Dialectologia. Special issue, 13 *(2024), 47-70.*
ISSN: 2013-2247

Received 24 February 2024

Accepted 9 May 2024

Published November 2024

DOI: 10.1344/DIALECTOLOGIA2024.2024B.4

INTELIGENCIA ARTIFICIAL Y ANÁLISIS DE RASGOS LINGÜÍSTICOS EN CORPUS DE TEXTOS HÍBRIDOS. EL CASO DEL CASTELLANO Y EL ASTURIANO

Cristina BLEORȚU & Miguel CUEVAS ALONSO*

Universidad "Ștefan cel Mare" / Suceava / Universidade de Vigo

cristina.bleortu@usm.ro / miguel.cuevas@uvigo.es

ORCID: 0000-0002-1645-7932 / 0000-0001-7656-2374

Resumen

El objetivo de este trabajo es explicar cómo los algoritmos lingüísticos pueden simplificar eficientemente las tareas realizadas por los lingüistas en el análisis de corpus. Para ilustrarlo, utilizaremos el corpus La Pola Siero, un conjunto de datos de un área situada en el norte de España; será el punto de partida previo al abordaje de corpus más amplios. Recopilado en 2014, se caracteriza por su naturaleza híbrida, pues incorpora textos que presentan rasgos tanto del asturiano central como del castellano. Aunque es común presentar algoritmos basados en redes neuronales, hemos optado por utilizar un clasificador bayesiano ingenuo. Esta decisión, aunque pueda considerarse algo anticuada por algunos investigadores, está justificada por varias razones: este clasificador abordará de manera efectiva la complejidad del corpus mencionado y arrojará luz sobre la relevancia y utilidad de los clasificadores bayesianos ingenuos en entornos lingüísticos específicos.

Palabras clave: clasificador bayesiano ingenuo, asturiano central, castellano, algoritmo lingüístico

* Lingvistică computațională și științe umaniste digitale, Universidad Stefan cel Mare de Suceava, Corp, birou, str. Universității, Suceava 720229 / Facultad de Filología y Traducción, Pavillón A, Despacho 56, As Lagoas, Marcosende, 36310, Vigo.

INTEL·LIGÈNCIA ARTIFICIAL I ANÀLISI DE TRETS LINGÜÍSTICS EN CORPUS DE TEXTOS HÍBRIDS. EL CAS DEL CASTELLÀ I L'ASTURIÀ

Resum

En aquest article, el nostre objectiu és explicar com els algoritmes lingüístics poden simplificar eficientment les tasques realitzades pels lingüistes en l'anàlisi de corpus lingüístics. Per il·lustrar-ho, utilitzarem el corpus La Pola Siero, un conjunt de dades d'un poble del nord d'Espanya, com a punt de partida. Recollit el 2014, aquest corpus es caracteritza per la seva naturalesa híbrida, que incorpora trets tant de l'asturià central com del castellà. Encara que és comú presentar algoritmes basats en xarxes neuronals, hem optat per utilitzar un classificador bayesià ingenu. Aquesta decisió, encara que alguns investigadors la puguin considerar una mica antiquada, està justificada per diverses raons: aquest classificador abordarà de manera efectiva la complexitat del corpus esmentat i donarà llum a la rellevància i la utilitat dels classificadors bayesians ingenus en entorns lingüístics específics.

Paraules clau: classificador bayesià ingenu, asturià central, castellà, algoritme lingüístic

ARTIFICIAL INTELLIGENCE AND ANALYSIS OF LINGUISTIC FEATURES IN HYBRID TEXT CORPORA. THE CASE OF SPANISH AND ASTURIAN

Abstract

In this paper, the aim is to explain how linguistic algorithms can efficiently simplify tasks carried out by linguists in the analysis of linguistic corpora. To illustrate this, the La Pola Siero corpus, a dataset from a town in northern Spain, will be used as the starting point. Collected in 2014, this corpus is characterized by its hybrid nature, incorporating features from both Central Asturian and Castilian. While it is common to present algorithms based on neural networks, the choice here is to use a Naive Bayesian classifier. This decision, although it may be considered somewhat outdated by some researchers, is justified for several reasons: this classifier will not only effectively address the complexity of the mentioned corpus but also shed light on the relevance and usefulness of Naive Bayesian classifiers in specific linguistic environments.

Keywords: Naive Bayesian classifier, Central Asturian, Castilian, linguistic algorithm

1. Introducción

Como indicábamos en una investigación previa (Bleorțu *et al.*, en prensa), en los últimos años hemos sido testigos del inicio de una nueva era caracterizada por los datos, los algoritmos y la inteligencia artificial (IA). Ha sido denominada como la Cuarta Revolución Industrial y se define por la convivencia entre los espacios materiales y electrónicos (Due & Toft 2021) y por la transformación de muchos de los sistemas que

nos rodean, lo que conlleva la modificación de nuestros comportamientos no solo en la vida ordinaria sino, también, en las maneras en las que se realiza investigación y en la forma en la que nos comunicamos (Jenks 2023). En esta nueva realidad, los datos se convierten en un componente esencial de nuestra vida personal y profesional, lo que ha llevado a nuevas maneras de gestión, almacenamiento, procesamiento y análisis, que implican el empleo de técnicas avanzadas para extraer información y patrones de los datos. Por esta razón, muchas veces se ha dicho que los datos son el nuevo petróleo del siglo XXI, ya que cada vez más empresas los utilizan para tomar decisiones rápidas y representan el motor de la economía (Gupta & Mamta 2024: xi), puesto que tienen aplicaciones en casi todos los campos (finanzas, marketing, medicina, historia, arte, filología, etc.) e, incluso, nos ayudan a salvar vidas (medicina de la precisión) y a ahorrar dinero y tiempo. En definitiva, los datos y su análisis han tenido tanto impacto que se ha llegado a decir que "son lo que marca la diferencia, lo que da la ventaja ante competidores, la nueva moneda" (Rodríguez 2018: 45).

Se debe poner de relieve que estas herramientas suponen un importante avance para el tratamiento de datos, especialmente cuando se trata de grandes cantidades, pues permiten abordar fenómenos complejos, incluidos los lingüísticos, desde una perspectiva más completa (Jindal *et al.* 2015; Orgueira-Crespo *et al.* 2021). Además, la IA se caracteriza por su capacidad de automejora recursiva (Coeckelbergh 2021), lo que significa que puede aprender de forma autónoma con el paso del tiempo, mejorando sus resultados de reconocimiento y clasificación.

Uno de los desafíos más significativos a los que nos enfrentamos los lingüistas en la actualidad es el procesamiento del lenguaje natural[1], y esto se vuelve aún más evidente cuando nos referimos a una lengua como el asturiano, donde la disponibilidad de corpus lingüísticos en línea es limitada y para la que, en muchas ocasiones, la hibridación con el castellano es su realidad más palpable. Con la

[1] La rapidez con la que se hace lingüística últimamente ha llevado a Kabatek (2018) a manifestarse a favor de una lingüística lenta (*slow linguistics*).

revolución de *big data*, los lingüistas también necesitamos desarrollar las herramientas necesarias para analizar corpus de complejos y grandes conjuntos de datos.

