalistic texts have been released and are publicly available at `HuggingFace` model hub[2].

`pylazaro`**.** Our five supervised models for lexical borrowing detection in Spanish are also available through a Python package called `pylazaro`[3][4]. `pylazaro` is open source and can be installed via `PyPI`[5]. The purpose of `pylazaro` is to offer a single Python interface that facilitates calling any of the supervised borrowing detection models for Spanish developed in the course of this thesis and that allows for switching between models smoothly.

[1] `https://github.com/lirondos/coalas`
[2] `https://huggingface.co/models?other=arxiv%3A2203.16169`
[3] `https://pylazaro.readthedocs.io/`
[4] `https://github.com/lirondos/pylazaro`
[5] `https://pypi.org/project/pylazaro/`

Borrowing vs codeswitching dataset. Reannotation of the LM2013 dataset by Lignos and Marcus (2013) that implements the borrowing vs codeswitching distinction[6].

`cascabel`. `cascabel`[7] (CAracterizing Spans for sequenCe lABELing) is an open-source Python package that implements the dimension-based analysis we performed on chapter 8. Given a dataset for a span identification task (such as the COALAS test set), `cascabel` provides information on how the spans in the given test set are distributed in terms of span shape, surrounding context and presence in the training set.

BLAS. BLAS[8] (Benchmark for Loanwords and Anglicisms in Spanish) is a collection of 1,836 test examples written by hand with the aim of exhaustively exploring the different linguistic characteristics that may affect anglicism retrieval and thus model perfomance, such as the orthographic shape of the anglicism in question or the varying contexts it may appear in.

ADoBo. We organized the first shared task on automatic detection borrowings held at IberLEF 2021. Information on the shared task (participation, results, etc) are available on the shared task website[9].

Observatorio Lázaro. An observatory of anglicism usage in the Spanish press[10]. Observatorio Lázaro is a data-driven approach to the study of lexical borrowing that monitors the anglicism usage in the Spanish media. The project is a practical application that showcases and implements the findings made in the course of this dissertation.

[6] `https://github.com/lirondos/borrowing-or-codeswitching`
[7] `https://github.com/lirondos/cascabel`
[8] `https://github.com/lirondos/blas`
[9] `https://adobo-task.github.io/2021.html`
[10] `https://observatoriolazaro.es/`

11.3 Limitations

The observations and resources we have presented in this thesis should not be considered a full depiction of either the process of borrowing or the Spanish language in general. We have identified four important considerations that any future systems that build off this research should be aware of.

Scope of the phenomenon. The scope of our research is quite narrow: our work has focused on the identification of unassimilated lexical borrowings (mainly from the English language) that appear in Spanish journalistic texts. This necessarily means that our observations and methods are restricted and can only be applied with some certainty to tasks that are similar in terms of the languages and domains involved. Projects that seek to, let's say, retrieve Spanish borrowings from Basque from social media texts may not find our approaches fully transferrable.

The process of borrowing. Borrowing is a complex phenomenon that can manifest at all linguistic levels (phonological, morphological, lexical, syntactic, semantic). This work is exclusively concerned with lexical borrowings. Furthermore, in this work we have taken a synchronic approach to borrowing: we deal with borrowings that are considered as such in a given dialect and at a given point in time. The process of borrowing assimilation is a diachronic process, and the notion of what is perceived as unassimilated can vary across time and varieties. As a result, our dataset and models may not be suitable to account for partially assimilated borrowings or even for unassimilated borrowings in a different time period.

Language variety. The datasets and models we have presented for lexical borrowing retrieval from journalistic texts are exclusively composed of European Spanish journalistic texts. In addition, the guidelines we have described were designed to capture a very specific

phenomena: unassimilated borrowings in the Spanish press. In fact, our annotation guidelines rely on sources such as *Diccionario de la Lengua Española*, a lexicographic source whose Spain-centric criteria has been previously pointed out Blanch (1995); Fernández Gordillo (2014). Consequently, the scope of our work is restricted to unassimilated borrowings in journalistic European Spanish. Our dataset and models may not translate adequately to other Spanish-speaking areas or genres.

The preeminence of written language. In our work, the notion of what a borrowing is is heavily influenced by how a word is written. According to our definition, a word like *meeting* will be considered unassimilated, while the Spanish form *mitin* will be considered assimilated. These preferences in writing may indirectly reveal how well-established a loanword is or how foreign it is perceived by the speaker. But it is questionable that these two forms necessarily represent a difference in pronunciation or linguistic status in the speaker's mental lexicon. How a word is written can be helpful for the purpose of detecting novel anglicisms in written text, but ideally one would not establish a definition of borrowing solely based on lexicographic, corpus-derived or orthotypographic cues. These are all valuable pieces of information, but they only represent an indirect evidence of the status that the word holds in the lexicon. After all, speakers will identify a word as an anglicism (and use it as such), regardless of whether the word is written in text or used in speech.

On the other hand, the lexicographic fact that a word came from another language may not be enough as a criterion to establish the notion of borrowing. Speakers use words all the time without necessarily knowing where they came from or how long ago they were incorporated into the language. The origin of the word may just be a piece of trivia that is totally irrelevant or unknown to the speaker at the time of speaking, so the etymological origin of the word might not be enough to account for the differences among borrowings. In fact, what lies at the core of the unassimilated versus assimilated distinction is the awareness of speakers when they use a certain word (Poplack et al., 1988). The notion of what a borrowing is

lies within the brain of the speaker. In this work we are only indirectly observing that status through written form, and written language is unreliable evidence to the speakers' linguistic knowledge or behavior (Linell, 2005). Writing is an unsatisfactory proxy to what happens inside the mind of the speakers, which is ultimately where language resides. Therefore our definition of borrowing and assimilation cannot be regarded as perfect or bullet-proof.

11.4 Future Work

Although we have shown that in-context lexical borrowing detection can successfully be framed as a span-identification task, our results indicate that the task is far from being solved. Our models produced a maximum F1 score of 84% when performing on COALAS test set, with a recall of just 78%. And scores plummet below 40% in terms of recall when models were confronted with BLAS. These numbers leave ample room for improvement, especially in terms of recall, which has shown to be crucial in term extraction and lexicographic work.

In addition, our exploration of LLMs was limited to prompting one small model (8B-Llama3) on an almost zero-shot fashion. Although our results can serve as a starting baseline, future work should consider performing a thorough assessment of how different LLMs and prompting techniques work in various lexical borrowing detection scenarios, and whether larger models can produce better results.

Another possible avenue for future work could encompass transferring our dimension-guided evaluation to dataset creation: in chapters 8 and 9 we proposed an evaluation framework that takes into account the variety of shapes and contexts in which a borrowing may appear through a dimension analysis. However, these dimensions were only considered for evaluation purposes, and their applicability to the training phase was never explored. In other words, we wonder whether our dimensions from chapter 8 could guide the construction of a training set for lexical borrowing detection (and of other span-identification tasks, to that matter), so that the different shapes and contexts in which a span may appear would be sufficiently represented in the training data. We wonder whether models trained on such

a dimension-guided training set would need less training data, perform more evenly across dimensions and be more robust to real-world scenarios.

Of course, the question of how to build such a dimensionally-guided balanced corpus in a timely manner naturally arises, given that borrowings are by definition sparse. We could always take the training split of COALAS and enrich it by applying the systematic transformations we explored for BLAS (transforming all the corpus to uppercase, adding and removing quotation marks, etc.), but that would leave the sentence position issue unsolved.

A possible course of action would be to create an ensemble with all the models from chapter 4, each of the models contributing to the annotation pile with the type of borrowings it excels at retrieving. In fact, a natural continuation of this work could be experimenting with an ensemble of models that takes into account the performance across dimensions of each model, so that if a borrowing candidate is spotted, not only the majority vote is considered but also the accuracy of the model that identified the borrowing at retrieving that type of spans. In other words, in a dimension-based ensemble, if a sentence-initial span is retrieved by a model, we could decide how much credibility we should give the finding based on the performance that the model in question displayed previously for spans of that type.

In a broader sense, the final aim of this area of work should be guided towards the creation of models that can rely exclusively on monolingual data and identify borrowing candidates in a language-agnostic fashion. Although the topic is niche and definitely not mainstream in the realm of NLP, developing technology for multilingual borrowing detection could contribute to having a better understanding of language contact in general and the process of borrowing in particular.

11.5 Publications

The following publications have been derived from this dissertation:

- Álvarez Mellado, E., Espinosa Anke, L., Gonzalo Arroyo, J., Lignos, C., and Porta Zamorano, J. (2021). Overview of ADoBo 2021: Automatic Detection of Unassimi-

lated Borrowings in the Spanish Press. *Procesamiento del Lenguaje Natural*, 67:277–285

- Álvarez Mellado, E. and Lignos, C. (2022a). Borrowing or Codeswitching? Annotating for Finer-Grained Distinctions in Language Mixing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3195–3201

- Álvarez Mellado, E. and Lignos, C. (2022b). Detecting unassimilated borrowings in Spanish: An annotated corpus and approaches to modeling. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3868–3888, Dublin, Ireland. Association for Computational Linguistics

- Álvarez Mellado, E. and Gonzalo, J. (2024). Characterizing Spans for Sequence Labeling: A Case on Anglicism Detection. *Procesamiento del Lenguaje Natural*, 73:235–246

Part V

Appendices

Appendix A

Annotation guidelines

Objective

This document proposes a set of guidelines for annotating emergent unassimilated lexical borrowings, with a focus on English lexical borrowings (or anglicisms). The purpose of these annotation guidelines is to assist annotators to annotate unassimilated lexical borrowings from English that appear in Spanish newswire, i.e. words from English origin that are introduced into Spanish without any morphological or orthographic adaptation.

This project approaches the phenomenon of lexical borrowing from a synchronic point of view, which means that we will not be annotating all words that have been borrowed at some point of the history of the Spanish language (like arabisms), but only those that have been recently imported and have not been integrated into the recipient language (in this case, Spanish).

Tagset

We will consider two possible tags for our annotation: `ENG`, for borrowings that come from the English language (or anglicisms), and `OTHER` for other borrowings that comply with the following guidelines but that come from languages other than English.

Defining an unassimilated lexical borrowing

In this section we provide an overview of what words will be considered as unassimilated lexical borrowings for the sake of our annotation project.

Definition and scope

The concept of *linguistic borrowing* covers a wide range of linguistic phenomena. We will first provide a general overview of what lexical borrowing is and what will be understood as an anglicism within the scope of this project.

Lexical borrowing is the incorporation of single lexical units from one language (the donor language) into another language (the recipient language) and is usually accompanied by morphological and phonological modification to conform with the patterns of the recipient language (Haugen, 1950; Onysko, 2007; Poplack et al., 1988).

Anglicisms are lexical borrowings that come from the English language (Gómez Capuz, 1997; Pratt, 1980; Rodríguez González, 1999; Núñez Nogueroles, 2017b). For our annotation project, we will focus on direct, unassimilated, emerging anglicisms, i.e. lexical borrowings from the English language into Spanish that have recently been imported and that have still not been assimilated into Spanish, that is, words like *smartphone*, *influencer*, *hype*, *lawfare* or *reality show*.

Although this project focuses on lexical borrowings from English, we will also consider borrowings from other languages that comply with these guidelines. Borrowings from the English language will be annotated with the tag `ENG`, while borrowings from other languages shall be annotated with the tag `OTHER`:

```
... financiados a través de la plataforma de [crowdfunding](ENG) del
club [gourmet](OTHER) que tengas más cerca
```
[1]

[1]Examples in these guidelines will display the lexical borrowing that should be labeled between square brackets, with the the corresponding tag in parentheses. Examples with no words marked with brackets will illustrate cases where no lexical borrowing should be tagged.

Other types of borrowings, such as semantic calques, syntactic anglicisms or literal translations will be considered beyond the scope of these annotation project and will not be covered in these guidelines.

Types of lexical borrowing

Lexical borrowings can be adapted (the spelling of the word is modified to comply with the phonological and orthographic patterns of the recipient language, as in *fútbol* or *tuit*) or unadapted (the word preserves its original spelling: *millennial*, *newsletter*, *like*). For this annotation project, we will be focusing on unassimilated lexical borrowings: this means that adapted borrowings will be ignored and only unadapted borrowings will be tagged.

Multiword borrowings

Lexical borrowings can be both single-token units (*online*, *impeachment*), as well as multiword expressions (*reality show*, *best seller*). Multiword borrowings will be labeled as one entity.

```
imagina ser un ‘[tech bro]’ con millones de dólares (ENG)
```

The annotation should however distinguish between a multiword borrowing and adjacent borrowings. A phrase like *signature look* is a multiword borrowing (the full phrase has been borrowed as a single unit) and should be annotated as such.

```
para recrear su [total look] (ENG)
```

However, a phrase like *look sporty* follows the NAdj order that is typical of Spanish grammar (but impossible in English): these are in fact two separate borrowings (*look* and *sporty*) that have been borrowed independently and happen to be collocated in a phrase. The annotation should capture these nuances:

```
un [look] (ENG) [sporty] (ENG) perfecto
```

Origin of the borrowings

Establishing the origin of a certain borrowings can sometimes be tricky, as the language of origin can sometimes be disputed. Additionally, certain borrowings might have originated in a certain language, but may have reached the recipient language through another language.

In order to establish the origin of borrowings, the origin attributed by reference dictionaries and institutions (Real Academia Española, 2020, 2022) will be followed.

This means that words like *junior* and *senior* (whose frequency and perhaps even their pronunciation may have changed due to the influence of English) will still be considered as latinisms, as DLE registers their adaptated versions (*júnior* and *sénior*) as such (and mentions no English influence). Similarly, the word *barista* might have entered the Spanish language via English, but RAE's Observatorio de Palabras considers it of Italian origin (and should therefore be annotated with OTHER label).

What is not an unassimilated lexical borrowing

In the previous section we provided an overview of what words will be considered as an unassimilated lexical borrowing for the sake of our annotation project. In this section we will cover what an unassimilated lexical borrowing is *not*.

There are several phenomena that are close enough to unassimilated borrowing and that can sometimes be mistaken with. In this section we will list what phenomena will not be considered as unassimilated lexical borrowings (and are therefore beyond the scope of our annotation project), as well as provide guidelines in order to distinguish these cases and adjudicate them.

We will focus on three main phenomena: assimilated borrowings, proper names and code-mixed inclusions.

Figure A.1 summarizes the decision steps that can be followed when deciding if a certain word should be labeled or not as a lexical borrowing.

Assimilated vs unassimilated borrowings

This annotation project aims to capture unassimilated lexical borrowings. As a general rule, all unadapted lexical borrowings should be tagged. This means that direct borrowings that have not gone through any morphological or orthographic modification process should be labeled.

Lexical adaptation, however, is a diachronic process and, as a result, what constitutes

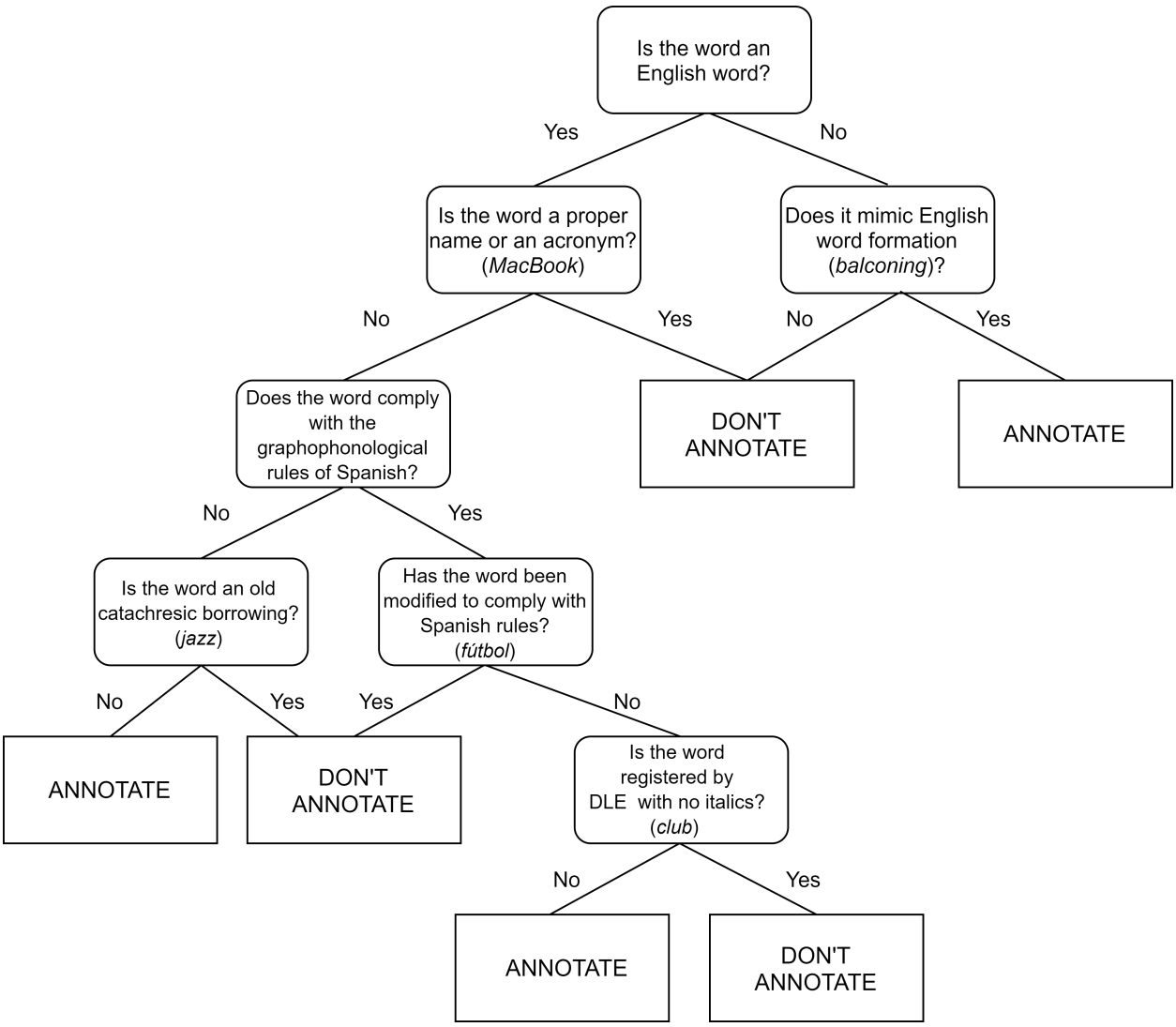

Figure A.1: Decision steps to follow during the annotation process to decide whether to annotate a word as an anglicism.

an unadapted borrowing is not clear-cut. The following guidelines define what borrowings will be considered as unassimilated (and therefore should be tagged) versus those that have already been integrated into the recipient language (and therefore should not be tagged).

