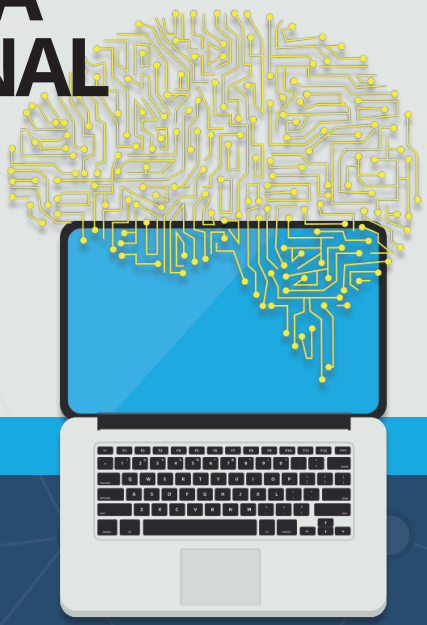
Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020.

Molina Mejia, Jorge Mauricio, Valdivia Martin, Pablo y Venegas Velásquez, René Alejandro.

LIBRO DE RESÚMENES

00
03

CONGRESO INTERNACIONAL DE **LINGÜÍSTICA COMPUTACIONAL** Y DE **CORPUS**



III Congreso Internacional de Lingüística Computacional y de Corpus
Una mirada desde las tecnologías del lenguaje y las Humanidades Digitales
(CILCC 2020)
y
V Workshop en Procesamiento Automatizado de Textos y Corpus
(WOPATEC_2020)

Editores
Jorge Mauricio Molina Mejía, Ph.D.
Pablo Valdivia Martin, Ph.D.
René Alejandro Venegas Velásquez, Ph.D.

Universidad de Antioquia / University of Groningen
Medellín 2020

Agradecimientos

El III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 tuvo lugar del 21 al 23 de octubre de 2020 en la ciudad de Medellín (Colombia). El evento que, en esta ocasión, reunió a las comunidades internacionales de los campos de la Lingüística Computacional, la Lingüística de Corpus y de las Humanidades Digitales nos reunimos con el fin de compartir e intercambiar conocimientos. En esta tercera ocasión, nuestra reunión se vio marcada por un contexto internacional complejo debido a la pandemia conocida bajo el nombre de COVID-19; por lo que el congreso se realizó bajo la modalidad virtual. En este apartado, queremos agradecer a todos aquellos estamentos que hicieron posible que el CILCC 2020 fuera una realidad.

En primer lugar, agradecer a las dos universidades organizadoras del evento, es decir, a la alianza de la Universidad de Antioquia y de la *University of Groningen*. Debemos exaltar la labor de ambos centros de educación superior para convocar y organizar este gran evento internacional en una época tan particular, como la actual.

En segundo lugar, infinitas gracias a las entidades patrocinadoras del CILCC 2020, tales como:

- La Fundación Universidad de Antioquia.
- Las Vicerrectorías de Docencia e Investigación de la Universidad de Antioquia.
- El Tecnológico de Antioquia.
- El *NAACL - The North American Chapter of the Association for Computational Linguistics*.

Gracias a ellas por el soporte institucional y económico que nos permitió asegurar la participación de los ponentes y asistentes al congreso de manera gratuita.

En tercer lugar, un agradecimiento muy especial para las instituciones que nos apoyaron durante la organización y desarrollo del congreso. La lista está conformada por:

- La Facultad de Comunicaciones de la Universidad de Antioquia.
- Ude@ Educación Virtual.
- La Universidad Nacional de Colombia.
- El *COLAB Laspau* (*Affiliated with Harvard University*).
- La Universidad Distrital Francisco José de Caldas.
- *Converging Horizons* - Chile.
- La Universidad de Medellín.
- La Universidad del Valle.
- La Cátedra Pablo Valdivia de Comunicación, Humanidades y Tecnología del Doctorado en Comunicación de la Universidad de la Frontera y la Universidad Austral de Chile.
- El *Netherlands Research School for Literary Studies*.

- WoPATeC (Workshop de Procesamiento Automático de Textos y Corpus) de la ALTL (Asociación Latinoamericana de Tratamiento del Lenguaje).

En cuarto lugar, agradecer a nuestros conferencistas invitados, a los profesores Dra. Malvina Nissim (University of Groningen), Dr. Eric Mazur (Harvard University), Dr. Pablo Valdivia (University of Groningen), Dr. Tony Berber Sardinha (Catholic University of Sao Paulo), Dr. Juan Carlos Tordera Yllescas (Universitat de València), Dr. Alfonso Ureña López (Universidad de Jaén), Dr. Felipe Bravo (Universidad de Chile) y Dra. Bárbara Poblete (Universidad de Chile). Infinitas gracias por compartir con nosotros su experticia y su conocimiento en cada uno de sus campos de investigación. Verdaderamente fue un gran honor para nosotros haber podido contar con ustedes.

En quinto lugar, queremos agradecer, muy en el alma, a todo nuestro comité científico que tuvo como tarea evaluar una gran cantidad de propuestas de investigación. A estos colegas, qué, de forma desinteresada, sacrificaron gran parte de su tiempo libre y más aún; no podemos más que decirles que sin su apoyo este congreso no habría sido posible.

En sexto lugar, muchas, pero muchas gracias en verdad a todos los colegas que creyeron en este evento y que enviaron sus propuestas para ser evaluadas, y a los que luego de ser aceptadas las presentaron en el CILCC 2020 y en WoPATeC 2020. No podemos más que señalar que sus ponencias le dieron un inmenso nivel científico a nuestros eventos.

En séptimo lugar, agradecemos de todo corazón a los miembros del comité de organización, tanto a los profesores como a los estudiantes de la Universidad de Antioquia y de Groningen por todo su trabajo: Andrés Felipe Grajales, Esther Andela, Gabriel Quiroz, Karen Rocha, Laura Quintero, Lirian Ciro, María Adelaida Zapata, María Camila Cardona, María Isabel Marín, María Victoria Delgado, Maribel Betancur, Melisa Rodríguez, Yóselin Uribe y Vanessa Zuleta. Quiero que sepan que sin su labor y apoyo incondicional, el congreso no hubiera sido más que un sueño. Un especial reconocimiento a los estudiantes y profesores, miembros del semillero de investigación *Corpus Ex Machina*, simple y llanamente ustedes son los mejores.

Finalmente, nuestro agradecimiento al público presente en cada una de las jornadas que tuvimos. Gracias, no solamente por estar ahí presentes, sino por su interés e interacción durante las diferentes sesiones o en las conferencias plenarias y talleres. Nos alegra inmensamente saber que muchos de ustedes estuvieron allí en cada una de nuestras ponencias y demás espacios del congreso.

Contenido

III Congreso Internacional de Lingüística Computacional y de Corpus "Una mirada desde las tecnologías del lenguaje y las Humanidades Digitales"

Presentación CILCC 2020

Por cuarta vez, luego de Cali en 2016, Bogotá 2017, y Medellín 2018, la comunidad de investigadores colombianos e iberoamericanos de los campos de la Lingüística Computacional y de Corpus volvemos a reunirnos con el fin de discutir y presentar nuestros trabajos y de unirnos, una vez más, como comunidad científica. En esta ocasión, además, contamos con la participación de colegas de nuestro ámbito de trabajo que presentan sus propuestas en lengua inglesa.

En esta edición, tan especial, la Universidad de Antioquia se ha hermanado con la Universidad de Groningen, que, desde el campo de las Humanidades Digitales propone, por primera vez, un congreso en el que se ven representados los tres grandes campos del conocimiento.

Es así, que para este año 2020, el III Congreso Internacional de Lingüística Computacional y de Corpus además acoge todo un conjunto de contribuciones científicas y académicas realizadas desde el ámbito de las Humanidades Digitales y que converge en la intersección de múltiples miradas complementarias y enriquecedoras.

Por todo ello, queremos dar las gracias muy especialmente a la Universidad de Antioquia, a la Universidad de Groningen y al resto de entidades que han colaborado en la organización y la preparación de este congreso internacional en el que hemos tratado transformar en un impulso creativo y positivo las dificultades y desafíos particulares que la pandemia de la Covid-19 está presentando para la actividad universitaria e investigadora.

En este sentido, con el objetivo de ofrecer una respuesta de progreso y desde la ciencia a cuestiones comunes sobre cómo generar conocimiento creativamente y transferirlo en un contexto tan demandante como en el que nos encontramos, consideramos oportuno ofrecer un workshop con el profesor Eric Mazur de la Universidad de Harvard que sin duda nos ayudará a repensar de qué manera podemos integrar el resultado de nuestras investigaciones de manera activa en nuestras prácticas docentes dentro del marco de la pandemia. Por otro lado, tenemos las ponencias plenarias de los profesores Malvina Nissim, Tony Berber Sardinha y Juan Carlos Tordera Yllescas, los cuales desde sus campos de conocimiento nos ayudarán a vislumbrar los actuales avances en el procesamiento del lenguaje natural y las metodologías basadas o conducidas por corpus.

Sin duda alguna, la tecnología, en sus múltiples formas y posibilidades, no es una opción sino un elemento que ha llegado para quedarse y que está revolucionando no sólo la manera en la que hacemos ciencia sino también cómo la enseñamos. Por ello, hemos tratado de dar cabida a la mayor pluralidad de acercamientos posibles desde el máximo rigor académico. Para así, de esta manera, ofrecer a los participantes de este congreso una constelación de los últimos avances que se están realizando desde los ámbitos de la Lingüística de Corpus, la Lingüística Computacional y las Humanidades Digitales. Bienvenidos a un universo de ideas inspirador y necesario desde las fronteras de la innovación y de la ciencia.

Para finalizar, queremos también agradecer a aquellas personas que nos enviaron sus propuestas en forma de resumen. En este sentido, poner de manifiesto que para el evento recibimos, en total, 108 resúmenes que tienen que ver con alguno de los subtemas de las tres grandes disciplinas que aquí se tratan. Luego de un riguroso proceso de evaluación se aceptaron los siguientes números de ponencias: 16 en área de la Lingüística Computacional y Procesamiento del Lenguaje Natural, 37 en lo que respecta a la Lingüística de Corpus, y 15 en el campo de las Humanidades Digitales; para un consolidado de 68 ponencias, lo que equivale a una tasa de aceptación del 62,96%. Por otro lado, los colegas de la ALTL (Asociación Latinoamericana de Tratamiento del Lenguaje) se han unido a nuestro evento con 17 ponencias más. Todo esto, sumado a las 5 conferencias plenarias y dos talleres nos da un gran total de 92 presentaciones, gracias a las cuales nuestra comunidad académica se verá altamente beneficiada tanto a nivel de la transmisión del conocimiento, como del intercambio de ideas.

Medellín - Groningen, 21-23 de octubre de 2020

Dr. Jorge Mauricio Molina Mejía, Universidad de Antioquia (Colombia).
Dr. Pablo Valdivia Martin, University of Groningen (Holanda).

Comité científico III CILCC 2020

Coordinadores

Jorge Mauricio Molina Mejía, Universidad de Antioquia (Colombia)

Pablo Valdivia Martin, University of Groningen (Netherlands)

Miembros

Ana María Agudelo Ochoa, Universidad de Antioquia (Colombia)

Bell Manrique Losada, Universidad de Medellín (Colombia)

Carlos A. Mayora Pernía, Universidad del Valle (Colombia)

Carlos M. Zapata Jaramillo, Universidad Nacional de Colombia (Colombia)

Carmina Gregori-Signes, Universitat de València (España)

Didier Schwab, Université Grenoble-Alpes (France)

Diego A. Burgos, Wake Forest University (United States of America)

Duván A. Gómez Betancur, Tecnológico de Antioquia (Colombia)

Fabio A. González Osorio, Universidad Nacional de Colombia (Colombia)

Fernán A. Villa Garzón, Universidad Nacional de Colombia (Colombia)

Gabriel A. Quiroz Herrera, Universidad de Antioquia (Colombia)

Georges Antoniadis, Université Grenoble-Alpes (France)

George E. Dueñas Luna, Universidad Nacional de Colombia (Colombia)

Jerid Francom, Wake Forest University (United States of America)

Joaquim Llisterri Boix, Universitat Autònoma de Barcelona (España)

Juan Albá Duran, University of Groningen (Netherlands)

Juan Carlos Tordera Yllescas, Universitat de València (España)

Juan Rafael Orozco Arroyave, Universidad de Antioquia (Colombia)

Juan Camilo Vásquez Correa, Friedrich Alexander Universität (Germany)

Laura M. Quintero Montoya, Universidad de Antioquia (Colombia)

Lirian Astrid Ciro, Universidad del Valle (Colombia)

Malvina Nissim, University of Groningen (Colombia)

María E. Guapacha Chamorro, University of Auckland / Universidad del Valle (New Zealand / Colombia)

María Isabel Marín, Tecnológico de Antioquia (Colombia)

Miguel Fuster Márquez, Universitat de València (España)

Norman D. Gómez Hernández, Johannes Gutenberg-Universität Mainz (Germany)

René A. Venegas Velásquez, Pontificia Universidad Católica de Valparaíso (Chile)

Sergio Jiménez Vargas, Instituto Caro y Cuervo (Colombia)

Sonia Ordoñez Salinas, Universidad Distrital Francisco José de Caldas (Colombia)

Tomás Arias Vergara, Universidad de Antioquia (Colombia)

Comité de Organización

Andrés Felipe Grajales Ramírez, Universidad de Antioquia (Colombia)

Bell Manrique Losada, Universidad de Medellín (Colombia)

Daniel Taborda Obando, Universidad de Antioquia (Colombia)

Esther Andela, University of Groningen (Netherlands)

Gabriel Ángel Quiroz Herrera, Universidad de Antioquia (Colombia)

George Enrique Dueñas Luna, Universidad Nacional de Colombia (Colombia)

Jorge Mauricio Molina Mejía, Universidad de Antioquia (Colombia)

José Luis Pemberty Tamayo, Universidad de Antioquia (Colombia)

Karen Paola Rocha Torres, Universidad de Antioquia (Colombia)

Laura Marcela Quintero Montoya, Universidad de Antioquia (Colombia)

Lirian Astrid Ciro, Universidad del Valle (Colombia)

Maria Adelaida Zapata Granados, Universidad de Antioquia (Colombia)

María Camila Cardona Tobón, Universidad de Antioquia (Colombia)

María Isabel Marín Morales, Tecnológico de Antioquia (Colombia)

Maribel Betancur Serna, Universidad de Antioquia (Colombia)

Melisa Rodríguez Bermúdez, Universidad de Antioquia (Colombia)

Pablo Valdivia Martin, University of Groningen (Netherlands)

René Alejandro Venegas Velásquez, Pontificia Universidad Católica de Valparaíso (Chile)

Sonia Ordoñez, Universidad Distrital Francisco José de Caldas (Colombia)

Yósselin Uribe Ortiz, Universidad de Antioquia, (Colombia)

V Workshop en Procesamiento Automatizado de Textos y Corpus (WOPATEC_2020)

Presentación WoPATeC 2020

La crisis social vivida en Chile durante el segundo semestre, así como los actuales eventos mundiales referidos a la pandemia, imposibilitaron el desarrollo de nuestro WOPATEC-2019. Debido a ello, y en la búsqueda de crear alianzas latinoamericanas para continuar con el espíritu de nuestro Workshop, pudimos concretar un vínculo que nos permite seguir activos en nuestro afán de formar comunidad en torno al procesamiento automatizado de textos y corpus. En este sentido, no podemos estar sino más que agradecidos de haber recibido el apoyo de los miembros del comité organizador del 3er Congreso Internacional de Lingüística Computacional y de Corpus (CILCC2020), pues nos han dado la oportunidad de realizar nuestro encuentro vía virtual en el marco de su encuentro anual. Asimismo, agradecemos a la Universidad de Antioquía y la Universidad de Groningen, y en especial al profesor Jorge Molina Mejía de la Universidad de Antioquia, Colombia y el profesor Pablo Valdivia Martin de la Universidad de Groningen, Holanda, por compartir sus experiencias, sus recursos técnicos y administrativos con nuestro WOPATEC-2020.

WOPATEC es un espacio académico de encuentro interdisciplinar en el que se reflexiona sobre el análisis automatizado de la información de los textos desde áreas tales como la lingüística de corpus, lingüística computacional, semántica computacional, ingeniería lingüística y procesamiento del lenguaje natural. Sus objetivos principales son fomentar y promover la excelencia en la investigación de los textos y los corpus textuales, a través del análisis y procesamiento automatizado de ellos en sus diversos soportes tecnológicos para contribuir a su conocimiento teórico y aplicado. Esta orientación y objetivos han permitido una potente combinación con CiLCC-2020, demostrando que en Latinoamérica existe un interés en aumento por estas áreas de investigación y desarrollo.

Esta quinta versión de WOPATEC-2020, organizada conjuntamente por el Núcleo de Investigación en Procesamiento del Lenguaje Aplicado (#NIPLA), la Asociación Latinoamericana de Tratamiento del Lenguaje (ALTL) y CiLCC-2020 se configura como una actividad de afianzamiento entre los especialistas en procesamiento automatizado de textos y visualización de aplicaciones y emprendimientos en diversos dominios disciplinares y temáticos, sumando a ello la humanidades digitales fuertemente representadas en CilCC-2020.

En esta oportunidad el programa científico de Wopatec-2020 incluye las conferencias del Dr. Alfonso Ureña de la Universidad de Jaén, y presidente de la Sociedad Española de Procesamiento del Lenguaje Natural, y la de la Dra. Bárbara Poblete de la Universidad de Chile. Asimismo, contamos con el workshop desarrollado por el Dr. Felipe Bravo de la Universidad de Chile. Aprovechamos de agradecerles su generosidad y desinteresada participación.

Además, en esta ocasión se presentaron resúmenes para demostraciones de software y comunicaciones sincrónicas y videograbadas. Así en las 16 presentaciones participaron más de 30 expositores, provenientes de Argentina, Australia, Cuba, Chile, Francia y México. Los temas tratados en este WOPATEC-2020 corresponden a identificación automatizada de hiperonimia, programas para análisis y apoyo a la escritura de textos, detección de neologismos, identificación y clasificación automatizada de textos y procesamiento del lenguaje natural a nivel semántico y discursivo.

Todo ello se complementa con las 4 conferencias, las más de 100 comunicaciones y las 9 presentaciones de doctorandos de CiLCC-2020, lo que nos permite tener un panorama amplio y rico en diversidad de investigaciones, materializado en este libro de resúmenes.

Esta primera experiencia conjunta entre WOPATEC-2020 y CiLCC-2020, sin duda, abre un espacio de colaboración y amistad académica, que esperamos se mantenga en el tiempo y se expanda a toda Latinoamérica y el mundo, convirtiéndose prontamente en un referente para el desarrollo de los intereses compartidos en nuestras disciplinas.

Por último, agradecemos el permanente patrocinio y apoyo del Instituto de Literatura y Ciencias del Lenguaje, los Programas de Postgrado en Lingüística, la Facultad de Ingeniería y la Escuela de Ingeniería Informática de la Pontificia Universidad Católica de Valparaíso, Chile. Así como el compromiso de los miembros de la comisión organizadora.

Para mayor información visite www.wopatec.cl o escriba a: wopatec@gmail.com

Comisión organizadora WOPATEC-2020

Dr. René Venegas, Pontificia Universidad Católica de Valparaíso (Chile).

Dr.(c) Rodrigo Alfaro, Pontificia Universidad Católica de Valparaíso (Chile).

Dr. Rogelio Nazar, Pontificia Universidad Católica de Valparaíso (Chile).

Dr. Pedro Alfaro, Pontificia Universidad Católica de Valparaíso (Chile).

Dr. Héctor Allende, Pontificia Universidad Católica de Valparaíso (Chile).

Ponencias plenarias / Keynote speakers

Active Learning and Perusall: Sharing Practices for a Successful Learning Engagement

Workshop

The Covid-19 crisis has been a disruptive force that has accelerated changes and transformations already present in our societies before the start of the pandemic. In regard to the field Education, the digital transformation, caused by the quick transition from traditional face-to-face teaching to remote and online learning, has impacted how we organize teaching in our institutions and has forced us to re-think the very foundations of our pedagogies.

Eric Mazur and Pablo Valdivia have extensive experience in constructive aligned course designs. They have combined active learning strategies with peer instruction and collaborative reading in Perusall. Their educational innovations provide efficient and effective new learning experiences adapted to meet our students' novel competence needs and facilitate a quality learning engagement in remote teaching and learning.

This webinar comprises two parts. In the first part, Eric Mazur will elaborate on flipped learning principles in the context of remote education and how Perusall is a useful tool for facilitating interaction. In the second part, Pablo Valdivia will expand on a study case where Perusall had a crucial role in developing students' intrinsic motivation and resulted in a co-creation student-driven learning experience.

Eric Mazur[1] *Harvard University*, United States
[http://ericmazur.com/]

Pablo Valdivia[2] *University of Groningen*, Netherlands
[p.valdivia.martin@rug.nl]

[1] **Eric Mazur** is the Balkanski Professor of Physics and Applied Physics and Area Chair of Applied Physics at Harvard University, Member of the Faculty of Education at the Harvard Graduate School of Education, and Past President of the Optical Society.
Mazur is a prominent physicist known for his contributions in nanophotonics, an internationally recognized educational innovator, and a sought-after speaker. In education he is widely known for his work on Peer Instruction, an interactive teaching method aimed at engaging students in the classroom and beyond. In 2014 Mazur became the inaugural recipient of the Minerva Prize for Advancements in Higher Education. He has received many awards for his work in physics and in education and has founded several successful companies. Mazur has widely published in peer-reviewed journals and holds numerous patents. He has also written extensively on education and is the author of Peer Instruction: A User's Manual (Prentice Hall, 1997), a book that explains how to teach large lecture classes interactively, and of the Principles and Practice of Physics (Pearson, 2015), a book that presents a groundbreaking new approach to teaching introductory calculus-based physics.
Mazur is a leading speaker on optics and on education. His motivational lectures on interactive teaching, educational technology, and assessment have inspired people around the world to change their approach to teaching.

[2] **Pablo Valdivia** is Full Professor and Chair of European Culture and Literature (University of Groningen), Associate in Applied Physics at Harvard Paulson School of Engineering and Applied Sciences (Harvard University), Academic Director of the Netherlands Research School for Literary Studies (OSL), among others important associations and relations with academic institutions around the world. Before joining the University of Groningen in 2016, he worked at the University of Amsterdam, The Cambridge Foundation Villiers Park and the University of Nottingham. He obtained a Research MA degree on "Research in European Literature and Theatre" awarded by UNED (Spain). In 2007, he received his PhD degree on "Philosophy of Hispanic Studies" at the University of Nottingham.
His research deals primarily with the Humanities, Reading Science, Cultural Analytics and Technology, and the notions of Culture, Literature and Crisis from an interdisciplinary transnational perspective. He is an expert on Cultural Narratives and Conceptual Metaphors. He carries multidisciplinary research with special emphasis in the fields of Cultural Analytics, Artificial Intelligence, Reading Science, University Innovation, Data Science, Applied Physics, Social Sciences and Cognitive Sciences.
In 2018, Prof. Dr. Pablo Valdivia was awarded "Lecturer of the Year" Faculty of Arts (University of Groningen).

Análisis de sentimientos

Conferencia

Alfonso Ureña[3]

[laurena@ujaen.es]

Universidad de Jaén, España

En esta charla se presenta la importancia del análisis de sentimientos y el porqué de su investigación. Se motiva el interés suscitado en los últimos tiempos con casos reales. Asimismo, se profundiza en las tareas específicas y componentes de la minería de opiniones, exponiendo el estado actual de la investigación, los retos y nuevas tareas. Seguidamente se aborda la tarea de análisis de sentimientos, la minería de emociones, una tarea de gran relevancia actualmente en el ámbito del Procesamiento del Lenguaje Natural. Para ello, se presentan las principales aplicaciones, centrándonos en la experimentación de sistemas de alerta temprana en el ámbito de la detección del discurso del odio y de las enfermedades mentales. Concretamente en un sistema específico para abordar la xenofobia y la misoginia a través de las redes sociales, en el que se muestran los diferentes enfoques para su detección. Del mismo modo, se muestra un sistema específico para la detección de la anorexia en las redes sociales. Seguidamente se presentan diferentes recursos lingüísticos generados para el

[3] Prof. **L. Alfonso Ureña López** is full professor in the Department of Computer Science at University of Jaén (Spain). He is author of over 200 publications on various topics of Natural Language Processing (NLP). Editor-In-Chief of the Journal of Procesamiento del Lenguaje Natural. Also, he is President of Spanish Society for Natural Language Processing (SEPLN). Programme Chair and keynote speaker of several major international conferences. Dr. Ureña is Director of the Research Institute in Information and Communication Technologies (University of Jaén).

análisis de sentimientos, especialmente para el español. Se concluye la charla con las nuevas tendencias.

Automatizando la extracción de conocimientos desde las redes sociales: ¿cómo generar valor en un mundo incierto?

Conferencia

Bárbara Poblete[4]
[bpoblete@uchile.cl]
Universidad de Chile, Chile

En esta charla hablaré sobre diversos temas relacionados al valor de la información que se puede extraer desde las redes sociales, en particular para el manejo de desastres y para comprender el mundo real. Además, hablaré sobre nuestra investigación en relación con noticias falsas e identificación de lenguaje de odio.

[4] **Bárbara Poblete** es Profesora Asociada del Departamento de Computación (DCC) de la Universidad de Chile e investigadora asociada del Instituto Milenio Fundamentos de los Datos (IMFD). Tiene un PhD de la Universitat Pompeu Fabra en España y fue investigadora en Yahoo Labs por seis años, donde realizó investigación aplicada a la industria. Actualmente, se desempeña en las áreas de minería de datos, análisis de redes sociales y recuperación de información en la Web. Su investigación en el tema de credibilidad de información en Twitter fue la primera en abordar este tema y es ampliamente reconocida. Su trabajo ha sido difundido por medios destacados como Scientific American Magazine, The Wall Street Journal, Slate Magazine, BBC News, entre otros. Cuenta con más de 60 publicaciones científicas. Más información https://www.barbara.cl/

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Breve recorrido de las gramáticas formales utilizadas en Lingüística computacional: logros y retos del análisis formal

Conferencia

Juan Carlos Tordera Yllescas[5]
[juan.tordera@uv.es]
Universitat de València, España

El objetivo de la presente ponencia es presentar un breve recorrido de las principales gramáticas formales que se gestaron en el siglo XX y que han sido utilizadas en el campo de la Lingüística computacional (LC). En concreto, se contextualizará las aportaciones de cada gramática dentro del contexto de la Historia de la Lingüística y se realizará un pequeño análisis comparativo de distintas cuestiones lingüísticas, como es el tratamiento de los elementos exigidos/subcategorizados, los elementos dependientes no acotados/desplazados, el tratamiento del significado oracional (Sintaxis <-> Semántica), entre otras cuestiones.

Aunque la Gramática Generativa (Transformacional) se haya mostrado una gramática compleja de aplicar para la LC, nuestro análisis comenzará con esta Las aportaciones chomskianas fueron de las primeras en ser aplicadas computacionalmente y, sobre todo, supusieron el punto de

[5] Doctor en Lengua española por la Universitat de València (UV). Ha trabajado en la Universidad Católica de Valencia (UCV) y en distintos departamentos de la UV: en el departamento de Filología española y, actualmente, en el de Didáctica de la lengua y la literatura.
Sus líneas de investigación abarcan diferentes campos de la Ciencia Cognitiva y la Lingüística Aplicada. Ha publicado diversos trabajos relacionados con la Lingüística computacional (LC), especialmente con las gramáticas formales utilizadas en LC, entre ellos, algunos libros como Introducción a la Gramática Léxico-Funcional. Teoría y aplicaciones (2008), Lingüística computacional. Tratamientos del habla (2011), Lingüística computacional. Análisis, generación y traducción automática (2011) o El abecé de la Lingüística computacional (2012) en la editorial Arco/Libros. En los últimos años su interés se ha dirigido hacia el campo de la Lingüística clínica, así como también a la sintaxis, la lingüística teórica y la enseñanza del español como lengua extranjera.

arranque sobre las que se construyeron el resto de las gramáticas que, en su formulación, no pudieron obviar nunca el punto de vista chomskiano. En este sentido, abordaremos las denominadas Gramáticas de Unificación, como son la Gramática Sintagmática Generalizada (*Generalized Phrase Structure Grammar*, GPSG) y su continuación con la Gramática Sintagmática orientada al núcleo (*Head-driven Phrase Structure Grammar*, HPG), o la Gramática Léxico-Funcional (*Lexical-Functional Grammar*, LFG). Serán contrastadas estas gramáticas con otras también utilizadas en la LC, pero cuyo peso de la tradición chomskiana es menor, como ocurre con la Gramática Funcional de Dik, o la Gramática de Papel y Referencia de Foley, Van Valin o LaPolla, entre otras.

Entendemos que este repaso histórico nos permitirá entender mejor cómo algunos problemas que se nos presentan en el análisis no son tanto problemas del objeto de estudio analizado, sino que son problemas creados desde la propia teoría. De esta manera, oraciones como *Cogimos el coche* son aceptadas como oraciones gramaticales desde las distintas teorías gramaticales, pero cuesta más explicar la agramaticalidad de oraciones como **Cogimos que vendrá mañana* para determinadas teorías. Entendemos que este análisis nos permite identificar qué teorías pueden ser más eficientes, al proponer menos mecanismos sintácticos creados *ad hoc* para explicar la mayor cantidad de fenómenos posibles.

Sesgo en modelos de Word Embeddings

Workshop

Felipe Bravo[6]
[fbravo@dcc.uchile.cl]
Universidad de Chile, Chile

Los *word embeddings* son vectores de palabras entrenados sobre grandes colecciones de documentos que hoy en día son ampliamente usados para el procesamiento de lenguaje natural. Diversos estudios han mostrado que los modelos de *word embeddings* exhiben sesgos estereotípicos de género, raza y religión, entre otros criterios. Varias métricas de equidad se han propuesto para cuantificar automáticamente estos sesgos. Aunque todas las métricas tienen un objetivo similar, la relación entre estas no es clara. Dos problemas impiden una comparación entre sus resultados: la primera es que operan con parámetros de entrada distintos, y la segunda es que sus salidas son incompatibles entre sí. Esto implica que un modelo de *word embedding* que muestra buenos resultados con respecto

[6] Felipe Bravo Márquez es profesor asistente en el Departamento de Ciencias de la Computación de la Universidad de Chile e Investigador Joven del Instituto Milenio Fundamento de los Datos. Realizó su doctorado en el grupo Machine Learning de la Universidad de Waikato, Nueva Zelanda, donde también trabajó como Research Fellow durante dos años. Actualmente mantiene un puesto de Investigador Asociado Honorífico en este grupo. Anteriormente, recibió dos títulos profesionales en ingeniería en computación e ingeniería industrial, y un magíster en ciencias de la computación en la Universidad de Chile. Trabajó durante tres años como ingeniero de investigación en Yahoo! Labs Latin America. Sus intereses de investigación y experiencia se centran en la adquisición de conocimientos e información a partir del lenguaje natural, abarcando las áreas del procesamiento del lenguaje natural (NLP), el aprendizaje automático (ML), la inteligencia artificial (AI) y la recuperación de información (IR). En su investigación, ha desarrollado varios métodos de NLP y ML para el análisis de opiniones y emociones en medios de comunicación social, que han sido publicados en conferencias y revistas de prestigio como por ejemplo, IJCAI, ECAI, JMLR y Knowledge-based Systems. Ha formado parte del comité de programa en conferencias importantes en procesamiento de lenguaje natural e inteligencia artificial, tales como ACL, EMNLP, NAACL, IJCAI y ECAI.

a una métrica de equidad, no necesariamente mostrará los mismos resultados al usar una métrica diferente. En esta charla presentaremos a WEFE, *the word embeddings fairness evaluation framework*, un marco teórico para encapsular, evaluar y comparar diversas métricas de equidad. Nuestro marco toma como entrada una lista de modelos de *word embeddings* pre-entrenados y un conjunto de pruebas de sesgo agrupadas en distintos criterios de equidad (género, raza, religión, etc.). Luego ranquea los modelos según estos criterios de sesgo y comprueba sus correlaciones entre los *rankings*.

Junto al desarrollo del marco, efectuamos un estudio de caso que mostró que rankings producidos por las métricas de equidad existentes tienden a correlacionarse cuando se mide el sesgo de género. Sin embargo, esta correlación es considerablemente menor para otros criterios como la raza o la religión. También comparamos los rankings de equidad generados por nuestro estudio de caso con rankings de evaluación de desempeño de los modelos de *word embeddings*. Los resultados mostraron que no hay una correlación clara entre la equidad y el desempeño de los modelos. Finalmente presentamos la implementación de nuestro marco teórico como librería de Python, la cual fue publicada como software de código abierto.

The Power of Weak Signal in NLP

Conference

Malvina Nissim[7]
[m.nissim@rug.nl]
University of Groningen, Netherlands

For language processing, we love supervised models: solid and accurate. What we love less is the effort to acquire labels, and the fact that porting these models to new domains, let alone new languages, is tricky (unless new labels are acquired, but we said we don't love it much). The balance between the advantages of fully blown signal and the disadvantages of limited portability is as delicate as it is fundamental.

In this talk, I will address these issues discussing three case studies in three different classification tasks in the field of language processing. In two of them, I will show how we can make up for missing signal, but we can successfully exploit what we get only if we use it in some indirect

[7] Malvina Nissim is Professor at the University of Groningen, The Netherlands, with a Chair in Computational Linguistics and Society. She has extensive experience in sentiment analysis and author identification and profiling, as well as in modelling the interplay of lexical semantics and pragmatics, especially regarding the computational treatment of figurative language and modality. She is the author of 100+ publications in international venues, is member of the main associations in the field, and annually reviews for the major conferences and journals. She has co-chaired the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), and was the general chair of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM 2018). She is also active in the field of resource and system evaluation, as both organiser and participant of shared tasks, and is interested in the philosophy behind them. She graduated in Linguistics from the University of Pisa, and obtained her PhD in Linguistics from the University of Pavia. Before joining the University of Groningen, she was a tenured researcher at the University of Bologna (2006-2014), and a post-doc at the Institute for Cognitive Science and Technology of the National Research Council in Rome (2006) and at the University of Edinburgh (2001-2005). In 2017, she was elected as the 2016 University of Groningen Lecturer of the Year.

way. In the other one, I will show how we can get rid of too much signal to enhance portability, while preserving modelling power.

Tracking discourses around the coronavirus pandemic

Conference

Tony Berber Sardinha[8]
[tonycorpuslg@gmail.com]
Catholic University of Sao Paulo, Brazil

The primary goal of this study is to detect the macro discourses around the coronavirus pandemic in news sources in English from the patterns of shared lexis across texts. Discourses are 'ways of looking at the world, of constructing objects and concepts in certain ways, of representing reality' (Baker & McEnery, 2015, p. 5). Corpus-based research on discourse has largely resorted to small corpora and techniques such as keyword analysis and concordancing. In contrast, in this study a very large corpus is analyzed, following a lexical multi-dimensional perspective (Berber Sardinha, in press; Kauffmann & Berber Sardinha, in press; Fitzsimmons-Doolan, 2014), which enables modeling the variation such that each text is shaped simultaneously by the different dimensions.

[8] Professor with the Graduate Program in Applied Linguistics and the Linguistics Department, the Catholic University of Sao Paulo, Brazil. His recent publications include Multidimensional Analysis: Research Methods and Current Issues (2019, Bloomsbury, co-edited with Marcia Veirano Pinto), Multidimensional Analysis: 25 Years on (2014, John Benjamins, co-edited with Marcia Veirano Pinto), Working with Portuguese Corpora (2015, Bloomsbury, co-edited with Telma Sao Bento Ferreira and Cristina Meyer), Metaphor in Specialist Discourse (2015, John Benjamins, co-edited with Berenike Herrmann).
His main research interests are multi-dimensional analysis, the use of corpora for historiography and historical discourse analysis, the development of corpus methods for metaphor retrieval, the application of corpus techniques in forensic linguistics, corpus-based analysis of translated texts, and the interface between corpus linguistics and language teaching, and corpus linguistics and Digital Humanities. He is on the board of several journals and book series such as the International Journal of Corpus Linguistics, Corpora, Applied Corpus Linguistics, Metaphor and the Social World, Studies in Corpus Linguistics, Register Studies, and Metaphor in Language, Cognition, and Communication.

The corpus for this study was curated from the Coronavirus corpus (english-corpora.org), which comprises texts from news sources from 21 different English-speaking countries published from January 1, 2020 through September 30, 2020, by 8,123 news providers. The corpus includes ca. 658,000,000 tokens and 839,689 texts. Lemmas for nouns, verbs, adjectives, and adverbs were extracted, and their text dispersion calculated. Keywords were extracted by comparing the count of texts in which the lemmas occurred in the coronavirus corpus with the counts of texts in which the same words occurred in the iWeb corpus. The iWeb corpus comprises more than 22 million texts collected in 2018, thereby providing a 'pre-covid' sample. A log-likelihood (LL) index was computed, and the lemmas whose LL value as significant at $p<.05$ were selected. From those, the 300 lemmas with the highest LL values were chosen to be entered in the factor analysis.