En este sentido, la aplicación de la IA permite reconocer patrones lingüísticos que incluso serían difíciles de percibir por un lingüista avezado, pues, muchas veces, es necesario un volumen de datos (textos) tan grande que el especialista apenas los podría abordar superficialmente por sí mismo. Ofrece herramientas poderosas que nos sirven para desentrañar las complejidades del discurso, permitiendo un análisis más profundo de textos escritos y orales. Esta capacidad de procesar y analizar grandes volúmenes de datos lingüísticos a una velocidad y escala sin precedentes abre nuevas perspectivas en diversos campos, desde la gramática o la sociolingüística hasta la comunicación política y los estudios de medios. No obstante, debemos tener mucho cuidado con algunas de las problemáticas de esta nueva era: 1) la veracidad de los datos; 2) el volumen; 3) la velocidad y 4) la variación y los atributos fundamentales de este nuevo periodo, tal como señala Laney (2001), porque es importante que los datos que analizamos sean de buena calidad, fiables y bastante estructurados.

La integración de disciplinas como la informática y la lingüística ha llevado a la aparición del primer traductor y analizador morfosintáctico en línea del español-asturiano y asturiano-español [2] y también al primer sistema de anotación morfosintáctica automática FreeLing (cf. Lluís Padró 2012). No obstante, queda mucho camino por recorrer; el algoritmo que presentamos en este trabajo podría contribuir a este reto apasionante frente al que nos sitúa el desarrollo del PLN y de la IA.

En nuestra opinión, una de las dificultades más interesantes del procesamiento del lenguaje natural está relacionada con el tratamiento, análisis, clasificación y procesado automático de textos de carácter híbrido, esto es, aquellos que muestran características lingüísticas de dos lenguas o variedades que se encuentran en contacto, puesto que presentan formas intermedias cuya detección es difícil de automatizar. Este hibridismo puede ocurrir en diferentes niveles de la lengua, incluyendo el léxico,

[2] <Eslema: https://eslema.it.uniovi.es/comun/traductor.php>. El Seminariu de Filoloxía Asturiana <https://sfa.grupos.uniovi.es/>, coordinado por Ramón d'Andrés, es el responsable de estas herramientas.

la morfosintaxis, el componente fónico, la semántica e, incluso, la pragmática, pues no deja de ser un reflejo de contactos culturales y sociales entre hablantes (Weinreich 1968, Sanchez-Stockhammer 2012). El análisis de este tipo de textos es interesante en sí mismo, pues permite analizar la pervivencia de fronteras entre lenguas y/o variedades, cuestionar las ideas tradicionales sobre la identidad lingüística y la cultural, ofrecer vías de investigación nuevas o alternativas para el entendimiento de las dinámicas de las lenguas en contacto.

Así pues, el objetivo de este trabajo es ofrecer una primera versión de un algoritmo que permita el análisis de corpus híbridos diferenciando las formas de las lenguas en contacto. Aunque en este primer trabajo, será probado con textos híbridos asturiano-castellano, con un grado distinto de interferencias de las dos lenguas en 24 discursos recogidos, la finalidad es que sirva como base para el análisis de otros corpus de características semejantes (francés-español, portugués-español, rumano-español, etc.).

2. El corpus

El corpus de La Pola Siero[3] constituye un conjunto de datos compuesto por 24 entrevistas realizadas oralmente y transcritas manualmente (216.000 palabras) que fueron recogidas en 2014 como parte de la tesis de doctorado de Bleorțu (2018 [publicada en 2021]). Cabe destacar que los informantes para cada una de ellas han sido seleccionados en función de las tres variables sociales clásicas que se tienen en cuenta en la mayoría de los estudios sociolingüísticos[4]: el género (hombres y mujeres), la edad (divididos en tres franjas etarias: 18-37, 38-57 y mayores de 58) y el nivel de estudios (estudios secundarios y superiores):

[3] Pueden consultarse las entrevistas en el siguiente enlace: <https://dlf.uzh.ch/sites/poladesiero/entrevistas/>.

[4] Se siguió en gran parte la metodología del proyecto PRESEEA, coordinado por Francisco Moreno Fernández: <https://preseea.uah.es/>; la autora de esta investigación participó en él como investigadora y realizó su trabajo fin de máster sobre el habla de Oviedo.

	Generación 18-37		Generación 38-57		Generación > 58	
Estudios secundarios	2 H	2 M	2 H	2 M	2 H	2 M
Estudios superiores	2 H	2 M	2 H	2 M	2 H	2 M

Figura 1. La muestra estratificada del corpus

Es necesario señalar algunas cuestiones relevantes de la situación sociolingüística del área en la que se obtuvieron los datos. Debido al devenir histórico, nos encontramos ante una zona en la que se produce contacto lingüístico; se debe indicar que "cohabitan dos lenguas románicas: (a) el asturiano y (b) el castellano; así pues, se trata de una sociedad en la que las dos lenguas tienen algún tipo de vigencia social, esto es, son usadas en determinados contextos de acuerdo con normas explícitas o implícitas" (Bleorțu 2021: 36). Ahora bien, a partir de su análisis del corpus, esta autora realiza una serie de matizaciones que ayudan a comprender las interrelaciones que, entre las dos lenguas, se producen en esta área y que revelan una compleja y dinámica gama de discursos lingüísticos en el marco de un contínuum en el que se manifiestan entrelazados elementos de ambas.

Desde el punto de vista sociolingüístico, se debe destacar, también, que algunos hablantes optan por un castellano de Asturias cercano al estándar, mientras que otros presentan variedades de castellano asturianizado en el que se observan interferencias con el asturiano central en distintos grados; además, se identifican hablantes que emplean el asturiano dialectal de la zona en cuestión, así como otros cuya variedad asturiana se acerca más a la estandarizada para esta lengua, algo que parece deberse a su participación en el proceso de normalización del asturiano, pues trabajan en la Academia de la Llingua Asturiana o publican libros en asturiano normalizado[5] (Bleorțu, 2021, Bleorțu & Cuevas-Alonso 2023a y 2023b).

[5] Para más información sobre la situación compleja de Asturias, cf. Andrés (2002).

Esta situación viene dada, en parte, por la complejidad de los contextos comunicativos en los que se ve inmerso el hablante, su pertenencia a grupos y subgrupos que son parte de un mismo macrogrupo social, a sus jerarquías y normas socioculturales y a las actitudes propias de los grupos a los que pertenece. Emergen, también, dinámicas relativas al grado de vinculación de los hablantes con el estándar castellano, a la búsqueda de un refuerzo de la identidad lingüística, al prestigio de las variantes en contacto, a la aprobación social, a la paulatina pérdida de la identificación entre grupo de pertenencia y lengua, etc. Como resultado de todo ello podemos observar que, incluso entre hablantes con similitudes sociales, afloran distintos rasgos lingüísticos y se manifiestan variaciones relevantes en los textos producidos, algo que suele pasar en nuestros días, tal como observa Blommaert (2012) en su manual de sociolingüística de la globalización.