Adapted borrowings

Words that have already gone through orthographical or morphological adaptation (such as *fútbol*, *líder*, *tuit* or *espóiler*) will be considered assimilated and therefore should not be labeled. Partial adaptations (such as *márketing*, where an accent has been added) will also be excluded.

Borrowings that have not been adapted but whose original spelling complies with graphophonological rules of Spanish (and are therefore unlikely to be further adapted, such as *bar*, *fan*, *web*, *internet*, *club*, *set* or *videoclip*) will be tagged as a borrowing or not not depending on how recent or emergent they are. In order to determine which unadapted, graphophonologically acceptable borrowings are to be annotated, the latest online version of the *Diccionario de la lengua española*[2] (Real Academia Española, 2020) will be consulted. If the DLE dictionary already registers the word with that meaning and with no italics or quotation marks, then it will be considered assimilated and therefore should not be tagged.

This means that a word like *set* (when used to refer to a collection of things, a television studio or a part of a tennis match) will be considered assimilated because it is already registered in DLE dictionary with no italics, and therefore should not be labeled as `ENG`. On the other hand, a word like *nude*, although its spelling also complies with Spanish graphophonological rules, will be considered an unassimilated borrowing because it has not been registered yet in the dictionary, and should therefore be tagged as such.

```
ganó el primer set

los tonos ‘[nude]’ (ENG)
```

[2] `https://dle.rae.es/`

It should be noted that this guideline only applies to lexical borrowings that comply with graphophonological rules of Spanish. Unadapted lexical borrowings that do not comply with graphophonological rules of Spanish (such as *show*, *look*, etc) will be tagged as borrowing, regardless of whether the word is included in the dictionary or not.

It is important to emphasize that, in order for an unadapted graphophonologically-compliant borrowing to be considered assimilated it should be registered in the dictionary both without italics and with the corresponding meaning. For instance, a word like *top* (that is graphophonologically acceptable in Spanish) is registered in DLE with no italics, but it is only registered with the meaning of a piece of clothing. The word *top* as referring to the upper part of something (as in *top 5*) is not registered. Consequently, the borrowing *top* will be considered assimilated when referring to the piece of clothing, but unassimilated when used to talk about the best elements of a ranking or the upper part of something.

`un top estampado`

`el [top] cinco de artistas (ENG)`

`la [top] desfiló (ENG)`

Similarly, the word *post* will not be considered a borrowing when used as a prefix of Latin origin, but will be labeled with `ENG` when used to refer to something that is published on a social media platform.

`el mundo post pandemia`

`un [post] de Facebook (ENG)`

Additionally, assimilated borrowings can still be part of new unassimilated borrowings, in which case they will be labeled as such:

`un [boys club] (ENG)`

Words derived from foreign lexemes

Words derived from foreign lexemes that do not comply with Spanish orthotactics but that have been morphologically derived following the Spanish paradigm (such as *hacktivista*, *randomizar*, *shakespeariano*) will be considered assimilated and should therefore not be labeled as a borrowing.

Compound names where one of the lexemes is a borrowing will be labeled as a borrowing or not according to the degree of independence among the lexemes. A verb+noun compound (as *caza-clicks*) will not be labeled as a borrowing, because the elements are not independent from one another. However, noun-noun compounds where each of the lexemes work can work independent from one can be labeled as borrowings:

```
una casa-[loft] (ENG)
```

Similarly, prefixed borrowings will be labeled as a borrowing, as long as the borrowing keeps independence from the prefix:

```
la ex [influencer] (ENG)
```

For prefixed borrowings, it should be checked whether the prefix can also be considered part of the borrowing:

```
los [nano influencers] (ENG)
```

Number inflection

Unassimilated borrowings may be incorporated as invariable in number *los master*, with the same plural inflection that they had in the donor language (*los pappardelle*) or may form a new plural that is non-existant in the donor language (*los pappardelles*). For number inflection, we follow the same criteria that DLE (Real Academia Española, 2020) follows: a non-Italian plural like *pizzas* is still regarded as unadapted (and therefore should be written italicized even when the true Italian plural would be *pizze*). Consequently, non assimilated

borrowings that have a non-cannonical plural inflection form will still be considered as an unassimilated borrowing and labeled as such.

```
una serie de animación de [mechas] (OTHER)
```

Pseudoanglicisms

Words that do not exist in English (or exist with a different meaning) but were coined following English morphological paradigm to imitate English words (such as *footing* or *balconing*) will be annotated as anglicisms.

```
la imagen del '[balconing]' y las excursiones etílicas (ENG)
```

```
practicaba [footing] por la calle (ENG)
```

Realia words

Borrowings that refer to culture-specific elements (often called *realia words*) that were imported long ago but that have remained unadapted will not be tagged as borrowing. This means that if a borrowing is not adapted (i.e. its form remained exactly as it came from the donor language) but refers to a particular cultural object that came via the original language, that has been registered for a while in Spanish dictionaries and is not perceived as new anymore, then it will not be tagged as a borrowing, even if does not comply with graphophonologic rules of Spanish.

The purpose of this guideline is to account for cultural terms such as *pizza*, *whisky*, *jazz*, *blues*, *banjo* or *sheriff*. These are all borrowings that are reluctant to be adapted or translated, even when they have been around in the Spanish language for long. The reason is that they refer to cultural inventions (the name was imported along with the object it refers to), and, given their cultural significance, they never competed with a Spanish equivalent and are seen as assimilated.

Therefore, unadapted borrowings that refer to cultural innovations (such as music, cooking, sport names etc) and that have been registered for long in the Spanish language[3] will not be tagged as emergent borrowings.

It should be noted that this only applies to borrowings that have been around enough time to be registered in dictionaries. A word like *hip hop* is a realia word, but it is still recent enough and has not been registered in the dictionary. In that case, it should be considered as unassimilated and tagged as such.

Latinisms

Borrowings that were introduced directly from Latin language (such as *deficit*, *curriculum*, etc) will not be considered emergent and therefore will not be tagged as a borrowing. However, it should be noted that unassimilated borrowings from other languages that happen to have a Latin etymology and and that are introduced with a distinct meaning (such as *adlib* or *premium* etc) will still be tagged as borrowings.

Borrowings vs names

Proper nouns

Non-Spanish proper nouns will not be tagged as borrowings. These include:

- person names: *Bernie Sanders*.
- organization names: *WikiLeaks*.
- product names: *Slack*.
- location names: *Times Square*.
- dates and celebrations: *St. Patrick's Day*, *Black Friday*.

[3]RAE dictionary `https://dle.rae.es/`, Mapa de diccionarios `https://webfrl.rae.es/ntllet/SrvltGUILoginNtlletPub` and CREA `http://corpus.rae.es/creanet.html` and CORPES `https://webfrl.rae.es/CORPES/view/inicioExterno.view` can be consulted

- event names: *Brexit*, *procés*.
- social and political movements: *Black Lives Matter*, *MeToo*.
- treaties and documents: *New Deal*, *Privacy Shield*, *French Tech Visa*.
- titles of cultural productions: *Stranger Things*.

Borrowings in proper nouns

Borrowings that appear as part of proper nouns or named entities (such as book titles or organization names, as in *Los Hermanos Podcast*) will not be labeled as borrowings.

Proper nouns in borrowings

Multiword borrowings and expressions can sometimes include proper nouns. Even when a proper noun in isolation cannot be considered a borrowing, proper nouns within a borrowed expression will be considered part of the borrowing, as long as the proper noun is part of the borrowing and is used following the grammar rules of the donor language (for example, in an English noun noun compound):

```
Tecnología [made in Spain] (ENG)

[Google cooking] (ENG)
```

Names of institutions and political roles

Non-Spanish names that refer to political institutions (such as *Parlament* or *Bundestag*) or to political roles and figures (*lehendakari*, *president*, *conseller*) will be excluded and will not be tagged as borrowings.

Words derived from proper nouns

Words derived from proper nouns (via metonymy or eponymy) will not be tagged as a borrowing, as long as the relation with the proper noun they come from is transparent to the speaker such as:

- products: *un iPhone*, *un whatsapp*, *un bizum*, *un Scalextric*, *el Satisfyer*.
- works of arts: *un monet*
- characters: *un frankestein*.

However, borrowings that originated from a proper noun in the donor language but entered the Spanish language as common nouns and are currently recognized as such, will be labeled as borrowings. In order to adjudicate which of these words are still used in Spanish as proper names and which are common nouns, dictionaries and other reference works can be consulted.

Names of peoples or languages

Names of peoples or languages (such as *inuit*) will not be labeled as borrowings, even if the word is borrowed from another language and is not registered in Spanish dictionaries.

Ficticious creatures

Unadapted names of fictitious creatures (such as *hobbit* or *troll*) will be labeled as a borrowing.

```
En un agujero en el suelo vivía un [hobbit] (ENG)
```

Scientific units

Unadapted borrowings that refer to widespread scientific units (such as *hertz*, *newton*, *byte*, etc) will be considered assimilated and should not be tagged as a borrowing

Names of species

Scientific names of a species (such as Latin names) will not be tagged as a lexical borrowing (*anisakis*). Names of fruit, vegetable and plant varieties (such as *manzana golden*, *patatas Kennebec* or *aguacate Hass*) will also be excluded.

Borrowings vs other code-mixed inclusions

Borrowing (using units from one language in another language) and codeswitching (intertwining segments of different languages in the same discourse) have frequently been described as a continuum (Clyne et al., 2003), with a fuzzy frontier between the two. As a result, it can be difficult to tell the difference between borrowing and other code-mixed inclusions. The following guidelines can assist annotators adjudicate edge cases.

When in doubt while dealing with code-mixed inclusion, the annotator may find it helpful to ask the following question as a rule of thumb: would it make sense to have this non-Spanish word registered in a dictionary of Spanish? If the answer is no (for instance, because the word reflects the literal quotation of what someone said or because the inclusions is metalinguistic usage rather than borrowing), then we are probably not in front of a borrowing but of another type of code-mixed inclusion (and should not be tagged as a borrowing).

Acronyms and acronym expansions

We consider acronyms to be a different phenomenon from borrowings. Consequently, acronyms will not be tagged as a borrowing, even if the acronym is of non-Spanish origin

```
un lector de CD
```

An acronym however may be tagged as a borrowing if it appears as part of a borrowed multiword expression, as in *CD player*, *peak TV*, *PC gaming*:

```
un [CD player] (ENG)
```

Acronym expansions, that is, the expansion of an acronym into the words that form the acronym (that is usually added in between brackets after an acronym has been introduced) will also not be considered a borrowing:

```
La técnica de PCR (protein chain reaction)
```

It is important to note that for a sequence to be considered as an acronym expansion it must appear after the acronym has been introduced and serve as a gloss to it (so that it expands what the letters in the acronym stand for). Usages where the full sequence is introduced in the text and later on acronymized for the sake of brevity can still be considered as borrowings.

```
Utilizaron técnicas de [Machine Learning] (también conocido como ML)
(ENG)
```

Digits

Similarly to proper nouns, digits in isolation cannot be considered borrowings. As a result, we cannot take for granted that digits within the surroundings of a borrowing will automatically be part of the borrowing.

```
[top ten] (ENG)
```

```
[top] 10 (ENG)
```

However, if the word order of the tokens makes it clear that the digit is part of a multi-token borrowing (because the order complies with the grammatical structure of an English noun-noun compound), we can label it as part of the borrowing:

```
los [10% banks] (ENG)
```

Metalinguistic usage

Non-Spanish words that appear to refer to the word itself in linguistic discourse and do not cover a lexical gap will not be tagged as a borrowing:

```
El término viene de la palabra ‘‘ghost’’, que en inglés es ‘fantasma’
```

It should be noted that the newer, less adapted, less transparent a new word is, the more likely that the speaker will be aware of the decoding difficulty it may pose to the reader and will decide to add some sort of metalinguistic strategy or awareness around it, in the form of metacomments, word-pointers, meaning explanations, etc (*known as*, *so called*). Borrowings with these types of signals will still be considered borrowings, as long as they are covering a lexical gap.

True metalinguistic usage where the foreign word covers no lexical gap but exclusively provides linguistic information (such as etymological information) will not be considered a borrowing.

Literal quotations

Words or sequences in languages other than Spanish that are reflecting literally what someone said or wrote (as in a quotation, a statement or a slogan) will not be considered a borrowing.

```
El eslogan ‘Make America Great Again’
```

```
Es uno de los primeros resultados de Google cuando alguien busca "remote work in Spain" (trabajo en remoto en España).
```

Expressions

In general terms, multiword borrowings will be tagged as borrowings. However, phrases and expressions that are not integrated into the sentence will be excluded. This means that autonomous expressions that are rather code switched sentences (rather than real borrowings) that work as a unit totally independently of the rest of the linguistic context (and that we would not expect to be registered in a dictionary) will not be considered or tagged as a borrowing.

La innovación y la competencia tan escasas en la radiotelevisión o peor aún en Internet ("the winners takes all" o "most").

Limitations of these guidelines

These guidelines are intended to assist annotators when labeling lexical borrowings. These guidelines, however, were created with a specific goal in mind (to capture unassimilated English lexical borrowings from a corpus of Spanish newswire) and may not be suitable if applied to a project with a different scope. These are some of the shortcomings and limitations that these guidelines may have.

Text genre

These guidelines were designed to specifically capture borrowings in a corpus of Spanish newswire. Newswire is a very specific genre of text that by no means represent the whole of a language (Plank, 2016).

Donor language

These guidelines were created with English lexical borrowings in mind, which are the most frequent source of borrowing today in the Spanish press. Although the criteria can be applied to other languages as well (and in fact the annotation tagset we propose includes the tag `OTHER` to account for borrowings from other languages other than English), a more fine-grained approach would require further guidelines.

Synchronic approach to borrowing

This project approaches emergent, unassimilated lexical borrowing in a synchronic fashion. The process of borrowing and the notion of assimilation is, however, time-dependent. A diachronic approach to lexical borrowing would require a wider scope, a different theoretical framework and an expanded set of criteria.

Geographic variety

The guidelines in this document were designed to capture borrowings used in Spanish newspapers, that is, written in the variety of Spanish that is spoken in Spain and may not be suitable to account for other dialects. For instance, according to the guidelines we have just introduced, a word like *living* (that is used heavily in some Latin American varieties to refer to the living room) would be considered unassimilated. It is arguable whether these criteria would be suitable for a project that tried to capture emergent lexical borrowings in Argentinian text, for example.

Appendix B

Data statement

We document the information concerning our dataset following the data statement format proposed by Bender and Friedman (2018).

Data set name: Corpus of Anglicisms in the Spanish Press (COALAS)

Data set developer: Elena Álvarez-Mellado

Dataset license: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY- NC-SA 4.0)

Link to dataset: `https://github.com/lirondos/coalas`

Curation rationale

The corpus consist of a collection of Spanish newspaper sentences. These sentences are annotated with unassimilated lexical borrowings.

Data was collected separately for the training, development, and test sets to ensure minimal overlap in borrowings, topics, and time periods (see Section B).

To focus annotation efforts for the training set on articles likely to contain unassimilated borrowings, the articles to be annotated were selected by first using a baseline model and were then human-annotated. To detect potential borrowings, the CRF model and data from Álvarez Mellado (2020a) was used along with a dictionary look-up pipeline. Articles that

contained more than 5 borrowing candidates were selected for annotation.

The main goal of data selection for the development and test sets was to create borrowing-dense, OOV-rich datasets, allowing for better assessment of generalization. To that end, the annotation was based on sentences instead of full articles. If a sentence contained a word either flagged as a borrowing by the CRF model, contained in a wordlist of English, or simply not present in the training set, it was selected for annotation. This data selection approach ensured a high number of borrowings and OOV words, both borrowings and non-borrowings. While the training set contains 6 borrowings per 1,000 tokens, the test set contains 20 borrowings per 1,000 tokens. Additionally, 90% of the unique borrowings in the development set were OOV (not present in training). 92% of the borrowings in the test set did not appear in training.

Language variety

The language of this corpus is Standard Spanish (ISO 639-1 es). Since all the sources in the dataset are Spanish newspapers, the dialect is standard European Spanish.

Speaker demographic

No detailed information was collected regarding the demographics of the authors of the collected sentences. However, we can infer that the authors of the text were Spanish journalists aged between 20-65.

Annotator demographic

The annotator was a 30-35 year-old female graduate student from Spain, who was trained in linguistics and computational linguistics, whose native language is Spanish and who had extensive professional proficiency in English.

Speech situation

The training set consists of a collection of articles appearing between August and December 2020 in *elDiario.es*, a progressive online newspaper based in Spain. The development set contains sentences in articles from January 2021 from the same source.

The test set consisted of annotated sentences extracted in February and March 2021 from a diverse collection of online Spanish media that covers specialized topics rich in lexical borrowings and usually not covered by elDiario.es, such as sports, gossip or videogames.

All the data came from written sources and was presumably edited according to the style guides of each source.

Text characteristics

All the sentences in this dataset come from journalistic texts. The data in the training set and development set come from a general online newspaper. The data in the test set come from a diverse collection of online media covering specialized topics, which include politics, feminism, economy, gossip, videogames, cinema & TV, technology, science, travel, parenthood, lifestyle, sports and automobiles.

Recording quality

N/A

Other

N/A

Provenance appendix

See table 3.5.

Appendix C

A practical comparison between evaluation approaches

In this appendix we empirically assess how the different evaluation strategies introduced in chapter 7 rank outputs differently. Our assessment will be based on the following real example:

Example 25. *Penélope Cruz enamora con su look total black en su llegada a San Sebastián.*[1][2]

In example 25 we have two anglicisms: the first one is *look*; the second one is the two-token borrowing *total black*. These are two distinct anglicisms that happen to be collocated, as they follow the expected word order in Spanish (noun-adjective). They should not be confused with other possible and erroneous analysis, such as being a one sole multiword anglicism made of three tokens or being three separate single-token anglicisms. Consequently, the goldstandard annotation of example 25 based on spans would yield one `ENG` span of length 1 for *look*, and a second `ENG` span of length 2 for *total black*. The token-classification annotation using BIO encoding would be `B-ENG B-ENG I-ENG`.

Although the example is minimal, its configuration of one single-word span followed by

[1] "Penélope Cruz makes a big splash with her total black look on her arrival to San Sebastián".