The method of analysis is based on the Multi-dimensional Analysis (MD) framework (Biber, 1988; Berber Sardinha & Veirano Pinto, 2019), whose general goal is to identify the major parameters underlying text variation in a language or domain. Two major strands of MD analysis exist: in a grammatical MD analysis, the goal is to uncover the functional parameters of variation, whereas in a lexical MD analysis, as employed here, the goal is typically to detect such constructs as topics, themes or discourses, encoded through the lexical choices.

A factor analysis was conducted in SAS University Edition, yielding four factors upon rotation. The factors were interpreted by considering the major discourses signaled by the loading variables, which was aided by reading the texts having outstanding scores, and by concordancing the texts for the loading variables. The overall themes and macro discourses reflected in the factors were considered in the interpretation of the dimensions.

The five provisional dimensions are the following: Dim. 1: Economic vs Human concerns; Dim. 2: Monitoring the spread of the disease; Dim. 3: Symptoms; Dim. 4: New social rules; Dim. 5: Hospitalization. Variation across the texts was measured with respect to the countries in which the texts were published, the time of their publication, and the news source that published the texts. The ANOVAS showed significant yet small effects for country and time period, but moderate effects for source of publication. A Discriminant Function Analysis (DFA) using the keywords as variables showed it was possible to predict the country of publication with variable success (from 53% for texts published in Jamaica to 13% for texts published in Singapore). A sample of the texts was taken to represent liberal and conservative news sources (3100 texts each). The DFA showed it was possible to predict the political leaning of the texts successfully (75% for liberal, 55% conservative).

At the moment, the findings suggest five major large-scale discourses shaping the way the world communicates news and opinions about the coronavirus pandemic in English. Five considerations can be drawn from the findings. First, the most salient dimension, dimension # 1, exposes the difficulty in bridging the gap between those in favor of opening up the economy and those in favor of protecting people from exposure to the virus: the two discourses never meet in the texts; rather these two positions are argued in separate texts, as if it were not possible to protect the economy while at the same time protecting human lives. Secondly, the dimensions reveal a global discourse on the pandemic, which crosses national borders. At least for English, the discourses around the pandemic circulate freely around the world, and their origins are difficult to pin down. Thirdly, the time of publication is generally a poor predictor of the discourse. This might be due to the short time span covered by the sample. So far we do not see these discourses being tied to any particular period of time. Fourthly, the news source of the story is a moderate predictor for some

of the discourses (eg. dim. 1 and dim. 5): particular news sources show a preference for particular discourses. And finally, it is possible to predict all of the countries and even the political leanings of some sources using the individual keywords rather than the dimensions. The individual words are better predictors than the broad discourses captured by the dimensions.

References

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2019). *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic.

Berber Sardinha, T. (in press). Discourse of Academia from a Multi-dimensional Perspective. In E. Friginal & J. Hardy (Eds.), *The Routledge Handbook of Corpus Approaches to Discourse Analysis*. London: Routledge.

Kauffmann, C., & Berber Sardinha, T. (in press). Brazilian Portuguese literary style. In E. Friginal & J. Hardy (Eds.), *The Routledge Handbook of Corpus Approaches to Discourse Analysis*. London: Routledge.

Author 2019.

Author in print-a.

Author in print-b.

Baker, P. & Egbert, J. (2016) (Eds.). *Triangulating Methodological Approaches in Corpus Linguistic Research*. London: Routledge.

Baker, P., & McEnery, T. (2015). Introduction. In P. Baker & T. McEnery (Eds.), *Corpora and discourse studies: Integrating discourse and corpora* (pp. 1-20). Basingstoke: Palgrave Macmillan.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social order analysis. *Behavior Research Methods* 49(3), pp. 803-821.

Fitzsimmons-Doolan, S. (2014). Using lexical variables to identify language ideologies in a policy corpus. *Corpora*, 9(1), 57-82.

Riedl, M., & Biemann, C. (2012). Text Segmentation with Topic Models. *Journal for Language Technology and Computational Linguistics*, 27(47-69), 13-24.

Lingüística Computacional y Procesamiento del Lenguaje Natural

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

A deep-learning model for discursive segmentation shows high accuracies for sentences classification in biomedical scientific papers

Juan Pavez, Sebastián Rodríguez & Eduardo N. Fuentes[9]
[jp@writewise.cl | sebastian.rodriguez.p@pucv.cl | ef@writewise.io]
Artificial Intelligence for Scientific Writing Group, Santiago, Chile

One of the main challenges for researchers when writing scientific papers is to coherently structure and organize the content, specifically at rhetorical-discursive level. Modeling these types of text is difficult and new computation approaches are necessary. Currently, language model pre-training that learned word representations from a large amount of unannotated text has been shown to be effective for improving many natural language processing (NLP) tasks. Recent models have focused on learning context dependent word representations, such as: 1) **E**mbeddings from **L**anguage **Mo**dels (ELMo) (Peters *et al.*, 2018); 2) **G**enerative **P**retrained **T**ransformer (GPT) (Radford *et al.*, 2018); 3) **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) (Devlin *et al.*, 2019). Specifically, BERT which consists of a transformer architecture (Vaswani *et al.*, 2017) that produces contextualized word representations has shown state-of-the-art performance on several NLP benchmarks. Despite these advantages, BERT has been trained and tested mainly on datasets containing general domain texts (e.g. Wikipedia). Therefore, its performance in other genre types of text, such as biomedical scientific papers, is not optimal. Recently, BioBERT- the first domain-specific BERT based model pretrained on biomedical corpora (PubMed) – has shown to outperform previous models on biomedical NLP tasks (Lee *et al.*, 2019). However, little research has been performed at rhetorical-discursive level using these state-of-the-art language models and applied them to the challenging task of identification of rhetorical-discursive steps (i.e. functional linguistic unit that fulfills a communicative purpose in a sentence). Therefore, the aim of this study was to test the accuracy of BioBERT on rhetorical- discursive steps classification in biomedical scientific papers.

Methods

1) Rhetorical-discursive model: The rhetorical-discursive model of Swales (1990) was used and adapted it to biomedical scientific papers. In this model the linguistic units is the following: macromoves (MM) contains moves (M) which contains steps (S). The present study focused on the analysis of rhetorical-discursive steps in each macromoves (MM). **2) Corpus and annotation system for discursive segmentation:** We used a corpora of 65 biomedical scientific papers from which 17,000 sentences were tagged in total for all MM by experts in scientific writing using prodigy (https://prodi.gy/). **3) Models:** The main model used in this work was BioBERT (Lee *et al.*, 2020).

[9] Our group is focus on research in deep learning applied to natural language processing for scientific papers. We applied this research to develop a unique software that guides scientists on how to write scientific articles. Specifically, we are training deep neural networks that "are learning" the structure and content of well-written papers. This thanks to a unique state-of-the-art models including different type of Transformers.

BioBERT is a BERT-type of model pre-trained on a large-scale of biomedical corpora collected from PubMed abstracts and PMC full-text articles. Tagged sentences of the training corpora were separated into their MM classification. Then, a separate BioBERT model was trained with few iterations (4-8) in the sentences of each MM to classify each sentence into one of the S of an specific MM. We also tested other models including some baselines such as Support Vector Machine and Random Forests and more complex deep-learning models such as ELMo, BERT, SciBERT. **4) Model testing and tool development:** A tool was developed and integrated in a modern web application in proven technologies: Vue.js and Django for frontend and backend development, respectively. Therefore, the users were able to interact with a graphic interface developed for writing and reviewing the structure of scientific papers. The usefulness of the tool was tested with over 100 researchers deploying the module in a productive operational environment (Technology Readiness Level 9). Opinions from the users were collected after use of the tool.

Results

1) Rhetorical-discursive guide characterization in biomedical scientific papers: To characterize biomedical scientific papers corpus, we used a rhetorical-discursive guide adapted from Swales (1990). We found five prototypical MM, specifically: Abstract (MM0), Introduction (MM1), Method (MM2), Results (MM3), Discussion (MM4). In detail we found: 1) Four S in MM0: Introduction (S1), Methods (S2), Results (S3), Conclusion (S4); 2) Five S in MM1: Background (S1), State of the Art (S2), Problem (S3), Aim/Methods (S4), Main Results (S5); 3) Three S in MM2: Methodological Framework (S1), Method specifics (S2), Data processing/Statistics (S3); 4) Three S in MM3: Result Context (S1), Results Specifics (S2), Specific Conclusions (S3); 5) Four S in MM4: Discussion Background (S1), Discussion Specifics (S2), Take home message (S3), Significance/Perspectives (S4). **2) Prodigy allows quick model training with few examples:** To tag rhetorical discursive steps in biomedical papers we used Prodigy, an annotation tool powered by active learning that allows fast labelling of text corpus. With this tool we were able to tag thousands of sentences in a few weeks. The data labelled with Prodigy was used to fine-tune huge models such as BioBERT obtaining very good results. **3) BioBERT predicts with high accuracy rhetorical-discursive steps:** To classify rhetorical- discursive steps in biomedical papers in each MM, we used BioBERT obtaining very good results in terms of accuracy, recall and F1. Specifically, the average accuracy was: MM0 = 0.84 (four classes); MM1 = 0.85 (five classes); MM2 = 0.80 (three classes); MM3 = 0.89 (three classes); MM4 = 0.81 (four classes). We also compared BioBERT with other models and we consistently found that modern pre-trained deep learning models outperforms shallow models. Shallow models showed in average less than 0.55 accuracy for MM in general and deep learning models showed in average 0.7 accuracy for MM in general. Altogether, these results demonstrate that BioBERT predicts with high accuracy rhetorical-discursive steps using just few thousand tagged sentence in different MM. **4) BioBERT could help structuring scientific papers:** To test whether or not BioBERT helps better structuring scientific papers, we developed a specific tool with a graphic interface that users can interact with. This tool is part of a module of the scientific writing software WriteWise (https://web.writewise.io/). This module helped users to visualize the structure of a paper. Specifically, users were able to: (1) visualize where they position each sentence in a manuscript [e.g. the 3 first sentences of the first paragraph are communicating Background]; (2) identify the number of sentences communicating the same

underlying information [e.g. 8 sentences in total are communicating Background in the Introduction]. The users were also able to re- structure their papers considering their own writing style and communicational purpose of each sentence and also journals´ writing style. According to the opinion of 100 researchers 74% considered the tool extremely useful and highly novel.

Conclusions

In summary, we found that modern pretrained deep neural network architectures - specifically fine-tuned BioBERT- show optimal results for classification of rhetorical-discursive steps in biomedical scientific papers. We showed that by labelling a relatively small amount of sentences (in the order of thousands) we can obtain very good results using the previously mentioned method. This is thanks that these models have been pretrained in huge amounts of text data, this allows them to learn complex words and sentences representations that then can be used to quickly learn new tasks. Finally, we consider that this work shows the importance of considering the particularities of genre, and that domain-specific models are fundamental to obtain good results. Further research regarding the application of BioBERT or other recent state-of-the-art transformer-based models (e.g. GPT-3) to improve scientific papers writing structure for biomedical text remained to be investigated.

Keywords: rhetorical-discursive steps, biomedical papers, deep-learning models, BioBERT

References

Beltagy *et al.* 2019. *SciBERT: A Pretrained Language Model for Scientific Text* (https://arxiv.org/abs/1903.10676).

Devlin *et al.* 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (https://arxiv.org/pdf/1810.04805).

Lee *et al.* 2019. *BioBERT: a pre-trained biomedical language representation model for biomedical text mining* (https://arxiv.org/abs/1901.08746).

Neumann *et al.* 2019. SciSpacy. (https://arxiv.org/pdf/1902.07669).

Peters *et al.* 2018. *Pretrained deep contextualized word embedding* (*ELMo*) (https://arxiv.org/pdf/1802.05365).

Radford *et al.*, 2018. *Language Models are Unsupervised Multitask Learners* (https://arxiv.org/pdf/1904.02679.pdf).

Swales, 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

Vaswani *et al.*, 2017. *Attention Is All You Need*. (https://arxiv.org/abs/1706.03762)

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Aplicaciones para el tratamiento informático de un corpus bilingüe de fraseología

Joseph García Rodríguez[10]
[Joseph.Garcia@uab.cat]
Universidad Autónoma de Barcelona, España

El desarrollo de la lingüística computacional ha permitido crear herramientas sumamente útiles para el tratamiento informático de una cantidad ingente de datos (Pérez Hernández y Moreno Ortiz, 2010). Las aplicaciones que se encuentran a nuestra disposición son capaces de crear modelos relacionales que dan respuesta a las diferentes necesidades de los investigadores. En el caso de la lingüística, las bases de datos (BD) contribuyen al análisis y la comparación de distintos fenómenos que solo pueden observarse a través del manejo de dichos programas.

En el contexto de la fraseología bilingüe, la función que desempeñan estas plataformas puede llegar a ser de especial relevancia si se persigue la finalidad de conocer en profundidad las semejanzas y divergencias que presentan las unidades de dos sistemas lingüísticos (García Rodríguez, 2019). Los estudios contrastivos se basan en la recopilación de expresiones y su volcado en BD con el fin de observar patrones comunes y disimilitudes entre dos o más idiomas (García Rodríguez, 2020). Aun así, algunas de las aplicaciones más utilizadas para la clasificación, examen y visualización de los datos, como Access (e incluso Excel), contienen ciertas limitaciones cuando se trabaja con varios parámetros de comparación (Muñoz Álvarez, 2016).

Precisamente, en las últimas décadas se han creado programas de diversa índole que facilitan un análisis relacional profundo de las unidades lingüísticas objeto de estudio. Este es el caso de Qlik, una plataforma que permite gestionar múltiples datos y a la vez diseñar aplicaciones personalizadas para trabajar con ellos. De hecho, algunos ya la consideran como un nuevo modelo de generación de análisis y visualización de los datos (Femenía Millet, Pérez Díez y Olmos Vila, 2019), puesto que cuenta, por un lado, con un motor asociativo único y, por otro, con inteligencia artificial sofisticada, cuya finalidad es que el sistema pueda prever acciones y automatizar procesos debido a la interacción que se genera entre la aplicación y el especialista.

Teniendo en consideración todo lo anterior, en el estudio que se pretende desarrollar se utilizará dicha plataforma para analizar un corpus de unidades fraseológicas (UFS) del español y el catalán compilado, inicialmente, con Access y, posteriormente, volcado en el gestor de datos que utiliza Qlik. Las expresiones que componen el corpus contienen información diversa, a saber, fuente de extracción, lengua de la UF, concepto que subyace en el fraseologismo, categorización cognitiva, marca gramatical y registro de uso, entre otras. La plataforma Qlik permite crear diferentes muestras de resultados a partir de los datos anteriores. De este modo, es posible conocer automáticamente y mediante diferentes tipos de gráficos, la relevancia que contiene cada ítem en el corpus. Además, dicho programa permite crear aplicaciones, en función de las necesidades del usuario.

[10] Doctor en Filología Española por la UAB y profesor asociado en la misma Universidad. Especialista en fraseología, lexicología, lexicografía y español lengua extranjera.

En el contexto que aquí nos concierne, se ha elaborado una muestra de diccionario electrónico de fraseología bilingüe en español y catalán a través de las posibilidades que brinda la plataforma. En tal sentido, Qlik no solo ayuda a analizar, ordenar y visualizar datos, sino que también es capaz de crear modelos aplicativos que sirvan como base para la consulta de la información contenida en la BD (en este caso, un corpus bilingüe de UFS). Por todo ello, y según las características que debe contener una obra lexicográfica de este calibre (Leffa, 1993; Gouws, Schweickard y Wiegand, 2003; Müller-Spitzer, 2003), se presenta un prototipo inicial con la intención de que pueda contribuir al futuro desarrollo de una novedosa obra fraseográfica en formato electrónico para los idiomas mencionados.

En suma, la presente investigación tiene como objetivos principales los siguientes: (1) demostrar la utilidad de la plataforma Qlik para el análisis y la presentación de datos lingüísticos y (2) avanzar en el desarrollo de una aplicación interactiva y dinámica que simule un diccionario bilingüe de fraseología (español-catalán).

Palabras clave: aplicaciones informáticas, fraseología bilingüe, fraseografía, corpus

Bibliografía

Femenía Millet, Olga; Pérez Díez, Álvaro; Olmos Vila, Rafael (2019): "La mejora del proceso de enseñanza mediante la visualización de datos", *Revista Digital de Educación y Formación del Profesorado*, 16.

García Rodríguez, Joseph (2019): *Las unidades fraseológicas del español y el catalán con elementos de la naturaleza: estudio cognitivo-contrastivo y propuesta de un diccionario electrónico de fraseología bilingüe*. Bellaterra: Universidad Autónoma de Barcelona. Disponible en línea: https://www.educacion.gob.es/teseo/mostrarRef.do?ref=1259292#.

García Rodríguez, Joseph (2020): *La fraseología del español y el catalán: semántica cognitiva, simbolismo y contrastividad* (Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation. Heraugegeben von Gerd Wotjak). Frankfurt am Main: Peter Lang.

Gouws, Rufus H.; Schweickard, Wolfgang; Wiegand, Herbert E. (2003): "Lexicography through the ages: From the early beginnings to the electronic age". En Gouws, Rufus *et al.* (eds.), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin: De Gruyter, pp. 1-24.

Leffa, Vilson J. (1993): "Using an Electronic Dictionary to Understand Foreign Language Texts", *Trabalhos Em Linguistica Aplicada*, 21, pp. 19-29.

Müller-Spitzer, Carolin (2003): "Textual structures in electronic dictionaries compared with printed dictionaries: A short general survey". En Gouws, R./U. Heid, W. Schweickard/Wiegand, H. E. (eds.), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlín: De Gruyter, pp. 367-381.

Muñoz Álvarez, Óscar (2016): *Implantación de Qlik Sense® en una e-commerce*. Barcelona: Servicio de publicaciones UOC.

Pérez Hernández, Chantal y Moreno Ortiz, Antonio (2009): "Lingüística computacional y lingüística de corpus. Potencialidades para la investigación textual". En *Teoría y literatura artística en la sociedad digital*. Nuria Rodríguez Ortega (dir.). TREA: Gijón, pp. 67-96.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Applying the Linguistic Economy Principle to Programming Languages

Michael Dorin[11]
[mike.dorin@stthomas.edu]
University of St. Thomas, USA
Sergio Montenegro[12]
[sergio.montenegro@uni-wuerzburg.de]
Universität Würzburg, Germany

Although the importance of readability is routinely presented in introductory programming language classes, there is still doubt among software engineers about what makes source code too complicated for review. The sheer quantity of software metrics makes it evident that there is no current consensus of what makes source code complicated. This is a problem because more time and effort are spent on the writing and maintaining of source files that are difficult to read. Using concepts employed by natural language linguists, we analyze the syntax and semantics of source code and define a new approach for measuring code complicatedness. The principle of linguistic economy describes humans' desire for spoken communication to use the least amount of effort. The natural instinct of a programmers is to create easier code when possible, so we use this principle to measure complicatedness of program source code. By identifying the most common software phrases and sentences, we show how the linguistic economy principle applies to source code. Our work evaluated more than 100 million lines of code, and the results show that this methodology provides a measurement of code readability and can identify software modules laden with hard to read code.

Introduction

Measuring software complicacy has been a challenging topic for decades, and this is made evident by the many different software measurement metrics that have been created [1]. One can find dozens of supported metrics included in the Understand tool made by Scientific Toolworks [2]. There are many aspects of source code that impact quality, such as architecture, style, logic, semantics, and syntax. For this project, we focus on the complicacy of syntax, specifically by examining statements and expressions. Statements and expressions are fundamental for building software as they control data assignments and the logical decisions software makes during execution.

[11] Michael Dorin has nearly 30 years of software development experience and has worked in many engineering environments. His background includes engineering work in public safety communications, medical devices, telephony, and aircraft navigation. Michael teaches for the University of St. Thomas and is currently working towards a PhD through the University of Würzburg.

[12] Sergio Montenegro received a master's degree in computer science from the Technical University of Berlin, followed by a PhD from the same university. He has been programming satellites for the last 20 years. Currently he is a professor for aerospace information technology at the University of Würzburg (Germany). His research interests include dependability and simplification for control software.

The approach used in this paper is inspired by the Principle of Least Effort developed by the linguist George Zipf. For the Principle of Least Effort, Zipf hypothesized that humans, in language and other endeavors, will select actions to perform based on the most efficient way to complete a task [3]. Zipf illustrated this principle by describing artisans who must survive by performing work with a set of tools in the most economical manner possible. For the artisans to use their tools with a maximum economy, they must arrange them such the sum of their usage efforts will be minimum. For software artisans, statements and expressions will be tools. The artist will endeavor to use the most straightforward tools most often and organize them, so they are nearby. As a result, we can expect to find the artisans' tools arranged such that the tool furthest from the artisan would be the tool requiring the most effort to use. As Zipf believed the Principle of Least Effort is fundamental to human behavior we should be able to find it in all our daily actions. One way to demonstrate this is by analyzing speech and the concept of Linguistic Economy. From the perspective of this paper, Linguistic Economy can be explained as speech being performed using a set of tools, analogous to the artisan, and those tools of speech are the vocabulary [3]. Thus, when humans speak, the most commonly used words are the most comfortable and efficient words for communication and understanding. In this paper, we present that for a programmer, the most commonly used statements and expressions (tools) are the most efficient for making understandable source code.

We believe a limitation of traditional metrics is that their creation is based on software development treated like a mathematical practice or an engineering discipline. Treating software development as mathematical cannot completely account for human behavior. Applying the work of George Zipf can overcome this limitation and bridge the gap between pure logical software analysis and human behavior. By identifying the most common and easy to understand software statements and expressions, we show how the Principle of Linguistic Economy applies to programming resulting in an accurate measure of understanding and readability. We show that complicated code is undesirable for review and ultimately causes developers more problems.

Conclusion

The results support the measurement of statement and expression complicacy using concepts from the Principle of Least Effort and Linguistic Economy. Unpleasant to review files were shown to have a greater number of low-ranking statements, indicating that Zipf's methods can apply to code complicacy. The results also reflect that projects with statements and expressions that we identify as less understandable are less desirable to review and likely contain more faults. Since statements and expressions are a core part of a program source, it is crucial to the quality of a project that they are readable and understandable. Our study demonstrated that these linguistics concepts could be applied to program quality. This knowledge can be employed to make more readable, higher-quality programs.

Keywords: Linguistic Economy, Corpus, Programming Languages, Comprehension

Bibliography

[1] Fenton, Norman E., and Martin Neil. "Software metrics: successes, failures and new directions." *Journal of Systems and Software* 47.2-3 (1999): 149-157.

[2] Scientific Toolworks, Understand.https://scitools.com/feature/metrics/. Accessed: 2020-7-31

[3] George, K. "Zipf. 1949. Human Behavior and the Principle of Least Effort." (1949,1974).

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

CRF aplicado al POS Tagging para el español

Álvaro Ricardo Cujar Rosero[13], Raúl Ernesto Gutiérrez de Piñerez Reyes[14], Robinson Andrés Jiménez Toledo[15] & Franklin Eduardo Jiménez Giraldo[16]

[rcujar@umariana.edu.co | raul.gutierrez@correounivalle.edu.co | rjimenez@umariana.edu.co | fjimenez@umariana.edu.co]

Universidad Mariana, Universidad del Valle, Colombia

El Procesamiento de Lenguaje Natural comprende todas aquellas tareas que permiten el procesamiento automático de texto, entre las que se incluyen análisis morfológico, sintáctico, semántico, entre otras [1]. Dentro de ellas se encuentra el etiquetado de categorías (POS Tagging).

En el POS Tagging se realiza la asignación de la categoría lingüística a cada palabra que contiene un texto. Así, en la frase 'El escritorio es azul', donde la palabra 'el' corresponde con un determinante; 'escritorio' es un sustantivo; 'es' corresponde con un verbo y 'azul' es un sustantivo. El problema que afronta esta tarea es la ambigüedad de las palabras, cuando existen dos o más palabras que tienen igual escritura, pero diferente significado.

Numerosos modelos se han utilizado para la realización de esta tarea [2], [3], [4], [5]. Se destacan los modelos de Cadenas Ocultas de Markov (HMM), Modelo de Markov de Máxima Entropía (MEMM), entre otras.

Los campos aleatorios condicionales (CRF) corresponden con un modelo gráfico no dirigido, que permite el etiquetamiento de secuencias [6]. Este modelo, a diferencia del modelo dirigido HMM, permite la inclusión de características en los datos de entrada, además de tener en cuenta, para la clasificación de un estado, a todos los estados.

CRF ha sido utilizado en algunos estudios, a saber, en [7] se hace uso del tageador CRF++ Tagger aplicado al Assamese, el cual es un lenguaje de la zona nororiental de la India, alcanzando porcentajes no tan altos en la tarea del POS Tagging, debido a que no existe un corpus ampliamente anotado. Igualmente, en [8] se hace realiza una investigación utilizando algunos lenguajes de la India, en la que se obtienen resultados de exactitud para el Telegú, Hindi y Bengalí, de 77.37%, 78.66% y 76.08%, respectivamente, en la tarea del POS Tagging utilizando CRF y Aprendizaje Basado en Transformaciones. La tarea del POS Tagging ha sido objeto de estudio en [9] en donde se aplica CRF obteniendo un porcentaje de exactitud de 69,40 para el Hindi y en [10], utilizando el corpus Kannada, se alcanza una exactitud de 96.86%.

AnCora es un corpus con diferentes niveles de anotación, desarrollado en el CLiC-UB (Centre de Llenguatge i Computaci'o, Universitat de Barcelona) y en el TALP, Software Department, Universitat Politécnica de Catalunya [11]. El corpus AnCora contiene 500000 palabras anotadas multinivel, con información para las tareas de análisis morfológico, análisis sintáctico y semántico, entre otras.

[13] Magister en Ingeniería de Sistemas y Computación.

[14] Doctor en Ingeniería.

[15] Magister en Docencia Universitaria.

[16] Magister en Software Libre.

En la presente investigación, de tipo cuantitativa y descriptiva, se utilizó el corpus AnCora, tanto para el entrenamiento como para el testeo del modelo CRF, para el español; se utilizaron características como, el tiempo, el género, el número y la persona. Se tiene en cuenta características como, si la palabra comienza con mayúscula, si la palabra contiene algún símbolo, si la palabra contiene algún dígito, el tamaño de la palabra, obteniendo los siguientes resultados: exactitud 96.39%, precisión 95.73%, recuperación 95.89%, score FB1 95,81%.

Bibliografía

[1] Cortez, A., Vega, H., Pariona, J. (2009). Procesamiento de lenguaje natural. Revista de Ingeniería de Sistemas e Informática vol. 6, N.º 2, Julio - Diciembre 2009.

[2] Kanakaraddi, S.G., Nandyal, S.S. (2018). Survey on Parts of Speech Tagger Techniques, 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, 2018, pp. 1-6.

[3] Purkayastha, B., S., Roy, B., C. (2016). A Study on Different Part of Speech (POS) Tagging Approaches in Assamese Language. International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.

[4] Francis, M. (2015). A comprehensive survey on parts of speech tagging approaches in dravidian languages. Proceedings of 30th The IIER International Conference, Beijing, China, 26th July 2015, ISBN: 978-93-85465-57-4.

[5] Fernando, S., Ranathunga, S. (2018). Evaluation of Different Classifiers for Sinhala POS Tagging, Moratuwa Engineering Research Conference (MERCon), Moratuwa, 2018, pp. 96-101.

[6] Lafferty, J., McCallum, A., Pereira, F. (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 18th International Conference on Machine Learning 2001 (ICML 2001), pages 282–289, 2001.

[7] Barman, A., K., Sarmah, J. Sarma, S. (2013). POS Tagging of Assamese Language and Performance Analysis of CRF plus plus and fnTBL Approaches. Proceedings of the 2013 UKSim 15th International Conference on Computer Modelling and Simulation, 2013.

[8] Avinesh, P., V., S., Karthik, G. (2007). Part of speech tagging and chunking using conditional random fields and transformation based learning. In Proceedings Shallow Parsing for South Asian Languages (SPSAL) workshop at IJCAI At Hyderabad India, 2007.

[9] Awasthi, P., Rao, D., Ravindran, B. (2006). Part Of Speech Tagging and Chunking with HMM and CRF. NLP Association of India (NLPAI) Machine Learning Contest, 2006.

[10] Suraksha, N., M., Reshma, K., Kumar, K., M., S. (2017). Part-of-speech tagging and parsing of Kannada text using Conditional Random Fields (CRFs). International Conference on Intelligent Computing and Control (I2C2), Coimbatore, 2017, pp. 1-5.

[11] Taulé, M., Martí M.A. (2009). Recasens M. Ancora: Multilevel annotated corpora for catalan and spanish. Proceedings of 6th International Conference on language Resources and Evaluation, 2009.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Detección y corrección automática en textos especializados: análisis de patrones de errores en un corpus del dominio medico

Jésica López-Hernández[17] & Ángela Almela[18]
[jesica.lopez@um.es | angelalm@um.es]
Universidad de Murcia, España

Las tareas de detección y corrección automática de errores constituyen un elemento esencial en las tecnologías del procesamiento del lenguaje natural (Martin y Jurafsky, 2009). No obstante, la efectividad de los correctores ortográficos, como la mayoría de aplicaciones que se construyen para procesamiento de textos, se ve condicionada en gran medida por las características del dominio donde se van a aplicar. Los lenguajes de especialidad presentan particularidades que dificultan en gran medida las tareas de detección y corrección automática, como el uso de terminología especializada o la creación de neologismos. En el caso del lenguaje médico, y más concretamente de los informes clínicos, estas particularidades resultan cruciales, debido a que los textos en muchas ocasiones contienen abreviaturas, ambigüedades y errores ortográficos (Ruch, Baud y Geissbühler, 2003). Son diversos los estudios que señalan la presencia de errores lingüísticos en informes clínicos y presentan propuestas de corrección ortográfica automática para este dominio. La mayor parte se han centrado en el inglés (Lai, Topaz, Goss y Zhou, 2015; Fivez, Suster y Daelemans, 2016; o Wong y Glance), aunque también existen investigaciones para el sueco (Dziadek, Henriksson y Duneld, 2017) o el húngaro (Siklósi, Novák y Prószéky, 2016).

En el desarrollo de sistemas automáticos de detección y corrección resulta de gran interés profundizar en la naturaleza de los errores lingüísticos que ocurren en los informes clínicos, para detectarlos y tratarlos adecuadamente. Sin embargo, actualmente no existen datos cuantitativos sobre patrones de error en textos en español que procedan del ámbito biosanitario, y tampoco hay una revisión sistemática sobre la naturaleza de los mismos. Por lo tanto, es necesario llevar a cabo un estudio de errores lingüísticos que nos permita saber qué tipos de errores tienden a ocurrir en este dominio, cuáles son sus propiedades y cómo podemos proporcionar una base de conocimiento lingüístico a los métodos de detección y corrección existentes para este propósito. Además, no es posible elaborar una tipología universal de errores, sino que los tipos de errores identificados pueden variar según el alcance y el contexto de uso (Díaz, 2005), de ahí la necesidad de llevar a cabo una tipificación para el lenguaje médico.

---

[17] Jésica López-Hernández es investigadora predoctoral financiada por el Ministerio de Educación de España a través del Programa Nacional de Formación del Profesorado Universitario (FPU). Ha cursado un máster sobre lingüística teórica y aplicada y ha trabajado como lingüista computacional. Actualmente está realizando una investigación sobre detección y corrección ortográfica automática en el dominio médico en el Departamento de Informática y Sistemas de la Universidad de Murcia (España).

[18] Ángela Almela es profesora de lingüística computacional y de corpus en la Universidad de Murcia (España). Su investigación y publicaciones se centran principalmente en lingüística aplicada y aplicaciones forenses de la lingüística computacional, incluyendo análisis contrastivo entre corpus en inglés y español. Entre sus publicaciones se incluyen artículos en revistas científicas como Ibérica, Vigo International Journal of Applied Linguistics y Journal of Information Science.

En consecuencia, el objetivo de este trabajo es el análisis de errores contenidos en un corpus de informes médicos en lenguaje natural, mediante técnicas de extracción de información y explotación estadística de corpus. Se han tenido en cuenta parámetros como frecuencia de aparición y tipo de error (omisión, sustitución, adición o trasposición), posición del error en la palabra, longitud de la palabra o contexto en el que se produce. Para ello, se toman como referencia algunas contribuciones sobre clasificación y análisis de errores realizadas para otros dominios e idiomas, como las de Yannakoudakis y Fawthrop (1983); Pollock y Zamora (1983); Mitton (1987); Kukich (1992); Ramírez y López (2006); Baba y Suzuki (2012); Gimenes, Roman y Carvalho (2015); o Siklósi, Novák y Prószéky (2016); entre otros.

El corpus objeto de análisis está compuesto por una colección de informes clínicos electrónicos en español, pertenecientes a la especialidad médica de Urgencias, y contiene 631576 palabras. La tipificación de los patrones de error proporcionados va a contribuir al desarrollo de un módulo basado en conocimiento lingüístico, actualmente en progreso, que podrá mejorar el rendimiento de los sistemas de detección y corrección de errores para el dominio biomédico.

Palabras clave: análisis de errores, detección automática de errores, corrección ortográfica automática, corpus biomédico, procesamiento del lenguaje natural.

Bibliografía

Baba, Y. y Suzuki, H. (2012). How are spelling errors generated and corrected? A study of corrected and uncorrected spelling errors using keystroke logs. En H. Li, C. Lin, M. Osborne, G. G. Lee y J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*: *Short Papers* (pp. 373–377). Jeju Island: Association for Computational Linguistics (ACL).

Díaz Villa, A. (2005). Tipología de errores gramaticales para un corrector automático. *Procesamiento del Lenguaje Natural,* 35, 409-416.

Dziadek, J., Henriksson, A. y Duneld, M. (2017). Improving terminology mapping in clinical text with context-sensitive spelling correction. *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, 235, 241–245.

Fivez, P., Suster, S. y Daelemans, W. (2016). Unsupervised context-sensitive spelling correction of clinical free-text with word and character N-Gram embeddings. En K. Bretonnel Cohen, D. Demner-Fushman, S. Ananiadou y J. Tsujii (Eds.), *Proceedings of the BioNLP 2017 Workshop* (pp. 143-148). Vancouver: Association for Computational Linguistics (ACL).

Gimenes, P., Roman, N. y Carvalho, A. (2015). Spelling error patterns in Brazilian Portuguese. *Computational Linguistics,* 41(1), 175-183.

Kukich, K. (1992). Technique for automatically correcting words in text. *ACM Computing Surveys*, 24(4), 377-439.

Lai, K. H., M. Topaz, F. R. Goss, y L. Zhou. (2015). Automated misspelling detection and correction in clinical free-text records. Journal of Biomedical Informatics, 55, 188–195.

Martin, J. H., y Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River: Pearson/Prentice Hall.

Mitton, R. (1987). Spelling checkers, spelling correctors, and the misspellings of poor spellers, *Information Processing & Management*, 23(5), 495-505.

Pollock, J. J. y Zamora, A. (1983). Collection and characterization of spelling errors in scientific and scholarly text*, Journal of American Society of Informatics and Science*, 34(1), 51–58.

Ramírez, F. y López, E. (2006). Spelling error patterns in Spanish for word processing applications. En *Proceedings of Fifth international conference on Language Resources and Evaluation LREC'06* (pp. 93-98). Genoa: European Language Resources Association.

Ruch, P., Baud, R. y Geissbühler, A. (2003). Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record, *Artificial Intelligence in Medicine*, 29(2), 169-184.

Siklósi, B., Novák, A. y Prószéky, G. (2016). Context-aware correction of spelling errors in Hungarian medical documents. *Computer Speech & Language,* 35, 219-233.

Wong, W. y Glance, D. (2011). Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes, *Artificial Intelligence in Medicine*, 53, 171-180.

Yannakoudakis, E. y Fawthrop, D. (1983). The rules of spelling errors. *Information processing and management*, 19(12), 101-108.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

El desafío de la lingüística computacional: las lenguas visogestuales

Eva Freijeiro & Inmaculada C. Báez[19]
[evafreijeiro@gmail.com | cbaez@uvigo.es]
Grupo de investigación GRILES, Universidad de Vigo, España

La lingüística computacional, hasta ahora, ha venido trabajando fundamentalmente con la generación del lenguaje natural proveniente de voz y de texto, pero en muy pocos casos, por no decir en ninguno, se ha planteado la generación de lenguaje natural a partir de los gestos de lenguas visogestuales, a partir de los signos de las lenguas naturales de las comunidades sordas.