La complejidad de estas expresiones lingüísticas ha llevado a asociar los textos con el concepto de identidad lingüística (Barnes 2016); la elección de la lengua y los rasgos lingüísticos adoptados reflejan no solo la diversidad del entorno sino, también, la riqueza y la complejidad de la(s) identidad(es) lingüísticas de los hablantes y, basándonos en estas premisas y en otras variables (como la escolarización en asturiano), los hablantes del corpus fueron clasificados en tres grupos: (1) hablantes castellanos; (2) hablantes de asturiano que no fueron alfabetizados en esa lengua y, (3) hablantes de asturiano dialectal alfabetizados en dicha lengua. Cada grupo se caracteriza por la presencia de ciertos fenómenos morfosintácticos y léxicos particulares que muestran diferentes grados de hibridación asturiano-castellano (cf. Bleorțu 2021, Bleorțu & Cuevas-Alonso 2023a, b).

3. El algoritmo

Como se ha indicado anteriormente, la utilización de herramientas de análisis *Big Data* y de IA para el análisis lingüístico es relativamente reciente y tiene beneficios

importantes para la investigación, pues permiten el procesado, la organización y el uso de grandes cantidades de datos organizados y no estructurados de procedencias muy diversas, cuya recolección puede ser, también, automatizada a gran escala y de manera extensiva, incluso aplicando automáticamente criterios de inclusión y/o de detección de errores; una buena conceptualización del algoritmo que va a llevar a cabo su adquisición/procesamiento puede garantizar una elevada calidad, pues la finalidad es que tanto su obtención como su análisis automático produzca el menor número de errores (Gupta & Mamta 2024). Esto conlleva un aumento de la eficiencia, mediante la optimización de la automatización, una mayor precisión, más flexibilidad y permite un análisis más complejo y completo de los hechos lingüísticos.

La elaboración de un algoritmo de estas características implica diversas fases: 1) la clasificación de datos, 2) la realización de predicciones, 3) la optimización del algoritmo y 4) el procesamiento de (*big*) *data* basado en la información disponible. En la primera de ellas se organizan los datos en categorías en función de sus atributos o características comunes, de modo que sirvan para establecer a qué categoría pertenece una nueva observación a partir del entrenamiento previo del algoritmo con datos ya clasificados; establece, pues, la base para la comprensión del entorno. En la fase relativa a la predicción, un modelo de aprendizaje automático debe prever resultados futuros basándose en datos históricos. En tercer lugar, es necesaria la optimización del algoritmo con el fin de mejorar su rendimiento y su eficiencia, para lo que se seleccionan las características más relevantes para el modelo, se ajustan sus parámetros o se cambia su estructura; de este modo, se mejora la precisión, se reduce el tiempo de computación o se manejan mejor los datos. Finalmente, son necesarias técnicas para trabajar con grandes conjuntos de datos que son demasiado complejos para los sistemas de procesamiento tradicionales, de modo que sea posible extraer conocimiento significativo (Schutt & O'Neil 2013, Kelleher & Tierney 2018, Matter 2024). Cada una de estas fases y procesos contribuye de manera integral a la eficiencia y a la precisión del algoritmo, desempeñando papeles distintos pero interrelacionados en el logro de resultados óptimos.

Estas etapas, fundamentales para su desarrollo, se sustentan en tres procesos cognitivos esenciales: 1) el aprendizaje, que facilita la asimilación de patrones y regularidades, 2) el razonamiento o el uso de la lógica para identificar relaciones, deducir nuevas informaciones a partir de las conocidas o resolver problemas y 3) la corrección, es decir, la capacidad de identificar y corregir errores o desviaciones respecto a lo esperado o lo correcto, lo que posibilita la adaptación continua y la mejora del algoritmo. Este enfoque integral, que abarca tanto las etapas como los procesos cognitivos subyacentes, constituye una perspectiva clave en la concepción y desarrollo de algoritmos efectivos en diversos contextos.

3.1 El preprocesamiento

Antes de desarrollar un algoritmo, es necesario tomar una serie de decisiones que permitan abordar y preprocesar el corpus con garantías, de modo que el entrenamiento sea eficaz y óptimo. Para ello y desde el punto de vista lingüístico, partimos de un enfoque similar al empleado por Bleorțu (2021) en su anotación y análisis manual del corpus, siguiendo la metodología de Andrés (2017).

En primer lugar, al abordar un texto híbrido, optamos por considerar como referencia la variedad del asturiano central, esto es, aquella a la que pertenece el habla de La Pola Siero; para el castellano nos basamos en el estándar, variedad con cada vez mayor presencia e influencia debido a la importancia actual de los medios de comunicación, si bien con particularidades específicas de Asturias. En segundo lugar, tuvimos que decidir el conjunto de variables morfosintácticas empleadas para el entrenamiento del algoritmo. En este sentido, tomamos las 16 que fueron analizadas en Bleorțu (2021) y que llevaron a la caracterización de la construcción de los discursos híbridos en el área a la que pertenece el corpus:

1) Los plurales femeninos asturianos del paradigma *-a* / *-es* y plurales femeninos castellanos del paradigma *-a* / *-es* en los artículos, sustantivos, adjetivos y pronombres: *toes les families* / *todas las familias*.
2) Masculinos singulares de los sustantivos contables, de los adjetivos y los pronombres en *-u* / *-o*: *capital de concey**u*** / *capital de concej**o***.
3) La presencia (ast.) / ausencia del neutro de materia (cast.): *población dispers**o*** / *población dispers**a.***
4) Pronombre átono de 2ª persona (*vos* / *os*): *si **vos** digo* / *si **os** digo.*
5) Pronombres átonos asturianos / castellanos de complemento indirecto: *faen lo que-**yos** da* / *hacen lo que **les** da*.
6) Enclisis / proclisis de los pronombres: *dio**me*** / ***me** dio.*
7) Presencia de los demostrativos asturianos / castellanos: *es**i*** / *est**i*** // *es**e*** / *est**e***[6].
8) Posesivo asturiano prenuclear con artículo / posesivo prenuclear sin artículo: ***el mi** sobrín* / ***mi** sobrino.*
9) Ausencia / presencia de las contracciones[7]: *un acuerdo **col** gobierno* / *un acuerdo **con el** gobierno*.
10) Uso preferente de los diminutivos en *-ín* y *-ucu* // *-ito*, *-illo*: *grup**ín*** / *grup**ucu*** // *grup**ito*** / *grup**illo.***
11) La perifrasis *ir* + infinitivo: *vamos competir* / *vamos a competir*[8].
12) Pérdida / presencia de *-r* en el infinitivo cuando le sigue un enclítico: *voi decite* / *voy a decirte*[9].
13) Formas verbales del verbo ser en el presente de indicativo (*ye*[10], *yes* / *es*, *eres*): ***ye*** / ***yeres** guapu* // ***es*** / ***eres** guapo*.

[6] Somos conscientes de que los demostrativos asturianos pueden manifestarse no solo en asturiano sino, también, en castellano hablado. No obstante, creemos que el número de ocurrencias es más pronunciado en Asturias que en otras regiones.
[7] Una vez más reconocemos que se podría tratar, también, de un fenómeno típico del castellano hablado, pero creemos que el número de ocurrencias es mayor en Asturias por influencia del asturiano.
[8] Se da la misma situación que en la de las notas 6 y 7.
[9] Cf. la nota 7.
[10] *Cf.* nota 7.