[2] `https://www.abc.es/gente/penelope-cruz-enamora-look-total-black-llegada-20220916182706-vi.html`

one multiword span allows us to assess a myriad of possible error scenarios and combinations that are relevant for our task:

- A full single token span is missed.
- A full multiword span is missed.
- A multiword span is partially retrieved (one token is retrieved, one token is missed).
- A multiword span is splitted into two adjacent spans of length 1 each.
- Two adjacent spans are fused into one multiword span.
- Two adjacent spans are retrieved as two spans, but the boundary between them is ill-segmented.

We will now exhaustively explore all possible annotations that a model could produce for the sequence *look total black*, evaluate each of them according to the evaluation approaches we saw in section 7.2 and see how the same outputs will be ranked differently depending on the evaluation frame we adopt. We will focus on the following five evaluation approaches:

1. Strict span-based evaluation.
2. Lenient span-based evaluation.
3. Token-based evaluation with BIO encoding.
4. Token-based evaluation with IO encoding.
5. Token & separator evaluation.

For simplicity and to prioritize recall, we will restrict ourselves to the possible tagged outputs of the subsequence *look total black* only and disregard the rest of the sentence and the possible false positives that could occur in it. We will also omit forbidden tag sequences in BIO encoding (such as a span having an `I-ENG` label as first tag), because these can easily

look	*total*	*black*	Spans	TP	FP	FN	P	R	F1	Rank
			(0,1) (1,3)	2	0	0	1	1	1	1
			(1,3)	1	0	1	1	0.5	0.66	2
			(1,2) (2,3)	0	2	2	0	0	0	7
			(2,3)	0	1	2	0	0	0	7
			(1,2)	0	1	2	0	0	0	7
			-	0	0	2	0	0	0	7
			(0,1) (1,2) (2,3)	1	2	1	0.33	0.5	0.4	6
			(0,2) (2,3)	0	2	2	0	0	0	7
			(0,3)	0	1	2	0	0	0	7
			(0,1) (2,3)	1	1	1	0.5	0.5	0.5	4
			(0,1) (1,2)	1	1	1	0.5	0.5	0.5	4
			(0,2)	0	1	2	0	0	0	7
			(0,1)	1	0	1	1	0.5	0.66	2

Table C.1: Ranking of outputs on strict span level evaluation (blocks of color represent predicted spans).

be prevented by simple heuristics and are in fact automatically corrected by many sequence labeling processing libraries and evaluation scripts (Nakayama, 2018; Lester, 2020; Palen-Michel et al., 2021). In addition, to keep the example manageable, we will only consider the label `ENG` and ignore other language tags. This approach is in-line with Esuli and Sebastiani (2010), who argue that when labels are close in meaning (such as our tags `ENG` and `OTHER` are), we may consider all of them a subtag of a more general supertag and evaluate over spans ignoring labels.

Tables C.1, C.2, C.3, C.4 and C.5 display all possible outputs for the sequence *look total black* and how these are evaluated and ranked based on strict spans, lenient spans, BIO-encoded tokens, IO-encoded tokens and tokens & separators, respectively. These tables reveal some of the virtues and blind spots that we had already anticipated in section 7.2.

Different evaluation paradigms display different ranking granularities, with lenient span-based evaluation being the least granular (with a ranking that ranges from position #1 to #3) and the token and separator model being the most fine-grained (positions #1 to #8).

Table C.1 illustrates some of the shortcomings we documented for strict span-based eval-

look	*total*	*black*	Spans	TP	FP	FN	P	R	F1	Rank
			(0,1) (1,3)	2	0	0	1	1	1	1
			(1,3)	1	0	1	1	0.5	0.66	8
			(1,2) (2,3)	1	0	1	1	0.5	0.66	8
			(2,3)	1	0	1	1	0.5	0.66	8
			(1,2)	1	0	1	1	0.5	0.66	8
			-	0	0	2	0	0	0	13
			(0,1) (1,2) (2,3)	2	0	0	1	1	1	1
			(0,2) (2,3)	2	0	0	1	1	1	1
			(0,3)	2	0	0	1	1	1	1
			(0,1) (2,3)	2	0	0	1	1	1	1
			(0,1) (1,2)	2	0	0	1	1	1	1
			(0,2)	2	0	0	1	1	1	1
			(0,1)	1	0	1	1	0.5	0.66	8

Table C.2: Ranking of outputs on lenient span level evaluation (blocks of color represent predicted spans).

uation: in strict span-based evaluation, not detecting anything at all (ie., labeling all three tokens as `O`) will rank as low as only partially retrieving *total* or *black* or retrieving the missegmented span *look total* as a span. While only detecting *total* seems indeed as an unsatisfactory output, an evaluation approach that considers such result as bad as not detecting anything at all seems misguided. Moreover, span based evaluation will give more credit in terms of F1 score to the system that only detects the span *look* (position number #2 in the ranking) over the system that retrieves all three tokens but labels them as three separate spans (position number #6). What's more: under the strict span evaluation paradigm, detecting only the span *look* will yield the *same* recall score than retrieving *look*, *total* and *black* as three separate spans (R=0.5).

Similar issues happen when we evaluate over tokens with BIO encoding (table C.3): under this approach no detection whatsoever will rank as low as only detecting *black*. A ranking should distinguish between the perfect result and other less desirable (yet useful) outputs, and the worst result and other not very good (and yet better than the worst result) output. Both the strict span evaluation and BIO-encoded token evaluation fail in that regard.

look	*total*	*black*	B-ENG			I-ENG			P	R	F1	Rank
			TP	FP	FN	TP	FP	FN				
B-ENG	B-ENG	I-ENG	2	0	0	1	0	0	1.00	1.00	1.00	1
O	B-ENG	I-ENG	1	0	1	1	0	0	1.00	0.67	0.80	2
O	B-ENG	B-ENG	1	1	1	0	0	1	0.50	0.33	0.40	8
O	O	B-ENG	0	1	2	0	0	1	0.00	0.00	0.00	12
O	B-ENG	O	1	0	1	0	0	1	1.00	0.33	0.50	6
O	O	O	0	0	2	0	0	1	0.00	0.00	0.00	12
B-ENG	B-ENG	B-ENG	2	1	0	0	0	1	0.67	0.67	0.67	4
B-ENG	I-ENG	B-ENG	1	1	1	0	1	1	0.33	0.33	0.33	11
B-ENG	I-ENG	I-ENG	1	0	1	1	1	0	0.67	0.67	0.67	4
B-ENG	O	B-ENG	1	1	1	0	0	1	0.50	0.33	0.40	8
B-ENG	B-ENG	O	2	0	0	0	0	1	1.00	0.67	0.80	2
B-ENG	I-ENG	O	1	0	1	0	1	1	0.50	0.33	0.40	8
B-ENG	O	O	1	0	1	0	0	1	1.00	0.33	0.50	6

Table C.3: Ranking of outputs on token level evaluation with BIO encoding.

look	***total***	***black***	**TP**	**FP**	**FN**	**P**	**R**	**F1**	**Rank**
I-ENG	I-ENG	I-ENG	3	0	0	1	1	1	1
O	I-ENG	I-ENG	2	0	1	1	0.66	0.8	2
O	O	I-ENG	1	0	2	1	0.33	0.5	5
O	I-ENG	O	1	0	2	1	0.33	0.5	5
O	O	O	0	0	3	0	0	0	8
I-ENG	O	I-ENG	2	0	1	1	0.66	0.8	2
I-ENG	I-ENG	O	2	0	1	1	0.66	0.8	2
I-ENG	O	O	1	0	2	1	0.33	0.5	5

Table C.4: Ranking of outputs on token level evaluation with IO encoding.

At the other end of the spectrum, lenient span-based evaluation (table C.2) and token-based evaluation with IO encoding (table C.4) seem too generous. Lenient span-based evaluation will consider any partially matching output as good as the goldstandard: six of the systems were ranked #1 (that is, undistinguishable from the goldstandard), even when the six of them contained at least one mistake. This is not ideal.

On the other hand, IO-encoded token evaluation reserves the last position in the ranking only to the output where nothing was retrieved, which is adequate. However, token evaluation with IO encoding does not distinguish between systems that truly perform a perfect

look		*total*		*black*	TP	FP	FN	P	R	F1	Rank
ENG	O	ENG	ENG	ENG	4	0	0	1	1	1	1
O	O	ENG	ENG	ENG	3	0	1	1	0.75	0.86	3
O	O	ENG	O	ENG	2	0	2	1	0.5	0.66	6
O	O	O	O	ENG	1	0	3	1	0.25	0.4	10
O	O	ENG	O	O	1	0	3	1	0.25	0.4	10
O	O	O	O	O	0	0	4	0	0	0	13
ENG	O	ENG	O	ENG	3	0	1	1	0.75	0.85	3
ENG	ENG	ENG	O	ENG	3	1	1	0.75	0.75	0.75	5
ENG	ENG	ENG	ENG	ENG	4	1	0	0.8	1	0.88	2
ENG	O	O	O	ENG	2	0	2	1	0.5	0.66	6
ENG	O	ENG	O	O	2	0	2	1	0.5	0.66	6
ENG	ENG	ENG	O	O	2	1	2	0.66	0.5	0.57	9
ENG	O	O	O	O	1	0	3	1	0.25	0.4	10

Table C.5: Ranking of outputs on the token & separator evaluation.

segmentation and those that do not. Given that we are dealing with multiword spans and that boundary identification is an important issue, this also seems undesirable.

Compared to the discontents that these evaluation approaches produce, the granularity of the token & separator model shines (table C.5): with 8 different positions in the ranking, evaluation over tokens & separators distinguishes between a true perfect match (only reached by the goldstandard, unlike the IO-encoded token evaluation and the lenient span-based evaluation) and the truly worst case scenario ranked last (only reached when not retrieving absolutely anything, unlike the BIO-encoded token evaluation and the strict span evaluation).

The token & separator model assigns high positions in the ranking (while not #1) to correctly detected but incorrectly segmented spans. Interestingly, those same scenarios were considered perfect matches in the partial matching evaluation and in the IO-encoded evaluation, and worst case scenarios in strict-match evaluation. Under the token and separator model, the second to last position in the ranking is reserved to outputs that only retrieve a single token in the sequence. Fusing all three tokens into a single span and retrieving all

	Strict spans	**TS**	**Token (IO)**	**Token (BIO)**	**Partial spans**
Strict spans	1.00	0.38	**0.14**	0.65	**0.11**
TS		1.00	**0.91**	0.69	0.66
Token (IO)			1.00	0.50	0.85
Token (BIO)				1.00	0.40
Lenient spans					1.00

Table C.6: Spearman correlation coefficients between pairs of ranking orders produced by different evaluation approaches.

three tokens as three independent spans are ranked in positions #2 and #3 respectively. This seems like a desired evaluation behavior, because although these outputs are indeed incorrect, they are acceptable and preferable in terms of recall to other outputs.

As we have seen, these five evaluation approaches rank systems very differently to one another. Table C.6 displays the degree of correlation (calculated as the Spearman correlation coefficient) among the rankings we have produced. The two most similar rankings are the ones produced by the token & separator model and the evaluation on IO-encoded tokens, as these two evaluation paradigms produced almost indistinguishable rankings (correlation coefficient = 0.91). This is interesting, as the token & separator model, while being similar to the IO token encoding, preserves the ability to evaluate the segmentation of adjacent spans and produced double the amount of ranking positions than IO-encoded token evaluation (eight versus four). On the other hand, strict span evaluation and lenient span evaluation produced the most dissimilar among our rankings (correlation coefficient = 0.11), followed by the token evaluation on IO encoding and strict span evaluation (0.14) and the token & separator model and span evaluation (0.38).

Although the example we have based our analysis on revolved around lexical borrowing detection, our observations can be extended to other span identification tasks: the behavior of the evaluation approaches would have been analogous if our analysis had been applied to a NER sequence such as *London United Kingdom* (`B-LOC B-LOC I-LOC`) or the shallow parse of the phrase *Eat the rich* (`B-VP B-NP I-NP`).

Appendix D

Guidelines for the creation of examples in BLAS

In this section we document the criteria that guided the anglicism selection, writing process and punctuation transformations behind the examples in BLAS.

The first step when writing the examples that would compose BLAS was to select the anglicisms that would serve to evaluate a given linguistic attribute. Anglicisms were selected by hand from the anglicism dictionary *Nuevo diccionario de anglicismos* by Rodríguez González and Lillo Buades (2009) and from Observatorio Lázaro, an online resource that automatically compiles anglicisms used in the Spanish media (Álvarez Mellado, 2020b) (see chapter 10.

The following criteria were taken into account for the selection:

Frequency criteria With BLAS we wanted to evaluate how well models perform when they face anglicisms that are not seen in training. In other words, we wanted to ensure that most of the borrowings in BLAS would be unseen. However BLAS comes with no associated training data and, being an evaluation benchmark, we cannot anticipate or assume the type of data that future models evaluated on BLAS will be trained on. This means that we cannot guarantee that the anglicisms in BLAS will be unseen for the models.

To mitigate this limitation, we made a conscious effort to include borrowings in BLAS that are more niche and less widespread. Consequently, during the selection process we prioritized less obvious anglicisms over anglicisms that we would expect to find in a monolingual dictionary of Spanish, or that are so frequent that they could be expected to appear in a generalistic training set (while still keeping a balance in terms of domains represented).

We also consciously avoided extremely rare borrowings and nonce borrowings; to do so, we made sure that the selected borrowings appeared in the Spanish general media and newspapers. For borrowings retrieved from Observatorio Lázaro, this means that we looked for borrowings that had been documented more than 10 times and less than 50 in Spanish newspapers.

Sentence position criteria Because we wanted to assess performance on different sentence positions, we prioritized whenever possible to select borrowings that could work equally well mid sentence and in sentence initial position.

Topic and shape variety The selection was also curated to ensure that a variety of topics and shapes were covered, so that not all anglicisms belonged to the same domain or had similar character combinations.

Unassimilation criteria Our aim with BLAS is that it can serve as a stable benchmark that future models can be evaluated on. This wish clashes against a linguistic fact of life: the fact that a borrowing that is frequent enough may very easily end up being assimilated into the recipient language and thus not perceived as a borrowing anymore. This poses the risk of including borrowings in BLAS that will eventually become assimilated and will not serve our purpose anymore. This is particularly true for graphotactically compliant anglicisms (that is, anglicisms whose spelling does not violate the spelling rules of Spanish), as the fact of being compliant may cause that speakers rapidly forget that the word was borrowed, thus facilitating its assimilation process (while non-compliance preserves the speakers' perception

of word being foreign).

To account for this, for compliant borrowings we chose words where, even if their spelling did not violate the graphotactics of Spanish, there was no correspondence between the spelling of the word and its pronunciation according to the Spanish pronunciation expectations. In other words, a word like *barman* is a graphotactically compliant borrowing whose spelling matches the expectations of how the word is pronounced in Spanish (/ˈbaɾ.man/). This makes the borrowing *barman* a bad choice for BLAS, because due to its shape and pronunciation it makes it a good candidate for rapid assimilation. On the other hand, borrowings like *paper* (/ˈpeɪ.peɾ/) or *eyeliner* (/aɪ.ˈlaɪ.neɾ/), while still being fully compliant in their spelling, are currently pronounced in a way that violates the expectations of Spanish, which ensures that they are perceived as foreign by native speakers and that are far from becoming assimilated in the near future. Consequently, for graphotactically compliant borrowings we chose borrowings where there was no correspondence between its spelling and pronunciation.

Compliant and not-compliant spans

Compliant anglicisms are anglicisms whose spelling complies with the Spanish rules of spelling. On the other hand, non-compliant anglicisms are anglicisms whose spelling violates the rules of spelling. Non-compliant borrowings are easier to spot, because their spelling violates the expectations that Spanish speakers have about how a word should be written. As a result, in chapter 8 we had already pointed out that models tended to privilege non-compliant borrowings over compliant ones. To assess differences in performance between compliant and non-compliant borrowings, we selected the following anglicisms:

- 20 graphotactically compliant anglicisms (10 single word anglicisms, 10 multiword anglicisms): for example, *pipeline*, *medal race*.

- 20 graphotactically non-compliant anglicisms (10 single word anglicisms, 10 multiword anglicisms): for example, *cliffhanger*, *fact checker*.

- 10 graphotactically mixed-compliant anglicisms, where at least one token is compliant and one token is non-compliant (all multiword): for example, *tie break*.

Ambiguous spans

Ambiguous anglicisms are anglicisms that also exist as fully Spanish words (such as *pie*, that is a native word when it means "foot" but a borrowing pronounced as /paɪ/ when it refers to a dish). Ambiguous spans can easily confuse models, as they can be seen during training as native words or found as native words in Spanish lexicons. Our aim was to evaluate specifically how well models perform on borrowings that are prone to label mismatch.

As a result, we collected a set of anglicisms that are ambiguous with Spanish words (either single word items or multiword items), following the same procedure we described for compliant and non-compliant borrowings. The following borrowings were considered:

- 3 single-word ambiguous borrowings: for example, *pie*.
- 10 multiword ambiguous borrowings: for example, *global director*.
- 10 multiword mixed-ambiguous borrowings (where at least one of the words is ambiguous and one of the tokens is not): for example, *escape room*.

We identified very few candidates that were single-word ambiguous, but we still considered them to be interesting enough to include them in BLAS, as these examples can help us assess models' performance on an extremely difficult case.

For pure ambiguous borrowings (that is, not mixed ambiguous), we lifted the restriction that borrowings had to work in sentence initial position: given the limited number of ambiguous borrowings and in order to avoid populating our benchmark with unnatural examples (not all borrowings work equally well on sentence initial position and how natural a borrowing in sentence initial position sounds is given by the meaning of the borrowing), we focused exclusively on mid sentence position examples, where all borrowings work equally well.

Consequently, for pure ambiguous borrowings we only wrote one example per selected borrowing, in which the borrowing appeared mid-sentence. We still wrote two examples (one sentence initial, one mid-sentence) for mixed-ambiguous anglicisms, as mixed-ambiguous are not rare and we could easily find good candidates for mixed-ambiguous that could work on sentence initial position.

Adjacent spans

Finally, we wanted our benchmark to evaluate performance over sentences with adjacent borrowings, ie. two independent borrowings that are collocated next to one another in the sentence (as in *look total black*), as this type of sentences are prone to segmentation errors. We selected the following 20 pairs of borrowings that worked well collocated:

- 10 pairs of collocated borrowings, where both borrowings were single-word borrowings; for example, *marketing online* in *Una campaña de marketing online*[1].

- 10 pairs of collocated borrowings, where at least one of the borrowings in the pair was multiword; for example, *look total black* in *La actriz lució un look total black*[2].