Aunque no es muy fructífero, el impacto del aprendizaje automático (*machine learning*) también se puede detectar en la lingüística de las lenguas visogestuales. En este caso, el principal objetivo de la aplicación de la lingüística computacional se centra en favorecer el acceso a la información a las personas sordas con dificultades en lectoescritura de las lenguas orales. Para conseguirlo, se distinguen dos áreas principales:

a. Traducción

b. script (establecer el guion o las secuencias de comandos) que permitan el procesamiento de animaciones en avatares en 3D, a partir de datos de lexicón y expresiones faciales. Sin embargo, estos sistemas han sido construidos sin la aportación de los corpus anotados de LS que codifican detalladamente el movimiento humano (Lu y Huenerfauth, 2010 y Wagner, P., Cwiek, A., Samlowski, B. 2019). Sea cual sea el área de investigación, para poder avanzar en la lingüística computacional de las lenguas visogestuales son imprescindibles los corpus porque de ellos se van a extraer todos los datos que serán encriptados/ sistematizados/ etiquetados por el componente computacional.

Las cuestiones técnicas relacionadas con la naturaleza de los datos en lenguas visogetuales (calibración, protocolo de grabación, elementos del movimiento, dimensiones, etc.) avanzan rápidamente, pero se ven entorpecidas por la ausencia de estudios lingüísticos que acompañan al movimiento propio de estas lenguas. Es necesario desgranar las lenguas signadas, siguiendo o no modelos de análisis de lenguas orales, para poder atribuirles etiquetas y, por tanto, sistematización

[19] Eva Freijeiro e Inmaculada Báez somos dos miembros del grupo de investigación de lengua española y lenguas signadas (GRILES) http://www,GRILES.webs.uvigo.es. Estudiamos cuestiones relacionadas con la lingüística de las lenguas de signos, la enseñanza de la lengua oral a sordos y la lengua de signos a oyentes y estamos elaborando el corpus de lengua de signos española CORALSE. Pretendemos que CORALSE constituya una muestra representativa de la lengua de signos española y sea un conjunto accesible de datos lingüísticos reales almacenados con el fin de servir de base de estudio para investigadores de la lingüística general y de lingüística de las lenguas de signos. Esperamos que una vez disponibilizado en la web se nutra de la participación y colaboración de la comunidad lingüística de la LSE (sordos, CODA, lingüistas, intérpretes, e investigadores de la Lengua de signos española). Nuestros últimos trabajos han sido la dirección (Inmaculada Báez y Ana Mineiro) y defensa de la tesis doctoral de Eva Freijeiro Ocampo titulada "Lenguas de signos e implante coclear: armas comunicativas (de)construcción masiva", (6 de marzo de 2020) y la publicación del artículo (Báez, Bao y Veiga) "Differing Attitudes towards Spanish sign Languages in three Galician pre- and primary Schools" publicado en el volumen de Talia Bugel y Cecilia Montes-Alcalá titulado New Approaches to Languages Attitudes in the Hispanic and Lusophone World (John Benjamins , Public Company en 2020).

que ofrecerle a la máquina. De hecho y tal y como ya en 2008 señalaban Vogler y Goldstein nadie ha producido un sistema real para el análisis y el reconocimiento de alguna lengua signada porque no se ha conseguido abordar todas las posibles complicaciones que derivan de ellas, partiendo del punto de vista de la lingüística computacional.

Los objetivos de esta comunicación son tres: 1) promover la reflexión sobre los retos éticos y sociales a los que nos enfrentamos al incorporar la modalidad visogestual a la lingüística computacional y a las investigaciones sobre lenguaje natural, es decir, plantear los retos a los que nos enfrentamos al incorporar las lenguas de las comunidades sordas al mundo digital; 2) identificar buenas prácticas en la búsqueda de soluciones sostenibles para las lenguas signadas y 3) aprender a ver desde la óptica visogestual.

Para conseguirlo el primer paso consiste en a) revisar las características y ámbitos del lenguaje natural/ lingüística computacional y las herramientas tecnológicas aplicadas a las lenguas visogestuales, b) analizar el funcionamiento de las tecnologías digitales centradas en la comunicación visogestual.

Con respecto a este último punto, nuestra aportación consiste en la elaboración de un corpus de referencia de lengua de signos española (corpus CORALSE) que puede actuar como fuente de datos sobre la que aplicar modelos computacionales y de programación para favorecer los análisis de corpus multimodales, ya sean de lenguas visogestuales o de los componentes gestuales no lingüísticos.

Referencias

LU, Pengfei y HUENERFAUTH, Matt (2010). Collecting a Motion-Capture Corpus of American Sign Language for Data-Driven Generation Research. Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies, Los Angeles, California, junio 2010, pp. 89-97.

VOGLER, Christian y GOLDENSTEIN, Siome (2008). Toward computational understanding of sing language. Technology and Disability, 20, pp. 109-119.

WILBUR, Ronnie B. y MALAIA, Evguenia (2010). Contributions of sign language research to gesture understanding: what can multimodal computational systems learn from sign language research. International Journal of Semantic Computating, 2(1).

WAGNER, Petra., CWIEK, A., SAMLOWSKI, B. (2019). Exploiting the speech-gesture link to capture fine-grained prosodic prominence impressions and listening strategies. Journal of Phonetics, 76, UNSP 100911. doi: 10.1016 / j.wocn.2019.07.001

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Entrevista por medio de un chatbot con reconocimiento de voz enfocada a un modelo de negocio para la identificación de entidades

Laura Millán-Olivares[20], Moisés González-García[21] & Noé Alejandro Castro-Sánchez[22]

[ana.millan18ca@cenidet.edu.mx | moises.gg@cenidet.tecnm.mx | noe.cs@cenidet.tecnm.mx]

Centro Nacional de Investigación y Desarrollo Tecnológico, México

Resumen: El avance del Internet y de las nuevas Tecnologías de la Información y la Comunicación (TIC), ha sido el origen de importantes cambios que impactaron no sólo en la vida cotidiana de los individuos, sino también en organizaciones e instituciones sociales (Litwin, 2005). En consecuencia, estos cambios desembocaron en un paradigma que produjo modificaciones significativas en la forma de relacionarnos, de trabajar e incluso de enseñar y aprender.

El sector computacional ofrece sus servicios a una gran cantidad de clientes diariamente y a medida que un servicio es requerido por más clientes, las necesidades de ellos incrementan y las empresas se ven obligadas a invertir en mecanismos basados en las TIC para incrementar su productividad. Es por ello que promueven la creación y exploración de diferentes entornos y plataformas web que involucren la Inteligencia Artificial (IA), uno de estos mecanismos son los chatbots, que son agentes conversacionales que tienen el objetivo de simular una conversación de manera natural con los usuarios.

Los chatbots facilitan el incremento de eficiencia en las empresas (dado que realizan su tarea con la menor cantidad de recursos) y eficacia (logran el objetivo esperado) de su proceso de atención al usuario, siempre y cuando contengan las características que apoyen en la comprensión del servicio y alcance los resultados esperados y estos sean satisfactorios (Paredes, 2019). Para lograr esta

[20] Tecnológico Nacional de México campus CENIDET Ingeniero en Sistemas Computacionales, actualmente es estudiante de maestría de Ciencias Computaciones del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), ubicada en Cuernavaca, México. Se encuentra en quinto semestre, en el área de Ingeniería de Software. Su trabajo de investigación está orientado en la etapa de análisis del software el cual se enfoca en entender el dominio del problema, donde se tiene como objetivo obtener un conocimiento suficiente para poder comunicarse eficazmente con clientes y usuarios, e identificar los actores dentro de los procesos tomando como base el reconocimiento de la voz.

[21] Recibió en 2006, el grado de Doctor en Ciencias con especialidad en Ingeniería Eléctrica del CINVESTAV-IPN, México. Fue investigador en el Instituto de Investigaciones Eléctricas (IIE, 1977-1992), Cuernavaca México, y fue líder de la sección de análisis del centro de cómputo de la Secretaría de Salubridad y Asistencia (SSA, 1973-1977). Ingresó en 1988, al Departamento de Ciencias Computacionales del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Cuernavaca México, donde es miembro del Grupo de Ingeniería de Software. Sus áreas de interés actuales son el desarrollo dirigido por modelos (MDD), la inteligencia de software, y el desarrollo de software.

[22] Egresado del Instituto Politécnico Nacional en México, donde se especializó en el área de Procesamiento de Lenguaje Natural. Es miembro del Sistema Nacional de Investigadores, vocal de la Sociedad Mexicana de Inteligencia Artificial (SMIA) y se desempeña como profesor investigador en el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET). Sus trabajos de investigación se orientan a las áreas de minería de opiniones y análisis de sentimientos, procesamiento de lenguas indígenas, análisis del discurso, tecnología educativa y creación y procesamiento de recursos léxicos, impactando en áreas como seguridad, salud y educación.

eficiencia y eficacia las empresas recurren a la ingeniería de software que es una disciplina que ofrece métodos y técnicas para desarrollar y mantener software de calidad dentro de un dominio en específico.

Esta disciplina es llevada por analistas que se enfocan en el proceso de análisis de dominio, el cual utilizan para identificar, documentar características comunes y variables de sistemas en un dominio específico. Este proceso implica distintas fases y actividades con el objetivo de evidenciar qué entidades y salidas derivados de estas actividades, deben ser rastreables y consistentes. En este contexto, realizar un proceso de análisis de dominio sin soporte de herramientas aumenta los riesgos de falla, pero la herramienta utilizada debe soportar el proceso completo y no solo una parte del mismo (Lisboa, 2010).

Alcance

En el Tecnológico Nacional de México campus Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), se ha venido trabajando con el desarrollo de Atenea, un agente conversacional que interactúa con las personas a través de la voz y cuyo objetivo es obtener toda la información necesaria de un dominio de problema específico para elaborar un modelo de negocio de alto nivel de abstracción propuesto por Warnier Orr (Orr, 1981), en el que se señalan explícitamente las personas con las que el entrevistado interactúa (entidades del proceso) y los documentos que intercambian entre ellos (salidas). Esta información es recabada por medio de una entrevista al personal administrativo.

Resultados principales obtenidos

Las entrevistas se conforman por 12 preguntas que pueden repetirse según lo amerite el curso de la conversación. El chatbot se entrenó con 400 frases, conformándose su vocabulario por 5 grupos de palabras que se consideran equivalentes y 20 intenciones. En cada intención se definen ejemplos de declaraciones que pueden activarlo e indican lo que se puede extraer de la expresión y la manera en que éste debe de responder. (Cloud, 2020).

La entrevista se almacena en un historial que posteriormente es procesado por medio de una herramienta de análisis morfosintáctico denominada Freeling, con lo que es posible detectar las entidades y las salidas. El diagrama que se genera ayuda a identificar la secuencia de las salidas y conocer su punto de origen que se reconoce por medio del verbo “enviar” y su punto destino que se obtiene del verbo “recibir”. Se propone que esta secuencia y la identificación de las entidades sean el mínimo necesario para realizar el modelo de negocios de manera correcta y conocer el procesamiento del modelo de dominio específico.

Métodos utilizados

El chatbot fue creado en la plataforma DialogFlow de Google, que permite crear agentes inteligentes que reconocen el lenguaje natural mediante voz o texto de forma sencilla integrando aprendizaje automático. Para seleccionar la herramienta que mejor se adaptara a las necesidades de esta investigación y cumpliera el alcance de poder crear agentes conversacionales con reconocimiento de voz, se realizó un estudio de mapeo sistemático de herramientas utilizadas para la creación de chatbots. La revisión de la literatura estuvo centrada en artículos entre los años 2013 al 2018,

escritos en el idioma inglés y produjo la identificación de nueve herramientas más populares, pero solo tres de ellas han sido las más utilizadas por la facilidad, tecnología y documentación disponible. Algunas de las características que se tomaron en cuenta para comparar estas herramientas son: la tecnología que aplica, por ejemplo: redes neuronales, procesamiento de lenguaje natural (PLN), aprendizaje automático o aprendizaje basado en reglas por mencionar algunas; la disponibilidad de la documentación en internet, el idioma español, el lenguaje de programación que soporta, el tipo de licencia (enfocándose a licencias gratuitas), la posibilidad de reconocimiento de voz y por último, que no existiera un límite de interacciones.

Con la revisión literaria se concluyó que la herramienta que mejor se adapta a las necesidades de esta investigación fue la herramienta ofrecida por Google debido a que cumple con los criterios antes señalados. Adicionalmente, se ejecuta en una plataforma en la nube de Google, por lo que su ejecución y acceso a la información puede realizarse desde cualquier equipo, además, permite implementarse en cualquier plataforma externa, por ejemplo, el Asistente de Google y Messenger, entre otros. Asimismo, nos brinda la libertad de poder monitorear el proceso de la conversación y así lograr retroalimentar el chatbot, especialmente donde no sabe qué responder o preguntar, lo cual beneficia en un continuo mantenimiento a lo largo del tiempo.

Valor del trabajo

Facilitar la obtención directamente del usuario, de la descripción del problema y/o la especificación de requisitos de sistemas de software que soporten el desempeño de un negocio.

Conclusiones obtenidas de lo presentado

Se construyó el chatbot Atenea, con herramientas libres de reconocimiento de voz y la utilización de bibliotecas de proceso de texto, hasta obtener los datos suficientes para crear manualmente (por ahora), diagramas de entidades a nivel usuario.

Con el chatbot Atenea estamos en condiciones para experimentar y determinar las hipótesis siguientes. Los chatbots soportados por técnicas de IA, pueden facilitar que las empresas:

- Mejoren la eficiencia (dado que realizan su tarea con la menor cantidad de recursos) y
- Mejoren la eficacia (logran el objetivo esperado) de su proceso de atención al usuario.

Palabras clave: Chatbot, Reconocimiento de voz, Procesamiento de Lenguaje Natural, Modelo de negocio, Modelo Warnier Orr.

Referencias

Cloud, G., 2020. *Documentación.* [En línea] Available at: https://cloud.google.com/dialogflow/docs/basics?hl=es-419 [Último acceso: 03 marzo 2020].

García, A. E. E., 2009. 'Vida e Inteligencia Artificial'. *ACIMED,* 19(1), pp. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352009000100006.

Gomez, E. &. T., 2000. 'La conciencia: Capitulo 2'. En: *'El rompecabezas del cerebro'.* s.l.: s.n., p. 5.

Lisboa, L. &. G. V. &. L. D. &. A. E. &. M. S. &. F. R., 2010. *'A systematic review of domain analysis tools'*. ResearchGate. *Journal,* Volumen 52, pp. 1-13.

Litwin, E., 2005. Tecnologías educativas en tiempos de Internet. En: Buenos Aires: Amorrotu.

Orr, K. T., 1981. *'Structured Requeriments Definition. In: Structured Requeriments Definition'.*. United States of America: s.n.

Paredes, C. T. M., 2019. *'Impacto de los chatbots en la atención al cliente en la cooperativa de ahorro y credito El sagrario',* s.l.: Uniandes.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Identificación de los posibles fonemas vocálicos de la lengua tanimuka aplicando *Machine Learning*

Alejandro Montenegro Taborda[23] **& Wilson Andrés Guaqueta Mateus**[24]
[amontenegrot@unal.edu.co | waguaquetam@unal.edu.co]
Universidad Nacional de Colombia, Colombia

Pese a su auge en otros campos, la Inteligencia Artificial ha presentado un desarrollo escaso en la investigación lingüística de lenguas que no cuentan con mucha documentación. En ese sentido, la presente investigación buscó aplicar algoritmos de *Machine Learning* a datos acústicos procedentes de un hablante nativo de la lengua tanimuka (lengua en peligro de extinción de la Amazonía colombiana) para identificar los posibles fonemas vocálicos de esta lengua. Este proceso contó con la transcripción de vocales en archivos de audio que contenían información acústica de la lengua, y a partir de la extracción de los valores para los formantes (F1, F2 y F3) de cada uno de los segmentos analizados se creó un espacio de representación vocálica en el que se realizaron distintas agrupaciones de dichos datos. Lo anterior se realizó con el algoritmo de clasificación no supervisada *K-means*, el cual aprovecha la distancia media entre los distintos puntos del conjunto de datos (equivalente al *área de dispersión* en fonología), para agruparlos según un punto referente llamado centroide (equivalente al *fonema*). Los resultados muestran una identificación de los fonemas que en gran parte concuerdan con estudios gramaticales previos de la lengua, por ello se puede concluir que si bien el uso de este algoritmo no sustituye la labor del lingüista, sí puede usarse como una herramienta que, primero, aporte información de gran valor para los análisis a realizar en el plano fonológico, y, segundo, implementarse en el estudio de lenguas similares: minoritarias, de poco prestigio y con escasa documentación.

Palabras clave: k-means, fonología, fonema, carta de formantes, clasificación no supervisada, vocales, tanimuka

[23] Lingüista con interés y experiencia en el análisis de la investigación del lenguaje desde la Lingüística Computacional y la Neurolingüística

[24] Lingüista con interés y experiencia en la investigación de lenguas indígenas y minoritarias, en el campo de la pragmática, la semiótica y el análisis del discurso.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Implementación de un sistema de diálogo automático para un dominio específico a partir de un pequeño banco de preguntas y respuestas

Antonio Jesús Tamayo Herrera[25]
[antonio.tamayo@udea.edu.co]
Instituto Politécnico Nacional, México
María Adriana Ruíz Osorio[26] **& Gabriel Ángel Quiroz Herrera**[27]
[gabriel.quiroz@udea.edu.co]
Universidad de Antioquia, Colombia

Resumen

La inteligencia artificial (IA) provee técnicas que son utilizadas en el tratamiento del lenguaje natural, las cuales permiten, tanto desde el enfoque computacional como del lingüístico, investigar cómo el lenguaje puede ser utilizado para cumplir diferentes tareas (Sosa, 1997).

Un ejemplo de ello es la aplicación de modelos de diálogo como forma de comunicación entre dos o más personas, que incorporados en un sistema informático permiten la interacción hombre-máquina, usando el habla (o texto) para la solución de diferentes problemas cotidianos como responder de forma oportuna a las dudas presentadas por los usuarios para adquirir cualquier producto o servicio (Zapata y Mesa, 2009).

Para cualquier entidad, es un asunto de gran interés dar respuesta inmediata a las inquietudes de sus usuarios, por lo que dar solución a dichas inquietudes se convierte en un proceso, en muchos casos, repetitivo y poco óptimo, evidenciando la necesidad de ser automatizado.

Con el pasar del tiempo, en la búsqueda de automatizar y mejorar los procesos, se han implementado sistemas de diálogo, basados en preguntas y respuestas, que combinando información de dominios específicos y técnicas de programación, dieron origen a la creación de los chatbots, programas informáticos diseñados para mantener conversaciones prolongadas llegando a imitar diálogos no estructurados como el que se dá entre los humanos (Jurafsky y Martin, 2017).

[25] Es ingeniero de sistemas y magíster en ingeniería de la Universidad de Antioquia. Es candidato a doctor en ciencias de la computación del centro de investigación en computación del Instituto Politécnico Nacional en Ciudad de México, donde desarrolla su tesis en procesamiento de lenguaje natural, bajo la supervisión del Dr. Alexander Gelbukh. Ha trabajado como analista y desarrollador de software para diversos proyectos con universidades de Colombia, Noruega y EE.UU. Desde el año 2010, trabaja en proyectos de ingeniería lingüística con el grupo de investigación en traducción y nuevas tecnologías (TNT) de la Universidad de Antioquia. Actualmente es profesor de cátedra en los programas de ingeniería de sistemas de la Universidad de Antioquia y la Universidad de Medellín.

[26] Es ingeniera de sistemas de la Universidad de Antioquia. Desarrolló su trabajo de grado en el área del procesamiento de lenguaje natural, construyendo un sistema de diálogo automático basado en corpus, combinando diferentes técnicas lingüísticas y computacionales. Actualmente, continúa investigando en esta área del conocimiento.

[27] Profesor titular de traducción y lingüística de la Universidad de Antioquia. PhD en Lingüística. Desarrolla su docencia e investigación en las áreas de traducción, terminología y tecnologías del lenguaje.

En la actualidad, el uso de agentes conversacionales es muy común, pues se han convertido en instrumento de múltiples disciplinas. Por ejemplo, en la medicina se encuentran muchas aplicaciones de chatbots construidos usando inteligencia artificial, que ayudan en diferentes áreas, realizando autodiagnósticos de enfermedades de acuerdo a los síntomas presentados por el paciente (Divya *et al.*, 2018). Del mismo modo, se encuentran aplicaciones de estos sistemas de diálogo en la Pediatría, donde se pretende guiar a los padres en cuanto a la medicación de los niños (Comendador *et al.*, 2015).

El presente trabajo muestra la construcción de un sistema de diálogo llevado hasta producción en un ambiente w*eb,* el cual permite la interacción hombre-máquina a través de una comunicación por escrito, para dar respuesta a todas las inquietudes relacionadas con un proceso administrativo específico.

Para el desarrollo del sistema de diálogo se probaron diferentes estrategias de representación vectorial de los textos, entre ellas *bag-of-words* y *word embeddings*. De igual manera, se probaron diferentes modelos, dentro del enfoque de la recuperación de información, entre los cuales se encuentran aquellos basados en distancia y diferentes algoritmos de aprendizaje automático, como se puede ver en la tabla 1.

Modelo	Parámetros	Representación de los datos	Eficiencia (*Accuracy*)
Regresión logística	C=100	TF-IDF	0.88
Support Vector Machine	C=10; Gamma=0.1; kernel= "rbf"	TF-IDF	0.83
Feed Forward Neural Network	Número de capas ocultas: 3 Características: 537 Neuronas en la capa oculta: 40 Neuronas en la capa de salida: 45 Funciones de activación: relu y softmax	TF-IDF	0.82
Regresión logística	C=100	Bag-of-Words	0.82
K-Nearest Neighbors	k=1	TF-IDF	0.76
Regresión logística	C=100	Embeddings (fastText)	0.72
Support Vector Machine	C=10; Gamma=0.1; kernel= "rbf"	Bag-of-Words	0.72
Random Forest	100 árboles	Bag-of-Words	0.70
Random Forest	100 árboles	Embeddings (Wikipedia2Vec)	0.70
Random Forest	100 árboles	TF-IDF	0.69
K-Nearest Neighbors	k=1	Bag-of-Words	0.63

Tabla 1. Resultados de las diferentes estrategias y modelos para la recuperación de respuestas con la mejor configuración encontrada

El sistema de diálogo llevado a producción fue desarrollado bajo un enfoque basado en corpus, partiendo de un conjunto de 33 preguntas con sus respectivas respuestas (base de datos). Estas 33 preguntas y sus respectivas respuestas fueron recolectadas del histórico de correos electrónicos e interacciones propias del proceso administrativo que se buscó automatizar con la

presente investigación. Usando una estrategia de parafraseo para adicionar información relevante, se amplió la base de datos a 594 preguntas y 45 respuestas (clases), con 537 características o palabras únicas en el vocabulario. A partir de la representación de los textos usando el esquema *Term Frequency - Inverse Document Frequency* (TF-IDF), usando la estrategia de recuperación de información con el modelo de regresión logística, combinado con un sistema de reglas implementado con expresiones regulares, el sistema de diálogo logra dar respuesta a múltiples preguntas elaboradas en lenguaje natural para un dominio específico, alcanzando una eficiencia del 88% en dichas respuestas.

Palabras clave: sistema de diálogo (chatbot), agentes conversacionales, regresión logística, *Term Frequency - Inverse Document Frequency* (TF-IDF), reglas lingüísticas, interacción hombre-máquina.

Referencias

Comendador, B. E. V., Francisco, B. M. B., Medenilla, J. S., & Mae, S. (2015). Pharmabot: a pediatric generic medicine consultant chatbot. *Journal of Automation and Control Engineering*, *3*(2).

Divya, S., Indumathi, V., Ishwarya, S., Priyasankari, M., & Devi, S. K. (2018). A self-diagnosis medical chatbot using artificial intelligence. *Journal of Web Development and Web Designing*, *3*(1), 1-7.

Sosa, E. (1997). Procesamiento del lenguaje natural: revisión del estado actual, bases teóricas y aplicaciones (Parte I). *El profesional de la información*, *6*.

Jurafsky, D., & Martin, J. (2017). Speech and Language Processing: An *Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* New Jersey. Prentice Hall.

Zapata, C., & Mesa, J. (2009). Los modelos de diálogo y sus aplicaciones en sistemas de diálogo hombre-máquina: Revisión de la literatura. *DYNA, 76*(160), 293-303.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

La fraseología especializada de ríos con nombre propio: una exploración de su extracción automática

Juan Rojas García[28]
[juanrojas@ugr.es]
Universidad de Granada, España

Introducción

Ecolexicon (https://ecolexicon.ugr.es) es una base de conocimiento terminológica y multilingüe sobre el medioambiente, que supone la aplicación práctica del enfoque teórico de la Terminología Basada en Marcos (TBM) (Faber, 2012). El diseño flexible de EcoLexicon permite la contextualización de los datos por dominios especializados de conocimiento (v. gr., Ingenierías de Costas, Geología o Meteorología), situaciones comunicativas y áreas geográficas. Sin embargo, la representación de entidades geográficas con nombre propio (EGNP), como los ríos (v. gr., Río Yangtsé), requiere saber qué términos están relacionados semánticamente con cada EGNP dentro de un dominio especializado concreto, y cómo dichos términos se relacionan entre sí. En la TBM, un término es una unidad lingüística especializada, empleada dentro de una disciplina científica o técnica, que tiene significados específicos en ese campo de conocimiento y un potencial combinatorio concreto con otros términos.

Con el fin de representar en EcoLexicon tanto las estructuras conceptuales como la fraseología (entendida en sentido amplio) asociadas al uso lingüístico de las EGNP mencionadas en un corpus especializado de Ingenierías de Costas, en lengua inglesa y de pequeño tamaño (7 millones de palabras), los términos relacionados con cada EGNP y sus relaciones semánticas se han extraído con modelos semánticos distribucionales (MSD) y otras técnicas estadísticas. En este trabajo nos centramos en los ríos con nombre propio.

Si bien los MSD ayudan a identificar relaciones semánticas entre términos a partir de los datos de un corpus (Bernier Colborne y Drouin, 2016), la construcción de un MSD, apropiado para una tarea particular, precisa de una laboriosa optimización de sus parámetros de funcionamiento, como el tamaño de la ventana contextual alrededor de cada término, o la medida de asociación entre términos que se aplica. Aunque numerosos estudios han abordado la evaluación y la optimización de los MSD en corpus generales de gran tamaño, del orden de cientos, e incluso miles, de millones de palabras (Baroni *et al.*, 2014; Kiela y Clark, 2014; Lapesa *et al.*, 2014), la habilidad de los MSD para capturar relaciones semánticas en corpus

[28] Juan Rojas-Garcia holds degrees from the University of Malaga and the University of Granada in Telecommunication Engineering, Translation and Interpreting, and Teaching of Spanish as Native and Foreign Language. He is a member of the LexiCon research group at the University of Granada, and was awarded a research fellowship by the Spanish Ministry of Economy and Competitiveness to write his PhD on Terminology.

especializados de pequeño tamaño ha recibido poca atención entre los investigadores de la comunidad del procesamiento del lenguaje natural (Fabre *et al.*, 2014).

Objetivos

Consecuentemente, el objetivo de este trabajo es doble:

1. En primer lugar, se construye una colección de términos vinculados con ríos con nombre propio (RNP). Se trata de un repertorio de tripletas en las que se especifican las correspondientes relaciones semánticas entre cada término y un RNP. Para ello, se analizan las estructuras sintagmático colocacionales (en sentido amplio) de las oraciones donde se menciona un RNP en el corpus especializado referido más arriba.
2. En segundo lugar, se hallan combinaciones de parámetros en los MSD, basados en recuentos, que optimicen su funcionamiento para extraer términos, del mismo corpus, vinculados con los RNP mediante las tres relaciones semánticas takes_place_in, affects y located_at. Para ello, los modelos se evalúan con conjuntos de datos gold standard, esto es, datos de referencia con los que juzgar el funcionamiento de los modelos. Estos datos de evaluación son, precisamente, las tripletas que se obtienen del primer objetivo de investigación. Estos datos de evaluación no se pueden extraer de una fuente externa al corpus porque, actualmente, no existen recursos terminológicos u ontológicos que almacenen las relaciones semánticas que entablan los RNP en dominios especializados del conocimiento.

Materiales

Los RNP y sus términos relacionados se extraen de un corpus de textos escritos, en lengua inglesa, sobre Ingeniería de Costas, al que se aplica diversos MSD basados en recuentos. Este corpus, de 7 millones de palabras, está compuesto por textos especializados (artículos científicos, informes técnicos y tesis doctorales) y semiespecializados (libros de texto y enciclopedias sobre Ingeniería de Costas).

Los RNP que se mencionan en el corpus se reconocen automáticamente mediante la base de datos geográficos GeoNames (http://www.geonames.org).

Metodología

A nuestro corpus especializado se le aplican diversos MSD basados en recuentos. En este trabajo solo se emplean MSD basados en recuentos porque Asr *et al.* (2016), Sahlgren y Lenci (2016), y Nematzadeh *et al.* (2017) enfatizan que los modelos basados en recuentos superan a los predictivos en corpus de pequeño tamaño, inferiores a los 10 millones de palabras, aproximadamente. En los modelos se evalúan dos parámetros: (1) el tamaño de la ventana contextual (se prueba el rango entre 1 y 10 palabras); y (2) la medida de asociación entre términos (log likelihood estadístico, información mutua específica, y t score).

El repertorio de tripletas del primer objetivo de investigación actúa como datos de evaluación para los MSD, porque contiene parejas de términos relacionados semánticamente

que se han extraído de forma manual del mismo corpus. Uno de los términos de cada pareja es siempre un RNP, y el otro término es una entidad o un proceso. Tres terminólogos de nuestro grupo de investigación han anotado las relaciones semánticas en las tripletas. El índice de acuerdo entre los anotadores, medido con kappa de Cohen (κ), mostró un acuerdo alto entre todas las parejas de anotadores (κ>90 %, p valor<0.05).

Las relaciones semánticas que unen los términos en las tripletas son las que los RNP activan frecuentemente en el corpus, a saber: (1) takes_place_in (v. gr., SEA LEVEL RISE takes_place_in MISSISSIPPI RIVER DELTA); (2) affects (v. gr., MISSISSIPPI RIVER affects SOFT MUD); y (3) located_at (v. gr., BARRIER ISLAND located_at MISSISSIPPI RIVER MOUTH). Por tanto, se construyen tres conjuntos de datos gold standard, uno para cada relación semántica.

Gracias a los datos de evaluación, se juzga el desempeño de los MSD para extraer términos relacionados con los RNP mediante cada una de las tres relaciones semánticas anteriores. Para dicha evaluación, se emplea la figura de mérito Mean Average Precision (MAP) (Manning *et al.*, 2009: 158 162), que resume, con un único valor, la efectividad de un modelo cuando se mide su precisión (precision) a lo largo de diferentes niveles de exhaustividad (recall), esto es, MAP se aproxima al área bajo la curva de precision recall de un modelo.

Resultados

Para nuestro corpus especializado, en lengua inglesa y de pequeño tamaño, los resultados muestran, por una parte, que la medida de asociación log likelihood estadístico supera a la información mutua específica y al t score para todas las relaciones semánticas. Por otra, la detección de una relación específica depende del tamaño de la venta contextual del MSD. En concreto, un tamaño de ventana de 4 palabras para la relación takes_place_in, 3 palabras para affects, y 2 palabras para located_at (vid. Figura 1).

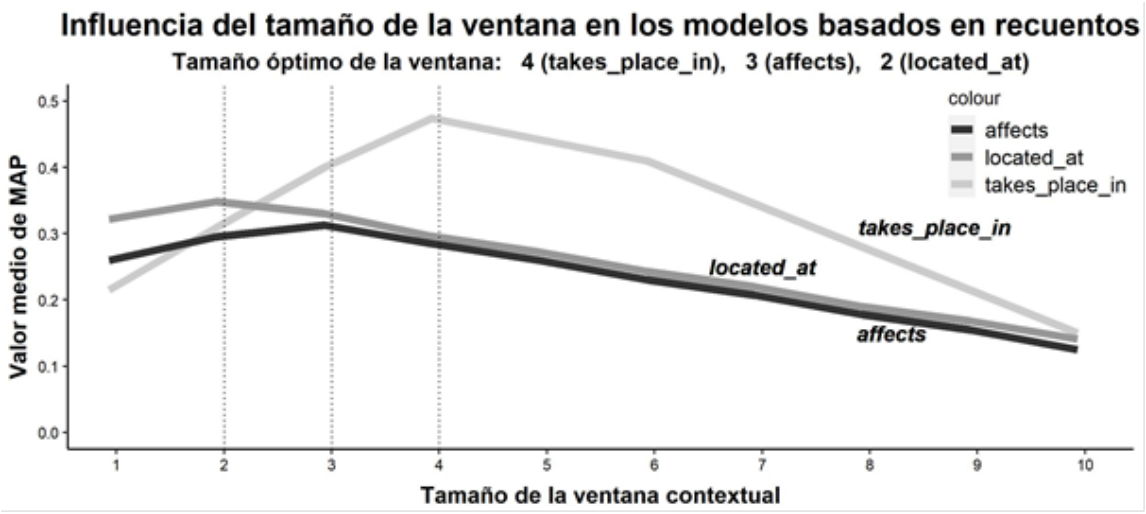


Figura 1: Valor medio de MAP en función del tamaño de la ventana contextual.

Asimismo, el método de extracción que se desarrolla para el inglés en este trabajo puede aplicarse a otras lenguas, ajustando, apropiadamente, el tamaño de la ventana contextual de los modelos para la detección de una relación específica.

Palabras clave: Río con nombre propio, Fraseología especializada, Representación conceptual, Modelo semántico distribucional basado en recuentos, Terminología.

Referencias

Asr, F., Willits, J., y Jones, M. (2016). Comparing predictive and co occurrence based models of lexical semantics trained on child directed speech. En Proceedings of the 38th Annual Conference of the CogSci (pp. 1092 1097). Filadelfia, EE. UU.: CogSci.

Baroni, M., Dinu, G., y Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context counting vs. context predicting semantic vectors. En Proceedings of the 52nd Annual Meeting of the ACL (pp. 238 247). Baltimore, EE. UU.: ACL, vol. 1.

Bernier Colborne, G., y Drouin, P. (2016). Evaluation of distributional semantic models: A holistic approach. En Proceedings of the 5th International Workshop on Computational Terminology (pp. 52 61). Osaka, Japón: Coling.

Boas, H.C. (2005). Semantic frames as interlingual representations for multilingual lexical databases. International Journal of Lexicography, 18(4), 445 478.

Faber, P. (Ed.) (2012). A cognitive linguistics view of Terminology and specialized language. Berlín/Boston: De Gruyter Mouton.

Fabre, C., Hathout, N., Sajous, F., y Tanguy, L. (2014). Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. En TALN-RECITAL 2014 Workshop (pp. 266 279). Marsella, Francia: ATAL.

Kiela, D., y Clark, S. (2014). A systematic study of semantic vector space model parameters. En Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (pp. 21 30). Gotemburgo, Suecia: ACL.

Manning, C.D., Raghavan, P., y Schütze, H. (2009). Introduction to information retrieval. Cambridge, Reino Unido: Cambridge University Press.

Nematzadeh, A., Meylan, S.C., y Griffiths, T.L. (2017). Evaluating vector space models of word representation, or, the unreasonable effectiveness of counting words near other words. En Proceedings of the 39th Annual Meeting of the CogSci (pp. 859 864). Londres: CogSci.

Sahlgren, M., y Lenci, A. (2016). The effects of data size and frequency range on distributional semantic models. En Proceedings of the 2016 Conference on EMNLP (pp. 975 980) Austin, EE. UU.: ACL.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

LISS: A Corpus of Literary Spanish Sentimental Sentences for Emotions Detection

Luis Gil Moreno Jiménez[29] **& Juan Manuel Torres Moreno**[30]
[luis-gil.moreno-jimenez@alumni.univ-avignon.fr | juan-manuel.torres@univ-avignon.fr]
Université d'Avignon, France

The LiSSS Corpus

Research in Natural Language Processing (NLP) focused in emotion classification has used corpora constituted mainly by text extracted from social media (Facebook or Twitter); this kind of text presents a very general language with grammar and orthographic errors [5, 7]. Other works have used encyclopedic documents like Wikipedia, journals (newspapers or magazines) [10, 2, 9] for the development and evaluation of classification models. Studies of literary corpora have been systematically left aside mainly because the level of literary discourse is more complex than other genres. In this work we introduce a new literary emotion annotated corpus in order to validate and evaluat the NLP algorithms for literary emotions classification task. Several corpora in Spanish have been made available to the scientificcommunity [3], but just some of them have been classified considering categories of emotions. For example SAB is a tweets corpus in Spanish [8]. The tweets of this corpus represent critics towards different comercial brands. Annotation was realized considering the emotion perceived in each tweet. The SAB corpus contains 4 548 annotated tweets in 8 emotions: [Trust, Satisfaction, Happiness, Love, Fear, Disaffection, Sadness, Anger]. Another data set concerning tweets is the TASS corpus [12]. It contains about 70K tweets related to different topics: Politics, Economy, Sport, Music, etc., classified automatically into the categories: [Positive, Negative, Neutral, None]. A global analysis (Word level) about emotions polarity is described in [1]. The corpus employed is composed of lexicons in 40 languages and a bi-polar annotation. Besides from above-mentioned corpora, a toy Literary Sentiment Sentences in Spanish corpus named LiSSS was

[29] Ingeniero en Tecnologías de la Información por la Universidad Tecnológica de la Selva, Chiapas, México Maestro en Ciencias de la Computación por el Centro de Investigación y Desarrollo Tecnológico, Cuernavaca, Morelos, México Doctorante en la Universidad de Avignon, Francia con la Tesis: "Creatividad Computacional: Generación Automática de Frases Literarias" Estancias de Investigación: Laboratorio de Informática de Avignon (Clasificación y Detección de Entidades Nombradas) Université d'Avignon, Francia; Laboratorio de Computación y Matemáticas (Generación Textual y Creatividad Computacional) Universidad del Estado de Rio de Janeiro, Brasil.