14) Algunas formas verbales de 2ª persona del singular y 3ª del plural en *-es*, *-en* / *-as*, *-an*: *tomab**es** una cervecina* / *tomab**as** una cervecina* // *los críos baj**en*** / *baj**an***.
15) Ausencia o presencia de la oposición entre un pretérito indefinido simple y un pretérito perfecto compuesto: *paróse* / *se ha parado*[11].
16) Algunas formas verbales asturianas / castellanas del presente del modo indicativo: ***paez*** *que* / ***parece*** *que* // ***sal*** *de ahí* / ***sale*** *de ahí*.

Una vez identificadas las variables relevantes de las dos lenguas en contacto y seleccionadas aquellas consideradas para el entrenamiento del algoritmo, pasamos al preprocesamiento de las transcripciones del corpus de La Pola Siero. Para ello, empleamos varias bibliotecas de Python, específicamente NLTK y Spacy. Lo que se realizó, en términos informáticos, fue una serie de actividades que se "pipelined" ('canalizaron') juntas, es decir, un conjunto de transformaciones que nos condujeron al producto final.

Así pues, durante esta fase, se eliminaron los signos de puntuación y se uniformizó el texto, cambiando las mayúsculas en minúsculas. Posteriormente, se realizó la anotación utilizando FreeLing,[12] una herramienta de etiquetado morfológico automatizado que facilitó la identificación y la clasificación de las variables señaladas. FreeLing realizó diversas tareas, incluyendo el análisis morfológico, el tratamiento de sufijos, la *retokenización* de los pronombres clíticos, la detección de palabras compuestas y el desglose de contracciones. Este proceso se llevó a cabo tanto para el asturiano como para el castellano, contribuyendo así a la aplicación del desarrollo del algoritmo.

También se realizó una revisión manual del etiquetado para asegurar su calidad. En este proceso se detectaron varios problemas: 1) la herramienta identificó el neutro

[11] Aunque este fenómeno está muy generalizado en el español norteño, los chicos jóvenes, por influencia probablemente de los medios de comunicación, emplean la distinción. También lo hacen algunas mujeres y algunos hablantes con estudios superiores. Por esta razón, decidimos utilizar esta variable para el estudio.
[12] Disponible en <https://nlp.lsi.upc.edu/freeling/index.php/node/4>.

de materia como masculino, por la concordancia del morfema correspondiente al neutro en asturiano con el del masculino en algunos casos del asturiano y con el del castellano, y 2) las perífrasis no siempre fueron identificadas, habida cuenta de que se trata de grupos de palabras unidos por una relación de sentido que, en ocasiones, escapa al etiquetador automático. Por lo que respecta al primer problema, se debe señalar que esta clasificación errada es normal cuando se entrena un algoritmo y un morfema aparece asociado a más de un rasgo. En el caso de las perífrasis se tuvo que trabajar con la generación de bigramas en Python, esto es, mediante la biblioteca KLNT, se obtuvo una lista de pares de palabras consecutivas, lo que ayuda a identificar palabras que a menudo se usan juntas, como en el caso de las perífrasis, en función de su probabilidad de coaparición.

Mediante Spacy, se dividió el texto en palabras, signos de puntuación, etc. (*tokemización*), se etiquetó la categoría léxica de cada palabra, se identificaron y clasificaron los nombres propios en categorías predefinidas, se realizó un análisis de dependencias, identificando relaciones entre "padres" e "hijos" en la estructura del árbol sintáctico y la lematización, se agruparon variantes morfológicas, se analizó la coincidencia de patrones y se vectorizaron las palabras para capturar significado semántico y similitudes morfosintácticas.

3.2 La clasificación de los datos y el algoritmo

Una vez que obtuvimos las transcripciones anotadas, y para llevar a cabo el análisis, se representó el texto como un continuum de rasgos. Para el entrenamiento escogimos un clasificador bayesiano ingenuo, es decir, un modelo de aprendizaje automático supervisado[13] que se basa en el supuesto de independencia condicional entre las características dado el valor de la clase. Aunque hubiera sido interesante

[13] En un algoritmo de aprendizaje supervisado, se identifican patrones en un conjunto de entrenamiento con el fin de establecer una correspondencia entre atributos que funcionarán como datos de referencia para realizar predicciones en un nuevo conjunto de datos. Este enfoque recibe su nombre "supervisado" debido a que el modelo tiene la capacidad de deducir información a partir de un algoritmo y un conjunto de datos previamente etiquetado, transfiriendo luego esas características a una predicción. Para más información, véase Moor (2006).

utilizar un algoritmo basado en redes neuronales, cuya principal ventaja es que sería capaz de aprender del conjunto de datos, decidimos emplear ese clasificador porque "No matter how good the learning algorithm is, it will be useless without a significant amount of relevant data. The efficiency and accuracy of these machine learning algorithms increase as we feed them with more relevant data. Data act as a blessing for machine learning algorithms" (Gupta & Mamta 2024: 69). En este sentido, entendimos que, para un corpus pequeño como el de La Pola Siero, el clasificador ingenuo es una herramienta relativamente sencilla que ofrece buenos resultados en trabajos como el que pretendíamos llevar a cabo; además, resulta particularmente beneficioso cuando no se dispone de una gran cantidad de datos de entrenamiento (Aung *et al.* 2011, Escudero *et al.* 2000, Fulmari & Chandak 2015, Wang *et al.* 2015). Es más, se trata de uno de los algoritmos de clasificación más efectivos, ya que puede realizar predicciones muy rápidamente (Gamallo *et al.* 2014, Gosal 2015), puesto que nos encontramos ante un modelo que extrae información de un corpus a partir de conjuntos de datos previamente etiquetados. Además, es posible incluir un elevado número de rasgos, potencialmente todos los que resulten necesarios para el proceso de elección de probabilidad (Nuñez Torres 2022), es decir, en la selección de modelos o distribuciones de probabilidad que mejor se ajustan a nuestro conjunto de datos o fenómeno observado para calcular, posteriormente y a partir de ellas, la probabilidad de que un elemento forme parte de una clase.

En este sentido, hemos creado el algoritmo partiendo de la entrevista 17 del corpus de La Pola Siero. Optamos por esta puesto que es la más representativa para la situación de contacto lingüístico que queríamos analizar; se hallaron muchísimos ejemplos de las variables que nos interesan, incluso de un fenómeno tan escaso como el neutro de materia. A partir de aquí, se generó una tabla de probabilidades para cada rasgo, puesto que en el caso de este algoritmo se considera que las ocurrencias del castellano y del asturiano son un conjunto de rasgos que presenta una determinada frecuencia, es decir, una probabilidad específica de manifestarse dentro de un discurso híbrido. De este modo, hemos creado la primera matriz, eso es, el procesamiento del

texto con relación a los contextos que tenía cada variable que nos interesaba, se establecieron las frecuencias de las 16 variables, tanto para el castellano como para el asturiano, y se tuvieron en cuenta los unigramas, cada palabra o token representados por cada variable y su frecuencia absoluta. Debemos tener en cuenta que el empleo de estas unidades se centra en la probabilidad de ocurrencia de palabras individuales en el corpus, sin tener en consideración la secuencia o el contexto en el que aparecen; ahora bien, son muy útiles para la clasificación de texto y la búsqueda de información. Después se generó un listado con los rasgos que mayor peso estadístico tenían dentro del corpus analizado.