Given that not all borrowings will work equally well collocated with other borrowings and in order to ensure natural sounding sentences, the selection of borrowings in this case was more lenient: more frequent borrowings were allowed in the selection and only mid-sentence position sentences were included, as what we were interested in assessing with these examples was not the retrieval capability of the models but their ability to correctly segment borrowings that may have been previously seen.

Similarly to ambiguous anglicisms, for adjacent anglicisms we only wrote one example per selected pair of borrowings, in which the borrowings appeared mid-sentence.

[1] "An online marketing campaign"
[2] "The actress wore a total black look"

Appendix E

Results over BLAS benchmark

Tables E.1-E.72 report experiments results on BLAS across all punctuation configurations (casing and quotation marks), span types and models.

We provide results under three evaluation paradigms:

1. Span based evaluation based on exact match (precision, recall, F1 score).
2. Token & separator evaluation model by Esuli and Sebastiani (2010) (precision, recall, F1 score).
3. An extended version of MUC error typology (see chapter 9, section 9.2.6).

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	20.0	33.33	100.0	18.75	31.58	2	8	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	ini	20.0	10.0	13.33	100.0	23.33	37.83	1	5	4	0	0	0	0	0	0	0
non compliant	multi	mid	75.0	60.0	66.67	100.0	66.67	80.0	6	2	2	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	100.0	3.33	6.45	0	9	1	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	50.0	10.0	16.67	100.0	13.33	23.52	1	8	1	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	45.0	62.07	100.0	41.3	58.46	9	11	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	60.0	75.0	100.0	60.0	75.0	12	8	0	0	0	0	0	0	0	0

Table E.1: Results obtained by CRF over BLAS with standard casing, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	20.0	10.0	13.33	100.0	23.33	37.83	1	5	4	0	0	0	0	0	0	0
non compliant	multi	mid	11.11	10.0	10.53	100.0	32.14	48.65	1	3	4	0	0	0	2	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	100.0	3.33	6.45	0	9	1	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	100.0	3.33	6.45	0	9	1	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	100.0	33.33	50.0	0	2	0	0	0	0	0	0	1	0
adjacent	multi	mid	33.33	10.0	15.38	82.61	43.18	56.72	2	11	0	0	0	3	0	1	0	0
adjacent	single	mid	62.5	25.0	35.71	78.57	55.0	64.71	5	9	0	0	0	3	0	0	0	0

Table E.2: Results obtained by CRF over BLAS with standard casing, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	75.0	30.0	42.86	75.0	28.12	40.9	3	7	0	0	0	0	0	0	0	1
compliant	multi	mid	100.0	20.0	33.33	100.0	18.75	31.58	2	8	0	0	0	0	0	0	0	0
compliant	single	ini	50.0	10.0	16.67	16.67	10.0	12.5	1	9	0	0	0	0	0	0	0	1
compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	ini	75.0	60.0	66.67	100.0	66.67	80.0	6	2	2	0	0	0	0	0	0	0
non compliant	multi	mid	66.67	60.0	63.16	86.96	66.67	75.48	6	2	2	0	0	0	0	0	0	1
non compliant	single	ini	83.33	50.0	62.5	62.5	50.0	55.56	5	5	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	66.67	20.0	30.77	85.71	20.0	32.43	2	8	0	0	0	0	0	0	0	1
mixed compliant	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	50.0	10.0	16.67	100.0	13.33	23.52	1	8	1	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	50.0	10.0	16.67	100.0	13.33	23.52	1	8	1	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	4
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	45.0	62.07	100.0	41.3	58.46	9	11	0	0	0	0	0	0	0	0
adjacent	single	mid	92.31	60.0	72.73	92.31	60.0	72.73	12	8	0	0	0	0	0	0	0	1

Table E.3: Results obtained by CRF over BLAS with lowercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.4: Results obtained by CRF over BLAS with titlecase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.5: Results obtained by CRF over BLAS with uppercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.6: Results obtained by CRF over BLAS with titlecase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.7: Results obtained by CRF over BLAS with uppercase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	66.67	20.0	30.77	66.67	18.75	29.27	2	8	0	0	0	0	0	0	0	1
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	50.0	10.0	16.67	16.67	10.0	12.5	1	9	0	0	0	0	0	0	0	1
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	62.5	50.0	55.56	100.0	60.0	75.0	5	2	3	0	0	0	0	0	0	0
non compliant	multi	mid	10.0	10.0	10.0	75.0	32.14	45.0	1	3	4	0	0	0	2	0	0	1
non compliant	single	ini	83.33	50.0	62.5	62.5	50.0	55.56	5	5	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	50.0	10.0	16.67	75.0	10.0	17.65	1	9	0	0	0	0	0	0	0	1
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	25.0	10.0	14.29	100.0	20.0	33.33	1	6	3	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	100.0	3.33	6.45	0	9	1	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	3
ambiguous	single	mid	0.0	0.0	0.0	100.0	33.33	50.0	0	2	0	0	0	0	0	0	1	0
adjacent	multi	mid	33.33	10.0	15.38	82.61	43.18	56.72	2	11	0	0	0	3	0	1	0	0
adjacent	single	mid	55.56	25.0	34.48	73.33	55.0	62.86	5	9	0	0	0	3	0	0	0	1

Table E.8: Results obtained by CRF over BLAS with lowercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.9: Results obtained by CRF over BLAS with titlecase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	85.71	13.04	22.64	0	18	0	0	0	1	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	66.67	10.0	17.39	0	18	0	0	0	1	0	0	0	0

Table E.10: Results obtained by CRF over BLAS with uppercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.11: Results obtained by CRF over BLAS with titlecase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.12: Results obtained by CRF over BLAS with uppercase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	57.14	40.0	47.06	100.0	53.12	69.38	4	3	3	0	0	0	0	0	0	0
compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	42.86	30.0	35.3	100.0	43.33	60.46	3	3	4	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	42.86	30.0	35.3	100.0	43.33	60.46	3	3	4	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	70.0	70.0	70.0	100.0	80.0	88.89	7	0	3	0	0	0	0	0	0	0
ambiguous	multi	mid	83.33	50.0	62.5	100.0	53.33	69.56	5	4	1	0	0	0	0	0	0	0
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	94.74	90.0	92.31	100.0	93.48	96.63	18	1	1	0	0	0	0	0	0	0
adjacent	single	mid	100.0	95.0	97.44	100.0	95.0	97.44	19	1	0	0	0	0	0	0	0	0

Table E.13: Results obtained by BETO over BLAS with standard casing, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	71.43	50.0	58.82	100.0	59.38	74.51	5	3	2	0	0	0	0	0	0	0
compliant	multi	mid	80.0	80.0	80.0	100.0	87.5	93.33	8	0	2	0	0	0	0	0	0	0
compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
non compliant	multi	ini	50.0	50.0	50.0	100.0	66.67	80.0	5	0	5	0	0	0	0	0	0	0
non compliant	multi	mid	80.0	80.0	80.0	100.0	86.67	92.86	8	0	2	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	66.67	40.0	50.0	100.0	46.67	63.64	4	4	2	0	0	0	0	0	0	0
mixed compliant	multi	mid	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	57.14	40.0	47.06	100.0	50.0	66.67	4	3	3	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	55.56	50.0	52.63	100.0	63.33	77.55	5	1	4	0	0	0	0	0	0	0
ambiguous	multi	mid	50.0	30.0	37.5	100.0	37.93	55.0	3	5	1	0	0	0	1	0	0	0
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	30.77	20.0	24.24	85.42	89.13	87.24	4	0	2	0	0	6	0	1	0	0
adjacent	single	mid	27.27	15.0	19.35	70.37	95.0	80.85	3	1	0	0	0	8	0	0	0	0

Table E.14: Results obtained by BETO over BLAS with standard casing, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	83.33	100.0	90.91	88.89	100.0	94.12	10	0	0	0	0	0	0	0	0	2
compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
compliant	single	ini	90.0	90.0	90.0	64.29	90.0	75.0	9	1	0	0	0	0	0	0	0	1
compliant	single	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	multi	mid	76.92	100.0	86.95	81.08	100.0	89.55	10	0	0	0	0	0	0	0	0	3
non compliant	single	ini	90.0	90.0	90.0	75.0	90.0	81.82	9	1	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	54.55	60.0	57.15	100.0	75.86	86.27	6	0	3	0	0	0	1	0	0	0
ambiguous	multi	mid	62.5	50.0	55.56	88.89	53.33	66.66	5	4	1	0	0	0	0	0	0	2
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	94.74	90.0	92.31	100.0	93.48	96.63	18	1	1	0	0	0	0	0	0	0
adjacent	single	mid	82.61	95.0	88.37	82.61	95.0	88.37	19	1	0	0	0	0	0	0	0	4

Table E.15: Results obtained by BETO over BLAS with lowercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	100.0	6.25	11.76	0	8	2	0	0	0	0	0	0	0
compliant	multi	mid	100.0	10.0	18.18	100.0	9.38	17.15	1	9	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	33.33	10.0	15.38	100.0	16.67	28.58	1	7	2	0	0	0	0	0	0	0
non compliant	multi	mid	33.33	10.0	15.38	100.0	16.67	28.58	1	7	2	0	0	0	0	0	0	0
non compliant	single	ini	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	100.0	3.33	6.45	0	9	1	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	100.0	3.33	6.45	0	9	1	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	5.0	9.52	100.0	2.17	4.25	1	19	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	2	18	0	0	0	0	0	0	0	0

Table E.16: Results obtained by BETO over BLAS with titlecase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.17: Results obtained by BETO over BLAS with uppercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	10.0	18.18	100.0	9.38	17.15	1	9	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	50.0	10.0	16.67	100.0	13.33	23.52	1	8	1	0	0	0	0	0	0	0
non compliant	multi	mid	50.0	10.0	16.67	100.0	13.33	23.52	1	8	1	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	2	18	0	0	0	0	0	0	0	0

Table E.18: Results obtained by BETO over BLAS with titlecase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.19: Results obtained by BETO over BLAS with uppercase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	72.73	80.0	76.19	93.1	84.38	88.53	8	1	1	0	0	0	0	0	0	2
compliant	multi	mid	80.0	80.0	80.0	100.0	87.5	93.33	8	0	2	0	0	0	0	0	0	0
compliant	single	ini	83.33	50.0	62.5	50.0	50.0	50.0	5	5	0	0	0	0	0	0	0	1
compliant	single	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
non compliant	multi	ini	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
non compliant	multi	mid	61.54	80.0	69.57	78.79	86.67	82.54	8	0	2	0	0	0	0	0	0	3
non compliant	single	ini	87.5	70.0	77.78	87.5	70.0	77.78	7	3	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
mixed compliant	multi	mid	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	60.0	60.0	60.0	100.0	73.33	84.61	6	0	4	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	55.56	50.0	52.63	100.0	63.33	77.55	5	1	4	0	0	0	0	0	0	0
ambiguous	multi	mid	37.5	30.0	33.33	84.62	37.93	52.38	3	5	1	0	0	0	1	0	0	2
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	16.67	10.0	12.5	83.67	89.13	86.31	2	0	2	0	0	7	0	1	0	0
adjacent	single	mid	20.0	15.0	17.14	61.29	95.0	74.51	3	1	0	0	0	8	0	0	0	4

Table E.20: Results obtained by BETO over BLAS with lowercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	50.0	10.0	16.67	100.0	12.5	22.22	1	8	1	0	0	0	0	0	0	0
compliant	multi	mid	100.0	10.0	18.18	100.0	9.38	17.15	1	9	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	25.0	10.0	14.29	100.0	20.0	33.33	1	6	3	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	100.0	6.67	12.51	0	8	2	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	66.67	10.0	17.39	0	18	0	0	0	1	0	0	0	0

Table E.21: Results obtained by BETO over BLAS with titlecase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.22: Results obtained by BETO over BLAS with uppercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	100.0	6.25	11.76	0	8	2	0	0	0	0	0	0	0
compliant	multi	mid	50.0	10.0	16.67	100.0	12.5	22.22	1	8	1	0	0	0	0	0	0	0
compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	100.0	6.67	12.51	0	8	2	0	0	0	0	0	0	0
non compliant	multi	mid	50.0	10.0	16.67	100.0	13.33	23.52	1	8	1	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	100.0	6.67	12.51	0	8	2	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	2	18	0	0	0	0	0	0	0	0

Table E.23: Results obtained by BETO over BLAS with titlecase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.24: Results obtained by BETO over BLAS with uppercase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	20.0	33.33	100.0	18.75	31.58	2	8	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	90.0	94.74	100.0	90.62	95.08	9	1	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	88.89	80.0	84.21	88.89	80.0	84.21	8	2	0	0	0	0	0	0	0	1
non compliant	multi	ini	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	66.67	20.0	30.77	100.0	23.33	37.83	2	7	1	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	75.0	30.0	42.86	100.0	33.33	50.0	3	6	1	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	87.5	70.0	77.78	100.0	73.33	84.61	7	2	1	0	0	0	0	0	0	0
ambiguous	multi	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	86.67	65.0	74.29	100.0	71.74	83.54	13	5	2	0	0	0	0	0	0	0
adjacent	single	mid	100.0	95.0	97.44	100.0	95.0	97.44	19	1	0	0	0	0	0	0	0	0

Table E.25: Results obtained by mBERT over BLAS with standard casing, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	40.0	57.14	100.0	37.5	54.55	4	6	0	0	0	0	0	0	0	0
compliant	multi	mid	88.89	80.0	84.21	100.0	84.38	91.53	8	1	1	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	88.89	80.0	84.21	88.89	80.0	84.21	8	2	0	0	0	0	0	0	0	1
non compliant	multi	ini	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	75.0	30.0	42.86	100.0	33.33	50.0	3	6	1	0	0	0	0	0	0	0
mixed compliant	multi	mid	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	60.0	30.0	40.0	100.0	36.67	53.66	3	5	2	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	50.0	50.0	50.0	100.0	65.52	79.17	5	1	3	0	0	0	1	0	0	0
ambiguous	multi	mid	80.0	40.0	53.33	100.0	43.33	60.46	4	5	1	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	42.86	30.0	35.3	89.47	73.91	80.95	6	2	4	0	0	4	0	0	0	0
adjacent	single	mid	41.67	25.0	31.25	73.08	95.0	82.61	5	1	0	0	0	7	0	0	0	0

Table E.26: Results obtained by mBERT over BLAS with standard casing, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	75.0	90.0	81.82	88.24	93.75	90.91	9	0	1	0	0	0	0	0	0	2
compliant	multi	mid	100.0	90.0	94.74	100.0	90.62	95.08	9	1	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
compliant	single	mid	88.89	80.0	84.21	88.89	80.0	84.21	8	2	0	0	0	0	0	0	0	1
non compliant	multi	ini	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	multi	mid	90.91	100.0	95.24	90.91	100.0	95.24	10	0	0	0	0	0	0	0	0	1
non compliant	single	ini	90.91	100.0	95.24	76.92	100.0	86.95	10	0	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	100.0	80.0	88.89	100.0	80.0	88.89	8	2	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	90.0	90.0	90.0	96.43	90.0	93.1	9	1	0	0	0	0	0	0	0	1
mixed ambiguous	multi	mid	87.5	70.0	77.78	100.0	73.33	84.61	7	2	1	0	0	0	0	0	0	0
ambiguous	multi	mid	62.5	50.0	55.56	75.0	50.0	60.0	5	5	0	0	0	0	0	0	0	3
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	86.67	65.0	74.29	100.0	71.74	83.54	13	5	2	0	0	0	0	0	0	0
adjacent	single	mid	94.74	90.0	92.31	94.74	90.0	92.31	18	2	0	0	0	0	0	0	0	1

Table E.27: Results obtained by mBERT over BLAS with lowercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	10.0	18.18	100.0	9.38	17.15	1	9	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	50.0	5.0	9.09	100.0	4.35	8.34	1	18	1	0	0	0	0	0	0	0
adjacent	single	mid	100.0	5.0	9.52	100.0	5.0	9.52	1	19	0	0	0	0	0	0	0	0

Table E.28: Results obtained by mBERT over BLAS with titlecase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.29: Results obtained by mBERT over BLAS with uppercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	100.0	3.12	6.05	0	9	1	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	100.0	3.12	6.05	0	9	1	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
non compliant	multi	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	100.0	3.33	6.45	0	9	1	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	2	18	0	0	0	0	0	0	0	0

Table E.30: Results obtained by mBERT over BLAS with titlecase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.31: Results obtained by mBERT over BLAS with uppercase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	72.73	80.0	76.19	87.1	84.38	85.72	8	1	1	0	0	0	0	0	0	2
compliant	multi	mid	88.89	80.0	84.21	100.0	84.38	91.53	8	1	1	0	0	0	0	0	0	0
compliant	single	ini	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
compliant	single	mid	88.89	80.0	84.21	88.89	80.0	84.21	8	2	0	0	0	0	0	0	0	1
non compliant	multi	ini	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	multi	mid	90.91	100.0	95.24	90.91	100.0	95.24	10	0	0	0	0	0	0	0	0	1
non compliant	single	ini	81.82	90.0	85.72	69.23	90.0	78.26	9	1	0	0	0	0	0	0	0	2
non compliant	single	mid	90.0	90.0	90.0	90.0	90.0	90.0	9	1	0	0	0	0	0	0	0	1
mixed compliant	multi	ini	100.0	80.0	88.89	100.0	80.0	88.89	8	2	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	80.0	88.89	100.0	80.0	88.89	8	2	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	40.0	40.0	40.0	100.0	58.62	73.91	4	1	4	0	0	0	1	0	0	0
ambiguous	multi	mid	44.44	40.0	42.1	73.68	46.67	57.14	4	4	2	0	0	0	0	0	0	3
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	46.15	30.0	36.36	89.19	71.74	79.52	6	3	3	0	0	4	0	0	0	0
adjacent	single	mid	38.46	25.0	30.3	70.37	95.0	80.85	5	1	0	0	0	7	0	0	0	1

Table E.32: Results obtained by mBERT over BLAS with lowercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	33.33	10.0	15.38	100.0	16.67	28.58	1	7	2	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	5.0	9.52	100.0	2.17	4.25	1	19	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	5.0	9.52	100.0	5.0	9.52	1	19	0	0	0	0	0	0	0	0