[30] Actualmente es Profesor asociado habilitado a dirigir investigación (HDR) en la Université d'Avignon (UAPV), en el equipo de investigación TALNE de Procesamiento de Lenguaje Natural (PLN), Laboratoire Informatique d'Avignon (LIA). Sus investigaciones conciernen el PLN, el resumen automático de documentos, generación automática de frases. compresión de frases, aprendizaje automático y conciencia artificial. Profesor asociado (2006-20) en el Departamento de Génie Informatique-génie logiciel de la Polytechnique de Montreal (Canadá) y Profesor Adjunto (2000-01) en el Departamento de Ciencias Aplicadas en la Universidad de Québec (Canadá).

introduced in [11]. We present an extended version of LiSSS corpus. This corpus consists exclusively of literary texts, which makes it more useful for studying the algorithms of automatic emotion classification or generation of literary text [6]. For classification, five categories of emotions were defined, instead of a binary category (positive-negative). This characteristic of LiSSS could be useful for more fine analysis. Our work was meticulous but on a modest scale. We decided to créate a small controlled corpus, exclusively composed of literary sentences in Spanish selected from universal literatura. The LISSS corpus was constituted manually with texts in Spanish from 211 authors, using Spanish original versions or human translated version for other languages. Each sentence in the corpus was read and manually classified using the following five categories: Anger (A); Love (L); Fear (F); Happiness (H); Sadness/Pain (S). Since the sentences or paragraphs (group of sentences) could belong to two or more emotions, we included the possibility of tagging the sentences using all the emotions expressed. The corpus currently has 705 sentences coded in utf8. Each line contains the information of the following fields:

ID The_sentence # Author

Each field is separated by a tab character. The ID field is composed of a sequential number (1,2,3,...) followed by a code (A, L, F, H, S) for each of the predefined emotions. If a sentence is ambiguous, it will have as many codes as categories it belongs to. Sentences were selected manually to maintain a balance between the five types of emotions. It should be noted that sentences are actually sometimes mini-paragraphs composed of several sentences. This was done to respect as much as possible the coherence of the idea expressed and the corresponding emotion. For example, sentence 80 of the emotion Love by Lope de Vega:

80L La raíz de todas las pasiones es el amor. De él nace la tristeza, el gozo, la alegría y la desesperación. The sentences were selected from quotes, stories, novels and some poems. Sentences in general language as well as those too short (N _ 3 words) or too long (N _ 50 words) were carefully avoided. There are _ 15 words per sentence. The final vocabulary is complex and aesthetic where certain literary figures like anaphora or metaphor can be observed. The LiSSS corpus is characterized in Tables 1 and 2. Table 2 represents the distribution of the 5 emotion categories. There are several overlapped emotions like AF, LS, etc. There are 249 sentences (_ 35%) considered as multi-emotional. The most important overlapping is detected between Love and Sadness/Pain, LS with 47 overlapped sentences. The emotion with the highest overlap degree corresponds to Sadness, with 42.6%.

	Sentences	Paragraphs	Words	Characters	Authors
LiSSS v0.700	705	49	13 565	75 294	210

Table 1: LiSSS corpus of literary sentences.

Emotion	A	F	H	L	S	Overlap%
A	102	10	1	14	3	22.1
F	10	124		12	7	19.0
H	1		114	4	26	21.9
L	14	12	4	127	47	37.7
S	3	7	26	47	113	42.6

Table 2: LiSSS corpus: multi-emotional distribution.

An example of this sort of sentences is the labeled with the identifier 14AL, Anger (A) and Love (L):

14AL Del amor al odio, solo hay mas amor (From love to anger, there's only more love) it belongs to both categories. Literary ambiguity represents a challenge to classification methods. The LiSSS corpus has the advantage of being homogeneous in terms of genre literary, but it is heterogeneous in terms of emotions classes. In others emotion corpora, the sentences may be in general language (sentences that give a fluency to the reading and provide a well cohesion between sentences). Likewise, the corpora with tweets are not able to be used with literary goals due to the presence of *noise*, emoticons and special characters. Another advantage of LiSSS corpus is that the presence of noise (cut phrases, pasted words, wrong syntax, etc.) was carefully avoided. It is suitable for testing the quality and performance of learning and classification text algorithms. The LiSSS corpus (base corpus and its POS tagged Freeling[31] version) is available on http://juanmanuel.torres.free.fr/corpus/lisss, under GPL3 licence.

Baseline test classification

The LiSSS corpus was tested with classical classification algorithms available in Weka[32]. We have employed: J48 al- gorithm; Naive Bayes (NB); Naive Bayes Multinomial (NBM); Support Vector Machine (SVM) and Recurrent Neural Networks (RNN) to compare their performance on this corpus using the default parameters configured in Weka. We have generated a learning corpus (independent of our test corpus) in order to train the algorithms, adapting it to the categories in the LiSSS corpus. This new corpus, called *CitasIn*, is composed of Spanish documents, mostly from the literary genre. A large number of documents belonging to different categories (friendship, lovers, beauty, success, hap- piness, laughter, enmity, deception, anger, fear, etc.) were retrieved from the website https://citas.in/temas/.[33]. These documents were classified into the 5 classes of the LiSSS corpus, from a manual mapping with their own categories. The generated corpus has an adequate size to be used in training tasks for automatic learning models. The disadvantage of *CitasIn* corpus is the presence of noise given by supporting

[31] http://nlp.lsi.upc.edu/freeling

[32] https://www.cs.waikato.ac.nz/ml/weka/

[33] All documents were downloaded, with the editor authorisation, on 25 March 2020 from the website: https://citas.in

sentences (with general language vocabu- lary) not belonging to the literary genre. The *CitasIn* corpus is characterized in Table 3. It is possible to reconstituting the corpus *CitasIn* using the class correspondence described in [11]. A classic pre-processing was applied. This pre- processing include the suppression of special symbols as well as the function words (using the Weka libraries and their stop-words lists); also, the capital letters were normalized. Finally a tokenization process was applied using Weka specific algorithms for Spanish.

The tests were performed using *CitasIn* as learning corpus and LiSSS corpus for testing. Results are shown in Table. The best model is the J48-tree algorithm, obtaining a weighed F-measure=0.520 (harmonic combination of Precision and Recall computed by Weka). This relatively low value shows the difficulty of classify multi-emotional literary corpora. In this corpus in particular, the emotions Sadness and Love were often confused. Another problem in the classification of this type of texts is to deal with the richness of the lexicon.

CitasIn	Sentences	Words	Word per sentence
Emotion	53 351	1 903 214	s 35.6
A	9 556	384 211	40.2
L	13 430	392 623	29.2
F	8 917	355 976	39.9
H	10 857	377 280	34.7
S	10 591	393 124	37.1

Table 3: CitasIn corpus: Sentences coming from different categories mapped into 5 emotions

Algorithm	A	F	H	L	S	mean F-score
NB	0.250	0.396	0.477	0.444	0.110	0.331
SVM	0.541	0.606	0.481	0.512	0.241	0.464
RNN	0.502	0.512	0.509	0.539	0.308	0.468
NBM	0.457	0.566	0.593	0.593	0.348	0.508
J48	0.514	0.524	0.631	0.615	0.340	0.520

Table 4: Evaluation of F-score models' performance per class.

Conclusion and future work

We have presented the LiSSS corpus of literary emotions in Spanish. The aim of this new corpus is to test Machine Learning algorithms on specialized literary corpus; no to train such algorithms. The sentences of literary texts are often multi-emotional and the semantic ambiguous. The overlap between the vocabulary of the classes causes the methods to be unable to correctly classify the corpus. The results show that the emotion classification on this type of documents is a difficult and complex exercise. We think that automatic classifiers can be enriched through the integration of other modules, using characteristics of language or style detection to achieve better classification [4, 6]. For future work, in order to enrich this corpus, we will introduce more authors, sentences and a pull of several annotators.

References

[1] Y. Chen and S. Skiena. Building sentiment lexicons for all major languages. In 52nd Annual Meeting of the ACL, volume 2, pages 383–389, 2014.

[2] I. da Cunha, M. T. Cabré, E. SanJuan, G. Sierra, J.-M. Torres-Moreno, and J. Vivaldi. Automatic specialized vs. non-specialized sentence differentiation. In CICLing, pages 266–276, 2011.

[3] I. da Cunha, J.-M. Torres-Moreno, and G. Sierra. On the development of the RST spanish treebank. In 5th Linguistic Annotation Workshop, LAW 2011, June 23-24, 2011, USA, pages 1–10. ACL, 2011.

[4] A. Edalat. Self-attachment: A holistic approach to computational psychiatry. In P. Érdi, B. Sen Bhattacharya, and A. Cochran, editors, Computational Neurology and Psychiatry. Bio-/Neuroinformatics, volume 6, pages 273–314. Springer, Cham, 2017.

[5] K. Howells and A. Ertugan. Applying fuzzy logic for sentiment analysis of social media network data in market- ing. In 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW, volume 120, pages 664–670. Elsevier, 2017.

[6] L.-G. Moreno-Jiménez, J.-M. Torres-Moreno, and R. Wedemann. Literary natural language generation with psychological traits. In 25th NLDB. NLDB, 2020. Accepted.

[7] S. Nadali, M. A. A. Murad, and R. A. Kadir. Sentiment classification of customer reviews based on fuzzy logic. In 2010 International Symposium on Information Technology, volume 2, pages 1037–1044. IEEE, 2010.

[8] M. Navas-Loro, V. Rodríguez-Doncel, I. Santana-Pérez, and A. Sánchez. Spanish corpus for sentiment analysis towards brands. In International Conference on Speech and Computer, pages 680–689, 2017.

[9] G. Sierra. Introducción a los Corpus Lingüísticos. UNAM Mexico., 2018.

[10] J.-M. Torres-Moreno. Automatic Text Summarization. Wiley, London, 2014.

[11] J.-M. Torres-Moreno and L.-G. Moreno-Jiménez. Lisss: A toy corpus of literary spanish sentences sentiment for emotions detection. arXiv, 2005.08223v1, 2020.

[12] J. Villena-Román, S. Lana-Serrano, E. Martínez-Cámara, and J. C. González-Cristóbal. Tass - workshop on sentiment analysis at sepln. Procesamiento del Lenguaje Natural, 50(0):37–44, 2013.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Modelo Semántico No Supervisado Para La Detección Automática De Sentimientos

Carlos Henríquez[34], Dixon Salcedo[35] & Diana Suárez[36]
[cnhenriquez@gmail.com | dsalcedo2@cuc.edu.co | dianam.suarezl@unilibre.edu.co]
Universidad del Magdalena, Universidad de la costa, Universidad Libre seccional Barranquilla, Colombia

Resumen: Este documento presenta un modelo no supervisado para el análisis de sentimientos basado en aspectos, que extrae automáticamente los aspectos de opinión y determina su polaridad asociada. El modelo utiliza ontologías y similitud semántica para la detección de aspectos explícitos e implícitos y aprendizaje automático no supervisado para la polaridad de los aspectos. El trabajo experimental se realizó utilizando un conjunto de datos bajo el dominio de restaurante con valores de F1 de 73.07 y exactitud de 84.8% y el dominio de hoteles 89,18 en F1 y exactitud de 88.46%.

Palabras clave: Análisis de sentimientos; Ontología; Similitud semántica; Aprendizaje automático; no supervisado.

Actualmente, la cantidad de datos producidos en todo el mundo es muy alta debido al uso masivo de redes sociales, servicios de mensajería, comercio electrónico, entre otros. Toda esta gama de datos es atractiva para diferentes establecimientos comerciales, industriales, académicos y de otro tipo; pero, la extracción y su respectivo procesamiento de forma manual

[34] Ingeniero de Sistemas, Especialista en estudios pedagógicos, Magister en Ingeniería de Software y PhD en Ingeniería de Sistemas e informática. Profesor de la Universidad del Magdalena. Investigador Senior ante Minciencias. Áreas de investigación: aprendizaje automático, procesamiento del lenguaje natural y análisis de sentimientos. Editor de la revista científica Prospectiva. Líder de grupo de investigación Sistemas inteligentes y nuevas categorías categorizado A. Director de proyectos de software como consultor y profesor. Consultor e instructor en tecnología JAVA (J2EE, J2SE, JavaCard, Android).

[35] Ph.D. en Ingeniería con énfasis en Telecomunicaciones. Magister en Software Libre, con énfasis Administración de Redes de Computadores y Sistemas. Ingeniero de Sistemas. Profesor tiempo completo del programa Ingeniería de Sistemas, adscrito al Departamento de Ciencias de la Computación y Electrónica, y es miembro del grupo de Investigación de Ingeniería del Software y Redes de la Universidad de la Costa-CUC. Investigador Junior ante Minciencias. Intereses de investigación son en el campo de calidad de servicio en redes Internet, ingeniería de tráfico, redes de computadores y protocolos de nueva generación.

[36] Ingeniero de Sistemas, Especialista en Ingeniería de software, Magister en Administración e innovación, doctoranda en ciencia y tecnología informática. Profesor de tiempo completo de la Universidad de Libre seccional Barranquilla, Investigador asociado ante Min ciencias. Áreas de investigación aprendizaje automático, procesamiento del lenguaje natural, recuperación y extracción de información.

hace que esta tarea sea muy compleja y difícil de realizar. Para poder analizar y administrar estos datos, existen grandes frentes de trabajo en búsqueda de nuevos modelos, técnicas y herramientas que permitan el análisis de textos automáticamente [1]. Investigaciones recientes han conllevado a una vertiente para el procesamiento del lenguaje natural (PNL), llamado análisis de sentimientos (AS). El AS busca analizar en textos, las opiniones, comentarios, actitudes y emociones de las personas hacia entidades tales como productos, servicios, individuos, así como sus características [2]. Para analizar los sentimientos, de acuerdo con[3] hay tres niveles: a nivel de documento, a nivel de oración y basados en aspecto. A nivel de documento tiene como meta clasificar el sentimiento de todo un documento en positivo o negativo, a nivel de oración obtiene el sentimiento de cada oración de un párrafo dentro de un texto. Finalmente, el enfoque basado en aspectos (ASBA) clasifica con respecto a las características específicas de cada una de las entidades. De acuerdo con [2], "*el nivel de documento y el nivel de frase no descubren qué es exactamente lo que le gusta y no le gusta a la gente, a diferencia del AS en el nivel de aspecto, que realiza el análisis en más detalle*", es decir, se centra en las características fundamentales de la opinión.

El ASBA, tiene como objetivo identificar las propiedades de una entidad y los sentimientos asociados a ella dentro de una expresión. Un aspecto es un atributo o componente de una entidad que puede aparecer explícitamente en la opinión o a través de una expresión no explicita. Por ejemplo, en la frase, "*El baño de la habitación está sucio*", el aspecto es "*baño*" y la entidad es "*habitación*", el sentimiento asociado es "*sucio*" que en este caso es una polaridad "*negativa*". En el AS a nivel de aspectos son aplicados básicamente dos procesos fundamentales, el primer proceso identifica y extrae los aspectos de una opinión y el segundo determina su polaridad (positiva, negativa o neutral). Además, se distinguen dos tipos de aspectos, el primero se refiere a los aspectos explícitos que son palabras en el documento que denotan directamente el objetivo de la opinión. El segundo, es el aspecto implícito, éste representa el objetivo de opinión de un documento, pero que no se especifica explícitamente en el texto.

En este documento se presenta un modelo de ASBA bajo un enfoque semántico y centrado en el concepto (significado) para la identificación de aspectos explícitos e implícitos y un enfoque no supervisado para determinar la polaridad de los aspectos identificados. En la siguiente figura se presenta el modelo propuesto:

En la capa 1, procesamiento del lenguaje , se permite la entrada de opiniones a través de un documento escrito, en donde se aplican técnicas de procesamiento de lenguaje natural como segmentación, normalización, etiquetado sintáctico y lematización [4]. La salida de esta capa es un conjunto de palabras etiquetadas y lematizadas. Por ejemplo, en el texto "*La carne salada*", la salida sería: *"la, D, el", "carne, N, carne", "salada, A, sal",* donde la palabra "*salada*" se identificó como adjetivo (A) y su raíz es "*Sal*".

En la capa 2, extracción de aspectos, se identifican y extraen los posibles aspectos de una entidad a partir de la salida de la capa anterior. Aquí se utiliza, un modelo semántico que verifica si un conjunto de aspectos candidatos se encuentran en la terminología de un dominio específico con la ayuda de una ontología y una base de datos léxica. Inicialmente, los aspectos candidatos son elegidos, a partir de su categoría gramatical (sustantivo), y se selecciona una ontología de dominio. Los aspectos candidatos se comparan con las clases e individuos de la ontología y los que coinciden se marcan como aspectos explícitos. Después del proceso anterior, los sustantivos de las opiniones que no se encontraron en la ontología experimentan un proceso de similitud semántica con las clases de la ontología. En esta propuesta, el cálculo de la similitud semántica se basa en el algoritmo de Wu & palmer [5] que considera la posición de los conceptos c1 y c2 en una taxonomía con respecto a la posición del concepto común más específico entre los dos (c1, c2) . Finalmente, para la extracción de aspectos implícitos se utilizan técnicas de doble propagación y matriz de coocurrencias entre aspectos explícitos y palabras de opinión para identificar posibles aspectos implícitos [6]. La salida de esta capa es un conjunto de palabras identificados como aspectos. Por ejemplo, en el texto "*La carne está salada. No volveré.*", la salida sería: "*carne*" aspecto explicito, y "*restaurante*" aspecto implícito hallado de la expresión "*No volveré*".

En la capa 3, identificación de sentimiento, se eligen expresiones que están relacionadas con los aspectos encontrados, para luego encontrar su polaridad. Para lograr esto, se utilizaron dos técnicas: ventana deslizante y reglas gramaticales. El proceso de ventana consiste en tomar la oración donde está el aspecto y establecer una ventana de palabras a la derecha e izquierda del aspecto seleccionado. Con esta longitud de ventana, el propósito es identificar expresiones de sentimiento que puedan afectar el aspecto. En la literatura, básicamente se han usado adjetivos como expresiones de sentimiento; para este modelo se ha definido que las expresiones cercanas al aspecto sean adjetivos y adverbios. Adicionalmente, las reglas gramaticales se utilizan para determinar si el sentimiento encontrado se ve afectado por la negación o la atenuación. El manejo de la negación de esta propuesta es negación simple y la atenuación consiste en descubrir la afectación del sentimiento por parte de adverbios generales como "*muy, bastante, demasiado, más*". Por ejemplo, en el texto "*La carne está muy salada*", la salida sería: "*carne*" aspecto explicito, "*salada*" expresión de sentimiento y "*muy*" como expresión de atenuación.

En la capa 4, clasificación de sentimientos, se utiliza una técnica basada en la medida de asociación conocida como punto de información mutua (PMI) [7]. Esta medida permite determinar la orientación semántica de las expresiones de sentimientos y los aspectos a través de la selección adecuada de semillas de sentimiento y un corpus de dominio. El PMI de dos palabras (x, y) se obtiene por la probabilidad de que las dos palabras aparezcan juntas divididas por las probabilidades de cada palabra individualmente (ver ecuación 1).

$$PMI(x,y) = \log 2\left(\frac{P(x,y)}{P(x)P(y)}\right) \qquad (1)$$

Para el modelo presentado, en el cálculo del PMI se tiene en cuenta el aspecto, la expresión de sentimiento y un conjunto de semillas. En el cálculo, se toma cada expresión de opinión x_i y su frecuencia $f(x_i)_A$ se calcula solo en el conjunto de opiniones en el que aparece el aspecto A. Lo mismo se hace para cada semilla $f(y_j)_A$ y las coincidencias entre los dos $f(x_i, y_j)_A$. Con estos valores, se obtiene un PMI mayor que cero. En el contexto del sistema propuesto, el $PMIP_A$ será el valor más alto de PMI entre la expresión de opinión y la semilla (ecuación 2). Formalmente se tiene un conjunto n de expresiones de opinión $\bar{X} = (S)$, un conjunto de m semillas $\bar{Y} = (SE)$ y el aspecto A. Entonces, el punto de información mutua positiva $PMIP_A$ dentro de un subconjunto de opiniones del corpus donde A está entre $\bar{X}\ y\ \bar{Y}$, será el valor más alto entre la concurrencia de cada semilla y_j y la expresión sentimental x_i.

$$PMIP_A(\bar{X}, \bar{Y}) = \max\left[\log 2\left(\frac{f(x_i y_j)_A}{f(x_i)_A f(y_j)_A}\right)\right] \quad (2)$$

El conjunto de semillas definidas para este trabajo consiste en cinco (5) palabras que representan una disposición emocional hacia lo positivo, lo negativo y lo neutral. Las semillas seleccionadas para positivo son "*excelente*" y "*bueno*", para negativo "*malo*" y "*pésimo*" y para neutral "*indiferente*". A medida que el modelo recibe más opiniones, éstas se almacenan en el corpus de opiniones que ajustan los valores del PMI. El resultado final de la capa de clasificación de sentimientos es un conjunto de aspectos con su expresión de sentimiento y su polaridad asociada que es el resultado final del modelo. Por ejemplo, en el texto "*La carne está muy salada*", la salida sería: "*carne*" aspecto explicito, "*salada*" expresión de sentimiento, "*muy*" como expresión de atenuación y la polaridad asociada "*negativo* "dependiendo de las opiniones del corpus.

El modelo presentado se implementó mediante la construcción de una aplicación (*AspectSA*) [8] bajo tecnología Java que integra diferentes herramientas y bibliotecas para la gestión del idioma español. Para validar el sistema construido, se llevaron a cabo una serie de experimentos en los dominios de restaurante y hoteles. Las medidas de evaluación que se utilizan en gran parte de los sistemas de análisis de sentimientos son: medida F1 para extracción de aspectos y exactitud (*accuracy*) para clasificación de sentimientos. El sistema *AspectSA* obtuvo un valor F1 de 73.07 y exactitud de 84.8% en el dominio de restaurantes [9] y para el dominio de hoteles obtuvo 89,18 en F1 y 88.46% bajo validación cruzada (*10-fold cross-validation*). Estos resultados son altamente competitivos con los sistemas similares existentes en español y ofrece mayor completitud en la extracción de aspectos.

Referencias

[1] F. A. R. Silva, "Analytical Intelligence in Processes: Data Science for Business," *IEEE Lat. Am. Trans.*, vol. 16, no. 8, pp. 2240–2247, 2018.

[2] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. 2015.

[3] B. Dalila, A. Mohamed, and H. Bendjanna, "A review of recent aspect extraction techniques for opinion mining systems," 2018, doi: 10.1109/ICNLSP.2018.8374382.

[4] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," 2013, doi: 10.1016/j.procs.2013.05.005.

[5] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994, pp. 133–138.

[6] L. Sun, S. Li, J. Li, and J. Lv, "A novel context-based implicit feature extracting method," in *Data Science and Advanced Analytics (DSAA), 2014 International Conference on*, 2014, pp. 420–424.

[7] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguist.*, vol. 16, no. 1, pp. 22–29, 1990.

[8] C. Henríquez and E. Buelvas, "AspectSA: Unsupervised system for aspect based sentiment analysis in Spanish," *Rev. Prospect.*, vol. 17, no. 1, pp. 87–95, 2019.

[9] M. Pontiki *et al.*, "SemEval-2016 Task 5: Aspect Based Sentiment Analysis," in *Semeval*, 2016, pp. 19–30.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Paralinguistic resources as clues to the disambiguation of meaning

Raquel Freitag, Paloma Cardoso & Bruno Pinheiro

[rkofreitag@uol.com.br | paloma-batistacardoso@hotmail.com | bpinnheiro@hotmail.com]

Federal University of Sergipe, Brazil

In natural languages, speakers can understand if a certain sentence is grammatical and if the propositional content is verifiable in the real world. In the interactional situation, the context and the knowledge shared between speakers are used to solve ambiguity. This is what happens in (1) and (2). In these examples, the same surface structure of construction can be related to more than one meaning (a and b), regarding polysemy.

(1) Acho que vai chover hoje. (I think [that] it is gonna rain today)
   a. 'tenho certeza' (I am sure)
   b 'pode ser' (It could be)

(2) Que menino bonitinho! (He is a cutie!)
   a. 'carinhoso' (affective)
   b. 'pejorativo' (pejorative)

Both in American English (Thompson & Mulac, 1991), and in Brazilian Portuguese constructions with cognitive verbs such as "acho que" (I think that) and "creio que" (I believe that) are reanalyzed into parenthetical constructions, with the pragmatical function of uncertainty and opinion/beliefs markers (Galvão, 1999, Freitag, 2003, Gonçalves, 2003). Uncertainty and opinion/beliefs have different meanings associated with the epistemic modality domain, and the ambiguity caused by the polysemy of "acho que" and "creio que" is resolved in interactional context by other contextual clues.

In the same direction, diminutive morphology in Brazilian Portuguese codifies values other than only size: it can be used in an appreciative and affective way (positive judgment) or in a depreciative and pejorative way (negative judgment) (Rio-Torto, 1993, Basílio, 1987, Bechara, 2009, Carvalho, 2009). In both cases, the different meanings of these constructions are selected in the interactional context, supposed by the shared knowledge and other contextual clues that cannot be systematized. In this paper, we claim that the disambiguation process is possible because of the paralinguistic resources associated with the propositional content.

Paralinguistic resources are elements that establish, maintain, and regulate the interaction between speaker/hearer by sound and visual characteristics. They work beyond linguistic structure but are important to communication because it disambiguates meanings (Traunmuller, 2000). Corrections, pauses, intonation, gestures, facial expressions, pupil dilatation, etc. are examples of paralinguistic resources. From a prosodic approach, uncertainty and opinion/beliefs can be differentiated by intonation, specifically by the fundamental frequency curve (F0): in uncertainty it is upward and in opinion/beliefs it is downward (Antunes, Aubergé, 2015). Besides that, when uncertain, the speaker tends to turn his head away, make pauses during speech and curve the eyebrows and lip lines, which configures doubt expression. (Antunes, Aubergé, Sasa, 2014).

In this study, facial expressions associated with the occurrence of the linguistic constructions are analyzed based on the premise that emotions have a physiological correlation with the facial expressions (an appraisal), and these are important clues to differentiate meanings because it indicates judgments about what is said. Some meanings are more sensitive to emotions, as judgments related to diminutive morphology use, while others are not, as meanings of uncertainty and opinion/beliefs.

The focus of the analysis is to identify facial expressions as a paralinguistic resource associated with the propositional content of parenthetic constructions “acho que” and “creio que”, and diminutive morphology (x-inho e x-zinho) to differentiate its meanings. The dataset was extracted from 20 sociolinguistics interviews documented by audio and video with university students. The occurrences of the target constructions were cut and submitted to the facial expression recognition process (Van Gent, 2016). The participant’s face features were detected by the algorithm ‘Haar Classificator’ (Viola & Jones, 2001).

This algorithm extracts 68 points for the face, located on the eyebrow line, mouth, nose, eyes, and chin. Its content was developed to recognize expressions of happiness, disgust, neutral, fear, and surprise. Scikit-Learn, a function Support Vector Machines (Vapnki, 2015), was applied: from every image cut, it identifies which emotions participants are probably expressing. By comparing the meanings associated with the context and the identified emotions, we can measure the effect of paralinguistic clues on the disambiguation of the “acho que” and “creio que”, and diminutive morphology (x-inho e x-zinho), thus resolving the polysemy.

References

Antunes, L. B., Aubergé, V. (2015) Análise prosódica da certeza e da incerteza em fala espontânea e atuada. *Diadorim*, 17(2), pp. 212-237.

Antunes, L. B., Aubergé, V.,Sasa Y. (2014) Certainty and uncertainty in Brazilian Portuguese: methodology of spontaneous corpus collection and data analysis. *Proceedings of the 7th Speech Prosody.* Dublin, Trinity College, pp. 110-114.

Basílio, M. (1987) *Teoria Lexical*. São Paulo, Ática.

Bechara, E. (2009) *Moderna Gramática Portuguesa*. Rio de Janeiro, Nova Fronteira.

Carvalho, M. C. G. (2009) *Sistematização funcional dos sufixos avaliativos no português do Brasil*. Dissertação Mestrado em Letras, Pontifícia Universidade Católica do Rio de Janeiro

Freitag, M. M. K. (2003) O papel da frequência de uso na gramaticalização de acho (que) e parece (que) marcadores de dúvida na fala de Florianópolis. *Veredas-Revista de Estudos Linguísticos*, 7(1-2), pp. 113-132.

Freitag, M. K., Cardoso, P. B., Pinheiro, B. F. M. (2020). Acho que é uma gripezinha: construções linguísticas como pistas de atitudes em tempos de pandemia. *Linguasagem* 35(esp), pp. 31-49.

Galvão, V. C. C. (1999). O achar no portugues no Brasil: um caso de gramaticalização. Dissertação de mestrado em Linguística, Universidade Estadual de Campinas.

Gonçalves, S. C. L. (2003). G*ramaticalização, modalidade epistêmica e evidencialidade*: um estudo de caso no português do Brasil. Doutorado em Linguística, Universidade Estadual de Campinas.

Rio-Torto, M. (1993) *Formação de palavras em português:* aspectos da construção de avaliativos. Tese de Doutoramento em Letras, Universidade de Coimbra

Van Gent, P. *Emotion recognition using facial landmarks*: a tech blog about fun things with Python and embedded electronics. <http://www.paulvangent.com/2016/08/05/emotion-recognition-using-facial-landmarks/>

Vapnik, V. N. (1995) *The nature of Statistical learning theory*. New York, Springer-Verlag.

Viola, P., Jones, M. (2001) Rapid object detection using boosted cascade of simple features. *In*: 5th Conference on Computer Vision and Pattern Recognition. *Proceeding* [...]. pp.1-9.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Una propuesta metodológica multidisciplinar para la identificación semiautomática de las metáforas

Virginia Mattioli[37]
[virginia.mattioli@pucv.cl]

Palabras clave: contexto distribucional; lingüística de corpus; lingüística computacional; teoría cognitiva de las metáforas; metáforas.

Resumen: En esta presentación se avanza una propuesta metodológica para la identificación sistemática y semiautomática de las metáforas que presenten un dominio conceptual determinado en un conjunto de textos en lengua española. Tras realizar una revisión de numerosas propuestas previas (Wilks, 1978, Fass, 1991, Mason, 2004, Gedigian *et al.*, 2006, Shutova *et al.*, 2010, Gutiérrez *et al.*, 2016, entre muchos otros) se desprende que, a pesar de la aproximación al problema desde distintas perspectivas y metodologías, tanto manuales como automáticas, aún no se ha conseguido diseñar un método que reúna las características requeridas para la lengua española. Por consiguiente, a partir de los postulados teóricos de la semántica distribucional y de la lingüística cognitiva, empleando las herramientas ofrecidas por la lingüística computacional y contando con el amplio alcance proporcionado por la lingüística de corpus, se pensó diseñar y comprobar una metodología original.

Según las teorías de la lingüística cognitiva, las metáforas son expresiones que conectan dos ámbitos cognitivos diferentes, empleando términos de un dominio fuente para referirse a un dominio conceptual distinto (Lakoff, 1993, 206-207). Por ejemplo, en la frase *hay que reducir los precios para derrotar la competencia*, se emplea el término DERROTAR que pertenece al dominio fuente GUERRA para referirse al dominio conceptual MERCADO, invocando así la metáfora "EL MERCADO ES UNA GUERRA". A raíz de la presencia de dominios o campos semánticos distintos en el mismo contexto oracional, esta propuesta parte de la idea según la cual las metáforas se pueden identificar mediante el contexto distribucional de los términos que las componen (es decir, considerando el abanico de elementos que pueden aparecer en el mismo contexto de un cierto término, de acuerdo con Harris, 1954, 146). Efectivamente, siendo expresiones que relacionan dos dominios distintos, es lógico pensar en la ausencia de una relación semántica evidente entre los términos que las constituyen.

[37] Virginia Mattioli is a Translation Studies and Linguistics scholar. She received her BA from the Autonomous University of Barcelona (UAB) and took an MA and PhD at the University Jaume I, with a thesis on a corpus-based comparison of translated and travel novels as for foreign words and otherness acceptance. She worked as an Associate Professor in Translation Studies at the Pontifical Catholic University of Valparaíso (PUCV) and a Research Assistant for the University of Birmingham in the ERC-funded project "Law and Language at the European Courts of Justice". Her research interests include cultural and literary translation, corpus-based studies and sociolinguistics.

Partiendo de estas premisas y basándose en teorías propias de la semántica distribucional (Harris, 1954, Lenci, 2008, Martí Antonín, 2018, entre muchos otros) se pretende diseñar una metodología que se compone de dos fases principales. En la primera, se observa el contexto de las palabras relacionadas con el dominio conceptual de las metáforas que se quieren identificar para determinar su contexto distribucional habitual en términos sintácticos y semánticos. Para hacerlo se siguen los pasos descritos detalladamente a continuación y ejemplificados basándose en la metáfora "EL MERCADO ES UNA GUERRA".

- Se escoge un dominio conceptual del cual buscar las metáforas, por ejemplo, MERCADO.
- Se escogen palabras representativas de dicho ámbito, como "negocio", "compra" o "tienda".
- Por cada una de ellas se observa el contexto distribucional usual.
  Considerando la palabra "negocio", se desprende que dicha palabra suele coocurrir, por ejemplo, con términos como adjetivos como "legal", "ilegal", "justo" o verbos como "fomentar", "crecer" o "incrementar", etc. Efectivamente, dichos adjetivos y verbos, que pertenecen al mismo campo semántico de la palabra "negocio", forman parte de su contexto distribucional usual como demuestra la frecuencia de aparición de expresiones como "comercio ilegal", "incrementar el comercio legal" o "fomentar el comercio".
  Este paso se repite por cada una de las palabras relacionadas con el dominio conceptual escogido ("negocio", "compra" y "tienda" relacionadas con el dominio conceptual MERCADO).
- Se agrupan los términos que componen el contexto distribucional de las palabras relacionadas con el dominio conceptual escogido según sus rasgos sintácticos y semánticos.
  De este modo, los términos del contexto distribucional de la palabra "negocio" que se identificaron, se reunirían bajo los grupos "adjetivos relacionados con el sistema jurídico" ("legal", "ilegal", "justo", etc.) y "verbos que indican crecimiento" ("fomentar", "crecer", "incrementar", etc.).

En la segunda fase, se implementaría un sistema de *scripts* para seleccionar de forma semiautomatizada los términos que presentan un contexto distribucional distinto a aquel determinado en la fase previa, es decir que no forman parte de ninguno de los grupos de términos creados porque no suelen coocurrir con las palabras del dominio conceptual escogido ya que, de acuerdo con la semántica distribucional, pertenecen a dominios o campos semánticos distintos y, por lo tanto, mantienen con estas una relación metafórica. Siguiendo con en el ejemplo anterior, se procedería de acuerdo con los siguientes pasos:

- Se implementa un sistema de *script* que elimine de forma semiautomática todas las ocurrencias de "adjetivos relacionados con el sistema jurídico" y "verbos que indican crecimiento".
- Tras eliminar todas dichas ocurrencias, se quedan aquellas que incluyen términos de un dominio completamente distinto a lo de MERCADO, como los verbos "florecer" o "brotar" que pertenecen al dominio NATURALEZA. Efectivamente, en expresiones como "el negocio florece" o "brota un nuevo comercio", ambos verbos se emplean de forma metafórica.

Para identificar de la forma más detallada posible el contexto distribucional de cada término relacionado con el dominio conceptual escogido, disminuir el ruido en los resultados y corregir otros posibles errores, durante el proceso de diseño de la metodología, esta se va aplicando a un corpus de dimensiones reducidas. De este modo, se procede de manera circular, alternando el diseño de la metodología con su aplicación al corpus, para mejorar gradualmente la propuesta inicial hasta obtener un método definitivo cuyo funcionamiento será comprobado mediante su aplicación a un corpus de dimensiones mayores.

A pesar de que el estudio está todavía en progreso, se considera que los resultados obtenidos serán útiles tanto para investigaciones futuras como para mejorar el estado actual de cada una de las disciplinas en las que se basa esta propuesta, tanto desde la perspectiva conceptual, como metodológica. Así, por un lado, los resultados colaborarán con el mejoramiento del estado de cada una de las disciplinas en las que se basa esta propuesta. En este sentido, se mostraría un ejemplo concreto de la posibilidad de aplicación de la lingüística de corpus a campos y conceptos hasta ahora raramente examinados con esta metodología, como el estudio de las metáforas; se incrementaría el número de investigaciones enfocadas en metáforas en textos españoles; y se ofrecerían algoritmos reutilizables en estudios futuros, ampliando las posibilidades de aplicación de la lingüística computacional. Por otro lado, la metodología obtenida será una herramienta útil para la identificación semiautomática de las metáforas en estudios futuros de todas las disciplinas que investigan la naturaleza y las implicaciones de las expresiones metafóricas (como la lingüística cognitiva, la semántica y la semiótica, entre otras).