Sin embargo, nos encontramos con un desafío específico al abordar el neutro de materia, puesto que contábamos con un número limitado de ejemplos, lo que generó una descompensación en la muestra (desbalance) que podría resultar en un sesgo del modelo de aprendizaje hacia las clases más representadas, con el consecuente rendimiento deficiente. Para superar esta limitación, utilizamos la técnica de sobremuestreo de la clase minoritaria mediante el uso de SMOTE (*Synthetic Minority Over-sampling Tecnique*); este método selecciona instancias cercanas en el espacio de características y genera nuevas instancias sintéticas como combinaciones lineales de aquellas de la clase minoritaria y sus vecinos más cercanos. Su ventaja radica en que reduce el riesgo de sobreajuste, es decir, de que el modelo aprenda a detectar las instancias replicadas. De este modo, se mejoró el equilibrio en la muestra y se fortaleció la capacidad del algoritmo para generalizar sobre esta variable, contribuyendo a la robustez y eficacia general del clasificador. Así, llevamos a cabo lo que siempre se recomienda cuando se debe entrenar un algoritmo, utilizar más datos: "[î]n învățarea profundă, nu sunt date mai bune decât mai multe date. Cu cât o rețea primește mai multe exemple pentru un anumit fenomen, cu atât poate discerne mai precis tiparele"[14] (Lee 2021: 32).

En un segundo paso, decidimos utilizar para el entrenamiento el 80% del corpus, esto es, 19 entrevistas. A partir de esos datos se generó una segunda matriz.

[14] "En el aprendizaje profundo, no hay mejores datos que más datos. Cuantos más ejemplos recibe una red para un fenómeno específico, más precisamente puede discernir los patrones" (traducción propia).

Aleatoriamente se generaron dos entrenamientos y una evaluación, cuyos resultados se presentan en el siguiente apartado.

3.3. Evaluación del algoritmo

Para la evaluación del algoritmo, se utilizaron las funciones que recomiendan Cahyo Untoro (2020): la precisión (*precision*), la tasa de verdaderos positivos o la sensibilidad (*recall*), la F-Score y la precisión global (*accuracy*) para las 16 variables, utilizando cuatro elementos: verdadero positivo (VP), verdadero negativo (VN), falso positivo (FP) y falso negativo (FN) y aplicando las siguientes fórmulas para cada una:

- Para la precisión:

$$Precisión = \frac{VP}{(VP + FP)}$$

- Para la tasa de verdaderos positivos o la sensibilidad:

$$Sensibilidad = \frac{VP}{(VP + VN + FP + FN)}$$

- Para F-Score:

$$F - Score = \frac{precisión * sensibilidad}{(precisión + sensibilidad)}$$

- Para la precisión global:

$$Precisión\ global = \frac{(VP + VN)}{(VP + VN + FP + FN)}$$

Los resultados que obtuvimos en la evaluación del resto del corpus de La Pola Siero (esto es, el 20% del corpus, 5 entrevistas) son los siguientes para las 16 variables que se tuvieron en cuenta:

Variables	Precisión	Sensibilidad	F-Score	Precisión Global
1 (los plurales femeninos)	92.50%	81.60%	85.50%	86.54%
2 (los masculinos singulares)	93.60%	83.40%	87.20%	88.07%
3 (neutro de materia)	72.30%	60.90%	76.79%	70%
4 (*vos* / *os*)	74.50%	62.80%	78.90%	72%
5 (complemento indirecto)	68.20%	60.20%	71.30%	67%
6 (enclisis / proclisis)	92.40%	82.70%	86.60%	82.70%
7 (demostrativos)	86%	86%	86%	85.99%
8 (posesivos)	87.30%	87.50%	87.40%	87.46%
9 (contracciones)	81.40%	81%	81.20%	81.20%
10 (diminutivos)	93.20%	92.70%	96.60%	94.17%
11 (perífrasis)	82.30%	78.90%	79.50%	80.23%
12 (infinitivos)	79.20%	78.50%	79.40%	79.03%
13 (el verbo *ser*)	91.40%	91.80%	88.60%	90.60%
14 (formas verbales de 2ª y 3ª)	93.20%	92.10%	91.80%	92.37%
15 (pretéritos)	79.80%	78.20%	79.10%	79.03%
16 (formas verbales del indicativo)	91.50%	90%	91%	91.17%

Figura 2. El entrenamiento de los datos con el clasificador bayesiano ingenuo

La tabla presenta las métricas de rendimiento para las 16 variables empleadas y podemos señalar lo siguiente (v. también los gráficos 1-4):

- En el caso de los plurales femeninos, que son muy comunes en el corpus, se observa un rendimiento sólido en términos de precisión global. No obstante, la sensibilidad es relativamente más baja, lo que indica que todavía hay algunos casos positivos que no están identificados correctamente.

- Los masculinos singulares presentan una situación parecida a los plurales femeninos.
- En las siguientes tres variables el modelo se debería ajustar para capturar mejor los casos positivos.
- En las demás variables, en general, se observa un rendimiento sólido, pero hay áreas donde el algoritmo se podría mejorar, especialmente aumentando la sensibilidad para capturar más positivos.