Table E.33: Results obtained by mBERT over BLAS with titlecase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.34: Results obtained by mBERT over BLAS with uppercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	10.0	18.18	100.0	9.38	17.15	1	9	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	50.0	10.0	16.67	50.0	10.0	16.67	1	9	0	0	0	0	0	0	0	1
non compliant	multi	ini	50.0	10.0	16.67	100.0	13.33	23.52	1	8	1	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.35: Results obtained by mBERT over BLAS with titlecase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.36: Results obtained by mBERT over BLAS with uppercase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	66.67	20.0	30.77	100.0	21.88	35.9	2	7	1	0	0	0	0	0	0	0
compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
compliant	single	mid	90.0	90.0	90.0	90.0	90.0	90.0	9	1	0	0	0	0	0	0	0	1
non compliant	multi	ini	60.0	30.0	40.0	100.0	36.67	53.66	3	5	2	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	50.0	20.0	28.57	100.0	26.67	42.11	2	6	2	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
ambiguous	multi	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	80.0	88.89	100.0	82.61	90.48	16	4	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	95.0	97.44	100.0	95.0	97.44	19	1	0	0	0	0	0	0	0	0

Table E.37: Results obtained by BiLSTM with unadapted embeddings over BLAS with standard casing, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	40.0	57.14	100.0	37.5	54.55	4	6	0	0	0	0	0	0	0	0
compliant	multi	mid	80.0	80.0	80.0	100.0	87.5	93.33	8	0	2	0	0	0	0	0	0	0
compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
compliant	single	mid	88.89	80.0	84.21	88.89	80.0	84.21	8	2	0	0	0	0	0	0	0	1
non compliant	multi	ini	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	60.0	30.0	40.0	100.0	36.67	53.66	3	5	2	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	50.0	20.0	28.57	100.0	26.67	42.11	2	6	2	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	66.67	60.0	63.16	100.0	70.0	82.35	6	1	3	0	0	0	0	0	0	0
ambiguous	multi	mid	50.0	20.0	28.57	100.0	26.67	42.11	2	6	2	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	10.0	5.0	6.67	82.61	82.61	82.61	1	2	1	0	0	8	0	0	0	0
adjacent	single	mid	57.14	40.0	47.06	76.92	100.0	86.95	8	0	0	0	0	6	0	0	0	0

Table E.38: Results obtained by BiLSTM with unadapted embeddings over BLAS with standard casing, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	58.33	70.0	63.63	85.71	75.0	80.0	7	0	3	0	0	0	0	0	0	2
compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
compliant	single	mid	90.0	90.0	90.0	90.0	90.0	90.0	9	1	0	0	0	0	0	0	0	1
non compliant	multi	ini	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	multi	mid	90.91	100.0	95.24	90.91	100.0	95.24	10	0	0	0	0	0	0	0	0	1
non compliant	single	ini	83.33	100.0	90.91	71.43	100.0	83.33	10	0	0	0	0	0	0	0	0	2
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	75.0	90.0	81.82	93.33	93.33	93.33	9	0	1	0	0	0	0	0	0	2
mixed compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
ambiguous	multi	mid	77.78	70.0	73.69	84.0	70.0	76.36	7	3	0	0	0	0	0	0	0	2
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	80.0	88.89	100.0	82.61	90.48	16	4	0	0	0	0	0	0	0	0
adjacent	single	mid	95.0	95.0	95.0	95.0	95.0	95.0	19	1	0	0	0	0	0	0	0	1

Table E.39: Results obtained by BiLSTM with unadapted embeddings over BLAS with lowercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	100.0	3.33	6.45	0	9	1	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	5.0	9.52	100.0	5.0	9.52	1	19	0	0	0	0	0	0	0	0

Table E.40: Results obtained by BiLSTM with unadapted embeddings over BLAS with titlecase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.41: Results obtained by BiLSTM with unadapted embeddings over BLAS with uppercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	10.0	18.18	100.0	9.38	17.15	1	9	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	5.0	9.52	100.0	2.17	4.25	1	19	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	2	18	0	0	0	0	0	0	0	0

Table E.42: Results obtained by BiLSTM with unadapted embeddings over BLAS with titlecase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.43: Results obtained by BiLSTM with unadapted embeddings over BLAS with uppercase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	58.33	70.0	63.63	85.71	75.0	80.0	7	0	3	0	0	0	0	0	0	2
compliant	multi	mid	80.0	80.0	80.0	100.0	87.5	93.33	8	0	2	0	0	0	0	0	0	0
compliant	single	ini	100.0	80.0	88.89	100.0	80.0	88.89	8	2	0	0	0	0	0	0	0	0
compliant	single	mid	88.89	80.0	84.21	88.89	80.0	84.21	8	2	0	0	0	0	0	0	0	1
non compliant	multi	ini	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	multi	mid	90.91	100.0	95.24	90.91	100.0	95.24	10	0	0	0	0	0	0	0	0	1
non compliant	single	ini	80.0	80.0	80.0	66.67	80.0	72.73	8	2	0	0	0	0	0	0	0	2
non compliant	single	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	75.0	90.0	81.82	93.33	93.33	93.33	9	0	1	0	0	0	0	0	0	2
mixed compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	37.5	30.0	33.33	100.0	44.83	61.91	3	3	3	0	0	0	1	0	0	0
mixed ambiguous	multi	mid	75.0	60.0	66.67	100.0	66.67	80.0	6	2	2	0	0	0	0	0	0	0
ambiguous	multi	mid	33.33	20.0	25.0	66.67	26.67	38.1	2	6	2	0	0	0	0	0	0	2
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	8.33	5.0	6.25	83.33	86.96	85.11	1	0	3	0	0	8	0	0	0	0
adjacent	single	mid	50.0	35.0	41.18	73.08	95.0	82.61	7	1	0	0	0	6	0	0	0	1

Table E.44: Results obtained by BiLSTM with unadapted embeddings over BLAS with lowercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	100.0	3.33	6.45	0	9	1	0	0	0	0	0	0	0
non compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	5.0	9.52	100.0	6.52	12.24	1	19	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	5.0	9.52	100.0	5.0	9.52	1	19	0	0	0	0	0	0	0	0

Table E.45: Results obtained by BiLSTM with unadapted embeddings over BLAS with titlecase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.46: Results obtained by BiLSTM with unadapted embeddings over BLAS with uppercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
non compliant	multi	ini	50.0	10.0	16.67	100.0	13.33	23.52	1	8	1	0	0	0	0	0	0	0
non compliant	multi	mid	50.0	10.0	16.67	100.0	13.33	23.52	1	8	1	0	0	0	0	0	0	0
non compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	100.0	3.33	6.45	0	9	1	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	10.0	18.18	100.0	8.7	16.01	2	18	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	2	18	0	0	0	0	0	0	0	0

Table E.47: Results obtained by BiLSTM with unadapted embeddings over BLAS with titlecase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.48: Results obtained by BiLSTM with unadapted embeddings over BLAS with uppercase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	30.0	46.15	100.0	28.12	43.9	3	7	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	90.0	94.74	100.0	90.62	95.08	9	1	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	multi	ini	66.67	20.0	30.77	100.0	23.33	37.83	2	7	1	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	66.67	20.0	30.77	100.0	23.33	37.83	2	7	1	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	100.0	6.67	12.51	0	8	2	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	75.0	60.0	66.67	100.0	66.67	80.0	6	2	2	0	0	0	0	0	0	0
ambiguous	multi	mid	83.33	50.0	62.5	100.0	53.33	69.56	5	4	1	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	94.12	80.0	86.49	100.0	84.78	91.76	16	3	1	0	0	0	0	0	0	0
adjacent	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	20	0	0	0	0	0	0	0	0	0

Table E.49: Results obtained by BiLSTM with codeswitch embeddings over BLAS with standard casing, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	30.0	46.15	100.0	28.12	43.9	3	7	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	90.0	94.74	100.0	90.62	95.08	9	1	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
non compliant	multi	ini	87.5	70.0	77.78	100.0	73.33	84.61	7	2	1	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	50.0	20.0	28.57	100.0	26.67	42.11	2	6	2	0	0	0	0	0	0	0
mixed compliant	multi	mid	80.0	80.0	80.0	100.0	86.67	92.86	8	0	2	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	100.0	6.67	12.51	0	8	2	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	55.56	50.0	52.63	100.0	63.33	77.55	5	1	4	0	0	0	0	0	0	0
ambiguous	multi	mid	25.0	10.0	14.29	100.0	20.0	33.33	1	6	3	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	25.0	15.0	18.75	84.62	95.65	89.8	3	0	1	0	0	8	0	0	0	0
adjacent	single	mid	53.85	35.0	42.43	76.0	95.0	84.44	7	1	0	0	0	6	0	0	0	0

Table E.50: Results obtained by BiLSTM with codeswitch embeddings over BLAS with standard casing, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	85.71	60.0	70.59	85.71	56.25	67.92	6	4	0	0	0	0	0	0	0	1
compliant	multi	mid	100.0	90.0	94.74	100.0	90.62	95.08	9	1	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	80.0	88.89	100.0	80.0	88.89	8	2	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	multi	mid	90.91	100.0	95.24	90.91	100.0	95.24	10	0	0	0	0	0	0	0	0	1
non compliant	single	ini	90.91	100.0	95.24	76.92	100.0	86.95	10	0	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	87.5	70.0	77.78	100.0	73.33	84.61	7	2	1	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	55.56	50.0	52.63	100.0	63.33	77.55	5	1	4	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	75.0	60.0	66.67	100.0	66.67	80.0	6	2	2	0	0	0	0	0	0	0
ambiguous	multi	mid	55.56	50.0	52.63	76.19	53.33	62.74	5	4	1	0	0	0	0	0	0	3
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	94.12	80.0	86.49	100.0	84.78	91.76	16	3	1	0	0	0	0	0	0	0
adjacent	single	mid	95.0	95.0	95.0	95.0	95.0	95.0	19	1	0	0	0	0	0	0	0	1

Table E.51: Results obtained by BiLSTM with codeswitch embeddings over BLAS with lowercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	4	16	0	0	0	0	0	0	0	0

Table E.52: Results obtained by BiLSTM with codeswitch embeddings over BLAS with titlecase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.53: Results obtained by BiLSTM with codeswitch embeddings over BLAS with uppercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	4	16	0	0	0	0	0	0	0	0

Table E.54: Results obtained by BiLSTM with codeswitch embeddings over BLAS with titlecase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	5.0	9.52	100.0	2.17	4.25	1	19	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	5.0	9.52	100.0	5.0	9.52	1	19	0	0	0	0	0	0	0	0

Table E.55: Results obtained by BiLSTM with codeswitch embeddings over BLAS with uppercase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	90.0	90.0	90.0	90.0	84.38	87.1	9	1	0	0	0	0	0	0	0	1
compliant	multi	mid	100.0	90.0	94.74	100.0	90.62	95.08	9	1	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	multi	mid	90.91	100.0	95.24	90.91	100.0	95.24	10	0	0	0	0	0	0	0	0	1
non compliant	single	ini	90.91	100.0	95.24	76.92	100.0	86.95	10	0	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	80.0	80.0	80.0	100.0	86.67	92.86	8	0	2	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	77.78	70.0	73.69	100.0	76.67	86.79	7	1	2	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	55.56	50.0	52.63	100.0	63.33	77.55	5	1	4	0	0	0	0	0	0	0
ambiguous	multi	mid	14.29	10.0	11.77	54.55	20.0	29.27	1	6	3	0	0	0	0	0	0	3
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	25.0	15.0	18.75	84.62	95.65	89.8	3	0	1	0	0	8	0	0	0	0
adjacent	single	mid	53.33	40.0	45.71	74.07	100.0	85.1	8	0	0	0	0	6	0	0	0	1

Table E.56: Results obtained by BiLSTM with codeswitch embeddings over BLAS with lowercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	2	18	0	0	0	0	0	0	0	0

Table E.57: Results obtained by BiLSTM with codeswitch embeddings over BLAS with titlecase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0

Table E.58: Results obtained by BiLSTM with codeswitch embeddings over BLAS with uppercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	5.0	9.52	100.0	5.0	9.52	1	19	0	0	0	0	0	0	0	0

Table E.59: Results obtained by BiLSTM with codeswitch embeddings over BLAS with titlecase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	5.0	9.52	100.0	2.17	4.25	1	19	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	5.0	9.52	100.0	5.0	9.52	1	19	0	0	0	0	0	0	0	0

Table E.60: Results obtained by BiLSTM with codeswitch embeddings over BLAS with uppercase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	90.0	94.74	100.0	90.62	95.08	9	1	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	80.0	88.89	100.0	81.25	89.66	8	2	0	0	0	0	0	0	0	0
compliant	single	ini	71.43	50.0	58.82	45.45	50.0	47.62	5	5	0	0	0	0	0	0	0	2
compliant	single	mid	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
non compliant	single	ini	88.89	80.0	84.21	88.89	80.0	84.21	8	2	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	88.89	80.0	84.21	100.0	83.33	90.91	8	1	1	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	80.0	88.89	100.0	80.0	88.89	8	2	0	0	0	0	0	0	0	0
ambiguous	multi	mid	70.0	70.0	70.0	87.5	70.0	77.78	7	3	0	0	0	0	0	0	0	3
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	55.0	70.97	100.0	58.7	73.98	11	9	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	30.0	46.15	100.0	30.0	46.15	6	14	0	0	0	0	0	0	0	0

Table E.61: Results obtained by Llama3 over BLAS with standard casing, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	20.0	33.33	100.0	25.0	40.0	2	8	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	30.0	46.15	100.0	28.12	43.9	3	7	0	0	0	0	0	0	0	0
compliant	single	ini	33.33	10.0	15.38	14.29	10.0	11.77	1	9	0	0	0	0	0	0	0	2
compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
non compliant	multi	mid	66.67	40.0	50.0	100.0	46.67	63.64	4	4	2	0	0	0	0	0	0	0
non compliant	single	ini	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	50.0	20.0	28.57	87.5	23.33	36.84	2	7	1	0	0	0	0	0	0	1
mixed compliant	multi	mid	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
ambiguous	multi	mid	33.33	10.0	15.38	60.0	10.0	17.14	1	9	0	0	0	0	0	0	0	2
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	66.67	10.0	17.39	100.0	15.22	26.42	2	17	1	0	0	0	0	0	0	0
adjacent	single	mid	60.0	15.0	24.0	71.43	25.0	37.04	3	15	0	0	0	1	0	0	0	1

Table E.62: Results obtained by Llama3 over BLAS with standard casing, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	80.0	88.89	100.0	81.25	89.66	8	2	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	90.0	94.74	100.0	90.62	95.08	9	1	0	0	0	0	0	0	0	0
compliant	single	ini	80.0	80.0	80.0	57.14	80.0	66.66	8	2	0	0	0	0	0	0	0	2
compliant	single	mid	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	100.0	100.0	100.0	100.0	100.0	10	0	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	90.0	90.0	90.0	100.0	93.33	96.55	9	0	1	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
ambiguous	multi	mid	70.0	70.0	70.0	87.5	70.0	77.78	7	3	0	0	0	0	0	0	0	3
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	55.0	70.97	100.0	58.7	73.98	11	9	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	25.0	40.0	100.0	25.0	40.0	5	15	0	0	0	0	0	0	0	0

Table E.63: Results obtained by Llama3 over BLAS with lowercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	40.0	57.14	100.0	37.5	54.55	4	6	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	20.0	33.33	100.0	18.75	31.58	2	8	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
ambiguous	multi	mid	50.0	10.0	16.67	75.0	10.0	17.65	1	9	0	0	0	0	0	0	0	1
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	15.0	26.09	100.0	19.57	32.73	3	17	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	4	16	0	0	0	0	0	0	0	0

Table E.64: Results obtained by Llama3 over BLAS with titlecase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	40.0	57.14	100.0	37.5	54.55	4	6	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
compliant	single	mid	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	ini	75.0	30.0	42.86	50.0	30.0	37.5	3	7	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	15.0	26.09	100.0	10.87	19.61	3	17	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	15.0	26.09	100.0	15.0	26.09	3	17	0	0	0	0	0	0	0	0

Table E.65: Results obtained by Llama3 over BLAS with uppercase text, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	60.0	75.0	100.0	56.25	72.0	6	4	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	40.0	57.14	100.0	43.75	60.87	4	6	0	0	0	0	0	0	0	0
compliant	single	ini	71.43	50.0	58.82	45.45	50.0	47.62	5	5	0	0	0	0	0	0	0	2
compliant	single	mid	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
non compliant	single	ini	88.89	80.0	84.21	88.89	80.0	84.21	8	2	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	100.0	80.0	88.89	100.0	80.0	88.89	8	2	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	80.0	88.89	100.0	80.0	88.89	8	2	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
ambiguous	multi	mid	80.0	40.0	53.33	92.31	40.0	55.81	4	6	0	0	0	0	0	0	0	1
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	40.0	57.14	100.0	39.13	56.25	8	12	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	30.0	46.15	100.0	30.0	46.15	6	14	0	0	0	0	0	0	0	0

Table E.66: Results obtained by Llama3 over BLAS with titlecase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology								
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	60.0	75.0	100.0	62.5	76.92	6	4	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	50.0	66.67	100.0	53.12	69.38	5	5	0	0	0	0	0	0	0	0
compliant	single	ini	71.43	50.0	58.82	45.45	50.0	47.62	5	5	0	0	0	0	0	0	0	2
compliant	single	mid	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	90.0	94.74	100.0	90.0	94.74	9	1	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
ambiguous	multi	mid	80.0	40.0	53.33	92.31	40.0	55.81	4	6	0	0	0	0	0	0	0	1
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	45.0	62.07	100.0	45.65	62.68	9	11	0	0	0	0	0	0	0	0
adjacent	single	mid	100.0	40.0	57.14	100.0	40.0	57.14	8	12	0	0	0	0	0	0	0	0

Table E.67: Results obtained by Llama3 over BLAS with uppercase spans, with quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology								
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	75.0	30.0	42.86	100.0	37.5	54.55	3	6	1	0	0	0	0	0	0	0
compliant	multi	mid	100.0	40.0	57.14	100.0	37.5	54.55	4	6	0	0	0	0	0	0	0	0
compliant	single	ini	50.0	20.0	28.57	25.0	20.0	22.22	2	8	0	0	0	0	0	0	0	2
compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
non compliant	multi	mid	66.67	40.0	50.0	100.0	46.67	63.64	4	4	2	0	0	0	0	0	0	0
non compliant	single	ini	80.0	40.0	53.33	80.0	40.0	53.33	4	6	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	83.33	50.0	62.5	100.0	53.33	69.56	5	4	1	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
ambiguous	multi	mid	50.0	10.0	16.67	75.0	10.0	17.65	1	9	0	0	0	0	0	0	0	1
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	66.67	10.0	17.39	100.0	15.22	26.42	2	17	1	0	0	0	0	0	0	0
adjacent	single	mid	75.0	15.0	25.0	83.33	25.0	38.46	3	15	0	0	0	1	0	0	0	0

Table E.68: Results obtained by Llama3 over BLAS with lowercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	10.0	18.18	100.0	9.38	17.15	1	9	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	30.0	46.15	100.0	30.0	46.15	3	7	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	75.0	30.0	42.86	90.0	30.0	45.0	3	7	0	0	0	0	0	0	0	1
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	5.0	9.52	100.0	6.52	12.24	1	19	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	50.0	10.0	16.67	0	18	0	0	0	1	0	0	0	1

Table E.69: Results obtained by Llama3 over BLAS with titlecase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	ini	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
compliant	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
non compliant	single	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	50.0	10.0	16.67	75.0	10.0	17.65	1	9	0	0	0	0	0	0	0	1
mixed compliant	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	10.0	18.18	100.0	10.0	18.18	1	9	0	0	0	0	0	0	0	0
ambiguous	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	10	0	0	0	0	0	0	0	1
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	0
adjacent	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	20	0	0	0	0	0	0	0	1

Table E.70: Results obtained by Llama3 over BLAS with uppercase text, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	10.0	18.18	100.0	9.38	17.15	1	9	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	20.0	33.33	100.0	18.75	31.58	2	8	0	0	0	0	0	0	0	0
compliant	single	ini	33.33	10.0	15.38	14.29	10.0	11.77	1	9	0	0	0	0	0	0	0	2
compliant	single	mid	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
non compliant	multi	mid	66.67	20.0	30.77	100.0	23.33	37.83	2	7	1	0	0	0	0	0	0	0
non compliant	single	ini	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
non compliant	single	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	80.0	40.0	53.33	100.0	43.33	60.46	4	5	1	0	0	0	0	0	0	0
mixed compliant	multi	mid	50.0	10.0	16.67	75.0	10.0	17.65	1	9	0	0	0	0	0	0	0	1
mixed ambiguous	multi	ini	100.0	20.0	33.33	100.0	20.0	33.33	2	8	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
ambiguous	multi	mid	33.33	10.0	15.38	60.0	10.0	17.14	1	9	0	0	0	0	0	0	0	2
ambiguous	single	mid	100.0	33.33	50.0	100.0	33.33	50.0	1	2	0	0	0	0	0	0	0	0
adjacent	multi	mid	100.0	30.0	46.15	100.0	30.43	46.66	6	14	0	0	0	0	0	0	0	0
adjacent	single	mid	25.0	5.0	8.33	62.5	25.0	35.71	1	15	0	0	0	2	0	0	0	1

Table E.71: Results obtained by Llama3 over BLAS with titlecase spans, without quotation marks.