Bibliografía:

Fass, D. (1991). met*: A method for discriminating metonymy and metaphor by computer. Computational Linguistics, 17(1), pp. 49-90.

Gedigian, M., Bryant, J., Narayanan, S. y Ciric, B. (2006, junio). Catching metaphors. En Proceedings of the Third Workshop on Scalable Natural Language Understanding, pp. 41-48. Association for Computational Linguistics.

Gutierrez, E. D., Shutova, E., Marghetis, T. y Bergen, B. (2016, agosto). Literal and metaphorical senses in compositional distributional semantic models. En Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 183-193.

Harris, Z. (1954). Distributional structure. Word, 10(23), pp. 146-162.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), Metaphor and Thought (pp. 202-251). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139173865.013

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. Italian journal of linguistics, 20(1), pp. 1-31.Martí Antonín, X. (2018).

Mason, Z. J. (2004). CorMet: a computational, corpus-based conventional metaphor extraction system. Computational linguistics, 30(1), pp. 23-44.Shoutova et al, X. (2010).

Wilks, Y. (1978). Making preferences more active. Artificial Intelligence, 11(3), pp. 197–223.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

UnderRL Tagger: Concepción y elaboración de un sistema de etiquetado semiautomático para Under-Resourced Languages

José Luis Pemberty Tamayo[38] **& Jorge Mauricio Molina Mejía**[39]
[jose.pemberty@udea.edu.co | jorge.molina@udea.edu.co]
Universidad de Antioquia, Colombia

Introducción

Existe un factor que se ha ido haciendo cada vez más importante, y este tiene que ver con el hecho de poder contar con recursos lingüísticos en diversas lenguas a nivel local y mundial, esto con el fin de proveer información fiable de todas estas diversas lenguas. En este sentido, la lingüística de corpus y la lingüística computacional han ido cobrando gran importancia en la comprensión y el estudio de las lenguas de forma general (Brezina, 2018; Mitkov, 2004; Parodi, 2010; Tognini-Bonelli, 2001).

Teniendo en cuenta la gran diversidad lingüística de Colombia (González & Rodríguez, 2010) y de otros países de Latinoamérica y del mundo, esto hace que sea difícil la tarea de crear recursos para el procesamiento automático en muchos casos (Besacier *et al.*, 2014; Maxwell & Hughes, 2006). La presente propuesta tiene por objetivo comentar los pasos de realización y mostrar el funcionamiento de la herramienta *UnderRL Tagger*, un etiquetador semiautomático a nivel de POS para *Under-Resourced Languages* que ha sido desarrollado en el seno del semillero de investigación Corpus Ex Machina de la Universidad de Antioquia en Medellín (Colombia).

Aspectos teóricos

El concepto de *Under-Resourced Languages* (U-RL) se toma de los trabajos de Krauwer (2003) y Berment (2004) y se define, principalmente, como una lengua que no cuenta con un equipo básico de procesamiento automático del lenguaje, así como con la disponibilidad de corpus orales y escritos debidamente etiquetados. Es lo que se podría denominar, en otras palabras, como

[38] Estudiante de último semestre del pregrado Letras: Filología Hispánica de la Facultad de Comunicaciones de la Universidad de Antioquia. Ha desarrollado su tarea investigativa en el semillero "Corpus Ex Machina" en las áreas de la lingüística computacional, el Procesamiento del Lenguaje Natural, la lingüística de corpus y la enseñanza/aprendizaje de lenguas asistidos por computador.

[39] Profesor de cátedra del área de lingüística, coordinador del Grupo de Estudios Sociolingüísticos y del semillero "Corpus Ex Machina", ambos adscritos a la Facultad de Comunicaciones y Filología de la Universidad de Antioquia. PhD en Informática y Ciencias del Lenguaje (Université Grenoble Alpes, Francia), con Maestría y Especialización en Ciencias del Lenguaje (Université Stendhal - Grenoble 3, Francia) y una Licenciatura en Enseñanza de Lenguas Extranjeras (Universidad de Antioquia, Colombia). Desarrolla sus tareas de docencia e investigación en las áreas de la Lingüística Computacional, la Lingüística de Corpus, la Lingüística Textual y la Enseñanza/Aprendizaje de Lenguas Asistidos por Computador (ELAO/ALAO).

"lenguas minoritarias poco dotadas" o "poco preparadas" (concepto que, infortunadamente, no se encuentra dentro de la literatura especializada en lengua española); es por esto, que se ha tomado la decisión de mantener el nombre de "*under-resourced languages*", mucho más cercano al concepto que se trata de buscar en este trabajo. Esto tiene que ver con el hecho de que una posible traducción al español no parece muy acertada (Pemberty, 2020, p.17-19). Por otra parte, es el nombre que comúnmente se le suele dar, incluso en autores de origen francés, como es el caso de Besacier *et al.* (2014).

Aspectos metodológicos

Para contribuir a subsanar la dificultad en el grupo de lenguas U-RL y teniendo en cuenta que el acceso a las tecnologías del lenguaje es primordial para la inclusión de los hablantes en el mundo globalizado y para la democratización de la información (Krauwer, 2003, p.4), el proyecto del *UnderRL Tagger* busca utilizar un sistema básico de aprendizaje de máquinas, en que se introduzcan manualmente diferentes categorías de etiquetado POS que la computadora utilizará para automatizar posteriormente el proceso de etiquetado (figura 1).

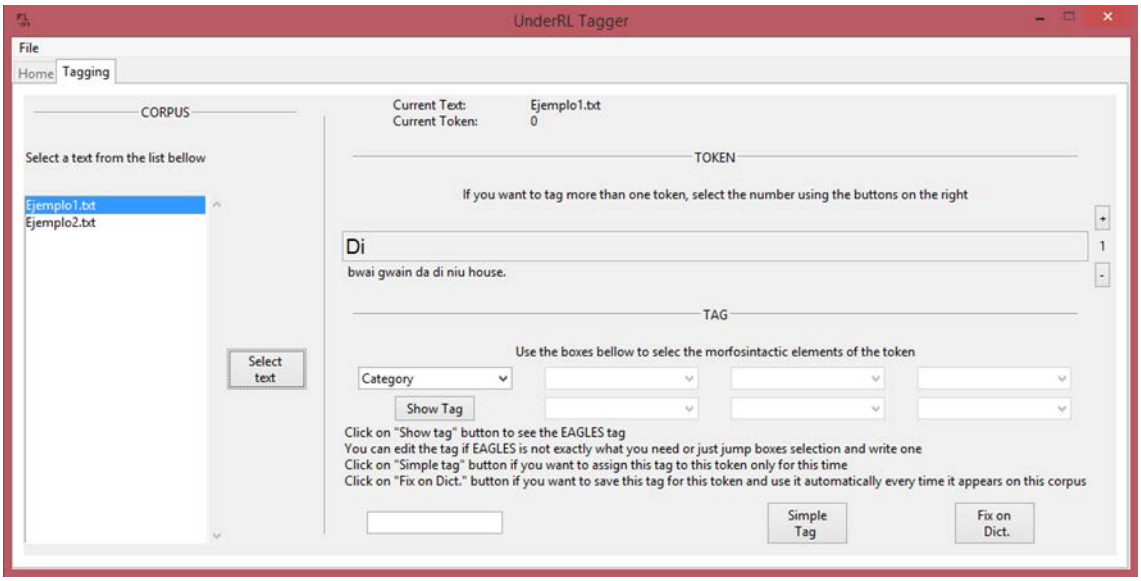


Figura 1. *UnderRL Tagger*, pantalla de etiquetado (Pemberty, 2020, p.32).

Este sistema está enfocado a las lenguas que ya cuentan con una producción escrita susceptible de ser digitalizada en el alfabeto normalmente accesible a los equipos informáticos en codificación de tipo UTF-8 y a los investigadores que se encuentren frente a la tarea de construir con esos textos un corpus etiquetado para futuras tareas de automatización e integración de la lengua en las tecnologías del lenguaje.

Para lograr este objetivo, se utiliza un sistema que presenta los textos en forma de tokens y permite asignar a cada unidad una etiqueta basada en el estándar EAGLES, con la opción de guardar esa etiqueta en un archivo de diccionario, que será posteriormente revisado cada vez

que el programa se encuentre con el mismo token, permitiendo reutilizar la etiqueta sin intervención del usuario.

Como puede deducirse de este proceso, la automatización será cada vez mayor en la medida en que se cuente con más información de los *tokens*, por lo que la asistencia del usuario será necesaria principalmente en fases iniciales del etiquetado.

Con el fin de obtener la herramienta de etiquetado aquí presentada, se parte de tres grandes fases metodológicas:

- **Proceso de modelado del sistema:** En esta etapa se tuvieron en cuenta los diferentes pasos que se deseaba que el sistema pudiera ejecutar, como es el caso de la recepción de textos en formato *.txt, y del tipo de formato UTF-8, acceso al sistema para lingüistas sin mayores conocimientos de informática, etc.
- **Diseño y ejecución de algoritmos:** Se diseñaron varios algoritmos que permiten la ejecución de los procesos de etiquetado, creación de un diccionario que guarda en el sistema las etiquetas creadas por el usuario, posibilidad de crear nuevas etiquetas, etc. Estos algoritmos han sido creados a partir del lenguaje *Python* y se encuentran descritos en Pemberty (2020) (véase, a manera de ejemplo, la figura 2).
- **Programación de *UnderRL Tagger*:** Finalmente, se procedió a efectuar la programación informática de la herramienta, teniendo en cuenta las dos fases anteriores. Para ello, se empleó el lenguaje de programación *Python*, esto debido a la facilidad de desarrollo que permite este tipo de lenguaje. Lo cual puede redundar en futuros desarrollos o mejoras que se puedan proponer desde otros equipos de investigación.

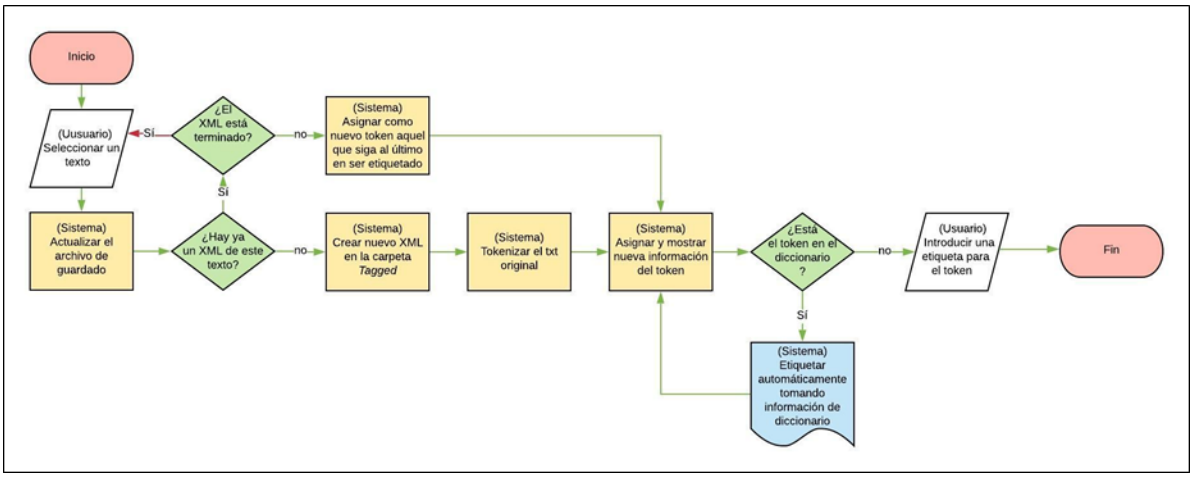


Figura 2. Procesamiento previo de un texto seleccionado (Pemberty, 2020, p.34).

Discusión

Durante la ponencia se explicarán tanto los detalles técnicos correspondientes al funcionamiento de la aplicación, como los ejemplos de uso y posibles aplicaciones en proyectos reales, lo que será útil para otros asistentes que tengan contacto con las U-RL en su campo de trabajo. Así mismo serán introducidos los elementos teóricos sobre lingüística de corpus, lingüística

computacional y etiquetado de corpus que han tenido importancia en el desarrollo del proyecto de investigación aquí esbozado.

Conclusiones

Finalmente, tenemos una herramienta que ha sido diseñada dentro de un trabajo de grado de pregrado que es funcional y que prontamente se pondrá a disposición de los investigadores de U-RL tanto de Colombia como de otros países, la idea es que esta herramienta sea de libre uso. Y que, más adelante, el desarrollo pueda mejorarse con la colaboración de investigadores de diferentes laboratorios y equipos de investigación.

Referencias

Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues « peu dotées »*, Tesis doctoral, Université Joseph-Fourier - Grenoble I. https://tel.archives-ou-vertes.fr/tel-00006313

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56, 85–100.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

González, M. S., & Rodríguez, M. L. (2010). *Lenguas indígenas de Colombia: Una visión descriptiva*. Instituto Caro y Cuervo.

Krauwer, S. (2003). The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. *Proceedings of SPECOM* 2003, 8–15.

Maxwell, M., & Hughes, B. (2006). Frontiers in linguistic annotation for lower-density languages. *Proceedings of the workshop on frontiers in linguistically annotated corpora* 2006, 29–37.

Mitkov, R. (2004). *The Oxford Handbook of Computational Linguistics*. OUP Oxford.

Parodi, G. (2010). *Lingüística de corpus: De la teoría a la empiria*. Iberoamericana.

Pemberty, J. L. (2020). *Concepción y elaboración de un sistema de etiquetado semiautomático para Under-Resourced Languages*. Trabajo de grado, Universidad de Antioquia.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins Publishing Company.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Use of Bayesian networks for the analysis of corpus of local problems related to the Sustainable Development Goals

Ernesto Llerena García[40] y Manuel Caro Piñeres[41]
[lle55re19@yahoo.com | manuelcaro@correo.unicordoba.edu.co]
Universidad de Antioquia, Colombia

Abstract

Bayesian networks are a widely used formalism for data analysis, modeling, and decision support in various domains. Currently, there is a need for techniques and tools that automatically build Bayesian networks from massive text or literature data. Collecting people's perception of the problems they face in their daily lives generates a great deal of textual information. Textual descriptions increase as new data collections are made. Due to the lexical differences between different regions of a country, it is necessary to constantly update the new modeled data.

Key words: Bayesian networks, United Nations, Sustainable Development, collecting data.

Introduction

The 17 Sustainable Development Goals (SDGs) are a plan of United Nations to achieve a better and more sustainable future for people and the planet by 2030. In these goals there are aspects related to poverty, hunger, good health and well being, quality education, clean water, clean energy among others. With just under ten years left to achieve the Sustainable Development Goals, world leaders at the SDG Summit in September 2019 called for a Decade of Action and delivery for sustainable development, and pledged to mobilize financing, enhance national implementation and strengthen institutions to achieve the Goals by the target date of 2030, leaving no one behind.

Thus, it was necessary the use of reliable technology for understanding people's needs all around the world, and during this decade achieve the 17 Sustainable Development Goals (SDGs) lead by United Nations. In that way, Bayesian network was used for collecting data

[40] Profesional en Idiomas por la Universidad de Antioquia. Magíster en etnolingüística por la Universidad de los Andes. Doctor en Lingüística por la Universidad de Antioquia. Profesor investigador grupo EduTLan, Universidad de Córdoba, Líneas de investigación: lexicografía, etnolingüística, sociolingüística y lingüística cognitiva. Productos académicos: Diccionario etnolingüístico de la lengua Ẽbẽra del Alto Sinú, El habla popular de los valles de los ríos Sinú y San Jorge (2011), Etnoliteratura digital (2016), Relatos ilustrados de la tradición indígena y afrocolombiana (2018). Actualmente sus temas de investigación se centran en jergas, lingüística cognitiva, etnoeducación y en el aprendizaje del español como segunda lengua.

[41] Licenciatura en Informática y Medios Audiovisuales. Máster en Nuevas Tecnologías aplicadas a la Educación, Universidad de Alicante (España). Doctorado Ingeniería Universidad Nacional De Colombia Sede Medellín. Profesor de planta en la Universidad de Córdoba. Profesor investigador y coordinador del grupo EduTLan, Universidad de Córdoba.

through a software created by EdutLan group which help to gather and analyze all the information needed to reach these goals.

Literature review

Bayesian networks are used for modeling knowledge in computational biology and bioinformatics, learning, medicine, biomonitoring, document classification, information retrieval, semantic search, image processing, data fusion, decision support systems, engineering, games and law.

Problem description

For decision-making at the governance level, it is necessary to know how non- compliance with the SDGs affects the well-being of the population. However, the SDGs are little known by the general population, so it is necessary to have techniques that can relate people's speech in relation to the language of the SDGs. To fulfill this purpose, it is necessary to collect a large number of descriptions of problems related to the SDGs in the communities.

Goal

The mail goal of this study is to describe the process of collecting, organizing, tagging and validating a corpus of more than 3,000 descriptions of problems related to compliance of the SDGs in three regions in Colombia.

Method

The descriptions were collected verbally for three years and contain regionalisms related to the SDGs from the Caribbean region, Antioquia and Bogotá. These corpora were taken through oral interviews with people (men and women) from diverse social level (mainly 1, 2, 3 social levels). The interviewer recorded the interview with a cell phone and instantly or when a wifi connection was able, all the information was gathered and analyzed. Thus, the system shows how people think about their necessities related to United Nations´s goals. This information will be used to promote prosperity while protecting the planet. United Nations recognizes that ending poverty must go hand-in-hand with strategies that build economic growth and address a range of social needs including education, health, social protection, and job opportunities, while tackling climate change and environmental protection.

Results

The main result of this study is a large digital corpus of descriptions of problems related to compliance of the SDGs in three regions in Colombia. The potential of the corpus was verified by evaluating the results of a Bayesian network algorithm. In the evaluation, the standard processing of the text by the algorithm produces a high rate of correct answers.

Bibliography

Naciones Unidas (2019). Informe de los objetivos de desarrollo sostenible. Nueva York: Naciones Unidas.

Devi, S., A.; Priya, M., V.; Akhila, P.; (2018). Analysis and prediction of student placement for improving the education standards. pp. 303-306.

Kutela, B., and Teng, H. (2019). Prediction of drivers and pedestrians' behaviors at signalized mid-block Danish offset crosswalks using Bayesian networks. Journal of Safety Research. pp.75-83.

Sandri, M.; Berchialla, P.; Baldi, I.; Gregori, D.; and De Blasi, R., A.(2014). Dynamic Bayesian Networks to predict sequences of organ failures in patients admitted to ICU. Journal of biomedical informatics. pp. 106-13.

Lingüística de Corpus

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

3DCOR: Creación de un glosario bilingüe (inglés-español) basado en corpus para la traducción de fichas técnicas de impresoras 3D

Angela Luque Giraldez[42] **& Míriam Seghiri Domínguez**[43]

[angelaluquegiraldez@gmail.com | seghiri@uma.es]

Universidad de Málaga, España

Palabras clave

Corpus virtual, traducción técnica, ficha técnica, impresoras 3D, glosario.

Resumen

En el mercado laboral de la traducción, la traducción técnica es la especialidad que mayor demanda genera (Bruno *et al.*, 2016). En esta misma línea, Rico y García (2016) coinciden en que las traducciones más demandadas pertenecen al área de la tecnología, que supone entre el 81 % y el 100 % de la facturación anual en el 31 % de los casos encuestados.

Por ello, podemos deducir que la traducción técnica es uno de los campos más demandados en el mercado profesional de la traducción, en particular, la traducción en el área de la tecnología. De hecho, es precisamente en esta área en la que recae nuestro interés.

Así, en los últimos tiempos, en el ámbito de la tecnología han irrumpido con fuerza las impresoras 3D. La impresión 3D es una de las tecnologías más novedosas y que ha avanzado a mayor velocidad hasta llegar al punto de poder construir impresoras que se adaptan a todo tipo de bolsillo para que cualquier usuario las pueda adquirir. Según estudios de TMT Deloitte (2019), la industria de la impresión 3D está creciendo a un ritmo aproximado del 12,5 % cada año y, además, pudo estimar que en 2019 las ventas superarían los 2 700 millones de dólares, como así ha sido, y prevé que en 2020 superarán los 3 000 millones.

En este contexto de alta demanda es donde nace la necesidad de traducir documentos relacionados con la venta de impresoras 3D —como pueden ser manuales de instrucciones o

[42] Estudiante en el programa de Doctorado de Literatura, Lengua y Traducción de la Universidad de Málaga. Titulada en el Máster de Traducción Especializada por la Universidad de Córdoba y en el Grado Traducción e Interpretación por la Universidad de Málaga. Sus principales líneas de investigación abarcan la traducción especializada, la lingüística de corpus y la aplicación de nuevas tecnologías a la traducción.

[43] Licenciada y Doctora en Traducción e Interpretación y Graduada en Derecho. En la actualidad es profesora titular en el Departamento de Traducción e Interpretación de la Universidad de Málaga. Además, ha impartido docencia en la Universidad norteamericana de Dickinson College y en la Universidad británica de Cambridge, así como en las universidades españolas de Córdoba y Murcia. Cuenta con una patente (N-Cor). Sus líneas de investigación abarcan la traducción especializada, la lingüística de corpus y las tecnologías de la traducción y la interpretación (Premio extraordinario de doctorado, Premio en Tecnologías de la Traducción, Premio María Zambrano y Premio George Campbell).

fichas técnicas, por ejemplo—, pues cada vez son más las empresas tecnológicas que desean comercializarlas por el mundo. Sin embargo, la reciente irrupción en el mercado de este producto hace que escaseen los recursos terminológicos en torno a él de los que pueda hacer uso el traductor.

Ante la carencia de recursos, al traductor, por lo general, no le queda otro remedio que construírselos él mismo. Llegados a este punto, y como base para la construcción de recursos terminológicos de calidad para el traductor, no cabe duda de que el uso de corpus textuales supone un gran aliado (Seghiri, 2016 y 2017a).

Así, surge el objetivo principal del este trabajo, a saber, la creación de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus que sea de utilidad para la traducción fichas técnicas de impresoras 3D.

La extracción de los términos que integrarán el mencionado glosario se realizará a partir de un corpus virtual, representativo desde el puno de vista tanto cualitativo como cuantitativo, creado a tal efecto, al que denominaremos 3DCOR. De este modo, en el presente trabajo, aunaremos el formato preferido por los traductores (y en particular los noveles) durante la fase documental, como es el glosario (Corpas, 2015; Picton y Mathilde, 2015; y Frérot, 2015), pero cuya implementación se basará en el recurso ideal para los investigadores, como es el corpus (Seghiri, 2015).

Para alcanzar el objetivo principal de este trabajo, se han propuesto los siguientes objetivos secundarios:

1. Delimitación del objeto de estudio, la ficha técnica, dentro del campo de la traducción técnica.
2. Creación de un corpus virtual de fichas técnicas de impresoras 3D, representativo, fiable y de calidad, al que se le denominará 3DCOR.
3. Estudio y extracción terminológica del corpus a través de SketchEngine.
4. Creación del glosario bilingüe y bidireccional para la traducción de fichas técnicas.

Con fin de cumplir los objetivos planteados, se proponen las siguientes actividades:
Para cumplir el objetivo n.º 1, estudiaremos el género textual de la ficha técnica, sus características y descripción en el marco de la traducción técnica, así como la demanda de este en el mercado de la traducción.

Más tarde, para abarcar el objetivo n.º 2, crearemos un corpus virtual formado por bitextos (inglés-español) de fichas técnicas de impresoras 3D. Para ello, seguiremos una metodología propuesta por Seghiri (2007a y 2017b), que se divide en dos fases bien diferenciadas: en primer lugar, aquella destinada a asegurar la representatividad cualitativa de la muestra compilada y, en segundo lugar, se encuentra la fase de aseguramiento de la representatividad cuantitativa. Así, en la primera fase, se establecerá el diseño del corpus y se seguirá un protocolo de compilación. Este protocolo de compilación se iniciará con un proceso de localización de texto en páginas web especializadas en venta y/o producción de impresoras 3D, seguido de la descarga automática de estos utilizando herramientas específicas. Estos textos se transformarán a formato ASCII con la ayuda de programas como Abbyy FineReader, para, posteriormente, ser

almacenados en una estructura de carpetas que se creará a tal efecto con una codificación ideada para facilitar su manipulación. Tras seguir estos pasos, procedemos a abordar la segunda fase, para comprobar la representatividad cuantitativa de la muestra compilada. Para ello, nos serviremos del programa ReCor[44].

Una vez complicado nuestro corpus, seguiremos con el objetivo n.º 3 a través de la explotación del corpus con el programa Sketch Engine con el que podremos analizar y estudiar el corpus compilado, así como compararlo con un corpus de referencia. Todo ello, con objeto de llevar a cabo una extracción terminológica.

Finalmente, se abordará el objetivo n.º 4 para el que exportaremos una lista de los términos más empleados en inglés y sus equivalentes en español con el fin de generar un glosario bilingüe y bidireccional (inglés-español/español-inglés) orientado a la traducción de fichas técnicas de impresoras 3D.

El corpus y glosario resultantes de este estudio, no solo podrás ser útiles para la traducción de fichas técnicas de impresoras 3D, sino para abordar la traducción de géneros análogos (como los manuales de instrucciones) o temas (impresoras láser) e, incluso, para su utilización por parte de otros profesionales como intérpretes o ingenieros desde un punto de vista tanto monolingüe como multilingüe.

Bibliografía

Bruno, L. V.; Luque, I.; Ferreyra, L. E. (2016). La traducción de textos técnicos (Tesis). Universidad Nacional de Córdoba.

Corpas Pastor, G. y Seghiri, M. (2007). Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor. Procesamiento del Lenguaje Natural, 39, 165-172.

Corpas Pastor, G. (2015). Corpora in Translation and Interpreting: a Means to all Ends. Trabajo presentado en el IV Congreso Internacional CULT (Corpus Use and Learning toTranslate), Alicante.

Deloitte Technology, Media & Telecommunications (2019). Predicciones TMT 2019. Recuperado de https://www2.deloitte.com/content/dam/Deloitte/ec/Documents/technology-mediatelecommunications/Deloitte-ES-TMT-Trends-2019-Folleto.pdf

Frérot, C. (2015). Major Achievements and New Perspectives in Using Corpora and Corpus Technology for Translation Purposes. Trabajo presentado en el IV Congreso Internacional CULT (Corpus Use and Learning to Translate), Alicante.

[44] Esta herramienta fue diseñada por Corpas Pastor y Seghiri (2007), por la que recibieron el Premio de Tecnología de la Traducción de España en 2008. ReCor sirve para determinar el tamaño mínimo de un corpus indicando a través de gráficas si se ha cubierto la terminología básica, es decir, si hay un momento en el que no entran palabras nuevas (*types*) al corpus.

Picton, A. y Mathilde, F. (2015). Corpora in Translation: addressing the Gap between the Scholars' and the Translators' Point of View. Trabajo presentado en el IV Congreso Internacional CULT (Corpus Use and Learning to Translate), Alicante.

Rico Pérez, C. Y García Aragón, A. (2016). Análisis del sector de la traducción en España (2014-2015) (Tesis doctoral). Universidad Europea.

Seghiri, M. (2015): Determinación de la representatividad cuantitativa de un corpus ad hoc bilingüe (inglés-español) de manuales de instrucciones generales de lectores electrónicosb/ Establishing the quantitative representativiness of an E-Reader User's Guide ad hoc corpus (English-Spanish), en María Teresa Sánchez Nieto (Eds.). Corpus-based Translation and Interpreting Studies: From description to application (pp. 125- 146). Berlín: Frank & Timme.

Seghiri, M. (2016). Diseño de una plantilla electrónica de evaluación de sedes web científicas para la creación de recursos de enseñanza-aprendizaje (alemán-inglés-español) / Designing an evaluation template of scientific web sites for the implementation of teaching and learning materials (German-English-Spanish). Educatio siglo XXI, 34 (3), pp. 9-26. Recuperado de http://revistas.um.es/educatio/article/view/275741/200401

Seghiri, M. (2017a). Corpus e interpretación biosanitaria: extracción terminológica basada en bitextos del campo de la Neurología para la fase documental del intérprete (Corpora and medical interpreting: terminology extraction based on bitexts for the interpreter's documentation process). Panacea, 18 (46). Volumen monográfico de interpretación en el ámbito biosanitario. pp. 123-132. Recuperado de http://www.tremedica.org/panacea/IndiceGeneral/n46_tribuna-MSeghiri.pdf

Seghiri, M. (2017b): Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores. Babel, 63 (1), 43-64.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

A corpus-based study of words constrained to orality: the case of Spanish loanwords in Falkland Islands English

Yliana Rodríguez
[ylianarodriguez@gmail.com]
Universidad de la República, Uruguay
Leiden University, Netherlands
Adolfo Elizaincín
[aelizain@gmail.com]
Universidad de la República, Uruguay
Paz González
[p.gonzalez@hum.leidenuniv.nl]
Leiden University, Netherlands

Since the English occupation in 1833, the Falkland Islands have been continuously inhabited by an English-speaking community, converting this variety into the youngest of the "Inner Circle" (Kachru, 1985). Although the most used language in the Islands is English, Spanish is the second most spoken language, according to the census that the government of the Islands carried out in 2016, and it was also spoken in the early days of the settlement in the 19th century when beef livestock farming was one of the economic engines of the Islands. Such business was mainly managed by gauchos from South America, who disappeared when the Islands were handed over to the sheep-raising industry, but their earlier presence is still remembered in the many place names of Spanish origin (Boumphrey, 1967), as well as in the lexicon of Falkland Islands English. An example of this is the word used in the Islands to refer to the rural area: “camp” (derived from the Spanish word “campo”) instead of the word "countryside" or a variant of it. In the past, Spanish names were applied both to horse tack and to the different types of horses found on the Islands (Boumphrey 1967). Not much has transpired in the scientific literature about the interference of these American actors in the history of the Islands. Much less about the linguistic inheritance they left in the English of the archipelago. Unfortunately, due to their register and context of use, these Spanish loanwords have remained mostly in orality. The very little corpora that exists of the Falklands variety (the local newspaper and books written by locals) rarely include these words. In our quest to know these words -and given the lack of corpora for our purpose- we decided to resort to texts in which these words are mentioned as part of travelers’ and locals’ observations on Falkland Islands speech.

The present work aims at presenting a novel methodological approach to the study of loanwords. In this particular case, this technique is used to unveil which Spanish words were borrowed into Falkland Islands English. The loanwords are analyzed in terms of their

occurrence, frequency, appearance in dictionaries and the semantic fields they have penetrated. Resorting to texts in which loanwords are mentioned has proven useful as a means of compensating for the lack of corpora in which these words appear. Even though the words are being used metalinguistically (they are mentioned by the authors but are not in actual use), we believe this technique has proven useful to the extent that it has allowed us to create a point of departure in terms of the groups of Spanish words that have been used in the Falklands before further research is carried out (e.g. fieldwork). We have attempted to account for the volume of words that Rioplatense Spanish-speaking gauchos would lend to Falkland Islands English. As already pointed out, in the scarce bibliography that mentions the existence of these loanwords, it is said that they belong to the equestrian world. However, we noted that they are not confined to such realm. We observed that Spanish loanwords are mainly -though not exclusively- related to horse tack and horse types: it is clear from our data that most words are tightly connected to gauchos argot and not exclusively with their equestrian duties. These findings coincide with linguistics pioneer Edward Sapir (1912)'s observation that the complete vocabulary of a language can be considered as an inventory of all the ideas, interests, and occupations that take up the attention of the community.

Keywords

Contact linguistics, corpus linguistics, loanwords, Rioplatense Spanish, Falkland Islands English

References

Boumphrey, R. S. (1967). *Place-Names Of The Falkland Islands*. The Falkland Islands Journal.

Kachru, B. (1985). Standards, codification and sociolinguistic realism: the English language in the Outer Circle. In Randolph Quirk, and Henry Widdowson (eds.), *English in the World: Teaching and Learning the Language and Literatures*, 11–30. Cambridge: Cambridge University Press.

Sapir, E. (1912). Language and environment. *American Anthropologist* 14: 226-242.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Análisis lingüístico de la evolución del humor en las niñas de Educación Primaria

Laura María Aliaga Aguza[45]
[aliagaaguza@gmail.com]
Universidad Internacional de La Rioja, España

Históricamente las mujeres no mostraban su humor en público, pues lo relegaban al ámbito privado; por este motivo, se pensaba que estas no tenían sentido del humor y no entendían los chistes (Galindo, 2017: 102). Actualmente, esta concepción está cambiando y se puede hablar de un humor típicamente masculino y típicamente femenino (Ruiz Gurillo, 2019). Esto puede ser debido a que los comportamientos entre hombres y mujeres difieren desde antes del nacimiento, ya que el género se empieza a desarrollar desde que los padres conocen el sexo del bebé (Jule, 2008) y conlleva que cada uno adopte roles diferentes a la hora de ver el mundo. Este hecho da lugar a que hombres y mujeres comprendan el mundo de forma diferente y lo utilicen condicionados por su género. En este sentido podemos señalar que existe un determinismo humorístico que hace que el ser humano genere humor a través de lo que conoce (Aliaga, 2014). En otras palabras, los aspectos tanto vitales como culturales determinan la identidad humorística de cada individuo.

La cuestión que se nos plantea ante esta situación es si existen divergencias de roles desde el inicio de la adquisición de este fenómeno metapragmático, alrededor de los seis años; y, por ende, diferencias en la producción y en la apreciación del humor, dado que a esta edad poseen un "menor nivel de 'contaminación' por las convenciones y prejuicios sociales" (Timofeeva, 2014: 197). Partimos de la hipótesis de que las niñas adquieren antes que los niños la conciencia tanto metalingüística como metapragmática y, en consecuencia, la metahumorística. No obstante, una vez que ambos géneros han alcanzado el nivel de maduración necesario, son los niños los que recurren más a menudo al humor en su día a día.

Para llevar a cabo este estudio manejaremos el corpus *CHILDHUM*, creado por el grupo GRIALE de la universidad de Alicante a través de varios proyectos de investigación vinculados (http://dfelg.ua.es/griale), donde se recogen narraciones humorísticas de tres grupos etarios en la primera fase de la educación obligatoria, concretamente, desde 4º hasta 6º de Primaria. Una etapa donde los niños "se encuentran en pleno proceso de adquisición de la conciencia metapragmática y su humor lingüístico evoluciona acorde al tránsito hacia un procesamiento lógico de la información" (Timofeeva, 2017: 11). Dicho corpus aún se encuentra en construcción.

[45] Laura Mª Aliaga Aguza es Doctora en Filología Hispánica por la Universidad de Alicante. Actualmente es PDI de la Universidad Internacional de La Rioja. Asimismo, es colaboradora honorífica de la Universidad de Alicante y forma parte del grupo GRIALE de dicha universidad. Sus líneas de investigación se centran en la Enseñanza de Español como Lengua Extranjera, la Didáctica de la Lengua y Literatura, la fraseología, el análisis del humor verbal y su aplicación a la clase de ELE, así como a aspectos relacionados con el género.

Actualmente consta de 448 redacciones humorísticas de tres grupos etarios (8, 10 y 12 años) de cinco colegios de la Comunidad Valenciana. La recogida de datos se ha elaborado en dos fases y se espera seguir con su ampliación a través de la Península. En la primera fase se obtuvieron 148 redacciones humorísticas de alumnos de 4º de Primaria en el curso escolar 2012 – 2013, etapa en la que "el niño ha alcanzado la madurez cómica siendo capaz de controlar sus impulsos y hacer reír a sus compañeros" (Aliaga, 2014: 225). La segunda fase se realizó dos cursos escolares posteriores (2014 – 2015) a estudiantes de 2º y 6º de Primaria de los mismos centros de la primera fase del estudio. De este modo, los discentes que compusieron la primera parte del corpus, ahora estarían en 6º. En esta fase se obtuvieron 140 redacciones de 8 años y 160 del grupo etario de 12.

El procedimiento que se llevó a cabo para la recogida de datos en ambas fases se realizó en varias etapas: en primer lugar, el grupo GRIALE se puso en contacto con los colegios seleccionados para el estudio y se estableció una fecha para la recogida de las redacciones. En segundo lugar, se procedió al etiquetado de las redacciones. De esta manera, cada redacción tiene un código para poder identificarlas. Primero se indica el grupo etario (8, 10, 12). A continuación, el año escolar en que fueron recogidas las muestras (12 – 13; 14-15) junto al colegio al que pertenece (A = Elda, B = Beneixama, I = Torrevieja, M = La Canyada, S = Monforte del Cid), posteriormente el sexo (A niña y O niño) y por último el número de redacción. De este modo, tendríamos que 10, 12 – 13 SA1, sería una redacción escrita por una niña de 10 años del colegio Divina Aurora en el curso escolar 2012 - 2013. En tercer lugar, se realizó una lectura y volcado de la información, con la que se elaboró la base de datos inicial, la cual, actualmente se está renovando.