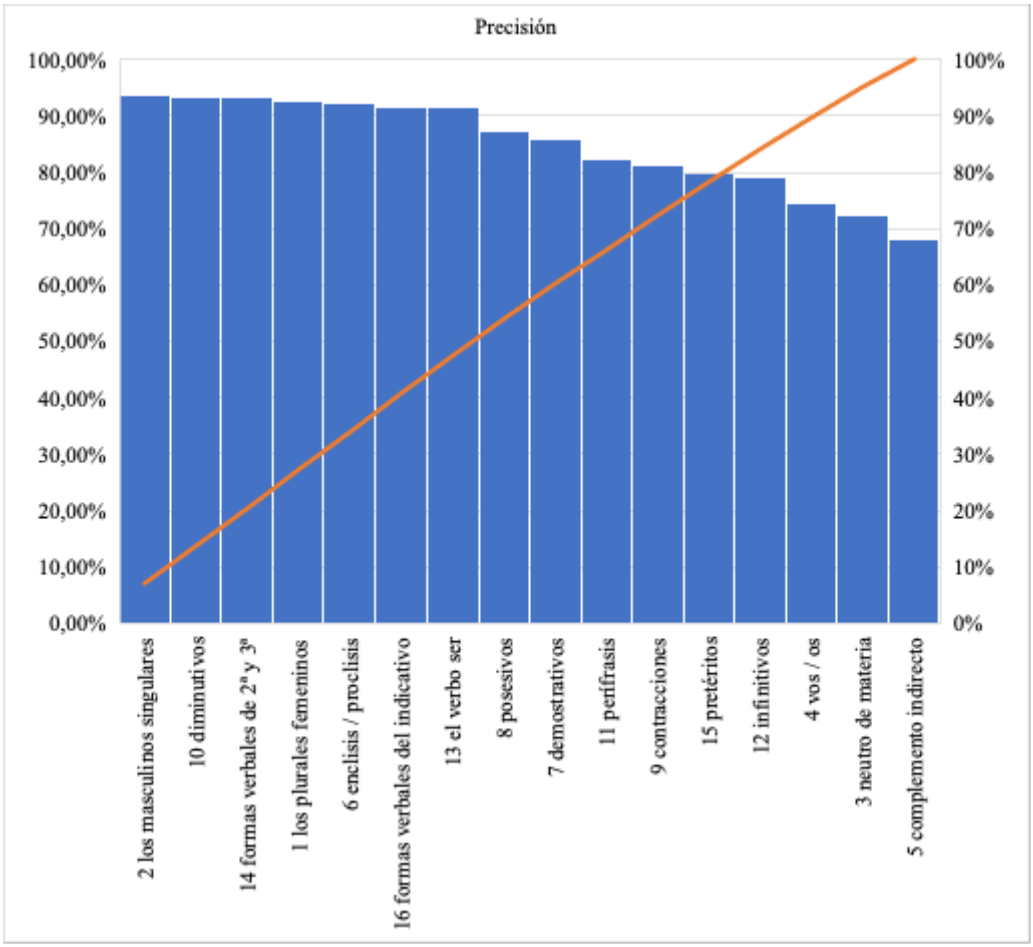

Gráfico 1. Variables ordenadas en función de la precisión

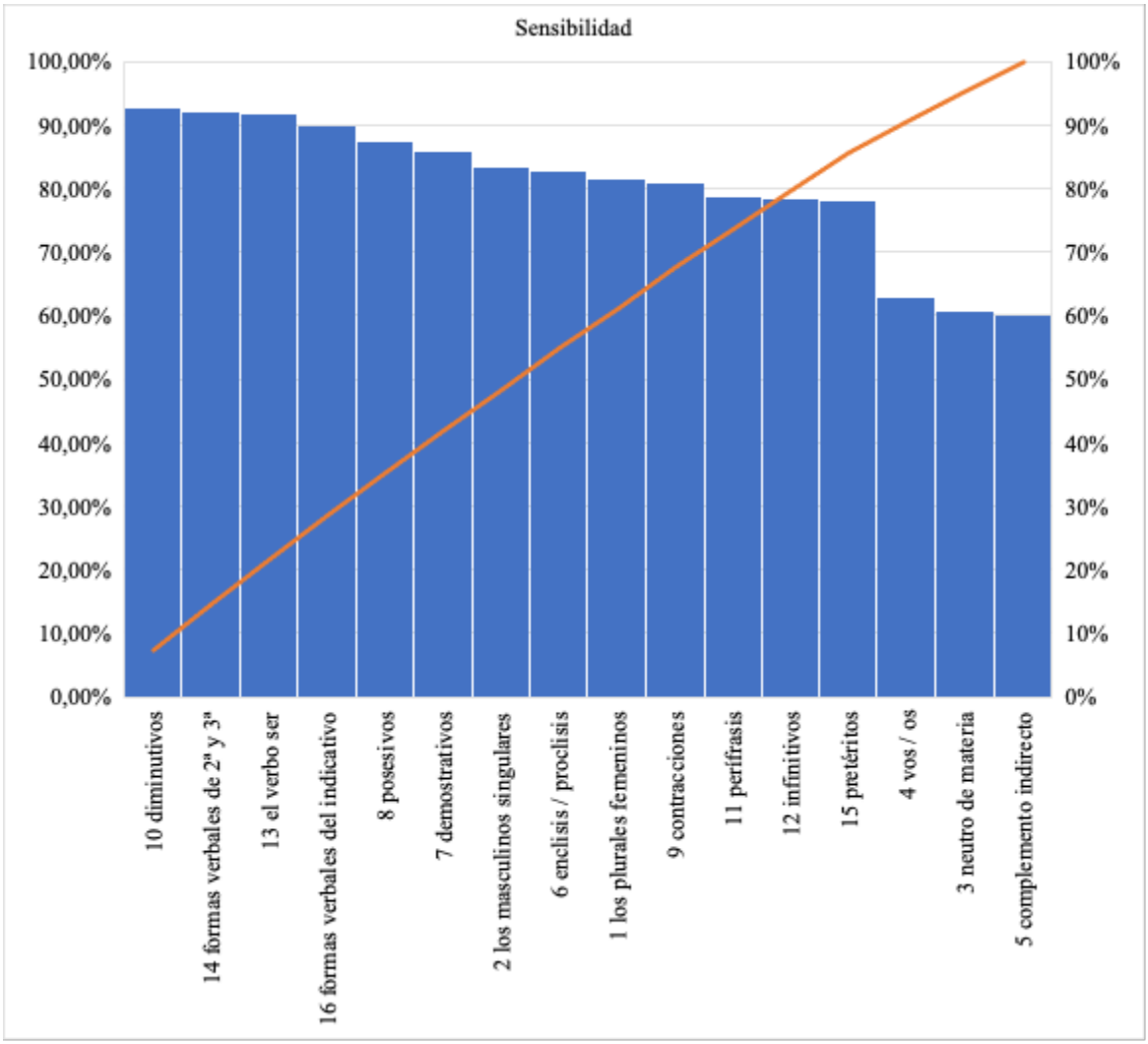

Gráfico 2. Variables ordenadas en función de la sensibilidad

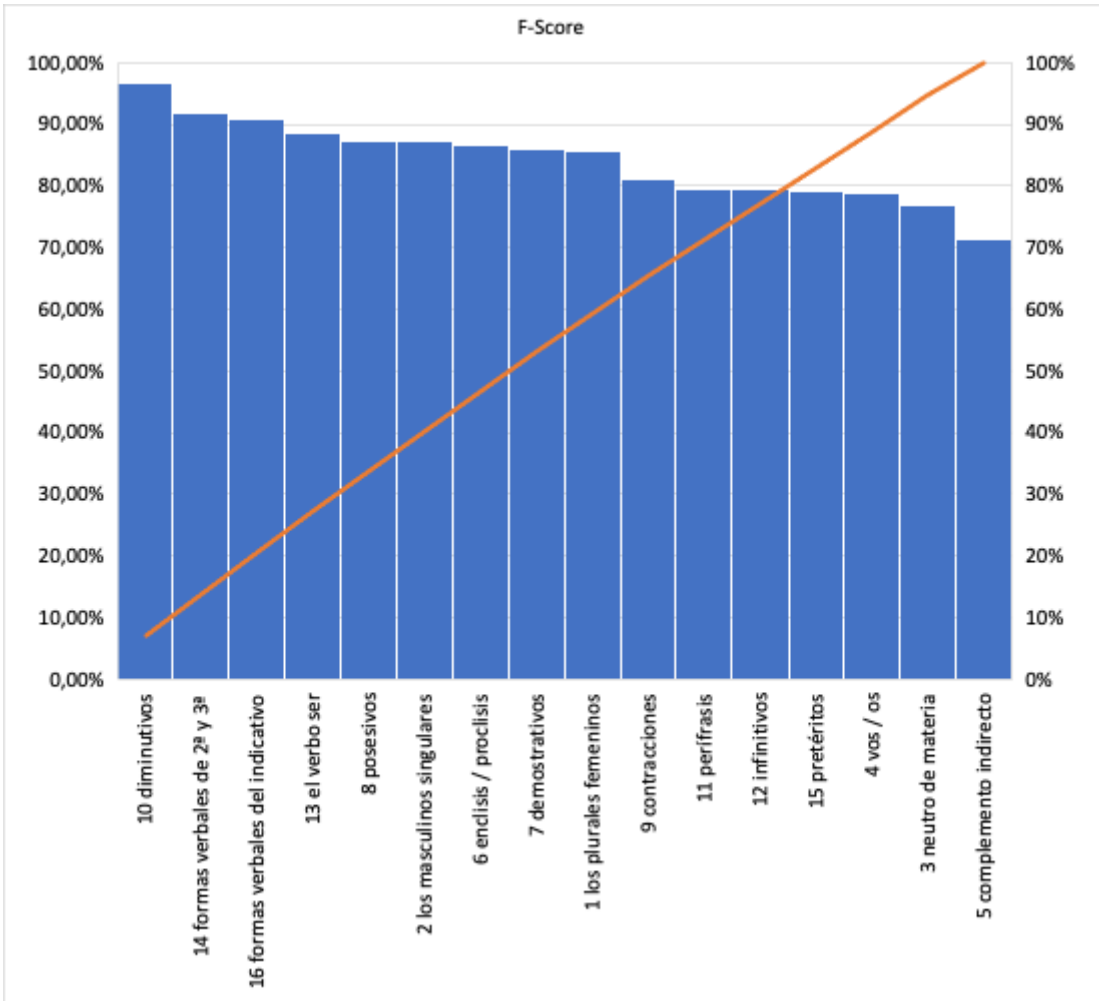

Gráfico 3. Variables ordenadas en función del F-Score

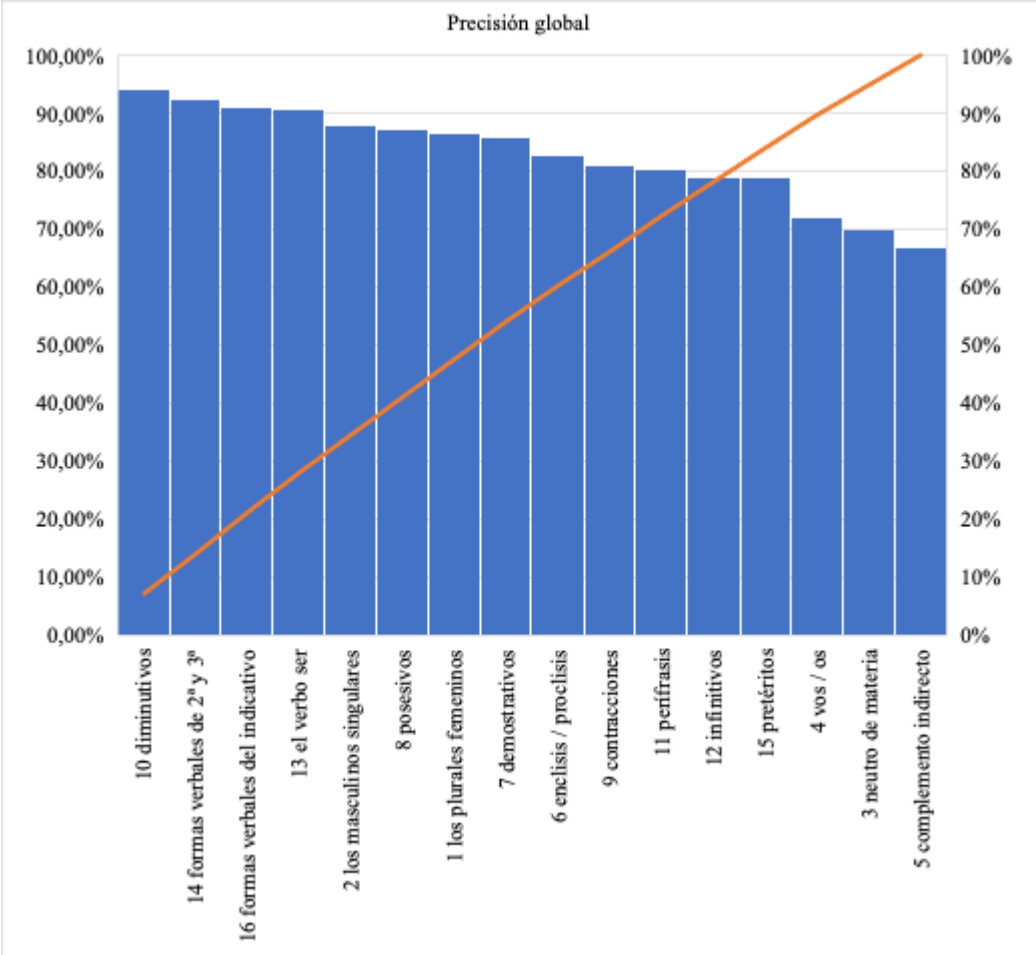

Gráfico 4. Variables ordenadas en función de la precisión global

4. Conclusiones

En esta investigación se ha presentado un algoritmo de aprendizaje automático supervisado para el análisis de la situación compleja de contacto lingüístico de La Pola Siero, una villa del norte de España en la que coexisten dos lenguas, el asturiano y el castellano. Para el algoritmo se ha elegido un clasificador bayesiano ingenuo, cuya base es la extracción de información de un corpus a partir de datos previamente etiquetados y que se justifica por su simplicidad y eficacia cuando se trabaja con corpus pequeños como el nuestro. Antes de elaborar el algoritmo, se llevó a cabo un

preprocesamiento del texto: se uniformizaron las 24 entrevistas y se anotó el corpus con FreeLing, facilitándose, así, la identificación de las variables morfosintácticas. Para el desbalance de ciertas variables se ha empleado la técnica de sobremuestreo SMOTE. Una vez que se ha tenido el algoritmo, se ha evaluado a través de métricas estándar (precisión, sensibilidad, F-Score, precisión global), que proporcionan una evaluación cuantitativa del algoritmo, permitiendo una comprensión detallada de su desempeño variable por variable. Los resultados establecen una base robusta para futuras investigaciones y refinamientos.

La clasificación efectiva de datos lingüísticos tiene aplicaciones prácticas en la comprensión de patrones sociolingüísticos y puede contribuir al desarrollo de herramientas para el análisis automático de corpus similares. No obstante, para asegurarnos de que el algoritmo funciona bien habrá que emplear otros clasificadores para la evaluación, como los de regresión logística, los árboles de decisión (*Decision tree*), Random forest, etc. También en un futuro sería procedente dar otro paso, el del aprendizaje profundo, esto es, algoritmos basados en la representación de redes neuronales.