			Span evaluation			Token-separator evaluation			Extended MUC error typology									
Split	Length	Position	Span precision	Span recall	Span F1	TS precision	TS recall	TS F1	Correct	Missed	Overlap missed	Overlap spurious	Overlap s. m.	Fused	Split	Missegmented	Type	Spurious
compliant	multi	ini	100.0	30.0	46.15	100.0	28.12	43.9	3	7	0	0	0	0	0	0	0	0
compliant	multi	mid	100.0	60.0	75.0	100.0	62.5	76.92	6	4	0	0	0	0	0	0	0	0
compliant	single	ini	50.0	20.0	28.57	25.0	20.0	22.22	2	8	0	0	0	0	0	0	0	2
compliant	single	mid	100.0	60.0	75.0	100.0	60.0	75.0	6	4	0	0	0	0	0	0	0	0
non compliant	multi	ini	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
non compliant	multi	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
non compliant	single	ini	85.71	60.0	70.59	85.71	60.0	70.59	6	4	0	0	0	0	0	0	0	1
non compliant	single	mid	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
mixed compliant	multi	ini	100.0	40.0	57.14	100.0	40.0	57.14	4	6	0	0	0	0	0	0	0	0
mixed compliant	multi	mid	100.0	50.0	66.67	100.0	50.0	66.67	5	5	0	0	0	0	0	0	0	0
mixed ambiguous	multi	ini	100.0	70.0	82.35	100.0	70.0	82.35	7	3	0	0	0	0	0	0	0	0
mixed ambiguous	multi	mid	71.43	50.0	58.82	94.12	53.33	68.08	5	4	1	0	0	0	0	0	0	1
ambiguous	multi	mid	80.0	40.0	53.33	92.31	40.0	55.81	4	6	0	0	0	0	0	0	0	1
ambiguous	single	mid	0.0	0.0	0.0	0.0	0.0	0.0	0	3	0	0	0	0	0	0	0	0
adjacent	multi	mid	55.56	25.0	34.48	88.57	67.39	76.54	5	7	0	0	0	4	0	0	0	0
adjacent	single	mid	50.0	20.0	28.57	75.0	60.0	66.67	4	8	0	0	0	4	0	0	0	0

Table E.72: Results obtained by Llama3 over BLAS with uppercase spans, without quotation marks.

Appendix F

Predictive ability of BLAS

Tables F.1-F.5 report number of spans per type in the five external datasets used to evaluate the predictive ability of BLAS.

Table F.6 compares true and predicted scores and ranking positions based on BLAS results per attribute, along with Pearson and Spearman correlation coefficients.

Casing	Quotation marks	Type	Length	Position	Number
Standard	w/o quot.	non compliant	single	mid	688
Standard	w/o quot.	compliant	single	mid	141
Standard	w/o quot.	non compliant	multi	mid	123
Standard	w/o quot.	mixed compliant	multi	mid	107
Standard	with quot.	non compliant	single	mid	42
Standard	with quot.	non compliant	multi	mid	20
Standard	with quot.	mixed compliant	multi	mid	20
Standard	w/o quot.	adjacent	single	mid	14
Standard	w/o quot.	mixed ambiguous	multi	mid	13
Standard	w/o quot.	non compliant	single	ini	11
Standard	w/o quot.	compliant	multi	mid	9
Standard	with quot.	compliant	multi	mid	9
Spans are titlecase	w/o quot.	non compliant	single	mid	8
Standard	with quot.	compliant	single	mid	7
Standard	w/o quot.	non compliant	multi	ini	4
Text is lowercase	w/o quot.	non compliant	multi	mid	3
Standard	w/o quot.	ambiguous	multi	mid	3
Standard	w/o quot.	mixed compliant	multi	ini	3
Spans are titlecase	w/o quot.	compliant	single	mid	2
Spans are titlecase	w/o quot.	mixed ambiguous	multi	mid	2
Standard	w/o quot.	ambiguous	single	mid	2
Standard	w/o quot.	adjacent	multi	mid	2
Spans are uppercase	w/o quot.	non compliant	single	ini	1
Spans are titlecase	w/o quot.	non compliant	multi	mid	1
Spans are titlecase	with quot.	non compliant	multi	mid	1
Standard	w/o quot.	compliant	single	ini	1
Standard	w/o quot.	mixed ambiguous	multi	ini	1
Spans are titlecase	w/o quot.	mixed compliant	multi	mid	1

Table F.1: Number of spans per type in COALAS test set.

Casing	Quotation marks	Type	Length	Position	Number
Standard casing	with quot.	non compliant	single	mid	90
Standard casing	w/o quot.	non compliant	single	mid	79
Standard casing	with quot.	non compliant	multi	mid	30
Standard casing	with quot.	mixed compliant	multi	mid	22
Standard casing	with quot.	compliant	single	mid	19
Standard casing	w/o quot.	compliant	single	mid	17
Standard casing	with quot.	compliant	multi	mid	9
Standard casing	w/o quot.	non compliant	multi	mid	8
Standard casing	w/o quot.	mixed compliant	multi	mid	6
Spans are titlecase	with quot.	mixed compliant	multi	mid	5
Spans are titlecase	with quot.	non compliant	single	mid	4
Spans are titlecase	w/o quot.	mixed compliant	multi	mid	3
Standard casing	with quot.	non compliant	single	ini	2
Spans are titlecase	w/o quot.	non compliant	multi	mid	2
Standard casing	w/o quot.	compliant	multi	mid	2
Standard casing	w/o quot.	non compliant	single	ini	1
Spans are titlecase	with quot.	non compliant	multi	mid	1
Standard casing	w/o quot.	non compliant	multi	ini	1
Standard casing	with quot.	compliant	single	ini	1
Standard casing	w/o quot.	compliant	single	ini	1
Standard casing	with quot.	ambiguous	multi	mid	1
Standard casing	with quot.	ambiguous	single	mid	1
Spans are titlecase	with quot.	mixed ambiguous	multi	mid	1

Table F.2: Number of spans per type in COALAS development set.

Casing	Quotation marks	Type	Length	Position	Number
Standard casing	w/o quot.	non compliant	single	mid	67
Standard casing	with quot.	non compliant	single	mid	55
Standard casing	w/o quot.	compliant	single	mid	18
Standard casing	with quot.	compliant	single	mid	15
Standard casing	w/o quot.	non compliant	single	ini	10
Standard casing	w/o quot.	non compliant	multi	mid	9
Standard casing	with quot.	non compliant	multi	mid	8
Standard casing	with quot.	mixed compliant	multi	mid	7
Standard casing	w/o quot.	ambiguous	multi	mid	5
Spans are titlecase	w/o quot.	non compliant	single	mid	3
Standard casing	with quot.	compliant	multi	mid	3
Spans are titlecase	w/o quot.	mixed compliant	multi	mid	3
Standard casing	w/o quot.	mixed compliant	multi	mid	3
Standard casing	w/o quot.	mixed ambiguous	multi	mid	3
Standard casing	with quot.	mixed ambiguous	multi	mid	2
Standard casing	with quot.	non compliant	single	ini	1
Text is lowercase	w/o quot.	non compliant	single	mid	1
Spans are titlecase	w/o quot.	compliant	single	mid	1
Standard casing	with quot.	compliant	single	ini	1
Text is titlecase	w/o quot.	compliant	single	ini	1
Standard casing	with quot.	ambiguous	multi	mid	1
Standard casing	w/o quot.	mixed compliant	multi	ini	1
Standard casing	with quot.	mixed ambiguous	multi	ini	1

Table F.3: Number of spans per type in the 2020a development set.

Casing	Quotation marks	Type	Length	Position	Number
Standard casing	w/o quot.	non compliant	single	mid	77
Standard casing	with quot.	non compliant	single	mid	47
Standard casing	w/o quot.	compliant	single	mid	19
Standard casing	with quot.	compliant	single	mid	14
Standard casing	w/o quot.	non compliant	multi	mid	7
Standard casing	w/o quot.	mixed compliant	multi	mid	6
Standard casing	with quot.	mixed compliant	multi	mid	6
Standard casing	with quot.	non compliant	multi	mid	5
Standard casing	w/o quot.	non compliant	single	ini	4
Standard casing	with quot.	compliant	multi	mid	4
Standard casing	with quot.	mixed compliant	multi	ini	2
Spans are titlecase	with quot.	mixed compliant	multi	mid	2
Spans are titlecase	w/o quot.	non compliant	single	mid	2
Standard casing	w/o quot.	mixed ambiguous	multi	mid	2
Standard casing	w/o quot.	mixed compliant	multi	ini	1
Spans are titlecase	w/o quot.	mixed compliant	multi	mid	1
Spans are titlecase	with quot.	non compliant	single	mid	1
Text is lowercase	with quot.	non compliant	single	mid	1
Standard casing	with quot.	non compliant	single	ini	1
Standard casing	w/o quot.	non compliant	multi	ini	1
Standard casing	w/o quot.	compliant	multi	mid	1
Standard casing	w/o quot.	compliant	multi	ini	1
Text is lowercase	w/o quot.	compliant	single	mid	1
Standard casing	with quot.	compliant	single	ini	1
Standard casing	w/o quot.	compliant	single	ini	1
Spans are titlecase	w/o quot.	mixed ambiguous	multi	mid	1
Standard casing	with quot.	mixed ambiguous	multi	mid	1
Spans are titlecase	with quot.	mixed ambiguous	multi	mid	1
Standard casing	w/o quot.	ambiguous	multi	mid	1

Table F.4: Number of spans per type in the 2020a test set A.

Casing	Quotation marks	Type	Length	Position	Number
Standard casing	w/o quot.	non compliant	single	mid	43
Standard casing	w/o quot.	compliant	single	mid	25
Standard casing	with quot.	non compliant	single	mid	20
Standard casing	with quot.	compliant	single	mid	12
Standard casing	with quot.	mixed compliant	multi	mid	4
Standard casing	w/o quot.	non compliant	multi	mid	3
Standard casing	with quot.	non compliant	multi	mid	3
Standard casing	w/o quot.	non compliant	single	ini	2
Standard casing	with quot.	non compliant	single	ini	2
Standard casing	w/o quot.	mixed compliant	multi	mid	2
Standard casing	w/o quot.	ambiguous	multi	mid	2
Standard casing	w/o quot.	compliant	single	ini	1
Standard casing	with quot.	compliant	multi	mid	1
Spans are titlecase	with quot.	compliant	multi	mid	1
Standard casing	w/o quot.	compliant	multi	mid	1
Spans are titlecase	with quot.	non compliant	single	mid	1
titlecased text	with quot.	non compliant	single	ini	1
Spans are titlecase	w/o quot.	mixed compliant	multi	mid	1
Standard casing	with quot.	mixed ambiguous	multi	mid	1

Table F.5: Number of spans per type in the 2020a test set B.

Models	BLAS	COALAS (test set)						COALAS (dev set)						CALCS (test set A)						CALCS (test set B)						CALCS (dev set)					
	R	Pred. R		True R		True F1		Pred. R		True R		True F1		Pred. R		True R		True F1		Pred. R		True R		True F1		Pred. R		True R		True F1	
CRF	6.50	26.58	#6	44.31	#5	56.54	#5	33.13	#6	68.63	#5	71.31	#5	29.00	#6	65.57	#5	70.20	#5	24.76	#6	53.17	#5	61.19	#5	28.99	#6	64.84	#5	71.90	#5
BETO	23.55	86.87	#3	77.99	#4	82.36	#4	87.62	#3	84.05	#4	78.81	#4	83.58	#4	83.49	#3	82.71	#4	79.13	#4	81.75	#4	76.87	#3	81.60	#4	86.30	#1	87.30	#2
mBERT	23.80	86.86	#4	78.85	#3	83.64	#3	87.10	#4	84.31	#3	82.21	#3	83.86	#3	85.85	#1	84.26	#2	83.17	#3	84.13	#3	76.26	#4	82.73	#3	86.30	#1	86.30	#3
BiLSTM (unad)	23.55	87.86	#2	80.88	#2	85.14	#2	88.73	#2	89.05	#1	86.73	#1	84.85	#2	82.08	#4	83.86	#3	84.04	#2	84.92	#2	84.25	#1	83.06	#2	83.11	#4	85.25	#4
BiLSTM (cs)	23.75	92.96	#1	81.32	#1	85.76	#1	91.08	#1	86.57	#2	84.66	#2	89.43	#1	83.96	#2	84.96	#1	89.52	#1	85.71	#1	83.08	#2	87.53	#1	85.84	#3	87.44	#1
Llama3	36.37	42.93	#5	33.33	#6	39.73	#6	55.46	#5	53.27	#6	40.30	#6	48.53	#5	53.30	#6	29.62	#6	43.17	#5	40.48	#6	17.17	#6	48.49	#5	50.68	#6	30.41	#6
Correlation				0.94	0.89	0.89	0.89			0.79	0.83	0.61	0.83			0.85	0.66	0.64	0.89			0.91	0.94	0.69	0.83			0.84	0.41	0.63	0.71

Table F.6: Pearson and Spearman correlation coefficients over true recall scores, F1 scores and ranking obtained by six models over different datasets and the predicted recall scores and rankings obtained by extrapolating BLAS scores.

Dataset	**Recall correl. with BLAS prediction**	**Ranking correl. with BLAS prediction**	**Recall correl. with BLAS**	**Ranking correl. with BLAS**
COALAS test set	0.94	0.89	-0.04	-0.08
COALAS dev set	0.79	0.83	-0.23	-0.17
2020a test set A	0.85	0.66	-0.19	-0.03
2020a test set B	0.91	0.94	-0.10	-0.37
2020a dev set	0.83	0.43	-0.20	0.06
Average correl.	**0.86**	**0.75**	**-0.15**	**-0.12**
Median correl.	**0.85**	**0.83**	**-0.19**	**-0.08**

Table F.7: Pearson and Spearman correlation coefficients obtained from comparing recall score distributions and ranking between the true results obtained on a dataset, the predicted results obtained by extrapolating BLAS scores and BLAS results.

References

Agarwal, O., Yang, Y., Wallace, B. C., and Nenkova, A. (2021). Entity-Switched Datasets: An Approach to Auditing the In-Domain Robustness of Named Entity Recognition Models. arXiv:2004.04123 [cs].

Aguilar, G., AlGhamdi, F., Soto, V., Diab, M., Hirschberg, J., and Solorio, T. (2018). Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Aguilar, G., Kar, S., and Solorio, T. (2020). LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

AI@Meta (2024). Llama 3 model card.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alex, B. (2008). *Automatic detection of English inclusions in mixed-lingual data with an application to parsing*. PhD thesis, University of Edinburgh.

Alshammari, N. and Alanazi, S. (2021). The impact of using different annotation schemes on named entity recognition. *Egyptian Informatics Journal*, 22(3):295–302.

Álvarez Mellado, E. (2020a). Lázaro: An extractor of emergent anglicisms in Spanish newswire. Master's thesis, Brandeis University.

Álvarez Mellado, E. (2020b). Observatorio Lázaro - Observatorio del anglicismo en la prensa española. Online.

Álvarez Mellado, E., Espinosa Anke, L., Gonzalo Arroyo, J., Lignos, C., and Porta Zamorano, J. (2021). Overview of ADoBo 2021: Automatic Detection of Unassimilated Borrowings in the Spanish Press. *Procesamiento del Lenguaje Natural*, 67:277–285.

Álvarez Mellado, E. and Gonzalo, J. (2024). Characterizing Spans for Sequence Labeling: A Case on Anglicism Detection. *Procesamiento del Lenguaje Natural*, 73:235–246.

Álvarez Mellado, E. and Lignos, C. (2022a). Borrowing or Codeswitching? Annotating for Finer-Grained Distinctions in Language Mixing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3195–3201.

Álvarez Mellado, E. and Lignos, C. (2022b). Detecting unassimilated borrowings in Spanish: An annotated corpus and approaches to modeling. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3868–3888, Dublin, Ireland. Association for Computational Linguistics.