Específicamente, en este análisis atenderemos a la evolución que experimentan las mujeres desde el inicio de la adquisición de la conciencia metapragmática (grupo etario de 8 años) hasta su consolidación al final de la educación primaria (grupo etario de 12 años). De esta manera, podremos comprobar el desarrollo de uno de los aspectos metapragmáticos más difíciles de dominar, como es la conciencia metahumorística. Asimismo, el humor desempeña una de las funciones sociales del lenguaje, esto es, establece vínculos entre las personas (Ojeda y Cruz, 2004) y permite la formación de la identidad y se muestra a través de la espontaneidad plasmada en el discurso planificado de las redacciones, lo que facilita la construcción de la identidad, bien manteniendo o reforzando jerarquías y estereotipos, o bien subvirtiéndolos (Ruiz Gurillo, 2019).

Estas muestras de humor se dan en forma de marcas, tanto lingüísticas como no lingüísticas, indicadores lingüísticos (Ruiz Gurillo & Padilla García, 2009) y / o estrategias propias (Aliaga, 2020), que se encuentran en los textos para guiar al oyente hacia una interpretación humorística. Esta capacidad metapragmática del sujeto "se va adquiriendo con los años y forma parte de la educación comunicativa del niño" (Timofeeva, 2017: 6), que poco a poco va logrando un control intencionado del mensaje a través de los parámetros lingüísticos que determinan la hilaridad del mensaje (Crespo y Alvarado, 2010).

Palabras clave: Humor, niñas, educación primaria

Bibliografía

ALIAGA AGUZA, L. (2014): "Evolución del humor en la mujer" en MURA, A. y RUIZ GURILLO, L. (2014): *Feminismo/s,* 24, *Género y humor en discursos de mujeres y hombres*, 221 – 242.

ALIAGA AGUZA, L. (2020): *Análisis lingüístico del humor en el medio audiovisual: las estrategias humorísticas de la comedia de situación,* (tesis doctoral inédita). Universidad de Alicante.

CRESPO, N. Y ALVARADO, C. (2010): "Conciencia metapragmática y memoria operativa en niños escolares", *Literatura y Lingüística*, 21, 93 – 108.

GALINDO MERINO, Mª M. (2017): "La identidad de género a través del humor en niños y niñas de 9 – 10 años" en *CLAC, Círculo de Lingüística Aplicada a la Comunicación*, 70, Madrid, 99 – 118.

JULE, A. (2008): *A Beginner's Guide to Language and Gender.* Clevedon: MM Textbooks.

MURA, A. y RUIZ GURILLO, L. (Eds.) (2014): *Feminismo/s, 24. Género y humor en discursos de mujeres y hombres*, Alicante: Universidad de Alicante.

OJEDA Y CRUZ (2004): "Yo me parto" oralidad, humor, gramática y pragmática, un cóctel lúdico para el aula de E/LE, Las gramáticas y los diccionarios en la enseñanza del español como segunda lengua, deseo y realidad: Actas del XV Congreso Internacional de ASELE, Sevilla 22-25 de septiembre de 2004 / coord. por María Auxiliadora Castillo Carballo, 2005, 234-240.

RUIZ GURILLO, L. (2019): *Humor de género: del texto a la identidad en español,* Editorial Iberoamericana Vervuert: Madrid.

RUIZ GURILLO, L. Y PADILLA GARCÍA, X. A. (2009): *Dime cómo ironizas y te diré quién eres.* Peter Lang: Berlín.

TIMOFEEVA (2014): "El humor verbal en niños de educación primaria: desarrollo de la conciencia metapragmática", en MURA, A. y RUIZ GURILLO, L. (Eds) (2014): *Feminismo/s, 24. Género y humor en discursos de mujeres y hombres*, 195 – 219.

TIMOFEEVA (2017): "Metapragmática del humor infantil" en *CLAC, Círculo de Lingüística Aplicada a la Comunicación*, 70, Madrid, 5 – 19.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Bases metodológicas: la construcción de un corpus para la detección de mentiras y la evaluación de la credibilidad

Pedro Eduardo Hernández Fuentes[46]
[hfuentes.eduardo@gmail.com]
Centro de Investigación en Ciencias Cognitivas, México

El estudio de la detección de mentiras y la evaluación de la credibilidad ha sido tema de interés para muchos especialistas y se ha abordado desde diferentes disciplinas. Aunque se han aportado herramientas científicas para su estudio, todavía se suele caer en la falsa creencia de que existen claves determinantes, señales corporales universales (por ejemplo, el movimiento de los ojos izquierda-derecha) o indicadores fisiológicos que son prueba irrefutable de que un individuo está mintiendo.

Sin embargo, la revisión sistemática para analizar cuantitativamente los resultados de las investigaciones (metaanálisis) revela que la mayoría de los indicadores que los investigadores suelen examinar en la detección de mentiras no están relacionados con el engaño en absoluto (Vrij, Granhag y Porter, 2010). Se ha sugerido que las *microexpresiones* —movimientos faciales rápidos con duración menor a una quinta parte de segundo (Ekman, 2017)— son indicios más confiables del engaño que canales como las palabras o lo lingüístico (Ekman, 2015), pero existe una discusión al respecto, pues el acto de mentir o engañar puede generar emociones positivas, negativas o, incluso, éstas pueden no estar presentes y, por lo tanto, el análisis de ellas no es la mejor forma de determinar cuando una persona oculta una verdad (Burgoon, 2018).

Los estudios metaanalíticos también revelan que la información verbal es un indicador más confiable para identificar mentiras o evaluar la credibilidad de un testimonio (DePaulo *et al*, 2003; Vrij, 2018). De aquí que actualmente se han desarrollado investigaciones desde la lingüística forense, la sociolingüística, la psicolingüística y, mayoritariamente, la psicología para realizar aportaciones científicas al respecto. Por ejemplo, Hwang, Matsumoto y Sandoval (2016) se han preocupado por el desarrollo de técnicas adecuadas y más precisas para la detección del engaño desde el contenido verbal; Picornell (2013) se ha dedicado a estudiar la detección del engaño en declaraciones escritas de testigos y ha propuesto formas de buscar señales de engaño desde las características narrativas de los testigos; y Fitzpatrick y Bachenko (2009) intentó probar la precisión de algunas señales lingüísticas vinculadas con el engaño. Estos esfuerzos

[46] Licenciado en Lengua y Literaturas Hispánicas (UNAM) y maestrando en Ciencias Cognitivas (UAEM). Cuenta con un diplomado en análisis del discurso público (UNAM) y una certificación internacional en detección de mentiras y evaluación de la credibilidad (No Verbal). Ha colaborado en el Centro de Estudios Lingüísticos y Literarios (El Colmex), en la Academia Mexicana de la Lengua y en el Seminario Universitario de Estudios del Discurso Forense. Actualmente es miembro del Seminario de Lenguaje No Verbal (UNAM) y del Laboratorio de Lenguaje y Cognición (CINCCO-UAEM), donde desarrolla una investigación en detección de mentiras y evaluación de la credibilidad.

vuelven imperativa la necesidad de enfocarse en la construcción de un corpus lingüístico que posibilite el estudio de la detección de mentiras y la evaluación de la credibilidad.

Por ello, el proyecto que se presentará es el resultado de un trabajo interdisciplinario entre la lingüística y la psicología en el que la construcción del corpus funciona como eje para una futura investigación. En la presentación se expondrán los avances del proyecto desarrollado en el Laboratorio de Lenguaje y Cognición del Centro de Investigación en Ciencias Cognitivas (UAEM), en el que se enfatizará en la metodología seguida para la construcción de un corpus que permita estudiar la detección de mentiras y la evaluación de la credibilidad. Se ahondará en la explicación del método y descripción general para la construcción del corpus. Por esto, se presentará el tipo de estudio, el tipo de participantes, los materiales e instrumentos y el procedimiento.

A lo largo de la exposición se exaltarán los riesgos en la construcción de un corpus de este tipo. También se aspira a introducir al espacio académico un tema poco abordado desde terrenos científicos, porque se puede afirmar que no se han realizado trabajos suficientes que consideren la teoría lingüística para abordar el fenómeno en cuestión: la mayoría de las investigaciones se han planteado desde la psicología cognitiva. Existe una carencia también en la poca investigación realizada sobre el idioma español, si bien algunas propuestas recientes consideran este idioma como objeto de estudio —véase, por ejemplo, el trabajo de Hwang, Matsumoto y Sandoval (2016) o el de Vrij *et al.* (2020)— aún son pocos los esfuerzos.

En suma, aunque las investigaciones han acentuado la preponderancia del análisis del contenido verbal en contraste con el del comportamiento no verbal y del paralingüístico, se ha observado que existe una carencia en este sentido y no se le ha dado el valor suficiente a la construcción del corpus para que, en un futuro, se puedan estudiar los principales indicadores lingüísticos diferenciadores entre un discurso que pretende engañar a otro y uno que no. Este proyecto realizará una aportación desde este vacío.

Palabras clave

Mentiras, corpus lingüístico, metodología, credibilidad, español

Bibliografía

Burgoon, J. K. (2018). Microexpressions Are Not the Best Way to Catch a Liar. *Frontiers in Psychology*, 9, 1-5. DOI: 10.3389/fpsyg.2018.01672.

DePaulo *et al.* (2003). Cues to Deception. *Psychological Bulletin*, 129 (1), 74-118. DOI: 10.1037/0033-2909.129.1.74.

Ekman, P. (2015). *Cómo detectar mentiras. Una guía para utilizar en el trabajo, la política y la pareja*. (L. Wolfson, trad.). Barcelona: Paidós. (Original publicado en 2001).

_____. (2017). *El rostro de las emociones. Qué nos revelas las expresiones faciales*. (J. J. Serra, trad.). México: RBA. (Original publicado en 2003).

Fitzpatrick, E. y Bachenko, J. (2009). Building a forensic corpus to test language-based indicators of deception. En Gries, S. T., Wulff, S. y Davies, M. (editores), *Corpus-linguistic*

applications. Current studies, new directions (pp. 183-196). Amsterdam, New York: Rodopi.

Hwang, H. C., Matsumoto, D. y Sandoval, V. (2016). Linguistic Cues of Deception Across Multiple Language Groups in a Mock Crime Context. *Journal of Investigative Psychology and Offender Profiling*, 13, 56-69. DOI: 10.1002/jip.1442.

Picornell, I. (2013). Analysing Deception in Written Witness Statements. *Linguistic Evidence in Security, Law and Intelligence*, 1 (1), 41-50. DOI: 10.5195/lesli.2013.2.

Vrij, A., Granhag, P. A. y Porter, S. (2010). Pitfalls and Opportunities in Nonverbal and Verbal Lie Detection. *Psychological Science in the Public Interest*, 11 (3), 89-121. DOI: 10.1177/1529100610390861.

Vrij, A. (2018). Deception and truth detection when analyzing nonverbal and verbal cues. *Applied Cognitive Psychology*, 33 (2), 160-167. DOI: 10.1002/acp.3457.

Vrij, A. *et al.* (2020). The Efficacy of Using Countermeasures in a Model Statement Interview. *The European Journal of Psychology Applied to Legal Context*, 12 (1), 23-34. DOI: 10.5093/ejpalc2020a3.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Caracterización del corpus de léxico disponible de profesores de Español-Literatura en formación de pregrado en la Universidad de Las Tunas, Cuba

Grechel Calzadilla Vega[47], **Luis Daniel Sánchez Ravelo**[48]
[calzadillavega@gmail.com | luisdaniel@ult.edu.cu]
Universidad de Las Tunas, Cuba
Marlen Aurora Domínguez Hernández[49]
[marlen@fayl.uh.cu]
Universidad de La Habana, Cuba

En las últimas décadas se ha producido un aumento considerable de las investigaciones lingüísticas que, junto al desarrollo de modernos recursos informáticos para el procesamiento de información, han contribuido a la aparición de una amplia variedad de corpus electrónicos, que ofrecen posibilidades para la interpretación gramatical, semántica, léxica, discursiva o de otro tipo sobre el uso de la lengua, además de su aplicación. En este sentido, Cuba se insertó desde 2005 en los estudios de disponibilidad léxica, cuyos resultados aportan corpus de léxico disponible, contentivos de datos reales y fiables sobre el uso de la variedad cubana del español.

Aunque la mayor parte de las investigaciones realizadas en Cuba se inscribe entre los resultados del Proyecto Panhispánico de Disponibilidad Léxica (PPDL), la necesidad de conocer el léxico de personas instruidas, que se encuentra potencialmente disponible para ser usado en diferentes situaciones comunicativas, ha abierto el espectro de estos estudios a muestras de informantes que no responden al criterio seguido por el PPDL de no estar condicionados aún por los estudios universitarios o por el ámbito profesional; si bien se siguen las restantes pautas metodológicas estandarizadas para emprender investigaciones de este tipo en el país (Domínguez, 2014).

A esta línea de investigación responde la investigación doctoral La disponibilidad léxica del profesor de Español-Literatura en formación de pregrado (Calzadilla, 2019), realizada en la Universidad de Las Tunas, de la que se obtuvo un corpus total del léxico disponible (Tabla 1) integrado por 26 590 palabras y 6 232 palabras diferentes o vocablos.

[47] Doctora en Ciencias Pedagógicas. Licenciada en Filología, especialidad Lingüística Hispánica. Profesora de la Disciplina Estudios Lingüísticos y Jefa del Departamento de Español-Literatura de la Universidad de Las Tunas, Cuba.

[48] Licenciado en Educación Español-Literatura. Profesor de la Disciplina Estudios Lingüísticos, Departamento de Español-Literatura de la Universidad de Las Tunas, Cuba.

[49] Doctora en Ciencias Filológicas. Profesora Titular y Consultante de la Facultad de Artes y Letras de la Universidad de La Habana. Miembro de la Academia Cubana de la Lengua.

Centros de interés	Total de palabras	%	Promedio de palabras emitidas por informante	Total de vocablos	%
01: Partes del cuerpo	1 401	5.3	28.6	263	4.2
02: La ropa	1 445	3.9	21.3	329	5.3
03: Partes de la casa (sin los muebles)	848	3.2	17.3	164	2.6
04: Los muebles de la casa	874	3.3	18.0	103	1.6
05: Alimentos y bebidas	2 716	10.2	55.4	410	6.6
06: Objetos colocados en la mesa para la comida	1 619	6.1	33.0	154	2.5
07: La cocina y sus utensilios	1 554	5.8	32.0	194	3.1
08: La escuela: muebles y materiales	1 748	6.6	36.0	259	4.1
09: Iluminación, calefacción y medios de refrescar o enfriar una habitación o edificio	851	3.2	17.3	116	1.9
10: La ciudad	1 207	4.5	25.0	373	6.0
11: El campo	2 054	7.7	42.0	572	9.2
12: Medios de transporte	1 092	4.1	22.2	297	4.8
13: Trabajos del campo y del jardín	804	3.0	16.4	294	4.7
14: Los animales	1 490	5.6	30.4	364	5.8
15: Juegos y distracciones	1264	4.7	26.0	355	5.7
16: Profesiones y oficios	1 363	5.1	28.0	431	6.9
17: La familia	1 224	4.6	25.9	305	4.9
18: Tecnologías de la información y las comunicaciones	933	3.5	19.0	297	4.8
19: La educación	1 162	4.4	24.0	525	8.4
20: La cultura	1 311	4.9	27.0	427	6.8
Corpus total	**26 590**			**6 232**	

Tabla 1. Resultados globales en el corpus del léxico disponible

Los resultados alcanzados respecto del total de palabras y el promedio de respuestas por centros de interés permitieron comprobar que en 15 áreas temáticas el promedio de respuestas superó las 20 palabras, mientras que solo cinco se quedaron por debajo de esta cifra. En cuanto al número de palabras diferentes, que ofrece una valiosa información sobre la menor o mayor capacidad de convocatoria asociativa de los centros de interés, se obtuvo un corpus total que ascendió a 6 232 vocablos; el promedio de vocablos evocados por cada sujeto ante el conjunto global de estímulos fue de 127.1 y, al realizar el cálculo por centros de interés, el promedio fue de 311.6 vocablos.

De los 20 centros de interés nueve superaron el promedio de vocablos por áreas temáticas. Los 11 centros de interés restantes mostraron valores inferiores al promedio y, entre ellos, el centro de interés 09 alcanzó el resultado más bajo, con una diferencia de 195.6 unidades respecto

de la media, lo que está generalmente en correspondencia con los resultados alcanzados en estudios similares, como parte del PPDL, dada la caracterización de este tipo de realidades.

Entre todos, los centros de interés más productivos resultaron el 11, el 19, el 16, el 20 y el 05, que superaron en 260.4, 213.4, 119.4, 115.4 y 98.4 vocablos, respectivamente, la media por área temática. No obstante, resulta pertinente señalar que los centros de interés 17, 12, 18, 13, 01 y 08 contaron solo con una diferencia de 6, 14, 14, 17, 48 y 52 unidades, respectivamente, para alcanzar el promedio de vocablos por centro de interés. Los cinco centros de interés restantes reflejaron la mayor diferencia respecto del promedio de vocablos por centro de interés, en tanto contaron con las mayores diferencias, 117, 147, 157, 195 y 208 vocablos, respectivamente.

De forma general, si bien el tamaño de la muestra incide directamente en la cantidad total de palabras actualizadas, un análisis más particular permite destacar que existen marcadas divergencias en la productividad léxica de los diferentes centros de interés, sin duda reflejo de realidades peculiares tanto psicosociales como medioambientales. De este modo, los datos obtenidos permitieron comprobar la existencia de campos asociativos que reflejaron cuantitativamente que en ellos los informantes cuentan con mayores recursos léxicos, lo que acentúa las diferencias al realizar comparaciones, por ejemplo, entre el centro de interés 11, en el que se actualizó el mayor número de vocablos del corpus total (572), y el centro de interés 04, en el que se actualizó el menor número de vocablos (103). Los datos del primer centro de interés fueron significativamente superiores, incluso, al promedio de vocablos alcanzado (311.6), al que superaron en 261 vocablos, lo que se justificó por las múltiples opciones tanto para actualizar vocablos diferentes como en cuanto a sus relaciones de sentido, que pueden ser actualizadas sobre un tema ampliamente conocido y semánticamente productivo como lo es el campo; mientras que, en el otro caso, tanto los vocablos como las relaciones de sentido que se pueden actualizar respecto de los muebles de la casa resultan claramente inferiores.

Mientras el promedio de respuestas permite medir la individualidad, a las tendencias del conjunto de la muestra se accede a través de la densidad léxica y del índice de cohesión. El análisis de la densidad léxica y del índice de cohesión en cada uno de los 20 centros de interés (Tabla 2) reveló que las áreas temáticas semánticamente más compactas, o sea, de mayor concreción semántica, capacidad asociativa y densidad léxica fueron los centros de interés 01, 02, 11 y 14.

Un análisis de estos datos permitió concluir que la mayor o menor producción léxica obedeció, de forma general, a la naturaleza de cada área temática con la que los informantes se relacionaron con mayor o menor cercanía; o sea, cómo la influencia del entorno condicionó la mayor o menor productividad léxica y la mayor difusión en el grado de cohesión semántica en las respuestas dadas.

Centros de interés	Densidad léxica	Índice de cohesión
20: La cultura	105.129	0.21
19: La educación	84.854	0.17
18: Tecnologías de la información y las comunicaciones	80.103	0.16
17: La familia	73.362	0.14
16: Profesiones y oficios	67.490	0.13
15: Juegos y distracciones	66.243	0.13
14: Los animales	53.269	0.10
13: Trabajos del campo y del jardín	51.707	0.10
12: Medios de transporte	40.934	0.08
11: El campo	40.131	0.08
10: La ciudad	36.767	0.07
09: Iluminación, calefacción y medios de refrescar o enfriar una habitación o edificio	35.909	0.07
08: La escuela: muebles y materiales	35.605	0.07
07: La cocina y sus utensilios	32.359	0.06
06: Objetos colocados en la mesa para la comida	31.762	0.06
05: Alimentos y bebidas	31.624	0.06
04: Los muebles de la casa	31.414	0.06
03: Partes de la casa (sin los muebles)	30.702	0.06
02: La ropa	27.346	0.05
01: Partes del cuerpo	22.133	0.04

Tabla 2. Resultados de la densidad léxica y del índice de cohesión en cada centro de interés

De forma general, y de acuerdo con los resultados obtenidos, fue posible clasificar los centros de interés de forma decreciente en tres grupos: a) campos semánticamente más compactos (mayor concreción semántica, mayor capacidad asociativa, más cerrados) y con mayor densidad léxica: 01, 03, 04, 05, 06, 07, 08 y 09; b) áreas temáticas con una densidad media y con un grado medio de coincidencia en las respuestas: 02, 10, 11, 12, 14, 15, 16, 17, 18: y 20; y c) campos semánticamente más difusos, con menor densidad léxica y, por tanto, más abiertos: 13 y 19.

A partir del análisis realizado respecto de los valores obtenidos en los centros de interés 13 y 19, que contaron con los índices más bajos de cohesión, se concluyó que tal situación está en correspondencia con que, a mayor contacto con un tema o un campo asociativo se obtiene una mayor variedad de vocabulario, a la vez que una mayor difusión en cuanto al grado de cohesión.

O sea, que teniendo en cuenta, por ejemplo, el centro de interés 19: La educación, que resulta significativo por su relación con la formación profesional de los informantes, los datos demostraron que los informantes actualizaron una amplia gama léxica que favoreció la difusión.

Palabras clave: Lingüística de corpus; corpus de léxico disponible; densidad léxica; índice de cohesión léxica.

Bibliografía

Calzadilla, G. (2019). *La disponibilidad léxica del profesor de Español-Literatura en formación de pregrado* (tesis doctoral inédita). Universidad de Las Tunas, las Tunas, Cuba.

Domínguez, M. A. (2014). *Pautas metodológicas para el estudio estandarizado de la disponibilidad léxica en Cuba*. La Habana: Facultad de Artes y Letras, Universidad de La Habana.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

CorpusPAP: lexicografía e internet como corpus

María Betulia Pedraza Pedraza[50]
[marpedrazapedraza@gmail.com]
Universidad del Norte, Colombia

Resumen

La elaboración de corpus a partir de internet es una de las alternativas que hoy adquiere especial relevancia. Si bien, buena parte de los trabajos en lexicografía se valen del trasvase de datos de nomenclaturas provenientes de obras previas, cada vez cobra más fuerza la utilización de fuentes de internet, bien sea documentos tomados de los sitios web o directamente de resultados ofrecidos por los buscadores. Esta comunicación busca mostrar los alcances de utilizar internet como corpus en la elaboración de trabajos terminológicos y lexicográficos, en concreto, el *Corpus Panhispánico de la Administración Pública (CorpusPAP)*. Esta herramienta digital se compone de tres bancos de datos del dominio disciplinar de la administración pública de cuatro países de habla hispana, Argentina, Colombia, España y México, e integra un buscador personalizado de Google que facilita al usuario acceder a otras webs relacionadas. Para el caso del *CorpusPAP* internet como corpus ha supuesto dos puntos de vista: uno recoger y *barrer* las *websites* de instituciones y generar unos bancos de términos de la administración pública. Otro punto de vista tiene que ver con la utilidad que los datos recogidos puedan tener en la construcción de productos lexicográficos para usuarios reales. El *CorpusPAP* como propuesta de banco de datos en línea integra un producto lexicográfico que a su vez se alimenta del banco llamado *Diccionario Panhispánico de la Administración Pública (DiPAP)*. Ambos recursos se inscriben en los preceptos de la teoría funcional y de allí que refleje la metodología que Tarp y Fuertes-Olivera (2014) sostienen deberían seguir los proyectos lexicográficos: mejorar la metodología para la obtención de datos de modo que estos sean realmente significativos. El *DiPAP* presenta un concepto de diccionario colaborativo que dispone la información en la interfaz de usuario en cuatro bloques: información visible, información semioculta, información oculta e información asociada. Esta disposición busca evitar la sobrecarga informativa de modo que el usuario acceda a los bloques según sus necesidades.

Palabras clave

Administración pública, internet, lenguaje claro, lexicografía, terminología

[50] Doctora en Filología Española con mención internacional por la Universitat Autònoma de Barcelona donde colabora en el Grupo de investigación en lengua de la ciencia y de la técnica - Neolcyt. Lexicógrafa y lingüista, ha colaborado en varios proyectos lexicográficos en Harper Collins, Ediciones SM, en la Academia Colombiana de la Lengua y en la ANLE. Forma parte de comité científico de Research Institute United States Spanish - RIUSS y dirige el desarrollo de corpus de sus proyectos terminológicos

Bibliografía

CABRÉ, María Teresa (1999): *La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, pp. 152, 153.

DUARTE I MONTSERRAT, Carles (2002): «El llenguatge administratiu i jurídic: la necessitat d'un nou pas endavant». En *Revista de Llengua i Dret*, núm. 38, diciembre, Barcelona, pp. 11.

NIELSEN, Sandro (2014): «Database of legal terms for communicative and knowledge information tools». In Mairtin Mac Aodha, ed., *Legal Lexicography. A comparative perspective*. Farnham: Ashgate Pulishing, pp.155, 156, 161.

NIELSEN, Sandro (2015): «Data Presentation Structures in Specialised Dictionaries: Law Dictionaries with Communicative Functions». In: *Lexicographica: International Annual for Lexicography / Revue Internationale de Lexicographie / Internationales Jahrbuch für Lexikographie*, vol. 31, núm. 1, pp. 200 - 216. Disponible en: < https://doi.org/10.1515/lexi-2015-0010> [Consultado: 13/10/2016]

NIELSEN, Sandro y MOURIER, Lise (2004): «Design of a funtion-based internet accouting dictionary». In Gottlieb Henrik y Mogensen Jens (eds.). *Dictionary visions, research and practice. Selected papers from the 12th International Symposium on Lexicography, Copenhagen 2004*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

PEDRAZA PEDRAZA, María Betulia (2016): «Democratización del léxico de especialidad y política lingüística panhispánica: Retos de una propuesta lexicográfica en la administración pública». En *El diccionario en la encrucijada: de la sintaxis y la cultura al desafío digital*. España: Escuela Universitaria de Turismo Altamira y Asociación Española de Lexicografía Hispánica, pp. 476.

POSTIGO DE DE BEDIA, Ana María (1999): *Términos de la administración pública. teoría y aplicaciones*. Salvador de Jujuy, Argentina: Universidad nacional de Jujuy, Secretaría de ciencia y técnica y estudios regionales, pp. 19, 20, 29.

SARMIENTO GONZÁLEZ, Ramón (2005): «El lenguaje de la Administración». En *Revista de Llengua i Dret*, núm. 43, Barcelona, pp. 14, 21, 25.

TARP, Sven (2012): «Online dictionaries: today and tomorrow». En *Lexicographica,* vol. 28, núm. 1. pp. 256.

TARP, Sven. (2015): «La teoría funcional en pocas palabras». En *Estudios de Lexicografía*, núm. 4, pp. 36,37. Disponible en: < http://cc.au.dk/ fileadmin/04._La_teori___a_funcional_en_pocas_palabras.pdf> [Consultado: 05/05/2016]

TARP, Sven (2018): «Lexicography as an independent science». In *The Routledge Handbook of Lexicography.* Fuertes-Olivera, P. A. (ed.). UK & USA: Routledge, pp.19-33, 15 p. (Routledge Handbooks in Linguistics).

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

CLEC Colombian Learner English Corpus. Primer Corpus en Línea de Producción Escrita en Inglés en Colombia

María Victoria Pardo[51]
[mpardoenudea@gmail.com]
Universidad del Norte, Colombia
Antonio Tamayo, Manuel Gómez[52] **& Nicolás Henao**
[antonio.tamayo@udea.edu.co]
Universidad de Antioquia, Colombia

El objetivo de esta ponencia es presentar a la comunidad investigadora el Corpus Colombiano de Aprendices de Inglés CLEC (*Colombian Learner English Corpus*). Este corpus fue creado siguiendo los lineamientos de la Lingüística de Corpus Computacional (McEnery & Hardie, 2011) y de acuerdo con los parámetros de compilación de corpus de aprendices definidos como "colecciones electrónicas de datos naturales o casi naturales producidos por estudiantes extranjeros o de segunda lengua (L2) y reunidos según criterios de diseño explícitos" Granger (2002, p. 7), Gilquin (2015, p.1).

El CLEC fue creado por el grupo de investigación TNT de la universidad de Antioquia con el fin de conocer el nivel de la interlengua de estudiantes de inglés a nivel universitario por medio del análisis de errores. Este análisis se basa en las teorías fundamentales de Análisis de Errores de Corder (1981) quien afirma que si los errores son estudiados sistemáticamente podremos atender las necesidades de los aprendices. El análisis sistemático de errores requiere de herramientas adecuadas y modernas que garanticen su confiabilidad y es aquí donde la lingüística de Corpus Computacional provee una respuesta.

Posterior a la compilación del corpus se llevó a cabo la codificación de los errores de acuerdo al Manual de Etiquetado de Errores de la Universidad de Louvaina (Dagneaux *et al.*, 2005). El manual de codificación de errores cuenta con 8 categorías subdivididas en 56 tipos de error de la siguiente manera:

---

[51] Profesora de tiempo completo de inglés como lengua extranjera de la Universidad del Norte, Barranquilla, Colombia. Magister en traducción de la Universidad de Ottawa, Canada. Doctora en Lingüística Universidad de Antioquia. Vinculada al grupo de investigación TNT de la universidad de Antioquia. Intereses de investigación en corpus computacional en la enseñanza de lenguas, el desarrollo de materiales de aprendizaje y análisis de patrones de lengua y del discurso.

[52] Es ingeniero de sistemas y magíster en ingeniería de la Universidad de Antioquia. Es candidato a doctor en ciencias de la computación del centro de investigación en computación del Instituto Politécnico Nacional en Ciudad de México, donde desarrolla su tesis en procesamiento de lenguaje natural, bajo la supervisión del Dr. Alexander Gelbukh. Ha trabajado como analista y desarrollador de software para diversos proyectos con universidades de Colombia, Noruega y EE.UU. Desde el año 2010, trabaja en proyectos de ingeniería lingüística con el grupo de investigación en traducción y nuevas tecnologías (TNT) de la Universidad de Antioquia. Actualmente es profesor de cátedra en los programas de ingeniería de sistemas de la Universidad de Antioquia y la Universidad de Medellín.

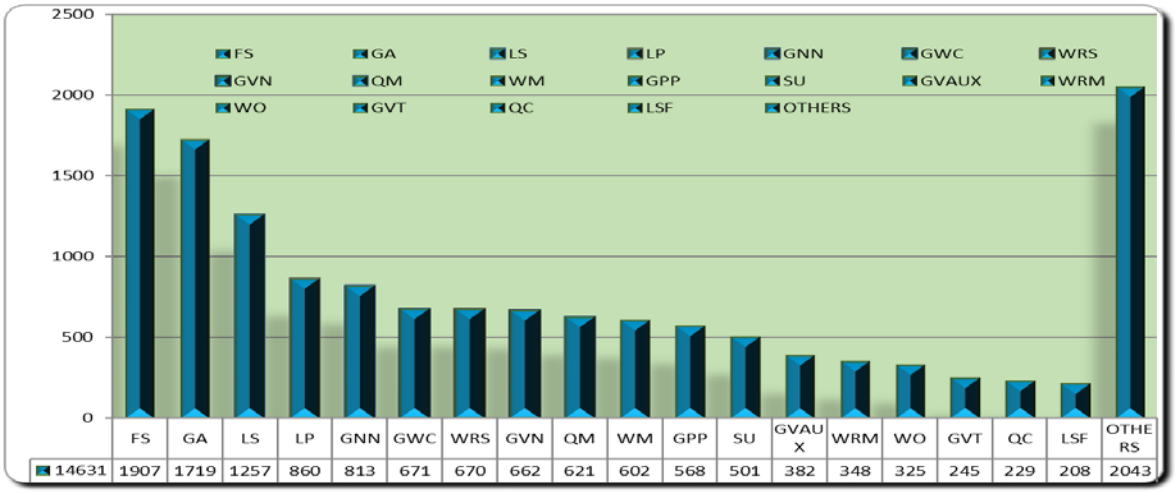
Category		Code (letter that stands for error category	Total
1.	Form errors	F	3
2.	Grammatical errors, i.e. errors that break general rules of English grammar	G	25
3.	Lexico-grammar errors, i.e. errors where the morpho-syntactic properties of a word have been violated.	X	9
4.	Lexical errors, i.e. errors involving the semantic properties of single words and phrases.	L	8
5.	Word Redundant, Word Missing and Word Order errors.	W	4
6.	Punctuation errors.	Q	4
7.	Style errors.	S	2
8.	Infelicities.	Z	1
		Error types	56

La extracción de errores se hizo con software especializado para el análisis lingüístico WordSmith Tools 5, Scott (2008) y Lancsbox, Brezina, McEnery, & Wattam (2015). Este software especializado es usado para encontrar patrones lingüísticos, frecuencias, concordancias, y en este caso, para obtener estadísticas de los errores del corpus.

El CLEC es una aplicación que compila 515 composiciones escritas de estudiantes de inglés como lengua extranjera a nivel universitario. Los textos fueron obtenidos de los estudiantes previamente clasificados en niveles B1.1, B1.2, B1.3-B2.1, B2.2 - B.2.3, de acuerdo con el Marco Común Europeo de Referencia (Europe, 2001). La clasificación de los participantes es un hecho que se da previo al inicio de los cursos de lengua y por lo tanto previo al inicio de la presente investigación, por esa razón no se pueden obtener los textos sin antes tener a los estudiantes en determinado nivel. El procedimiento de agrupación de los estudiantes depende del examen de clasificación que ellos presentan antes del inicio de nivel convirtiéndose así en el regulador de criterios para que un estudiante se encuentre en determinado nivel. Las composiciones escritas variaban de acuerdo a cada nivel, por ejemplo en los niveles básicos los estudiantes elaboraban párrafos cortos respondiendo una pregunta determinada. Su producción no sobrepasaba las 50 palabras, mientras que en los niveles avanzados producían ensayos de entre 50-700 palabras. En todo caso, parte de los análisis se hicieron del porcentaje de errores por cada 100 tokens para garantizar la proporcionalidad del análisis. La siguiente fue la distribución de los estudiantes/tokens/errores:

	Code	Level	Qty	Tokens	% Gender	% Token	# Of Errors	% Errors/Tokens
Male	1	**B1.1**	104	19,301	43%	41%	2,402	12.4%
Female	2	**B1.1**	135	27,592	56%	59%	3,226	11.7%
No gender	0	**B1.1**	1	176	0.4%	0.4%	23	13.1%
Male	1	**B1.2**	70	13,400	48%	46%	1,517	11.3%
Female	2	**B1.2**	74	15,605	51%	53%	1,754	11.2%
No gender	0	**B1.2**	2	437	1%	1%	44	10.1%
Male	1	**B1.3-B2.1**	8	1,784	47%	47%	279	15.6%
Female	2	**B1.3-B2.1**	9	1,977	53%	53%	302	15.3%
Male	1	**B2.2-B2.3**	56	33,725	50%	49%	2,526	7.5%
Female	2	**B2.2-B2.3**	56	35,328	50%	51%	2,558	7.2%

Los siguientes son los errores con mayor y menor incidencia:



Todos los textos del CLEC se recopilaron en el segundo semestre de 2015 y cada texto fue codificado con etiquetas de error.

La metodología del etiquetado fue como se describe a continuación:

Después de que se recopilaron los archivos, recibieron diferentes procesos porque estaban en diferentes formatos. Por ejemplo, y debido a que su trabajo final fue escrito a mano, para los niveles B1.1 a B2.1, el proceso comenzó con el escaneo seguido de la transcripción de los textos, asegurándose que las composiciones eran exactas a los originales. Luego, se convirtieron en textos TXT para hacer anotaciones de error. Los estudiantes de los niveles B2.2 a B2.3 hicieron directamente la versión digital, por lo que esos textos se convirtieron inmediatamente al formato TXT para el etiquetador de errores. Los archivos escritos a mano fueron en total 373 y el proceso de transcripción duró aproximadamente siete meses. Después de toda la preparación anterior, todos los archivos (515 en total) estaban listos para comenzar la anotación de errores. La anotación de errores se basa en taxonomías de error con categorías para la clasificación de error proporcionadas por los sistemas de anotación de error. En este caso particular, el investigador hizo cada anotación siguiendo las categorías del etiquetador de la Universidad de Louvain (Dagneaux *et al.*, 2005) previamente descritas. Cuando se detectó un

error, la etiqueta de error se colocó justo antes del error y la corrección siguió al error entre dos signos de dólar: $ corrección $ como lo indica el manual. Analicemos el siguiente ejemplo.

Nowadays we have seen (GADJN) differents $different$ (this error corresponds to the category of grammar and it is a pluralization of an adjective) Este error corresponde a la categoría grammatical y se refiere a la pluralización de un adjetivo en inglés.