Y así llegamos al final de nuestro homenaje, querido Ramón, donde La Pola Siero se convirtió en la musa de una investigación más. ¡Hasta la próxima; la lingüística seguirá siendo la protagonista de otra investigación junto al asturiano y al castellano!

Referencias

ANDRÉS, Ramón d' (2002) "L'asturianu mínimu urbanu. Delles hipótesis", *Lletres Asturianes*, 81, 21-38.

ANDRÉS, Ramón d' (2017) "Índiz d'asturianidá de dos testos falaos: un ensayu", *Archivum*, 67, 41-88.

AUNG, Nwe Than, Khin Mar SOE & Ni Lar THEIN (2011) "A word sense disambiguation system using naïve Bayesian algorithm for Myanmar language", *International Journal of Scientific & Engineering Research*, 9, 1-7.

BARNES, Sonia (2016a) "Negotiating local identity: rural migration and sociolinguistic perception in urban Asturias", *Lengua y migración*, 8, 2, 45-77.

BLEORȚU, Cristina (2021) *Aproximación al habla de La Pola Siero. Variación lingüística: descripción y percepción*, Uviéu: Academia de la Llingua Asturiana.

BLEORȚU, Cristina & Miguel CUEVAS-ALONSO (2023a) "Actitudes lingüísticas entre los jóvenes de La Pola Siero", en Avelino Corral Esteban (ed.), *The Asturian Language. Distinctiveness, Identity, and Officiality*, Berlín: Peter Lang.

BLEORȚU, Cristina & Miguel CUEVAS-ALONSO (2023b) "Las realizaciones de /-d/ y /-d-/ en La La Pola Siero", en Covadonga Lamar Prieto y Álvaro González Alba (eds.), *Digital Flux, Linguistic Justice and Minoritized Languages*, Berlín: de Gruyter.

BLEORȚU, Cristina, Miguel CUEVAS-ALONSO, Míriam VILLAZÓN VALBUENA & Covadonga LAMAR PRIETO (en prensa) "Lingüística e inteligencia artificial. El corpus de La Pola Siero", Star Scholars Press.

BLOMMAERT, Jan (2012) *Sociolinguistics of globalization*, Cambridge: Cambridge University Press.

CAHYO UNTORO, Meida, Mugi PRASEPTIAWAN, Mastuti WIDIANINGSIH, Ilham Firman ASHARI, Aidil AFRIANSYAH & OKTAFIANTO (2020) "Evaluation of Decision Tree, K-NN, Naive Bayes and SWM with MWNOTE on UCI Dataset", *Journal of Physics: Conference Series*, 1477, 1-9.

COECKELBERGH, Mark (2021) *Ética de la inteligencia artificial*, Barcelona: Cátedra.

DUE, Brian L. & Thomas TOFT (2021) "Phygital highlighting: Achieving joint visual attention when physically co-editing a digital text", *Journal of Pragmatics*, 177, 1-17.

ESCUDERO, G., L. MÁRQUEZ & G. RIGAU (2000) "A comparison between supervised learning algorithms for Word Sense Disambiguation", en *Actas del Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*. doi: 10.3115/1117601.1117609, 31-36.

FULMARI, Abhishek & Manoj CHANDAK (2014) An approach for Word Sense Disambiguation using modified naïve bayes classifier, *International Journal of Innovative Research in Computer and Communication Engineering Organization* 2(4), 3867-3870.

GUPTA, Brij & MAMTA (2024) *Big data: Management and analytics*, New Jersey / Beijing: World Scientific.

JENKS, Christopher Joseph (2023) *New frontiers in language and technology*, Cambridge: Cambridge University Press.

JINDAL, Rajni, Ruchika MALHOTRA & Abha JAIN (2015) "Techniques for text classification: Literature review and current trends", *Webology*, 12, 1-28

KELLEHER, John D. & Brendan TIERNEY (2018) *Data science*, Massachussets: The MIT Press.

LANEY, Doug (2001) *3D Data Management: Controlling Data Volume, Velocity, and Variety, Application Delivery Strategies*, Stanford: META Group Inc.

LEE, Kai-Fu (2021) *Superputerile inteligenței artificiale. China, Silicon Valley și noua ordine mondială*, București: Corint.

MATTER, Ulrich (2024) *Big data analytics: A guide to data science practitioners making the transition to big data*, New York: CRC Press, Taylor & Francis Group.

GAMALLO, Pablo, Susana SOTELO & José PICHEL (2014) "Comparing ranking-based and naive bayes approaches to language detection on tweets", artículo presentado en el Workshop TweetLID: Twitter Language Identification Workshop at SEPLN 2014, Girona, España, 16 de septiembre.

GOSAL, G. (2015) "A naïve bayes approach for Word Sense Disambiguation", *International Journal of Advanced Research in Computer Science and Software Engineering* 5(7), 336-340.

KABATEK, Johannes (2018) "Slow linguistics: a manifiesto", recuperado de https://www.rose.uzh.ch/de/seminar/wersindwir/mitarbeitende/kabatek.html, última consulta: 19/10/2023.

MOOR, J. (2006) "The Dartmouth College Artificial Intelligence conference: the next fifty years", *AI Magazine*, 27(4), 87-91. doi: 10.1609/aimag.v27i4.1911.

NUÑEZ TORRES, Fredy & María Beatriz PÉREZ CABELLO DE ALBA (2022) "Desarrollo de un sistema de aprendizaje automático supervisado para la desambiguación léxica automática utilizando DAMIEN (Data Mining Encountered)", *Rael (Revista Electrónica de Lingüística Aplicada)*, vol. 21(1), Jan.-Dec. 2022, pp. 150+. Gale Literature Resource Center.

ORGEIRA-CRESPO, Pedro, Carla MÍGUEZ-ÁLVAREZ, Miguel CUEVAS-ALONSO & Elena RIVO-LÓPEZ (2021) "An analysis of unconscious gender bias in academic texts by means of a decision algorithm", *PLoS ONE* 16(9): e0257903. https://doi.org/10.1371/journal. pone.0257903

OTTER, Daniel W., Julián R. MEDINA & Jugal K. KALITA (2021) "A survey of the usages of deep learning for natural language processing", *IEEE Transactions on Neural Networks and Learning Systems*, 32(2), 604-624. https://doi.org/10.1109/ tnnls.2020.2979670.

PADRÓ, Lluís & Evgeny STANILOVSKY (2012) "FreeLing 3.0: Towards Wider Multilinguality", *Proceedings of the Language Resources and Evaluation Conference* (LREC 2012) ELRA, Istanbul, Turkey, May, 2012, 2473-2479.

SANCHEZ STOCKHAMMER, Christina (2012) "Hybridization in language", en Philipp W. Stockhammer (ed.), *Conceptualizing cultural hybridization: A transdisciplinary approach*, Berlin / Heidelberg: Springer, 133-157.

SCHUTT, Rachel & Cathy O'NEIL (2013) *Doing data science*, Massachussetts: O'Reilly Media.

WANG Shasha, Liangxiao JIANG & Chaoqun LI (2016) "Adapting naive Bayes tree for text classification", *Knowledge and Information Systems*, 44(1), 77-89.

WEINREICH, Uriel (1968) *Languages in contact: Findings and problems*, New York: Mouton Publishers.