Andersen, G. (2012). Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In Furiassi, C., Pulcini, V., and Rodríguez González, F., editors, *The anglicization of European lexis*, pages 111–130. John Benjamins.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Ashok, D. and Lipton, Z. C. (2023). PromptNER: Prompting For Named Entity Recognition. arXiv:2305.15444 [cs].

Asociación de Academias de la Lengua Española (2010). Diccionario de americanismos. `https://lema.rae.es/damer/`.

Augenstein, I., Derczynski, L., and Bontcheva, K. (2017). Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Bali, K., Sharma, J., Choudhury, M., and Vyas, Y. (2014). "I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook. In Diab, M., Hirschberg, J., Fung, P., and Solorio, T., editors, *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, Doha, Qatar. Association for Computational Linguistics.

Balteiro, I. (2011a). Prescriptivism and descriptivism in the treatment of anglicisms in a series of bilingual Spanish-English dictionaries. *International Journal of Lexicography*, 24(3):277–305.

Balteiro, I. (2011b). A reassessment of traditional lexicographical tools in the light of new corpora: sports anglicisms in Spanish. *International Journal of English Studies*, 11(2):23–52.

Bar, K. and Dershowitz, N. (2014). The Tel aviv university system for the code-switching workshop shared task. In Diab, M., Hirschberg, J., Fung, P., and Solorio, T., editors, *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 139–143, Doha, Qatar. Association for Computational Linguistics.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., and Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bernier-Colborne, G. and Langlais, P. (2020). HardEval: Focusing on Challenging Tokens to Assess Robustness of NER. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1704–1711, Marseille, France. European Language Resources Association.

Blanch, J. M. L. (1995). Americanismo frente a españolismo lingüísticos. *Nueva revista de filología hispánica*, 43(2):433–440.

Bogunović, I. (2023). A Corpus-Based Approach To English Loanwords: Introducing The Database Of English Loanwords In Croatian. *FLUMINENSIA : časopis za filološka istraživanja*, 35(2):437–460. Publisher: Sveučilište u Rijeci, Filozofski fakultet, Odsjek za kroatistiku.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bowman, S. R. and Dahl, G. (2021). What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on Twitter. In Barzilay, R. and Johnson, M., editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Calude, A. S., Miller, S., and Pagel, M. (2020). Modelling loanword success – a sociolinguistic quantitative study of Māori loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory*, 16(1):29–66. Publisher: De Gruyter Mouton.

Cardellino, C. (2019). Spanish Billion Words Corpus and Embeddings. `https://crscardellino.github.io/SBWCE/`.

Cañete, J. (2019). Compilation of large Spanish unannotated corpora. `https://github.com/josecannete/spanish-corpora`.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Chen, S., Aguilar, G., Srinivasan, A., Diab, M., and Solorio, T. (2022). CALCS 2021 Shared Task: Machine Translation for Code-Switched Data. arXiv:2202.09625 [cs].

Chesley, P. (2010). Lexical borrowings in French: Anglicisms as a separate phenomenon. *Journal of French Language Studies*, 20(3):231–251. Publisher: Cambridge University Press.

Chesley, P. and Baayen, R. H. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, 48(6):1343.

Chinchor, N. and Sundheim, B. (1993). MUC-5 Evaluation Metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Chiruzzo, L., Agüero-Torales, M., Giménez-Lugo, G., Alvarez, A., Rodríguez, Y., Góngora, S., and Solorio, T. (2023). Overview of GUA-SPA at IberLEF 2023: Guarani-Spanish Code Switching Analysis. *Procesamiento del Lenguaje Natural*, 71:321–328.

Christen, P., Hand, D. J., and Kirielle, N. (2023). A review of the F-measure: Its History, Properties, Criticism, and Alternatives. *ACM Computing Surveys*, page 3606367.

Ciobanu, A. M. and Dinu, L. P. (2020). Automatic Identification and Production of Related Words for Historical Linguistics. *Computational Linguistics*, 45(4):667–704.

Clyne, M., Clyne, M. G., and Michael, C. (2003). *Dynamics of language contact: English and immigrant languages*. Cambridge University Press.

Colakovac, J. J. and Borucinsky, M. (2024). In the melting pot of web-crawled texts: The challenges of extracting English words from Croatian corpora. *International Journal of Applied Linguistics*, 34(1):166–182. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijal.12485.

Conklin, K. and Schmitt, N. (2012). The processing of formulaic language. *Annual review of applied linguistics*, 32:45–61.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Cooper, R., Crouch, D., Van Eijck, J., Fox, C., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., and others (1996). Using the framework. Technical report.

Cristea, A. M., Dinu, L. P., Georgescu, S., Mihai, M.-L., and Uban, A. S. (2021). Automatic Discrimination between Inherited and Borrowed Latin Words in Romance Languages. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2845–2855, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

De la Cruz Cabanillas, I. and Tejedor Martínez, C. (2012). Email or correo electrónico? Anglicisms in Spanish. *Revista española de lingüística aplicada*, (1):95–118.

De la Rosa, J. (2021). Adobo 2021: The futility of stilts for the classification of lexical borrowings in spanish. In *IberLEF@ SEPLN*, pages 947–955.

Derczynski, L., Nichols, E., van Erp, M., and Limsopatham, N. (2017). Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Deuchar, M. (2020). Code-Switching in Linguistics: A Position Paper. *Languages*, 5(2):22. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Deuchar, M. and Stammers, J. R. (2012). What is the "Nonce Borrowing Hypothesis" anyway? *Bilingualism: Language and Cognition*, 15(3):649–650. Publisher: Cambridge University Press.

Deuchar, M. and Stammers, J. R. (2016). English-Origin Verbs in Welsh: Adjudicating between Two Theoretical Approaches. *Languages*, 1(1):7. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diéguez, M. I. (2004). El anglicismo léxico en el discurso económico de divulgación científica del español de Chile. *Onomázein*, 2(10):117–141.

Ding, N., Xu, G., Chen, Y., Wang, X., Han, X., Xie, P., Zheng, H., and Liu, Z. (2021). Few-NERD: A Few-shot Named Entity Recognition Dataset. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

Dinu, L., Uban, A., Cristea, A., Dinu, A., Iordache, I.-B., Georgescu, S., and Zoicas, L. (2023). RoBoCoP: A comprehensive ROmance BOrrowing COgnate package and benchmark for multilingual cognate identification. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7610–7629, Singapore. Association for Computational Linguistics.

Dinu, L., Uban, A., Dinu, A., Iordache, I.-B., Georgescu, S., and Zoicas, L. (2024a). It takes two to borrow: a donor and a recipient. who's who? In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6023–6035, Bangkok, Thailand. Association for Computational Linguistics.

Dinu, L. P., Uban, A. S., Iordache, I.-B., Cristea, A. M., Georgescu, S., and Zoicas, L. (2024b). Pater incertus? there is a solution: Automatic discrimination between cognates and borrowings for Romance languages. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12657–12667, Torino, Italia. ELRA and ICCL.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Esuli, A. and Sebastiani, F. (2010). Evaluating Information Extraction. In Agosti, M., Ferro, N., Peters, C., de Rijke, M., and Smeaton, A., editors, *Multilingual and Multimodal Information Access Evaluation*, Lecture Notes in Computer Science, pages 100–111, Berlin, Heidelberg. Springer.

Ettinger, A., Rao, S., Daumé III, H., and Bender, E. M. (2017). Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Fazly, A., Cook, P., and Stevenson, S. (2009). Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1):61–103. Place: Cambridge, MA Publisher: MIT Press.

Fernández Gordillo, L. (2014). La lexicografía del español y el español hispanoamericano. *Andamios*, 11(26):53–89.

Freixa Aymerich, J. (2022). Garbell: l'avaluador automàtic de neologismes catalans (Garbell: the automatic catalans neologisms analyzer). *Terminàlia*, (26):7–16. Number: 26.

Fu, J., Liu, P., and Neubig, G. (2020). Interpretable Multi-dataset Evaluation for Named Entity Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Online. Association for Computational Linguistics.

Furiassi, C. and Hofland, K. (2007). The retrieval of false anglicisms in newspaper texts. In *Corpus Linguistics 25 Years On*, pages 347–363. Brill Rodopi.

Furiassi, C., Pulcini, V., and Rodríguez González, F. (2012). *The anglicization of European lexis*. John Benjamins Publishing.

Gambäck, B. and Das, A. (2016). Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855.

Garley, M. and Hockenmaier, J. (2012). Beefmoves: Dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In Li, H., Lin, C.-Y., Osborne, M., Lee, G. G., and Park, J. C., editors, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 135–139, Jeju Island, Korea. Association for Computational Linguistics.

Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Aremu, A., Bosselut, A., Chandu, K. R., Clinciu, M.-A., Das, D., Dhole, K., Du, W., Durmus, E., Dušek, O., Emezue, C. C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., Jhamtani, H., Ji, Y., Jolly, S., Kale, M., Kumar, D., Ladhak, F., Madaan, A., Maddela, M., Mahajan, K., Mahamood, S., Majumder, B. P., Martins, P. H., McMillan-Major, A., Mille, S., van Miltenburg, E., Nadeem, M., Narayan, S., Nikolaev, V., Niyongabo Rubungo, A., Osei, S., Parikh, A., Perez-Beltrachini, L., Rao, N. R., Raunak, V., Rodriguez, J. D., Santhanam, S., Sedoc, J., Sellam, T., Shaikh, S., Shimorina, A., Sobrevilla Cabezudo, M. A., Strobelt, H., Subramani, N., Xu, W., Yang, D., Yerukola, A., and Zhou, J. (2021). The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. In Bosselut, A., Durmus, E., Gangal, V. P., Gehrmann, S., Jernite, Y., Perez-Beltrachini, L., Shaikh, S., and Xu, W., editors, *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Gerding, C., Fuentes, M., Gómez, L., and Kotz, G. (2014). Anglicism: An active word-formation mechanism in Spanish. *Colombian Applied Linguistics Journal*, 16(1):40–54.

Gerding Salas, C., Cañete González, P., and Adam, C. (2018). Neología sintagmática anglicada en español: Calcos y préstamos. *Revista signos*, 51(97):175–192. Publisher: Pontificia Universidad Católica de Valparaíso.

Ghaddar, A., Langlais, P., Rashid, A., and Rezagholizadeh, M. (2021). Context-aware adversarial training for name regularity bias in named entity recognition. *Transactions of the Association for Computational Linguistics*, 9:586–604.

Gimeno Menéndez, F. and Gimeno Menéndez, M. (2003). *El desplazamiento lingüístico del español por el inglés*. Cátedra lingüística. Cátedra.

Goldberg, Y. and Elhadad, M. (2008). Identification of Transliterated Foreign Words in Hebrew Script. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 4919, pages 466–477. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Gómez Capuz, J. (1997). Towards a typological classification of linguistic borrowing (illustrated with anglicisms in romance languages). *Revista alicantina de estudios ingleses*, 10:81–94.

Gómez Capuz, J. (2004). *Los préstamos del español: lengua y sociedad*. Cuadernos de Lengua Española. Arco Libros.

Gorman, K. and Bedrick, S. (2019). We Need to Talk about Standard Splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., et al. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., and Villegas, M. (2021). Spanish language models.

Haspelmath, M. (2009). *II. Lexical borrowing: Concepts and issues*. De Gruyter Mouton. Pages: 35-54 Publication Title: Loanwords in the World's Languages Section: Loanwords in the World's Languages.

Haspelmath, M. and Tadmor, U. (2009). *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.

Haspelmath, M., Tadmor, U., and Forkel, R. (2014). The world loanword database (wold) 2009.

Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, 26(2):210–231.

Heigold, G., Varanasi, S., Neumann, G., and van Genabith, J. (2018). How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse? In Cherry, C. and Neubig, G., editors, *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 68–80, Boston, MA. Association for Machine Translation in the Americas.

Heinzerling, B. and Strube, M. (2018). BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hofland, K. (2000). A self-expanding corpus based on newspapers on the web. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. `https://spacy.io/`.

Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., Christodoulopoulos, C., Lasri, K., Saphra, N., Sinclair, A., et al. (2023). A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Jiang, S., Cui, T., Fu, Y., Lin, N., and Xiang, J. (2021). BERT4EVER at ADoBo 2021: Detection of borrowings in the Spanish language using pseudo-label technology. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings*.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. (2021). Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Kim, H. and Kang, J. (2022). How Do Your Biomedical Named Entity Recognition Models Generalize to Novel Entities? *IEEE Access*, 10:31513–31523. Conference Name: IEEE Access.

Koo, H. (2015). An unsupervised method for identifying loanwords in Korean. *Language Resources and Evaluation*, 49(2):355–373.

Kovatchev, V. and Lease, M. (2024). Benchmark transparency: Measuring the impact of data on evaluation. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1536–1551, Mexico City, Mexico. Association for Computational Linguistics.

Kudo, T. and Matsumoto, Y. (2000). Use of support vector learning for chunk identification. In *Fourth conference on computational natural language learning and the second learning language in logic workshop*.

Köllner, M. (2021). *Automatic Loanword Identification Using Tree Reconciliation*. PhD thesis. Accepted: 2021-08-20T10:22:47Z Publisher: Universität Tübingen.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In Knight, K., Nenkova, A., and Rambow, O., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Leidig, S., Schlippe, T., and Schultz, T. (2014). Automatic detection of anglicisms for the pronunciation dictionary generation: a case study on our German IT corpus. In *Spoken Language Technologies for Under-Resourced Languages*.

Lester, B. (2020). iobes: Library for span level processing. In Park, E. L., Hagiwara, M., Milajevs, D., Liu, N. F., Chauhan, G., and Tan, L., editors, *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 115–119, Online. Association for Computational Linguistics.

Levesque, H. J., Davis, E., and Morgenstern, L. (2012). The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561, Rome, Italy. AAAI Press.

Lignos, C. (2020). Nerpy: A framework for named entity recognition experiments. In *Proceedings of the LREC2020 Industry Track*, pages 36–39, Marseille, France. European Language Resources Association (ELRA).

Lignos, C. and Marcus, M. (2013). Toward web-scale analysis of codeswitching. Presented at the 87th Annual Meeting of the Linguistic Society of America.

Lillo, A. (2022). Anglicismos coloquiales en la toponimia española: Colloquial anglicisms in spanish toponymy. *Lebende Sprachen*, 67(1):133–167.

Lin, B. Y., Gao, W., Yan, J., Moreno, R., and Ren, X. (2021). RockNER: A Simple Method to Create Adversarial Examples for Evaluating the Robustness of Named Entity Recognition Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lin, H., Lu, Y., Tang, J., Han, X., Sun, L., Wei, Z., and Yuan, N. J. (2020). A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land? In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7291–7300, Online. Association for Computational Linguistics.

Linell, P. (2005). *The written language bias in linguistics : its nature, origins and transformations*. Routledge advances in communication and linguistic theory. Routledge, London, [2nd ed.]. edition.

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Lipski, J. M. (2005). Code-switching or borrowing? No sé so no puedo decir, you know. In *Selected proceedings of the second workshop on Spanish sociolinguistics*, pages 1–15. Cascadilla Proceedings Project Somerville, MA.

List, J. (2019). Automated methods for the investigation of language contact, with a focus on lexical borrowing. *Language and Linguistics Compass*, 13(10).

List, J.-M. and Forkel, R. (2021). LingPy. A Python Library for Historical Linguistics.

List, J.-M. and Moran, S. (2013). An Open Source Toolkit for Quantitative Historical Linguistics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Sofia, Bulgaria. Association for Computational Linguistics.

Lorenzo, E. (1996). *Anglicismos hispánicos*. Biblioteca románica hispánica: Estudios y ensayos. Gredos.

Losnegaard, G. S. and Lyse, G. I. (2012). A data-driven approach to anglicism identification in Norwegian. In Andersen, G., editor, *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, pages 131–154. John Benjamins Publishing.

Luján García, C. (2017). Analysis of the presence of Anglicisms in a Spanish internet forum: some terms from the fields of fashion, beauty and leisure. *Alicante Journal of English Studies*.

Luján-García, C. (2021). Anglicisms in the area of food and drinks among spanish speakers. why are anglicisms used to refer to food and drinks-related terms in spanish? *Analele Universităţii din Craiova. Seria Ştiinţe Filologice. Lingvistică*, (1-2):180–190.

Luján García, C. (2022). Are we concerned about eating healthily in spain? the increasing proliferation of anglicisms in the field of food and drink. In *45th AEDEAN Conference (Asociación Española de Estudios Anglo-Norteamericanos)*. Asociación Española de Estudios Anglo-Norteamericanos. AEDEAN.

Luján-García, C. (2023a). Anglicisms in spanish gastronomy: new words for new eating habits. *Sintagma: revista de lingüística: 35, 2023*, pages 51–69.

Luján-García, C. (2023b). '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. and Núñez-Nogueroles, E. E. (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.

Luján-García, C. (2023). Adults only or pets welcome: Use of anglicisms in the tourist domain in spanish digital press. *Lengua y Habla*, 27.

Luján-García, C. and Núñez-Nogueroles, E. E. (2022). Use of it English terms by Spanish university students: A case study. In *39º Congreso internacional de la Asociación Española de Lingüística Aplicada (AESLA)*.

Luján-García, C. I. (2021). Language contact in the media discourse: Anglicisms in section leisure in the newspaper la provincia. *Estudios de Linguistica Inglesa Aplicada (ELIA)*.

Lukichev, D., Kryanina, D., Bystrova, A., Fenogenova, A., and Tikhonova, M. (2023). Parameter-Efficient Tuning of Transformer Models for Anglicism Detection and Substitution in Russian. In *COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES"*. RSUH.

Ma, R., Wang, X., Zhou, X., Zhang, Q., and Huang, X. (2023). Towards building more robust NER datasets: An empirical study on NER dataset bias from a dataset difficulty view. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4616–4630, Singapore. Association for Computational Linguistics.

Mæhlum, P. and Ivanova, S. (2023). Phonotactics as an aid in low resource loan word detection and morphological analysis in sakha. In Ilinykh, N., Morger, F., Dannélls, D., Dobnik, S., Megyesi, B., and Nivre, J., editors, *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 111–120, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Maharjan, S., Blair, E., Bethard, S., and Solorio, T. (2015). Developing language-tagged corpora for code-switching tweets. In Meyers, A., Rehbein, I., and Zinsmeister, H., editors, *Proceedings of the 9th Linguistic Annotation Workshop*, pages 72–84, Denver, Colorado, USA. Association for Computational Linguistics.

Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. (1999). Performance Measures For Information Extraction. *Proceedings of DARPA broadcast news workshop*, pages 249–252.

Manning, C. (2006). Doing Named Entity Recognition? Don't optimize for F1. https://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html.