La extracción de errores se hizo utilizando el software WordSmith (Scott, 2008). Todos los errores se extrajeron de acuerdo con sus categorías, se colocaron en hojas de Excel y se enviaron a un revisor experto externo para verificar la exactitud y coherencia de los archivos en el etiquetado.

Este proyecto surgió como respuesta a la pregunta de investigación sobre cuáles serían los errores más comunes de los estudiantes universitarios en la adquisición del inglés como lengua extranjera en los diferentes niveles de adquisición, ya que en muchos casos hay errores prevalentes hasta niveles avanzados.

Con respecto a trabajos previamente desarrollados de autores como Granger, el presente trabajo es innovativo porque no solamente se hizo un análisis por niveles y categorías de error, además se hizo por estrato socioeconómico y por género. Este análisis propone el análisis de errores por medio de corpus computacional como una manera de establecer el nivel real de interlengua de los estudiantes y además, como resultado, para diseñar mejores materiales de aprendizaje.

La interfaz se desarrolló utilizando patrones de diseño de software para obtener una aplicación web sólida, amigable y compatible. La arquitectura de varios niveles tiene varias capas con un diseño orientado a objetos en el que la capa de servicio funciona con el lenguaje de programación Python con el marco Django. Este marco permite un desarrollo rápido y limpio con un diseño práctico, también incluye tres fortalezas principales: velocidad, seguridad y escalabilidad. Además de eso, Django tiene un patrón de diseño: patrón Modelo-Vista-Plantilla (MVT), diferente del patrón Modelo-Vista-Controlador (MVC). En el patrón MVT, la capa procesa qué datos se mostrarán, pero no cómo se mostrarán porque la capa de plantilla tiene la información sobre cómo se mostrarán los datos.

El resultado fue una aplicación web receptiva que realiza búsquedas completas y hace análisis sobre el corpus de errores etiquetado.

Referencias

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

Corder, S. (1981). *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J., & Thewissen, J. (2005). Error Tagging Manual Version 1.2. Centre for English Corpus Linguistics, Université Catholique de Louvain.

Europe, C. of. (2001). 1 The Common European Framework of Reference for Languages: Learning, teaching, assessment. *Common European Framework*. https://doi.org/10.1093/elt/cci105

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge M.I.T. Press.

Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3–33). Philadelphia: John Benjamins Publishing Company.

McEnery, A., & Hardie, A. (2011). *Corpus Linguistcis: Method, theory and practice*. Cambridge: Cambridge University Press.

Scott, M. (2008b). WordSmith Tools Version 5. Lexical analysis software. Liverpool.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Descortesía en las redes sociales: análisis de un corpus de comentarios de Facebook

Isabel García Martínez[53]
[isgarmar@gmail.com]
Universidad de Valencia, España

En las últimas décadas la interacción social se ha visto modificada por la interacción virtual y parte de este cambio tiene relación con las redes sociales. En este medio, la comunicación mediada por ordenador (CMO), los usuarios son más hostiles y ofensivos en comparación con las interacciones cara a cara (Moor, 2008: 7), por ello son más que frecuentes las manifestaciones de descortesía.

En este estudio se analiza un corpus de comentarios de usuarios en las páginas de Facebook de los principales partidos políticos españoles a través de la variable sexo. Este corpus se ha compilado durante la campaña electoral de noviembre de las elecciones generales españoles del año 2019. Para su elaboración se ha efectuado un análisis métrico de la campaña política con el programa Social Page Analyzer, así se ha podido determinar las dos jornadas en las que se han recibido más comentarios por parte de los usuarios de las páginas de los cinco partidos políticos españoles con más representación, es decir, Unidas Podemos, PSOE, Ciudadanos, PP y Vox. Posteriormente se han recopilado 2500 intervenciones reactivas, aisladas y anidadas, 500 comentarios de cada partido, a través de una web de extracción de comentarios.

El uso de la red social Facebook es bastante equitativo en cuestión de sexo, aunque las mujeres se encuentran en una posición ligeramente superior en España, siendo el 54% de los usuarios mujeres y el 46% los hombres, según el V Estudio sobre los usuarios de Facebook, Twitter e Instagram en España por parte de *The social media family (2019),* por ello nuestro principal objetivo es comprobar si se producen disparidades de comportamiento en función del sexo de los usuarios. Nos preguntamos qué sexo es el que más participa en las páginas de los partidos políticos más importantes de España en la red social Facebook y si el sexo implica un comportamiento diferente en cuanto a los actos de refuerzo de la imagen (*FFA - face flattering acts*) o los actos de ataque a la imagen (*FTA - face threatening acts*) en los comentarios, es decir, si se cumple el estereotipo de habla cortés e indirecta del sexo femenino frente al habla

[53] Licenciada en Traducción e Interpretación (Universidad de Málaga, 2009), Máster en Formación del Profesorado de Educación Secundaria (Universidad de Cantabria) y Máster Oficial en Enseñanza de Español como Lengua Extranjera (CIESE Comillas-UC). Es profesora asociada en la Universidad Europea del Atlántico y ha impartido cursos de lengua española y cultura hispánica en la Universidad Internacional Menéndez Pelayo (Santander), en el Centro de Idiomas de la Universidad de Cantabria y en el CIESE-Comillas. También ha participado en el proyecto Corpes XXI de la RAE y en el proyecto PRESEEA Santander. Actualmente realiza su doctorado en la Universitat de València.

descortés y directo masculino. Además, nos proponemos analizar estas diferencias en función de la filiación política.

Como resultados, se puede adelantar que las mujeres son más participativas en algunos partidos políticos, mientras que los hombres lo son en otros.

	Mujer	Hombre	Sin sexo
Unidas Podemos	51 %	49 %	0 %
PSOE	42,2 %	57,8 %	0 %
Ciudadanos	44 %	55 %	1 %
PP	53 %	47 %	0 %
Vox	35,8 %	64,2 %	0 %

Además, sí se produce un comportamiento diferente en función de la variable sexo y la filiación política en cuanto a la descortesía.

	FTA	FFA	FTA y FFA	No FTA no FFA
Unidas Podemos	44,8 %	45,6 %	7,2 %	2,4 %
PSOE	57,6 %	38,2 %	4 %	0 %
Ciudadanos	54 %	41 %	4 %	1 %
PP	52 %	38,2 %	7,8 %	2 %
Vox	37,2 %	52,4 %	8,4 %	2 %

Palabras clave

descortesía, Facebook, filiación política, sexo, redes sociales, comunicación mediada por ordenador

Bibliografía

Albelda Marco, M. (2004). “Cortesía en diferentes situaciones comunicativas. La conversación coloquial y la entrevista sociológica semiformal” en Pragmática sociocultural: estudios sobre el discurso de cortesía en español. Ariel Lingüística: Barcelona. págs. 109-134.

Bravo, D. (2004). "Tensión entre universalidad y relatividad en las teorías de la cortesía" en Pragmática sociocultural: estudios sobre el discurso de cortesía en español. Ariel Lingüística: Barcelona. págs. 15-37.

Briz Gómez, E. A. (2010). "La cortesía al hablar español" en III Jornadas de Formación e Profesores de ELE en China. Suplementos de SinoELE, 3.

Brown, P. y Levinson, S. (1987). Politeness. Some universals in language use. Cambridge: Cambridge University Press.

Chierichetti, L. (2014). "Descortesía en las páginas de Facebook de festivales de música" en Revista Normas, 4, págs. 27-48.

Culpeper, J. (1996). "Towards an anatomy of impoliteness" en Journal of Pragmatics 25, págs. 349-367.

Dalton, E. J. (2013). Impoliteness in computer mediated communication. Trabajo Fin de Máster. San Diego State University.

Díaz Pérez, J. C. (2012). Pragmalingüística del disfemismo y la descortesía. Tesis doctoral. Universidad Carlos III de Madrid.

Kaul de Marlangeon, S. y Cordisco, A. (2014). "La descortesía verbal en el contexto político-ideológico de las redes sociales" en Revista de Filología, 32. Págs. 145-162.

Kaul de Marlangeon, S. (2008). "Tipología del comportamiento verbal descortés en español", en Antonio Briz-Gómez *et al.* (eds.), Cortesía y conversación: de lo escrito a lo oral. Tercer coloquio internacional del programa Edice, Valencia/Estocolmo. Universidad de Valencia, págs. 254-266.

Kerbrat-Orecchioni, C. (2004). "¿Es universal la cortesía?" en Pragmática sociocultural: estudios sobre el discurso de cortesía en español. Ariel Lingüística: Barcelona. págs. 39-53.

Kienpointner, M. (1997). "Varieties of rudeness. Types and functions of impolite utterances" en Functions of Language, 4-2, págs. 251-287.

Goffman, E. (1970). Ritual de la interacción. Ed. Tiempo Contemporáneo: Buenos Aires.

Grice, H. P. (1975). "Logic and conversation" en Syntax and Semantic. Speech Acts. Nueva York: Academic Press, págs. 41-58.

Hernández Flores, N. (2004). "La cortesía como búsqueda del equilibrio de la imagen social" en Pragmática sociocultural: estudios sobre el discurso de cortesía en español. Ariel Lingüística: Barcelona. págs. 95-108.

Mak, B. C. N. y Chui, H. L. (2014). "Impoliteness in Facebook Status Updates: estrategia talk among colleagues "outside" the workplace" en Text&Talk, 34(2), págs. 165-185.

Mancera Rueda, A. (2009). "Manifestaciones de descortesía y violencia verbal en los foros de opinión digitales de los diarios españoles" en Discurso & Sociedad, Vol. 3(3), págs. 437-466.

Vigara Tauste, A. M.; Hernández Toribio, M. I. (2011). "Ciber(des)cortesía en los foros de opinión de la prensa escrita: un ejemplo" en ELUA: Estudios de Lingüística. Universidad de Alicante, 0(25), págs. 353-379.

Vivas Márquez, J. y Ridao Rodrigo, S. (2015). ""Lo siento pero me parecen horribles!!!": Análisis pragmalingüístico de la descortesía en la red social Facebook" en Revista de Filología, 33, págs. 217-236.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Disponibilidad y riqueza léxicas de un grupo de aprendientes de Español como Lengua Extranjera de niveles A2 y B2 de una institución universitaria de Medellín

Sindy Melissa Metaute Arango[54]
[sindy.metaute@udea.edu.co]
Universidad de Antioquia, Colombia
Diego Alexander Ortiz Rúa[55]
Universidad Nacional de Colombia, Colombia

El español se ha posicionado como una de las lenguas más habladas en el mundo y también como una de las más estudiadas desde el plano lingüístico por la difusión creciente que ha tenido en los últimos años. En la actualidad, en Colombia el Español como Lengua Extranjera (ELE) hace parte de la oferta educativa de diferentes Institutos de Educación Superior y de diversas escuelas privadas de español que se han creado en los últimos años.

Ahora bien, unos de los temas investigados alrededor del mundo en relación al español han sido la Disponibilidad Léxica (DL) y la Riqueza Léxica (RL). Los estudios de DL están relacionados principalmente con el Proyecto Panhispánico coordinado por Humberto López Morales. Por su parte, los estudios de Riqueza Léxica han estado direccionados a hablantes niños y jóvenes de ELM en diferentes países hispanohablantes.

Para hablar del contexto colombiano, según la bibliografía encontrada, hasta el momento se cuentan con tres estudios de DL en Español como Lengua Materna realizados por Mateus y Santiago (2006), Garzón y Penagos (2016), y Calderón y Mosquera (2018). De esta manera, según la bibliografía consultada no existen estudios que revisen la Disponibilidad Léxica del ELE de estudiantes que se encuentren realizando un curso de español en Colombia ni específicamente en la ciudad de Medellín. En relación a la Riqueza Léxica, según la bibliografía encontrada tampoco hay investigaciones realizadas en el país.

Tal y como lo plantea López (2014), la importancia y los aportes teóricos de los estudios de Disponibilidad Léxica y de Riqueza Léxica del Español como Lengua Extranjera residen en sus alcances como "conocer el vocabulario realmente disponible de los estudiantes, determinar la influencia de factores socioculturales en el aprendizaje del léxico y graduar el vocabulario para su enseñanza" (p.412), entre otros. A su vez, estos aportes contribuyen a la ampliación de la comunidad académica y discursiva sobre el tema en Colombia y al posicionamiento de la

[54] Filóloga Hispanista en formación, docente de Español como Lengua Extranjera en Learn Spanish Now, Medellín. Perteneciente al Semillero de Investigación en Pedagogía (SIP) de la Facultad de Educación y al Semillero de Investigación Español Histórico de Antioquia (SEHA) de la Facultad de Comunicaciones.
[55] Ingeniero químico y programador en Python.

Universidad de Antioquia como eje en el conocimiento de los fenómenos lingüísticos del ELE siendo además herramientas para próximos trabajos en diversas ramas de la lingüística.

De esta manera, el objetivo de esta investigación fue identificar la Disponibilidad y la Riqueza Léxicas en un corpus de pruebas de aprendientes de Español como Lengua Extranjera con niveles A2 y B2 de una universidad privada de la ciudad de Medellín. Como objetivos específicos, se pretendió identificar la DL de los centros de interés: lugares de la ciudad, alimentos-bebidas y medios de transporte en aprendientes de niveles A2 y B2 y comparar la Disponibilidad y Riqueza Léxicas en los centros de interés seleccionados entre aprendientes A2 y B2, y asimismo con un grupo control que es un grupo de hablantes nativos de español conformado por estudiantes universitarios de la Universidad de Antioquia.

Para la obtención de los datos, se utilizó la fórmula matemática de DL propuesta López Chávez y Strassburger Frías (2000) y para RL se empleó la fórmula propuesta por López Morales (2010). Para el procesamiento automático de las fórmulas y la organización y el análisis de los resultados, se crearon dos softwares en lenguaje Python.

En el corpus de esta investigación se contó con una muestra de 18 estudiantes que se encontraban realizando un curso de A2 y 14 estudiantes que se encontraban realizando un curso de B2, además de un grupo control conformado por 23 hablantes nativos de español que estudian Licenciatura en Literatura y Lengua Castellana en la Universidad de Antioquia. En cuanto a los instrumentos utilizados, estos correspondieron a dos muestras escritas. La primera correspondía al instrumento para recoger los datos para DL, de esta manera, cada estudiante recibió la indicación de escribir una lista de palabras relacionadas con cada centro de interés seleccionado (la ciudad, alimentos-bebidas y medios de transporte) en un tiempo máximo de dos minutos por centro de interés; y la segunda muestra escrita, es decir, la muestra para RL, consistió en un texto expositivo en relación con los centros de interés seleccionados para esta investigación.

En este trabajo tanto en DL, como en RL se encontraron errores léxicos y también interferencias de otras lenguas, mayormente del inglés, sin embargo, para el tamaño de los datos recogidos, estos no representaron mucha cantidad, lo que puede significar que los aprendientes de ELE tienen buenos niveles léxicos.

Además, en los tres centros de interés se encontraron vocablos que corresponden a un léxico muy contextualizado en la ciudad de Medellín, esto puede ser debido a que los estudiantes llevan a cabo un mecanismo semántico-cognitivo para recuperar el léxico.

Por otra parte, en cuanto a la DL, se encontraron diferencias significativas entre los grupos de A2 y B2, donde los aprendientes de B2 tienen un mayor léxico disponible, lo que concuerda con el Marco Común de Referencia Europeo, pero no se encontraron diferencias significativas entre B2 y el grupo control, lo que posiblemente se debe a que los aprendientes de B2 tienen un nivel más alto que se acerca al conocimiento del léxico de los nativos.

En cuanto a la riqueza léxica, se encontraron diferencias en la extensión, donde en B2 se escribieron textos más largos, posiblemente por la capacidad de los aprendientes de escribir con estructuras gramaticales más complejas y por lo tanto textos más extensos. Sin embargo, no se

encontraron diferencias significativas en RL entre ninguno de los tres grupos, y esto se evidencia al realizar la lectura de los textos, ya que si bien hay diferencias en las estructuras gramaticales y sintácticas, en cada uno de los niveles hay una selección y un uso adecuado del léxico, esto se debe posiblemente a los métodos de enseñanza en las instituciones de ELE, en las que se fortalece la producción escrita y, a que la riqueza léxica, no depende de la extensión de los textos.

Asimismo, se pusieron en relación los resultados encontrados en DL con los de RL y se encontró que existe una relación en la que los aprendientes con mayor número de palabras disponibles tenían a su vez mayor riqueza léxica.

Finalmente, con este trabajo se pudo evidenciar que es importante la elaboración de herramientas de lingüística computacional que permitan el análisis de los datos alrededor de los temas de DL y RL como el que se realizó para esta investigación y que los estudios del léxico son importantes, ya que desarrollar la competencia léxica es crucial para el aprendizaje de una lengua extranjera (Llach, 2017).

Referencias bibliográficas

Calderón, D; Mosquera, J. (2018). Disponibilidad léxica del español hablado en Granada (Meta). Tesis de Maestría. Tunja: Universidad Pedagógica y Tecnológica de Colombia.

Garzón, A. (2016). Disponibilidad léxica en estudiantes de primer semestre de pregrado en una institución universitaria de Villavicencio, Colombia. *Forma y Función*, 29 (1), pp. 63-84.

Llach, M. (2017). Aprendizaje de vocabulario en la Lengua Extranjera. *Palabras vocabulario léxico*, (1), pp.17-33.

López, H. (2010). Los índices de riqueza léxica y la enseñanza de lenguas. En: Del texto a la lengua: La aplicación de los textos a la enseñanza-aprendizaje del español L2-LE coord. por Javier de Santiago Guervós, pp. 15-28.

López, A. (2014). *Desarrollo de los estudios de disponibilidad léxica en Español Lengua Extranjera (ELE)*. En N. Contreras. La enseñanza del Español como LE/L2 en el siglo XXI (397-408).

Mateus, G; Santiago, W. (2006). Disponibilidad léxica en estudiantes bogotanos. *Folios*, *1*(24), pp. 3-26.

Strassburger, C; López, J. (2000). El diseño de una fórmula matemática para obtener un índice de disponibilidad léxica confiable. Anuario de letras: lingüística y filología, 38, pp. 227 – 251.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

El *Corpus de Estilo Indirecto Libre en Español (CEILE)*: proceso de elaboración y propuesta de explotación para un estudio gramatical del discurso narrativo

Noelia Estévez Rionegro[56]
[nrionegro@gmail.com]
Universidade de Vigo, España

Palabras clave
Lingüística de corpus, Análisis del discurso, Gramática, Estilo indirecto libre, CEILE

Resumen
En la narrativa, el deseo de innovación y ruptura con los procedimientos discursivos habituales lleva a los autores a explorar nuevos mecanismos para la reproducción de las palabras y los pensamientos de los personajes. De este modo, se produce el desarrollo de las secuencias de estilo directo y estilo indirecto, que desembocan en el estilo indirecto libre (Li, 1986).

Este último ha sido estudiado, principalmente, desde el ámbito de la Literatura y, aunque también existen estudios lingüísticos que lo exploran como construcción gramatical, siguen existiendo importantes lagunas en torno a la configuración sintáctica y discursiva de los enunciados. Además, en el aspecto gramatical, el estilo indirecto libre es objeto de disquisición entre los estudiosos del tema, y no solo ha recibido distintas denominaciones, sino que también ha sido caracterizado a partir de diversas teorías, en muchos casos, opuestas. En general, las investigaciones sobre la gramática del estilo indirecto libre abordan el análisis de las construcciones a partir de su relación con las demás formas de discurso reproducido, esto es, el estilo directo y el indirecto; sin embargo, mientras autores como Tobler (1894) o Lerch (1919) consideran el estilo indirecto libre una mezcla de los anteriores, otros como Verdín (1970), Domínguez (1971), Hernadi (1972), Voloshinov (1973) o Hickmann (1993) sostienen que no se trata de un simple mecanismo de mezcla o suma aritmética de dos formas, sino de una completamente nueva.

[56] Licenciada en Filología Hispánica y Filología Románica y doctora internacional en Lingüística con Premio Extraordinario por la Universidade de Santiago de Compostela. Ha participado en diferentes proyectos de investigación en Lingüística española y gallega en el Departamento de Lengua española de la Universidade de Santiago de Compostela, el Instituto da Lingua Galega, la Real Academia Española y el Centro Ramón Piñeiro para a Investigación en Humanidades. Ha impartido docencia en la Universidade de Santiago de Compostela, la Universidad Internacional de la Rioja y, en la actualidad, es profesora del Departamento de Lengua Española de la Universidade de Vigo.

Ante esta divergencia de opiniones, resultaría interesante realizar un estudio empírico y explorar el estilo indirecto libre a partir de un corpus de lengua real que recoja un número considerable de ejemplos, procedentes de distintas fuentes narrativas, y facilite la búsqueda de ciertos elementos característicos de la construcción. Dadas sus propiedades formales y semánticas, los enunciados de estilo indirecto libre no pueden seleccionares como tales con los patrones de búsqueda de los corpus existentes en español, por lo que parecía necesaria la creación de uno de elaboración propia. Así, el *Corpus de Estilo Indirecto Libre en Español* surge con la intención de contribuir a llenar modestamente una laguna en la Lingüística de corpus, al proporcionar un conjunto de enunciados de índole narrativo cuya anotación permite abordar el análisis a partir de ciertos patrones gramaticales que, además, comparte con el estilo directo y el estilo indirecto; lo que favorece, a su vez, el estudio contrastivo con otras formas de expresión del discurso referido.

El *Corpus de Estilo Indirecto Libre en Español* (CEILE) está formado por un conjunto de seiscientos cinco ejemplos de estilo indirecto libre, extraídos de siete obras representativas de la narrativa española contemporánea. Además, constituye la base documental que da lugar a su versión informatizada, en base de datos Acces, *Corpus Informatizado de Estilo Indirecto Libre en Español*, que permite al usuario realizar ordenaciones por palabras clave, señales demarcativas, etc.

Los datos del corpus se organizan en cinco bloques, que responden a la siguiente organización: la primera columna recoge el número de identificador del ejemplo; la segunda columna, la clave que identifica cada obra narrativa; la tercera columna, la página de la edición manejada en la que se encuentra el ejemplo; la cuarta columna, el ejemplo; y la quinta y última, el verbo que actúa como señal demarcativa (Girón Alconchel, 1989 y 2002) del acto comunicativo que supone la reproducción de un discurso verbal o mental (aunque no siempre lo hay y, en ese caso, la casilla correspondiente aparece en blanco).

Los elementos contextuales que señalan la introducción de un acto de habla en el discurso, como indica Girón Alconchel (1989 y 2002) cuando alude a señales demarcativas, son señales que sirven para anunciar la introducción de un discurso reproducido y son necesarias para poder interpretarlo como tal en el conjunto del texto. Son anotadas en el corpus solo cuando constituyen formas verbales, ya que su análisis podría revelar similitudes el estilo directo, el estilo indirecto y las variantes de ambos, donde la expresión introductora del discurso referido es, casi siempre, una forma verbal (Estévez-Rionegro, 2017).

El análisis de los enunciados del corpus, con especial atención a las señales demarcativas, permite analizar exhaustivamente la configuración de las construcciones de estilo indirecto libre y facilita el estudio contrastivo con otras formas de reproducción del discurso. Puede contribuir, además, a constatar o a cuestionar hipótesis defendidas con anterioridad en torno a la materia y a esbozar otras nuevas, por ejemplo, su posible conexión con las apenas estudiadas

construcciones de estilo directo con verbos atípicos, recogidas en el *Corpus de Estilo Directo Atípico en Español* (Estévez-Rionegro, 2020).

Relación de obras que conforman el *Corpus de Estilo Indirecto Libre en Español*:

Chirbes, Rafael (1996): *La larga marcha*. Barcelona: Anagrama.

Freire, Espido (1999): *Melocotones helados*. Barcelona: Planeta.

Grandes, Almudena (2002): *Los aires difíciles*, 7ª ed. Barcelona: Tusquets.

Mas, José (1993): *El lenguaje de las fuentes*, edición de 2003. Madrid: Cátedra.

Mayoral, Marina (1998): *Recuerda, cuerpo*. Madrid: Alfaguara.

Muñoz Molina, Antonio (1991): *El jinete polaco*, edición de 1992. Barcelona: RBA Editores.

Puértolas, Soledad (1989): *Queda la noche*, edición de 2001. Barcelona: Bibliotex.

Bibliografía

Domínguez, P. (1975). *El discurso indirecto libre en la novela argentina*. Porto Alegre: Pontificia Universidade Católica do Río Grande do Sul.

Estévez-Rionegro, N. (2020 a). Corpus de Estilo Indirecto Libre en Español (CEILE) *Linred: Lingüística en la red*, 17, pp. 141.

Estévez-Rionegro, N. (2020 b). Corpus de Estilo Directo Atípico en Español (CEDAE). *Linred: Lingüística en la red*, 17, pp. 1-18.

Estévez-Rionegro, N. (2017). *Las construcciones de estilo directo en español*. Estudio de corpus (tesis doctoral). Universidad de Santiago de Compostela, Santiago de Compostela.

Estévez-Rionegro, N. (2016). *Corpus Informatizado de Estilo Indirecto Libre en Español*. Número de asiento registral registral 03/2017/1093 del Registro de la Propiedad Intelectual.

Girón, J. L. (1989). Las formas del discurso referido en el "Cantar de Mio Cid". *Boletín de la Real Academia Española*, anejo XLIV.

Girón, J. L. (2002). Discurso indirecto libre y autobiografía en la Vida del capitán Contreras. En C. Saralegi y M. Casado (Eds.), *Pulchre, bene, recte. Estudios en homenaje al prof. Fernando González Ollé* (pp. 625-638). Barañáin: EUNSA Ediciones Universidad de Navarra.

Hernadi, P. (1972). *Beyond Genre. New directions in Literary Classification*. Londres: Cornell University Press.

Hickman, M. (1993). The boundaries of reported speech in narrative discourse: some develomental aspects. En J.A. Lucy (Ed.), *Reflexive language. Reported speech and metapragmatics* (pp. 481-498). Cambridge: Cambridge University Press.

Lerch, G. (1919). *Uneigentliche direkte Rede*. Munich: Diss.

Li, Ch. (1986). "Direct speech and indirect speech: A functional study". En F. Coulmas (Ed.), *Direct and Indirect Speech* (pp. 29-45) Berlín: Mouton de Gruyter.

Tobler, A. (1894). *Vermischte Beiträge zur französischen Grammatik II*. Leipzig: S. Hirzel.

Verdín, G. (1970). Introducción al estudio indirecto libre en español. *Revista de Filología Española*, anejo CXV.

Voloshinov, V. (1976). *El signo ideológico y la filosofía del lenguaje*. Buenos Aires: Ediciones Nueva Visión.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Estadística predictiva para el análisis de la oralidad. Diseño de una propuesta para la explicación funcional de los verbos cognitivos

María Amparo Soler Bonafont[57]
[m.amparo.soler@uv.es]
Universidad Internacional de Valencia, España

La explicación funcional, esto es, del uso semántico-pragmático, de algunas unidades epistémicas en los textos resulta aún hoy compleja. Tal es el caso, frecuente en la oralidad, de algunas formas verbales de primera persona del singular del presente de indicativo: *creo*, *pienso*..., conocidas como verbos cognitivos o de opinión (Fetzer y Johansson 2010, Fetzer 2014, González Ruiz 2015, Soler 2018). Estas formas verbales son subjetivas y, en algunas ocasiones, pueden manifestarse de manera integrada o parentética, desde el punto de vista morfo-sintáctico. No obstante, estas características que las identifican no son tan llamativas como otros de sus rasgos definitorios, los cuales dificultan el reconocimiento de sus usos: estos son su polisemia y su polifuncionalidad (Hartwell *et al.* 2017, Jansegers 2018, Soler 2019).

Los diferentes significados y funciones que pueden manifestar unidades como *creo* han sido estudiadas en diferentes géneros (tanto en español como en otras lenguas), entre los que destacan los de interacción oral, especialmente, la conversación y el debate parlamentario. Así bien, incluso en estos, *creo* y verbos semejantes a él manifiestan desde funciones intensificadoras hasta atenuantes (Cutting 2007; Fuentes Rodríguez 2010, 2016; De Hoop *et al.* 2018), a la vez que despliegan una gran variedad de valores semánticos, desde la creencia hasta el juicio (Soler 2019). Distinguir la multiplicidad de sus posibilidades semántico-pragmáticas no es tarea sencilla para el lingüista, que topa, desde hace más de un siglo, con un escollo adicional en estos verbos: la limitación de las herramientas lingüísticas tradicionales. Los pragmatistas se preguntan cómo definir los significados y significados en uso de unidades subjetivas como las que son objeto de este trabajo, para los que no son suficientemente explicativas las pruebas veritativo-condicionales ni las de la pragmática clásica. Por estos motivos, son cada vez más numerosos los estudios que realizan una aproximación cognitiva a estas formas, gracias a su concepción de la semántica y de la pragmática como un mero *continuum* (Achard 1998,

[57] MA. Soler es Licenciada en Filología Hispánica y Doctora en Estudios Hispánicos Avanzados por la Universitat de València. Actualmente es Profesora Doctora Contratada en la Universidad Internacional de Valencia (VIU). Sus áreas de investigación principales son la atenuación, la epistemicidad y los verbos doxásticos. En torno a estos aspectos, la profesora Soler ha realizado distintas publicaciones en editoriales (EUNSA, ELUA, etc.) y revistas de prestigio (Signos, RILI, Rilce, etc.) (inter)nacionales, así como presentado ponencias en distintas universidades. También es miembro del comité científico de diferentes revistas de semántica y pragmática (Oralia, Semas, ForoELE, etc.) y cuenta con distintas colaboradores en proyectos y universidades internacionales.

Jansegers y Gries 2010, Buceta 2014, Jansegers 2017, Boas y Ziem 2018), lo que ayuda a superar algunos obstáculos definitorios.

No obstante, y de acuerdo con diferentes estudios pragmáticos y sociolingüísticos recientes (Díaz-Campos y Gradoville 2011, González *et al.* 2014), la explicación cualitativa cognitiva queda incompleta si no se realiza un análisis riguroso de corpus, de tipo cuantitativo (Roldán 2005, Abdulrahim 2014, Milin *et al.* 2016).

El objetivo de esta investigación es, por consiguiente, realizar una propuesta de descripción de los usos semántico-pragmáticos de las formas performativas de verbos cognitivos como *creo* en la interacción oral, de corte cognitivo, y apoyado en el diseño de una propuesta estadística predictiva, a partir de un sistema de regresiones multinomiales programado para tal fin (a través de herramientas como STATA). Perseguimos que el modelo diseñado permita reconocer con un elevado grado de explicatividad ante qué significados y funciones pragmáticas de la unidad nos encontramos, una vez sistematizadas las principales circunstancias de aparición que los rodean. Para ello, realizamos un estudio de las formas en un corpus de conversaciones coloquiales (*COGILA*, *COJEM*, *Val.Es.Co. 2002*, *Val.Es.Co. 2.0*) y de discursos de debate parlamentario (*Congreso de los Diputados* del Gobierno de España y *Les Corts Valencianes*), sobre los que aplicamos las bases de la estadística (Abbhul y Mackey 2013, James *et al.* 2013). Se llega a obtener un diseño predictivo propio, si bien replicable en otros textos y géneros en los que aparezcan unidades epistémicas similares.

Palabras clave

creo, *verbos cognitivos*, *opinión*, *semántica cognitiva*, *estadística predictiva*

Bibliografía:

Abbhul, R. y M. Mackey (2013): "Experimental research design", en abbuhl, R., S. Gass y M. Mackey, *Research Methods in Linguistics*, Cambridge, Cambridge University Press.

Abdulrahim, D. (2014): "Annotating corpus data for a quantitative, constructional analysis of motion verbs in Modern Standard Arabic", en *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing* (*ANLP*), Doha, Qatar, 28-38.

Achard, M. (1998): "Representation of cognitive structures", *Cognitive Linguistics*, 15 (4), 588-594.

Boas, H. y A. Ziem (2018): "Constructing a constructicon for German. Empirical, theoretical, and methodological issues", en lyngfelt, B., L. Borin, K. Ohara y T. Timponi (eds.), *Constructicography: Constructicon development across languages*, Amsterdam/Philadelphia, John Benjamins, 183-228.

Buceta, O. (2014): "Construcciones del verbo 'creer'", *Factótum*, 12, 74-90.

Cutting, J. (ed.) (2007): *Vague Language Explored*, London, Palgrave MacMillan.

De Hoop, H., A. Foolen, G. Mulder y V. Van Mulken (2018): "*I think* and *I believe*: Evidential expressions in Dutch", en Foolen, A., H. de Hoop y G. Mulder (eds.), *Evidence for Evidentiality*, Amsterdam/Philadelphia, John Benjamins, 77-97.

Díaz-Campos, M. y M. Gradoville (2011): "An Analysis of Frequency as a Factor Contributing to the Diffusion of Variable Phenomena: Evidence from Spanish Data", en ortiz, L. (ed.), *Selected Proceedings of the 13th Hispanic Linguistics Symposium*, Cascadilla Proceedings Project.

Fetzer, A. (2014): "*I think*, *I mean* and *I believe* in political discourse. Collocates, functions and distribution", *Functions of Language*, 21/1, 67-94.

Fetzer, A. y M. Johansson (2010): "Cognitive verbs in context. A contrastive analysis of English and French argumentative discourse", *International Journal of Corpus Linguistics*, 15:2, 240-266.

Fuentes Rodríguez, C. (2010): "La aserción parlamentaria: de la modalidad al metadiscurso", *Oralia*, 13, 97-125.

Fuentes Rodríguez, C. (2016): "Atenuación e intensificación estratégicas", en Fuentes Rodríguez, C. (ed.), *Estrategias argumentativas y discurso político*, Madrid, Arco/Libros, 163-222.

González Ruiz, R. (2015): "Los verbos de opinión entre los verbos parentéticos y los verbos de rección débil: aspectos sintácticos y semántico-pragmáticos", *Círculo de Lingüística Aplicada a la Comunicación*, 62, 148-173.

González, J., P. Boeck y F. Tuerlinchx (2014): "Linear mixed modelling for data from a double mixed factorial design with covariates: a case-study on semantic categorization response times", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63, 2, 289-302.

Hartwell, L. M.; E. Esperança-rodier y A. Tutin (2017): "*I think we need...*: Verbal expressions of opinion in conference presentations in English and in French", *Romance Corpora and Linguistic Studies*, 4.1, 35-60.

James, G., D. Witten, T. Hastie y R. Tibshirani (2013): An Introduction to Statistical Learning: With Applications in R, Nueva York, Springer.

Jansegers, M. (2017): *Hacia un enfoque múltiple de la polisemia. Un estudio empírico del verbo multimodal "sentir" desde una perspectiva sincrónica y diacrónica*, Berlin/Boston, Mouton de Gruyter.

Jansegers, M. y S. Gries (2010): "Towards a dynamic behavioral profile: a diachronic study of polysemous 'sentir' in Spanish", *Corpus Linguistics and Linguistic Theory*,1-43.

Milin, P., D. Divjak, S. Dimitrijević y R. H. Baayen (2016): "Towards cognitively plausible data science in language research", *Cognitive Linguistics*, 27, 4, 507-526.

Roldán, A. (2005): "Applications of cognitive linguistics (CI) to languages for specific purposes (LSP)", en CARRIÓ, M. L. (coord.), *Perspectivas interdisciplinares de la lingüística aplicada*, Vol. 2, 325-332.

SOLER, M. A. (2018): "Algunos apuntes bibliográficos en torno a los verbos de opinión", ÁLVAREZ LÓPEZ, C. J. y M. R. MARTÍNEZ NAVARRO (coords.), *En busca de nuevos horizontes. Algunas líneas actuales en los estudios hispánicos*, Braga, Ediçoes Húmus, 59-70.

SOLER, M. A. (2019): *Semántica y pragmática de los verbos doxásticos en la interacción oral en español. Un estudio monográfico sobre la forma verbal* creo. Tesis doctoral. Universitat de València.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

ESTECNICOR: Explotación de un corpus de aprendices para la detección y clasificación de errores en la traducción de textos de automoción (inglés-español)

Míriam Pérez Carrasco[58] **& Míriam Seghiri Domínguez**[59]

[miriamperez@uma.es | seghiri@uma.es]

Universidad de Málaga, España

Como es sabido, la Lingüística de Corpus plantea como principal ventaja el estudio de la lengua a través de una colección de ejemplos de actos reales de habla, bien sean escritos u orales, en formato electrónico, elegidos y ordenados siguiendo criterios específicos para que puedan considerarse una muestra lingüística representativa de la lengua objeto de estudio (EAGLES, 1996; Sinclair, 2005). Dentro de los diversos tipos y subtipos de corpus existentes hoy en día, en esta ocasión nos centraremos en los denominados como «corpus de aprendices» o *learner corpora*. Se trata de colecciones de textos auténticos producidos por estudiantes —normalmente de idiomas— almacenados en formato electrónico para que puedan procesarse mediante programas de análisis de corpus. Este tipo de corpus es un recurso de gran valía para la investigación en la adquisición de una segunda lengua o incluso de las competencias propias específicas de disciplinas como la Traducción, pues hacen posible, por ejemplo, detectar las dificultades de aprendizaje de los discentes, los errores recurrentes cometidos por los estudiantes o las tendencias en cuanto a la terminología, las colocaciones, locuciones o fraseología que emplean en sus traducciones (Buendía Castro, 2020).