Mansikkaniemi, A. and Kurimo, M. (2012). Unsupervised Vocabulary Adaptation for Morph-based Language Models. In Ramabhadran, B., Khudanpur, S., and Arisoy, E., editors, *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 37–40, Montréal, Canada. Association for Computational Linguistics.

Martinović, V. (2023). Loanpydatahub/loanpy: Third stable release.

Matras, Y. and Sakel, J. (2007). *Grammatical borrowing in cross-linguistic perspective*, volume 38. Walter de Gruyter.

Mayhew, S., Tsygankova, T., and Roth, D. (2019). ner and pos when nothing is capitalized. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6256–6261, Hong Kong, China. Association for Computational Linguistics.

McClelland, J. (2021). A brief survey of anglicisms among Spanish dialects. *Colorado Research in Linguistics*.

Medina López, J. (1998). *El anglicismo en el español actual*. Cuadernos de lengua española. Arco Libros.

Mi, C. (2023a). Improving the Robustness of Loanword Identification in Social Media Texts. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):101:1–101:19.

Mi, C. (2023b). Loanword identification based on web resources: A case study on wikipedia. *Computer Speech & Language*, 81:101517.

Mi, C., Xie, L., and Zhang, Y. (2020). Loanword Identification in Low-Resource Languages with Minimal Supervision. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(3):43:1–43:22.

Mi, C., Yang, Y., Wang, L., Zhou, X., and Jiang, T. (2018a). A Neural Network Based Model for Loanword Identification in Uyghur. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mi, C., Yang, Y., Wang, L., Zhou, X., and Jiang, T. (2018b). Toward Better Loanword Identification in Uyghur Using Cross-lingual Word Embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3027–3037, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mikušová, N. (2020). Semi-Automated Methods Of Direct Anglicism Identification In Finnish Corpora. Master's thesis, University of Helsinki.

Miller, J., Pariasca, E., and Beltran Castañon, C. (2021). Neural Borrowing Detection with Monolingual Lexical Models. In Djabri, S., Gimadi, D., Mihaylova, T., and Nikolova-Koleva, I., editors, *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 109–117, Online. INCOMA Ltd.

Miller, J., Tresoldi, T., and List, J.-M. (2020a). Pybor, a python library for borrowing detection based on lexical language models. version 0.1. `https://github.com/lingpy/pybor`.

Miller, J. E. (2024). *Towards automatic detection of lexical borrowings in wordlists-with application to Latin American languages*. PhD thesis, Pontificia Universidad Catolica del Peru (Peru).

Miller, J. E. and List, J.-M. (2023). Detecting Lexical Borrowings from Dominant Languages in Multilingual Wordlists. In Vlachos, A. and Augenstein, I., editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2599–2605, Dubrovnik, Croatia. Association for Computational Linguistics.

Miller, J. E., Tresoldi, T., Zariquiey, R., Castañón, C. A. B., Morozova, N., and List, J.-M. (2020b). Using lexical language models to detect borrowings in monolingual wordlists. *PLOS ONE*, 15(12):e0242709.

Minett, J. W. and Wang, W. S.-Y. (2003). On detecting borrowing: Distance-based and character-based approaches. *Diachronica*, 20(2):289–330. Publisher: John Benjamins.

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., and Solorio, T. (2016). Overview for the second shared task on language identification in code-switched data. In Diab, M., Fung, P., Ghoneim, M., Hirschberg, J., and Solorio, T., editors, *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas. Association for Computational Linguistics.

Moreno Fernández, F. and Moreno Sandoval, A. (2018). Configuración lingüística de anglicismos procedentes de Twitter en el español estadounidense. *Revista signos*, 51(98):382–409. Publisher: Pontificia Universidad Católica de Valparaíso.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*. Cambridge University Press, Cambridge, UNITED KINGDOM.

Muñoz Ortiz, A. and Vilares, D. (2023). LyS A Coruña at GUA-SPA@IberLEF2023. Multi-Task Learning with Large Language Model Encoders for Guarani-Spanish Code Switching Analysis. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*. Accepted: 2023-09-13T07:34:38Z.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26.

Nadeau, D., Turney, P. D., and Matwin, S. (2006). Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity. In *Proceedings of the 19th international conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence*, AI'06, pages 266–277, Berlin, Heidelberg. Springer-Verlag.

Naik, A., Ravichander, A., Sadeh, N., Rose, C., and Neubig, G. (2018). Stress Test Evaluation for Natural Language Inference. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nakayama, H. (2018). seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. `https://github.com/doccano/doccano`.

Namysl, M., Behnke, S., and Köhler, J. (2021). Empirical error modeling improves robustness of noisy neural sequence labeling. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 314–329, Online. Association for Computational Linguistics.

Nath, A., Mahdipour Saravani, S., Khebour, I., Mannan, S., Li, Z., and Krishnaswamy, N. (2022). A Generalized Method for Automated Multilingual Loanword Detection. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nevěřilová, Z. (2016). Annotation of Czech Texts with Language Mixing. In Sojka, P., Horák, A., Kopeček, I., and Pala, K., editors, *Text, Speech, and Dialogue*, pages 279–286, Cham. Springer International Publishing.

Nouvel, D., Ehrmann, M., and Rosset, S. (2016). Evaluating Named Entity Recognition. In *Named Entities for Computational Linguistics*, pages 111–129. John Wiley & Sons, Ltd. Section: 6 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119268567.ch6.

Núñez Nogueroles, E. E. (2016). Anglicisms in CREA: a quantitative analysis in Spanish newspapers. *Language design: journal of theoretical and experimental linguistics*, 18:0215–242.

Núñez Nogueroles, E. E. (2017a). Typographical, orthographic and morphological variation of anglicisms in a corpus of Spanish newspaper texts. *Revista Canaria de Estudios Ingleses*, (75):175–190.

Núñez Nogueroles, E. E. (2017b). An up-to-date review of the literature on anglicisms in Spanish. *Diálogo de la Lengua, IX*, pages 1–54.

Núñez Nogueroles, E. E. (2018a). A comprehensive definition and typology of anglicisms in present-day Spanish. *Epos: Revista de filología*, (34):211–237.

Núñez Nogueroles, E. E. (2018b). A corpus-based study of anglicisms in the 21st century Spanish press. *Analecta Malacitana (AnMal electrónica)*, (44):123–159.

Núñez-Nogueroles, E. E. and Luján-García, C. (2022). Perceptions and reported use of information technology (it) anglicisms by Spanish university students. *Miscelánea: A Journal of English and American Studies*, 66:41–67.

Ok, H., Kil, T., Seo, S., and Lee, J. (2024). SCANNER: Knowledge-Enhanced Approach for Robust Multi-modal Named Entity Recognition of Unseen Entities. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7725–7737, Mexico City, Mexico. Association for Computational Linguistics.

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs). `http://www.chokkan.org/software/crfsuite/`.

Oncíns Martínez, J. L. (2012). Newly-coined anglicisms in contemporary Spanish. A corpus-based approach. In Furiassi, C., Pulcini, V., and Rodríguez González, F., editors, *The anglicization of European lexis*, pages 217–238.

Onysko, A. (2007). *Anglicisms in German: Borrowing, lexical productivity, and written codeswitching*, volume 23. Walter de Gruyter.

Palen-Michel, C., Holley, N., and Lignos, C. (2021). SeqScore: Addressing barriers to reproducible named entity recognition evaluation. In Gao, Y., Eger, S., Zhao, W., Lertvittayakumjorn, P., and Fomicheva, M., editors, *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 40–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Papay, S., Klinger, R., and Padó, S. (2020). Dissecting Span Identification Tasks with Performance Prediction. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4881–4895, Online. Association for Computational Linguistics.

Patro, J., Samanta, B., Singh, S., Basu, A., Mukherjee, P., Choudhury, M., and Mukherjee, A. (2017). All that is English may be Hindi: Enhancing language identification through automatic ranking of the likeliness of word borrowing in social media. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2264–2274, Copenhagen, Denmark. Association for Computational Linguistics.

Patzelt, C. (2011). The impact of English on Spanish-language media in the USA. In *Multilingual Discourse Production: Diachronic and Synchronic Perspectives*, volume 12, page 257. John Benjamins Publishing.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11). Publisher: Elsevier.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Peng, T. and Korobov, M. (2014). python-crfsuite. `https://github.com/scrapinghub/python-crfsuite`.

Pfaff, C. W. (1979). Constraints on Language Mixing: Intrasentential Code-Switching and Borrowing in Spanish/English. *Language*, 55(2):291–318. Publisher: Linguistic Society of America.

Phang, J., Févry, T., and Bowman, S. R. (2019). Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. arXiv preprint 1811.01088.

Plank, B. (2016). What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 13–20, Bochum, Germany.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching1. 18(7-8):581–618. Publisher: De Gruyter Mouton Section: Linguistics.

Poplack, S. (2012). What does the nonce borrowing hypothesis hypothesize? *Bilingualism: Language and Cognition*, 15(3):644–648. Publisher: Cambridge University Press.

Poplack, S. (2017). *Borrowing: Loanwords in the Speech Community and in the Grammar*. Oxford University Press.

Poplack, S. and Dion, N. (2012). Myths and facts about loanword development. *Language Variation and Change*, 24(3):279–315. Publisher: Cambridge University Press.

Poplack, S., Robillard, S., Dion, N., and Paolillo, J. C. (2020). REVISITING PHONETIC INTEGRATION IN BILINGUAL BORROWING. 96(1).

Poplack, S. and Sankoff, D. (1984). Borrowing: the synchrony of integration. *Linguistics*, 22(1).

Poplack, S., Sankoff, D., and Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1):47–104.

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Björkelund, A., Uryupina, O., Zhang, Y., and Zhong, Z. (2013). Towards robust linguistic analysis using OntoNotes. In Hockenmaier, J. and Riedel, S., editors, *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Pratt, C. (1980). *El anglicismo en el español peninsular contemporáneo*, volume 308. Gredos.

Pritzen, J., Gref, M., Schmidt, C., and Zuehlke, D. (2021). A Comparative Pronunciation Mapping Approach using G2P Conversion for Anglicisms in German Speech Recognition. In *Speech Communication; 14th ITG Conference*, pages 1–5.

Pugh, R. and Tyers, F. (2023). The ITML Submission to the IberLEF2023 Shared Task on Guarani-Spanish Code Switching Analysis. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Ratinov, L. and Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. In Stevenson, S. and Carreras, X., editors, *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Real Academia Española (2020). Diccionario de la lengua española, ed. 23.4.

Real Academia Española (2022). Fundéu. Fundación del Español Urgente. `http://www.fundeu.es`.

Real Academia Española (2011). *Ortografia de la Lengua Española*. Planeta.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Semantically equivalent adversarial rules for debugging NLP models. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Rodriguez, P., Barrow, J., Hoyle, A. M., Lalor, J. P., Jia, R., and Boyd-Graber, J. (2021). Evaluation examples are not equally informative: How should that change NLP leaderboards? In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Rodríguez González, F. (1999). Anglicisms in contemporary Spanish. An overview. *Atlantis*, 21(1/2):103–139.

Rodríguez González, F. (2002). Spanish. In Görlach, M., editor, *English in Europe*, chapter 7, pages 128–150. Oxford University Press.

Rodríguez González, F. (2018). Aspectos ortográficos del anglicismo. *Lebende Sprachen*, 63(2):350–373.

Rodríguez Medina, M. J. (2002). Los anglicismos de frecuencia sintácticos en español: estudio empírico. *RAEL. Revista electrónica de lingüística aplicada*.

Rodríguez González, F. and Lillo Buades, A. (2009). *Nuevo diccionario de anglicismos*. Biblioteca románica hispánica: Diccionarios. Gredos.

Rueda, A., Álvarez Mellado, E., and Lignos, C. (2024). CoNLL#: Fine-grained Error Analysis and a Corrected Test Set for CoNLL-03 English. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3718–3728, Torino, Italia. ELRA and ICCL.

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., and Agirre, E. (2023). NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Sanou, R. M. (2018). Anglicismos y redes sociales. *Cuadernos de la ALFAL*, 10:176–191.

Sarker, S. (2020). Codeswitch. `https://github.com/sagorbrur/codeswitch`.

Segura-Bedmar, I., Martínez, P., and Herrero-Zazo, M. (2013). SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Serigos, J. R. L. (2017a). *Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish*. PhD thesis, The University of Texas at Austin.

Serigos, J. R. L. (2017b). Using distributional semantics in loanword research: A concept-based approach to quantifying semantic specificity of Anglicisms in Spanish. *International Journal of Bilingualism*, 21(5):521–540. Publisher: SAGE Publications Sage UK: London, England.

Shirvani, R., Piergallini, M., Gautam, G. S., and Chouikha, M. (2016). The Howard University System submission for the Shared Task in Language Identification in Spanish-English Codeswitching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 116–120, Austin, Texas. Association for Computational Linguistics.

Søgaard, A., Ebert, S., Bastings, J., and Filippova, K. (2021). We need to talk about random splits. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In Diab, M., Hirschberg, J., Fung, P., and Solorio, T., editors, *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Spektor, Y. (2021). Detection and morphological analysis of novel russian loanwords. Master's thesis, City University of New York (CUNY).

Stammers, J. R. and Deuchar, M. (2012). Testing the nonce borrowing hypothesis: Counter-evidence from English-origin verbs in Welsh. *Bilingualism: Language and Cognition*, 15(3):630–643.

Stanislawek, T., Wróblewska, A., Wójcicka, A., Ziembicki, D., and Biecek, P. (2019). Named entity recognition - is there a glass ceiling? In Bansal, M. and Villavicencio, A., editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 624–633, Hong Kong, China. Association for Computational Linguistics.

Stewart, I., Yang, D., and Eisenstein, J. (2021). Tuiteamos o pongamos un tuit? Investigating the Social Constraints of Loanword Integration in Spanish Social Media. In Ettinger, A., Pavlick, E., and Prickett, B., editors, *Proceedings of the Society for Computation in Linguistics 2021*, pages 286–297, Online. Association for Computational Linguistics.

Suzuki, J., McDermott, E., and Isozaki, H. (2006). Training Conditional Random Fields with Multivariate Evaluation Measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 217–224, Sydney, Australia. Association for Computational Linguistics.

Søgaard, A., Ebert, S., Bastings, J., and Filippova, K. (2021). We Need To Talk About Random Splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Thomason, S. G. (2003). Social factors and linguistic processes in the emergence of stable mixed languages. In Matras, Yaron and Bakker, Peter, editors, *The Mixed Language Debate*, volume 145 of *Trends in Linguistics. Studies and Monographs [TiLSM]*. De Gruyter, Inc, Germany, 2003 edition. ISSN: 1861-4302.

Thomason, S. G. and Kaufman, T. (1992). *Language contact, creolization, and genetic linguistics*. Univ of California Press.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Treffers-Daller, J. (2023). The Simple View of borrowing and code-switching. *International Journal of Bilingualism*, page 13670069231168535. Publisher: SAGE Publications Ltd.

Tsvetkov, Y., Ammar, W., and Dyer, C. (2015). Constraint-Based Models of Lexical Borrowing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 598–608, Denver, Colorado. Association for Computational Linguistics.

Tsvetkov, Y. and Dyer, C. (2016). Cross-lingual bridges with models of lexical borrowing. *Journal of Artificial Intelligence Research*, 55:63–93.

Tu, J. and Lignos, C. (2021). TMR: Evaluating NER Recall on Tough Mentions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 155–163, Online. Association for Computational Linguistics.

Vajjala, S. and Balasubramaniam, R. (2022). What do we really know about State of the Art NER? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5983–5993, Marseille, France. European Language Resources Association.

Van Hout, R. and Muysken, P. (1994). Modeling lexical borrowability. *Language variation and change*, 6(1):39–62.

Van Rijsbergen, C. J. (1979). Evaluation. In *Information retrieval*. Butterworths.

Vélez Barreiro, M. (2003). *Anglicismos en la prensa económica española*. PhD thesis, Universidade da Coruña.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019a). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Wang, X., Liu, Q., Gui, T., Zhang, Q., Zou, Y., Zhou, X., Ye, J., Zhang, Y., Zheng, R., Pang, Z., Wu, Q., Li, Z., Zhang, C., Ma, R., Fei, Z., Cai, R., Zhao, J., Hu, X., Yan, Z., Tan, Y., Hu, Y., Bian, Q., Liu, Z., Qin, S., Zhu, B., Xing, X., Fu, J., Zhang, Y., Peng, M., Zheng, X., Zhou, Y., Wei, Z., Qiu, X., and Huang, X. (2021). TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing. In Ji, H., Park, J. C., and Xia, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online. Association for Computational Linguistics.

Wang, Z., Shang, J., Liu, L., Lu, L., Liu, J., and Han, J. (2019b). CrossWeigh: Training named entity tagger from imperfect annotations. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5154–5163, Hong Kong, China. Association for Computational Linguistics.

Weinreich, U. (1963). Languages in contact (1953). *The Hague: Mouton*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wu, T., Ribeiro, M. T., Heer, J., and Weld, D. (2019). Errudite: Scalable, reproducible, and testable error analysis. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763, Florence, Italy. Association for Computational Linguistics.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., and Chen, E. (2024). Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Zaitsev, K. and Minchenko, A. (2022). Automatic Detection of Borrowings in Low-Resource Languages of the Caucasus: Andic branch. In Serikov, O., Voloshina, E., Postnikova,

A., Klyachko, E., Neminova, E., Vylomova, E., Shavrina, T., Ferrand, E. L., Malykh, V., Tyers, F., Arkhangelskiy, T., Mikhailov, V., and Fenogenova, A., editors, *Proceedings of the first workshop on NLP applications to field linguistics*, pages 34–41, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Zenner, E., Speelman, D., and Geeraerts, D. (2012). Cognitive sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics*, 23(4):749–792.

Zheng, J., Ritter, A., and Xu, W. (2024). NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms. arXiv:2402.12261 [cs].

Zhou, L., Moreno-Casares, P. A., Martínez-Plumed, F., Burden, J., Burnell, R., Cheke, L., Ferri, C., Marcoci, A., Mehrbakhsh, B., Moros-Daval, Y., hÉigeartaigh, S. , Rutar, D., Schellaert, W., Voudouris, K., and Hernández-Orallo, J. (2023). Predictable Artificial Intelligence. arXiv:2310.06167 [cs].

Zurbuchen, L. and Voigt, R. (2024). A computational analysis and exploration of linguistic borrowings in French rap lyrics. In Fu, X. and Fleisig, E., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 200–208, Bangkok, Thailand. Association for Computational Linguistics.