En el presente trabajo nos ocuparemos, en concreto, de compilar un corpus bilingüe (inglés-español) de aprendices integrado por textos producidos por los estudiantes de la asignatura Traducción Científico-Técnica I, correspondiente al tercer curso del Grado en Traducción e

[58] Graduada en Traducción e Interpretación (2015), Máster en Profesorado de Educación Secundaria Obligatoria, Bachillerato, Formación Profesional y Enseñanza de Idiomas y Máster en Estudios Ingleses y Comunicación Multilingüe e Intercultural por la Universidad de Málaga. De 2017 a 2019 trabajó como Profesora Sustituta Interina en el Dpto. de Traducción e Interpretación de la Universidad de Málaga, donde actualmente continúa como Personal Investigador en Formación con un contrato FPU concedido por el Ministerio de Educación, Ciencia y Deporte. Sus líneas de investigación abarcan la traducción especializada, la lingüística del corpus y las nuevas tecnologías aplicadas a la traducción e interpretación.

[59] Licenciada y Doctora en Traducción e Interpretación y Graduada en Derecho. En la actualidad es profesora titular en el Departamento de Traducción e Interpretación de la Universidad de Málaga. Además, ha impartido docencia en la universidad norteamericana de Dickinson College y en la universidad británica de Cambridge, así como en las universidades españolas de Córdoba y Murcia. Cuenta con una patente (N-Cor). Sus líneas de investigación abarcan la traducción especializada, la lingüística de corpus y las tecnologías de la traducción y la interpretación (Premio extraordinario de doctorado, Premio en Tecnologías de la Traducción, Premio María Zambrano y Premio George Campbell).

Interpretación de la Universidad de Málaga. Así, los discentes tuvieron que enfrentarse a la traducción de fragmentos de texto extraídos de manuales de usuario del campo de la automoción y, en particular, de motocicletas. Hemos elegido esta temática pues en Europa son varias las directivas y normas que regulan, e incluso exigen, que este tipo de género textual sea traducido. De hecho, la *Directiva 2006/42/CE del Parlamento Europeo y del Consejo, de 17 de mayo de 2006, relativa a las máquinas y por la que se modifica la Directiva 95/16/CE* establece que «cada máquina deberá ir acompañada de un manual de instrucciones en la lengua o lenguas oficiales comunitarias del Estado miembro donde se comercialice y/o se ponga en servicio la máquina». Por su parte, la *Resolución del Consejo de 17 de diciembre de 1998 sobre las instrucciones de uso de los bienes de consumo técnico (98/C 411/01)* subraya que estas traducciones deben tener en cuenta «las características culturales distintivas de la zona en la que se usa el idioma correspondiente», lo que requiere, citando a la mencionada normativa, «que las traducciones sean hechas por expertos con la formación adecuada, que utilicen el idioma de los consumidores a los que está destinado el producto, y que, en la medida de lo posible, sean sometidas a una prueba de comprensión de los consumidores».

En este trabajo ilustraremos, en primer lugar, cómo se ha compilado el corpus de aprendices, al que denominaremos ESTECNICOR. Para ello, por una parte, explicaremos los criterios de diseño establecidos para la compilación de dicho corpus, basándonos en los criterios propuestos por Bowker y Pearson (2002), a saber, a) propósito de la compilación, b) tamaño del corpus, c) medio, d) temática, e) tipo textual, f) autoría, g) fecha de publicación, y h) lengua o lenguas de redacción. Por otra parte, desarrollaremos el protocolo de compilación de corpus establecido siguiendo los cuatro pasos propuestos por Seghiri (2006, 2008, 2010, 2011, 2015, 2017a y 2017b), a saber, búsqueda de la información, descarga, formato y almacenamiento.

Posteriormente, procederemos al etiquetado del corpus ESTECNICOR mediante el software TagAnt y procederemos a su explotación a través de un programa de concordancias, para lo que nos valdremos de un corpus de referencia de manuales originales, al que hemos llamado TECNICOR, diseñado para tal fin. Esto nos permitirá analizar y clasificar los errores detectados en las traducciones propuestas por los propios estudiantes. En este sentido, se hará especial hincapié en aquellos errores que tengan que ver con la terminología técnica propia del género textual en cuestión así como con la utilización de colocativos erróneos u otros aspectos gramaticales típicos de estos textos como son la alteración del orden de los elementos en la premodificación nominal múltiple o una interpretación incorrecta de este tipo de estructuras en inglés.

En definitiva, este estudio nos permitirá, como docentes, detectar y analizar los errores más frecuentes cometidos por los discentes y realizar, a partir de ello, propuestas para la mejora de sus destrezas traductológicas así como diseñar materiales pedagógicos para la enseñanza-aprendizaje de la traducción técnica inglés-español.

Palabras clave: corpus de aprendices, *learner corpora*, traducción técnica, manuales de usuario, automoción.

BIBLIOGRAFÍA

Bowker, Lynne y Pearson, Jenny (2002): *Working with Specialized Language: A practical guide to using corpora*. London: Routledge.

Buendía Castro, Miriam (2020): «Un estudio de caso basado en corpus sobre el uso de las colocaciones», en Miriam Seghiri (ed.), *La lingüística de corpus aplicada al desarrollo de la competencia tecnológica en los estudios de traducción e interpretación y la enseñanza de segundas lenguas*. Berlín: Peter Lang.

EAGLES (Expert Advisory Group on Language Engineering Standards) (1996): *Text corpora Working Group reading Guide. EAGLES Document EAG-TCWG-FR-2*. Versión de mayo de 1996. Recurso en línea <http://www.ilc.cnr.it/EAGLES/corpintr/corpintr.html> [07/06/2020].

Seghiri, Miriam (2006): *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad*. Tesis doctoral. Málaga: Universidad de Málaga.

Seghiri Domínguez, Miriam (2008): «Creating a virtual corpora step by step», en Purificación Sánchez Hernández *et al.* (eds.), *Researching and Teaching specialized languages: New contexts, new challenges*. La Manga, Murcia: Universidad de Murcia, Servicio de Publicaciones, Editum.

Seghiri Domínguez, Miriam (2010): «Metodología de diseño y compilación de un corpus representativo de seguros turísticos», en Rafael López Campos Bodineau, Carmen Balbuena Torezano y Manuel Álvarez Jurado (eds.), *Traducción y modernidad: textos científicos, jurídicos, económicos y audiovisuales*. Córdoba: Universidad de Córdoba, 59-70.

Seghiri, Miriam (2011): «Metodología protocolizada de compilación de un corpus de seguros de viaje: aspecto de diseño y representatividad», *Revista de lingüística teórica y aplicada*, 49(2), 13-30.

Seghiri, Miriam (2015): «Determinación de la representatividad cuantitativa de un corpus ad hoc bilingüe (inglés-español) de manuales de instrucciones generales de lectores electrónicos/Establishing the quantitative representativiness of an E-Reader User's Guide *ad hoc* corpus (English-Spanish)», en María Teresa Sánchez Nieto (ed.), *Corpus-based Translation and Interpreting Studies: From description to application*. Berlín: Frank & Timme, 125- 146.

Seghiri, Miriam (2017a): «Corpus e interpretación biosanitaria: extracción terminológica basada en bitextos del campo de la Neurología para la fase documental del intérprete (Corpora and medical interpreting: terminology extraction based on bitexts for the interpreter's documentation process)», *Panacea*, 18(46), 123-132.

Seghiri, Miriam (2017b): «Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores», *Babel. Revue Internationale de la Traduction,* 63(1), 43-64.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Normativas europeas

Directiva 2006/42/CE del Parlamento Europeo y del Consejo, de 17 de mayo de 2006, relativa a las máquinas y por la que se modifica la Directiva 95/16/CE. Disponible en línea <https://eur-lex.europa.eu/legal-content/es/TXT/?uri=CELEX:32006L0042> [último acceso: 08/06/2020].

Resolución del Consejo de 17 de diciembre de 1998 sobre las instrucciones de uso de los bienes de consumo técnicos (98/C 411/01). Disponible en línea <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:31998Y1231(02)&from=ES> [último acceso: 08/06/2020].

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Frecuencia, ubicación y adecuación de marcadores argumentativos no complejos de polifonía textual como indicadores de adquisición de lenguaje académico

Germán Alfredo Kohli[60], María Virginia Bruzzo[61] & Vanina Evelyn Zarza[62]

[kohligermanalfredo@hotmail.com | virginiabruzzo@gmail.com | evezarza7@gmail.com]

Universidad Nacional del Nordeste, Argentina

Sobre la base del estudio continuado de distintos corpus conformados por trabajos académicos de alumnos de nivel universitario hemos arribado a una precomprensión modelizante (Ladrière, 1978; Samaja, 2002 [1999]), que propone que los estudiantes universitarios de grado, tanto los ingresantes como los que cursan las últimas materias, no alcanzan, en la práctica, a desarrollar *plenamente* la capacidad escritural de presentar su propia voz a la hora de argumentar, ni la capacidad de sortear requerimientos técnicos académicos (e. g., cita y parafraseo adecuados). ¿Qué información aportaría realizar un seguimiento de los marcadores del discurso que indican polifonía en este contexto?, ¿la ubicación de dichos marcadores en el enunciado tendrá vinculación con el tipo de argumentación ineficiente que articulan los estudiantes?, ¿qué clase de marcadores prefieren usar los estudiantes y con qué frecuencia los articulan?, ¿qué grado de adecuación guardan los marcadores preferidos con la normativa española y académica?, ¿qué teoría de la adquisición podría explicar que pasados cinco o más años los estudiantes no incorporen ciertas prácticas en lo relativo a marcadores del discurso que indican polifonía?

En este sentido, la presente comunicación científica ilustra los resultados preliminares de una investigación situada que releva (a) la posición en el enunciado, (b) la frecuencia de uso y (c) la adecuación técnica de ciertos marcadores del discurso destinados a indicar polifonía en textos académicos recabados durante los años 2017, 2018 y 2019 y producidos por una población de estudiantes perteneciente a las carreras de Profesorado en Letras y Licenciatura en Letras de la Universidad Nacional del Nordeste (Argentina).

La muestra está constituida por un corpus de 50 documentos equivalentes en bruto a 330 carillas o cuartillas de tamaño A4 cuya escritura fue codificada y articulada mediante diferentes versiones (2007, 2010, 2013, 2016) del procesador de texto multiplataforma *Microsoft Word* (del paquete *Microsoft Office*), predominante en el contexto estudiado. El género académico en el que se enmarcó la práctica escritural que produjo la muestra es el género *ensayo* (Zunino y Muraca, 2012). La *situación* escritural (Camps Mundó y Castelló Badía, 2013) es la de

[60] Licenciado en Letras. Miembro del Equipo de Investigación del PI16H009.

[61] Licenciado en Letras. Miembro del Equipo de Investigación del PI16H009.

[62] Profesora en Lengua y Literatura.

evaluación no-presencial final (domiciliaria), de cuyo resultado depende la *regularización* de la materia, esto es: obtener el derecho a ser evaluado por un jurado de cara a la aprobación definitiva del Seminario de Literatura Contemporánea en Lenguas no Hispánicas.

La población escogida cumple con la condición de encontrarse próxima a obtener el título de *profesor* o *licenciado en letras*, rasgo distintivo útil para (a) definir y abordar una porción de población faltante orientada a completar una investigación preexistente que analizó la escritura polifónica de estudiantes *ingresantes* e *intermedios* de las mismas carreras durante el mismo periodo y (b) para construir un corpus que refleje el estado prácticamente acabado de la capacidad escritural de inminentes profesionales del ámbito de la educación, investigación y asesoramiento lingüístico (alcances de los títulos de *profesor* y *licenciado en letras*), todo lo cual permitiría especular acerca de la capacidad escritural real de los graduados.

Aunque los marcadores del discurso relevados se agrupan en torno de dos conjuntos, el de *marcadores no complejos* (signos puntuales) y de *marcadores complejos* (estructuras sintácticas), en esta oportunidad comunicaremos los resultados preliminares del relevamiento de los *marcadores no complejos*, conjunto que reúne signos compatibles con (a) las comillas inglesas de apertura y cierre (“”), (b) las comillas inglesas simples de apertura y cierre (‘’), (c) las comillas angulares de apertura y cierre («»), (d) recursos ortotipográficos (espaciado especial, margen especial o tipografía especial) y (e) signos inesperados (<<, >>, ´´, etcétera), a los que la población atribuye presumiblemente la misma función (marcar y delimitar discurso ajeno). Estos *marcadores no complejos*, en efecto, pueden ser naturalmente duales o dobles y constituir por tanto dos signos distintos (de apertura y cierre, por ejemplo). A su vez, también pueden manifestarse −prescindiendo de toda realización independiente− a través del tratamiento ortotipográfico mismo (por ejemplo, a través de la espacialidad del margen especial que indica una cita de más de 40 palabras o bien a través de la transición redonda-cursiva que la fuente deje ver).

Los resultados preliminares indican que, en cuanto a (a) la posición en el enunciado de los marcadores, la población tiende a articular polifonía especialmente al final de construcciones enunciativas (ya sean oraciones o párrafos), lo cual sería compatible con estrategias argumentativas que se valen de la voz de otro (fuente disciplinar/autoridad) para resolver una polémica instalada, para refutar una posición, para legitimar un punto de vista o bien para delegar una proposición compleja, todo lo cual resulta útil para evaluar si la población tiende a resolver pasajes críticos del texto echando mano de voces ajenas que, en ocasiones, llegan a reemplazar o sustituir el argumento mismo del estudiante que escribe. En lo relativo a (b) la frecuencia de uso, el número de *marcadores no complejos* de polifonía se ubica preliminarmente en torno de una tasa de 4 ocurrencias por carilla, lo cual es equivalente a 4 ocurrencias cada 332,48 palabras o a 1 *marcador no complejo* de polifonía cada 83 palabras. Finalmente, en cuanto a (c) la adecuación técnica de los marcadores relevados, existiría una relativa y variable compatibilidad con diversos formatos antiguos o contemporáneos de estilos estandarizados (APA, Chicago, MLA, Vancouver, Harvard, entre otros) pero una mínima compatibilidad con la normativa española básica, lo cual puede observarse en la prevalencia casi absoluta del uso

de comillas inglesas por sobre las angulares (propias del español), llegando a constituir una tasa de 1 comilla angular cada 214 comillas inglesas normales. Asimismo, se observan convivencias incoherentes de dos o más *recursos no complejos* orientados a marcar polifonía pero que destacan una sola voz (comillas y espaciado especial en simultáneo, por ejemplo, o bien comillas y cursivas en simultáneo).

Estos resultados, aunque preliminares, aportan sustantivamente a discusiones técnicas y teóricas. A nivel técnico, por ejemplo, los resultados resultan útiles para preguntarse acerca de las dificultades que entraña la convivencia entre estilos escriturales foráneos o estándares anglosajones de escritura académica (APA, Chicago, MLA, Harvard, entre otros) y la normativa propia del español. ¿Son aquellos estilos o estándares completamente aplicables a la redacción de un documento académico en español?, ¿deben diseñarse adecuaciones específicas?, ¿qué otras incompatibilidades se observan además de las que aquí se evidencian?, ¿debe Argentina desarrollar su propio criterio, como hizo Colombia, a través de la creación de las Normas ICONTEC-NTC 1486? Asimismo, la frecuencia de ocurrencias que podría, quizá, considerarse baja, también resultaría útil para pensar una posible clasificación de géneros textuales académicos en función de las voces que incorpora (de menor a mayor, por ejemplo: ensayo, monografía, tesis, etcétera).

Por otro lado, a nivel teórico, si interpretamos los resultados generado conforme lineamientos del modelo de alfabetización semiótica (Camblong y Fernández, 2012) y entendiendo *adquisición* como un proceso de subjetivización a través del cual el lenguaje captura al individuo (Lemos, 2000) en diferentes situaciones a lo largo de la vida, especialmente en el ámbito académico en el cual opera «[…] la escritura como una instancia de adquisición y […] cada género discursivo como una posibilidad de acceso diferente en relación con esa instancia de adquisición» (Desinano, 2009: 79), ¿qué conclusiones podemos comenzar a trazar? La primera conclusión podría sugerir que los estudiantes cuyas producciones fueron relevadas están atravesando (incluso al final de sus estudios) distintas *posiciones* (Desinano, 2009) del proceso de captura del discurso académico, pero que, en general, dicho proceso no está muy avanzado, lo cual explicaría la existencia de dificultades observadas al momento de articular discursos de alto nivel polifónico y los recursos propios que la polifonía demanda, sean éstos superficiales o no (concediendo que puedan ser llamados *superficiales*). Una conclusión de esta naturaleza sería, cuanto menos, llamativa, especialmente porque los estudiantes cuyas producciones fueren relevadas son los mismos estudiantes que, en muy poco tiempo, estarán ejerciendo la docencia y evaluando documentos académicos de alumnos en el sistema público. La segunda conclusión podría sugerir que buena parte de los estudiantes que conforman la población cuyas producciones fueron analizadas no llegó a desarrollar plenamente la capacidad escritural de presentar su propia voz y argumento, sobre todo si tenemos en cuenta que en un número considerable de casos los estudiantes evitaron o delegaron construcciones delicadas o críticas y prefirieron decirlas a través de otras voces, todo lo cual se refleja, al menos preliminarmente, en el crecimiento o aumento de la frecuencia de marcadores no complejos que indican polifonía al final de los enunciados, sección estratégica en que se resuelven tensiones argumentativas

eventualmente decisivas. Esta conclusión, además, se vería inicialmente avalada por el hecho de que, del grupo de ensayos desaprobados por los docentes evaluadores, muchos presentaron altas frecuencias de marcadores de polifonía precisamente al final de enunciados explicativos, lo cual indicaría una presunta correlación entre (a) la alta frecuencia de recursos no complejos de articulación polifónica ubicados al final del enunciado que cierra el párrafo o la oración, (b) la argumentación delegada y (c) la desaprobación o baja nota de ensayos cuya autoría pertenecería, según otros indicadores textuales, a estudiantes que se encontrarían aún atravesando las primeras etapas de la adquisición del lenguaje académico.

Palabras clave: marcadores del discurso, polifonía, argumentación, adquisición.

Bibliografía

Camblong, A. y Fernández, F. (2012). *Alfabetización semiótica en las fronteras. Volumen I. Dinámicas de las significaciones y el sentido*. Posadas: Editorial Universitaria Universidad Nacional de Misiones.

Camps Mundó, Anna y Castelló Badía, Montserrat (2013). "La escritura académica en la universidad", en *Revista de Docencia Universitaria*, volumen 11, número 1, enero-abril, pp. 17-36, disponible en [https://bit.ly/2zNCJxb], consultado el 2 de junio de 2020.

Desinano, N. (2009). *Los alumnos universitarios y la escritura académica. Análisis de un problema*. Rosario, Argentina: Homo Sapiens.

Ladrière, J. (1978). *El reto de la racionalidad. La ciencia y la tecnología frente a las culturas*. Salamanca: Ediciones Sígueme.

Lemos, C. de (2000). "Questioning the notion of development: the case of language acquisition", en *Culture & Psychology*, volumen 6, número 2, pp. 169-182.

Samaja, J. (2002 [1999]). *Epistemología y metodología: elementos para una teoría de la investigación científica*. Buenos Aires: Editorial Universitaria.

Zunino, C. y Muraca, M. (2012). "El ensayo académico", en Natale, Lucía (coordinadora). *En carrera: escritura y lectura de textos académicos y profesionales*. Los Polvorines: Universidad Nacional de General Sarmiento, pp. 61-77, disponible en [https://bit.ly/2YP7A4Y], consultado el 3 de junio de 2020.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Formulas rutinarias: la creación de un corpus del alemán coloquial actual

Karen Lorena Baquero Castro[63]
[klbaquero@universidadean.edu.co]
Universidad de Salamanca, España
Universidad Ean, Colombia

Las fórmulas rutinarias fueron incluidas desde 1973 en las investigaciones de fraseología del alemán por Burger, bajo la denominación de "*pragmatische Phraseme*"[64]. Sin embargo, a pesar de los avances teóricos realizados desde dicho momento en cuanto a sus múltiples funciones en la comunicación oral y escrita, aún no existe consenso entre los investigadores sobre las características definitorias y su clasificación. En la búsqueda de material auténtico que permita indagar sobre dichas unidades y aportar tanto a su construcción teórica como al desarrollo de una didáctica para el alemán[65] coloquial actual, se propone la creación de un corpus para dicha lengua.

El objetivo de esta ponencia consiste en mostrar el proceso que ha seguido la creación de este corpus, compuesto por las líneas de diálogo de la serie de televisión alemana *Türkisch für Anfänger*[66]. Inicialmente, se llevó a cabo el proceso de transcripción de la totalidad de los diálogos de la serie haciendo uso de herramientas de texto, con lo cual se conformó una base de datos del alemán oral compuesta por 12.911 líneas. Esta recopilación consideró la línea de diálogo, el lugar, el capítulo y el hablante. Con ello, se logró caracterizar el contexto de las expresiones lingüísticas dadas. En este proceso, el corpus se conformó a partir del video de todos

[63] Profesional en Lenguajes y Estudios Socioculturales con énfasis en alemán, de la Universidad de los Andes. Culminó su maestría en Estudios Contrastivos de Lengua, Literatura y Cultura Alemana con doble titulación de la Universidad de Salamanca y la Universität Leipzig. Fue becaria del DAAD, Colfuturo, ICETEX y Banco Santander. Ha sido profesora universitaria en el programa de Filología Alemana y de Lenguas Modernas desde 2009, inicialmente en la Universidad Nacional de Colombia y actualmente en la Universidad Ean. Enfoca su investigación en fraseología contrastiva del par de lenguas alemán-español. Es doctoranda de tercer año en el programa de Lenguas Modernas de la Universidad de Salamanca.

[64] A lo largo de la literatura se encuentran diferentes términos para referirse a dichas unidades: Pragmatische Idiome (Burger, 1973), Routineformeln (Coulmas, 1981; Burger, 1998; Stein, 1985; Gläser, 1986; Lüger, 1999; Sosa Mayor, 2006), Kommunikative Formeln (Fleischer, 1982), kommunikative Phraseologismen (Burger, 1998), Kommunikative Routineformeln (Hyvärinen, 2003).

[65] Dentro de los corpus del alemán escrito se cuentan, entre otros: das Deutsche Referenzkorpus - DeReKo (Instituts für Deutsche Sprache, 2018), Digitales Wörterbuch der deutschen Sprache – DWDS (Berlin-Brandenburgischen Akademie der Wissenschaften, s.f.), das Projekt deutscher Wortschatz (Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig, 1998) y el corpus Südtirol (Team Korpus Südtirol, s. f.). Para el alemán oral existe el Datenbank gesprochenes Deutsch - DGD2 (Deppermann & Schmidt, 2014) y el GeWiss (Herder-Institut - Universität Leipzig, s. f.).

[66] Serie de televisión alemana de comedia dramática, producida en los años 2006 a 2008.

los capítulos de la serie y con la indicación de su contexto. A continuación, se cargaron los datos en la herramienta MongoDB[67], con el fin de facilitar su uso, consulta, actualización y análisis. A través de esta plataforma, fue posible aislar datos con criterios específicos.

Con el corpus creado, se desarrolla actualmente una herramienta que posibilita una consulta rápida por contenido, su depuración y observación de estadísticas. Inicialmente, la herramienta que se ha usado para realizar estas tareas es la previamente mencionada y el objetivo final es desarrollar una aplicación sencilla que ofrezca una funcionalidad de consulta a través de un navegador web y facilite interactuar con el corpus de una forma más intuitiva. Además, se considera, en un desarrollo posterior del corpus, agregar las imágenes y el audio de las escenas que incluyen el uso de las fórmulas rutinarias. Dicha información permitirá, no solo identificar el contexto en el que tienen lugar las expresiones, sino observar aspectos de tipo gestual sobre el uso del objeto de estudio dentro del corpus.

El corpus configurado permitirá su aprovechamiento, tanto por parte de aprendices, como de profesores e investigadores de la lengua alemana, al facilitar la creación de aplicaciones didácticas y reflexiones pragmáticas y teóricas. Si bien el estudio presente se concentra en unidades de tipo fraseológico como las fórmulas rutinarias, este puede ser aprovechado en la investigación de diferentes objetos del lenguaje propios de la lengua coloquial. Con el uso de MongoDB, se puede obtener información específica con respecto a la recurrencia, lugar y usuario de las apariciones de diferentes tipos de unidades lingüísticas y no únicamente de las unidades objeto de estudio.

En esta ponencia también se evidenciará qué tipos de datos han sido recopilados y usados para la configuración del corpus, conforme al alemán oral y al modelo de las fórmulas rutinarias, actuales y propias del contexto coloquial de la lengua. Para ello, daremos un ejemplo de fórmulas rutinarias de saludo y cómo estas se pueden articular en las clases de alemán de acuerdo con el nivel del aprendiente según el Marco Común Europeo de Referencia. Así, en la ponencia se resaltarán tres fases: desarrollo y configuración del corpus, análisis de fórmulas de saludo del alemán de acuerdo con los datos del corpus y evaluación del corpus en relación con la enseñanza y el aprendizaje de dichas unidades.

Palabras clave

Corpus lingüístico, fórmulas rutinarias, fraseología, lengua extranjera, alemán coloquial

Bibliografía

Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig, Deutscher Wortschatz / Leipzig Corpora Collection, 1998, http://wortschatz.uni-leipzig.de/de [15.01.2020]

Berlin-Brandenburgischen Akademie der Wissenschaften, DWDS – Digitales Wörterbuch der deutschen Sprache, https://www.dwds.de [11.12.2019]

[67] MongoDB permite crear bases de datos no relacionales.

Burger, H., *Idiomatik des Deutschen (Germanistische Arbeitshefte)*. Tubinga: Niemeyer 1973.

Burger, H., *Phraseologie. Eine Einführung am Beispiel des Deutschen.* Berlín: Erich Schmidt Verlag 1998.

Coulmas, F., *Routine im Gespräch: zur pragmatischen Fundierung der Idiomatik.* Wiesbaden: Athenaion 1981.

Deppermann, A. & Schmidt , T., «Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik: Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD2)», *Mitteilungen des Deutschen Germanistenverbandes* 61(2014), 4–17. doi:10.14220/mdge.2014.61.1.4

Fleischer, W., *Phraseologie der deutschen Gegenwartssprache.* Leipzig: VEB Bibliographisches Institut 1982.

Gläser, R., *Phraseologie der englischen Sprache.* Tubinga: Niemeyer 1986.

Herder-Institut - Universität Leipzig, Gesprochene Wissenschaftssprache, https://gewiss.uni-leipzig.de [25.02.2020]

Hyvärinen, I., «Kommunikative Routineformeln im finnischen DaF-Unterricht», *Info DaF: Informationen Deutsch als Fremdsprache* 4 (2003), 335-351. https://www.academia.edu/1025371/Ein_Terrain_des_Fremdsprachenunterrichts_Deutsch_Interkulturelle_Kompetenz_in_der_Tourismusausbildung [10.02.2020]

Institut für Deutsche Sprache, Ausbau und Pflege der Korpora geschriebener Gegenwartssprache, 2018, http://www1.ids-mannheim.de/kl/projekte/korpora [26.01.2020]

Lüger, H., *Satzwertige Phraseologismen: Eine Pragmalinguistische Untersuchung.* Viena, Austria: Präsens Verlag 1999.

Sosa, I., *Routineformeln im Spanischen und im Deutschen. Eine pragmalinguistische Analyse.* Viena, Austria: Präsens Verlag 2006.

Stein, S. *Formelhafte Sprache. Untersuchungen zu ihren prag-matischen und kognitiven Funktionen im gegenwärtigen Deutsch*. Frankfurt am Main u. a.: Lang 1985

Team Korpus Südtirol, Korpus Südtirol, http://www.korpus-suedtirol.it/ [26.02.2020]

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Integración metodológica y diseño de la interfaz para el Corpus del Habla de Baja California

Jorge Lázaro[68], Rafael Saldívar Arreola[69] & Álvaro Rábago Tánori[70]

[jorgelazaroh@gmail.com | rafaelsaldivar@uabc.edu.mx | rabago@uabc.edu.mx]

Universidad Autónoma de Baja California, México

El objetivo de este trabajo es mostrar la metodología utilizada para el diseño del Corpus del Habla de Baja California. En la primera etapa se presentan 18 entrevistas del municipio de Mexicali, 18 entrevistas del municipio de Tijuana y 8 del municipio de Ensenada en versión textual y etiquetada. La metodología para las entrevistas es la misma que utiliza el PRESSEA (2014). El etiquetado se ha hecho mediante la adaptación del algoritmo *Log-linear Part-Of-Speech Tagger* de la Universidad de Stanford, utilizado por Rico-Sulayes, Saldívar, Rábago (2017). Esta adaptación ha logrado una precisión superior a cualquier otro en el etiquetado de español vía *Maximum Entropy tagger* (97.2%) y el *Maximum Entropy plus distributional similarity* (97.4%). El diseño de la interfaz está basado en el sitio que el Laboratorio de Estudios Fónicos ha dedicado al Corpus Sociolingüístico de la Ciudad de México (Butragueño y Lastra, 2011) y para la gestión de los documentos se ha utilizado GECO (Sierra, Solórzano, Curiel, 2017).

En este trabajo se presenta una versión completa del Corpus del Habla de Baja California (CHBC), ya como un repositorio con muestras de registro oral. Dichas muestras están distribuidas de manera representativa, siguiendo la metodología del Proyecto para el Estudio

[68] Licenciado en Lengua y Literaturas Hispánicas por la Universidad Nacional Autónoma de México y Doctor en Comunicación Lingüística y Mediación Multilingüe (Mención Doctor Europeo) por la Universitat Pompeu Fabra. Miembro del Grupo de Ingeniería Lingüística de la UNAM, donde fue investigador posdoctoral. Ha sido profesor en la Facultad de Filosofía y Letras y en la Facultad de Ingeniería de la UNAM. Actualmente es Investigador Titular A TC en la Facultad de Idiomas de la Universidad Autónoma de Baja California. Miembro del Sistema Nacional de Investigadores (SNI) del Consejo Nacional de Ciencia y Tecnología (CONACyT).

[69] Licenciado en Docencia del Idioma Inglés por la Universidad Autónoma de California. Maestro en Educación en la Universidad Pedagógica Nacional. Doctor en Lingüística por la Universidad Autónoma de Querétaro (UAQ). Trabaja las líneas de investigación lingüística de corpus y lexicología. El enfoque predominante de los trabajos realizados y dirigidos es la convergencia de estos dos campos en variedades diastráticas y diafásicas del habla fronteriza de Baja California y California, EEUU, así como algunos trabajos lexicográficos. Actualmente es Subdirector de la Facultad de Idiomas de la UABC, campus Mexicali. Miembro del Sistema Nacional de Investigadores (SNI).

[70] Doctor en Lingüística por la Universidad Autónoma de Querétaro. Maestro en Lingüística Hispánica por la Universidad Nacional Autónoma de México. Por veinte años, Profesor-Investigador de tiempo completo de la Facultad de Idiomas, en Universidad Autónoma de Baja California. Ha realizado diversos estudios en torno a verbos en español, así como estudios léxicos, sintáctico-semánticos y traductológicos basados en corpus. Coordina el Corpus del Habla de Baja California y participa con proyectos de integración de corpus internacionales como el Proyecto para el estudio Sociolingüístico del Español de España y América.

Sociolingüístico del Español de España y de América (PRESEEA), entre las principales entidades municipales del estado. Actualmente se cuenta con un avance de 18 entrevistas de Mexicali, 18 de Tijuana y 8 Ensenada, debidamente etiquetadas mediante la adaptación del algoritmo *Log-linear Part-Of-Speech Tagger* de la Universidad de Stanford, utilizado por Rico-Sulayes, Saldívar, Rábago (2017). Esta adaptación ha logrado una precisión superior a cualquier otro en el etiquetado de español vía *Maximum Entropy tagger* (97.2%) y el *Maximum Entropy plus distributional similarity* (97.4%).

Para el registro regional se modificó la metodología de PRESEEA de tal manera que se utiliza el mismo formato, pero en la sección de las áreas de interés se plantean preguntas orientadas a obtener palabras propias de la región a través del estímulo directo. En la primera etapa se recabaron las entrevistas de los informantes en el municipio de Mexicali y se llevó a cabo la transcripción y edición y etiquetado de los datos. Posteriormente, se llevaron a cabo las entrevistas en los municipios de Ensenada siguiendo la misma metodología. Finalmente, se elaboraron las entrevistas en el municipio de Tijuana en continuidad con la metodología planteada.

La interfaz del corpus del Habla de Baja California, basada en el sitio que el Laboratorio de Estudios Fónicos ha dedicado al Corpus Sociolingüístico de la Ciudad de México (Butragueño y Lastra, 2011), se encuentra en el Sistema de Gestión de Corpus del Grupo de Ingeniería Lingüística de la UNAM (Sierra, Solórzano, Curiel, 2017). El sitio está habilitado para que el usuario pueda realizar consultas sobre el habla de la región, así como de muestras de audio de cada entrevista, con el fin de contribuir con cualquier proyecto de investigación que involucre una o más ramas derivadas de la lingüística, sea de orden sociocultural, discursivo, pragmático, léxico, sintáctico, semántico, psicolingüístico y fonético entre otros.

Palabras clave

lingüística de corpus, etiquetado gramatical, interfaz de corpus, lexicología, Baja California

Referencias

Martín Butragueño, P. y Lastra, Y. (coords.), (2011). *Corpus sociolingüístico de la ciudad de México. Vol. 1: Hablantes de instrucción alta*. México: El Colegio de México.

PRESEEA (2014-) Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. Alcalá de Henares: Universidad de Alcalá.

Rico-Sulayes, A., Saldívar-Arreola, R., & Rábago-Tánori, Á. (2017). Etiquetado gramatical por entropía máxima y rasgos de similitud distribucional en un corpus subregional del español. *INGENIERÍA Y COMPETITIVIDAD*, 19(2).

Sierra, Gerardo; Solórzano Soto, Julián & Curiel Díaz, Arturo. (2017). "GECO, un Gestor de Corpus colaborativo basado en web". *Linguamática*, 9(2), 57-72.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Investigación y modalización verbal: ¿cómo se construye la retórica de la investigación en los artículos de investigación?

Nathalia Lamprea Abril[71], Andrea Torres Perdigón[72] & Diana Moreno Rodríguez[73]

[mnathalialamprea@javeriana.edu.co | atorresp@javeriana.edu.co | mo-diana@javeriana.edu.co]

Pontificia Universidad Javeriana, Colombia

Palabras clave: lingüística de corpus, humanidades digitales, investigar, artículo de investigación, modalización, efectos de personalización y despersonalización.

La presente ponencia se inscribe en el proyecto de investigación *Concepciones de la investigación en estudios relacionados con el lenguaje*, desarrollado en la Pontificia Universidad Javeriana desde 2018 hasta la fecha por un equipo de cuatro profesoras investigadoras. Este proyecto parte del supuesto siguiente: lo que se entiende y valora como investigación puede ser heterogéneo dependiendo de las áreas disciplinares y de los momentos históricos. Nuestro objetivo es entonces indagar, a través de uno de los géneros textuales dominantes de lo que se considera investigativo –el artículo de investigación–, cómo se elaboran esas diversas concepciones de investigación en la materialidad textual y lingüística de los textos,

[71] Profesora investigadora del Departamento de la Pontificia Universidad Javeriana. Doctora en ciencias del lenguaje de la Universidad de Nantes; magíster en enseñanza de lenguas extranjeras Universidad Pedagógica Nacional y en análisis del discurso y didáctica de FLE/FLS de la Universidad de Nantes. Licenciada en Humanidades con énfasis en español y Lenguas extranjeras de la Universidad Pedagógica Nacional. Actualmente, coordinadora del área de francés del Departamento de lenguas de la Pontificia Universidad Javeriana. Con intereses investigativos relacionados con los estudios del lenguajes, la didáctica de las lenguas y las culturas y las ciencias sociales.

[72] Profesora asistente e investigadora del Departamento de Lenguas de la Pontificia Universidad Javeriana, en Bogotá. Es doctora y magíster en Estudios Románicos Hispánicos por la Universidad París Sorbona y profesional en Estudios Literarios por la Pontificia Universidad Javeriana. Es autora del libro La littérature obstinée: le roman chez Juan José Saer, Ricardo Piglia et Roberto Bolaño (2015), del manual Escribir el trabajo de grado. Cómo redactar documentos de investigación formativa (2018) y coautora de Aprender a escribir, leer y argumentar en la universidad (2018). Actualmente es coordinadora del Centro de Escritura de la Pontificia Universidad Javeriana.

[73] Profesora investigadora del Departamento de Lenguas de la Pontificia Universidad Javeriana. Candidata a Doctora en Ciencias Sociales, Niñez y juventud, de la Universidad de Manizales-CINDE; Magíster en Docencia, de la Universidad de La Salle; Licenciada en Humanidades y Lengua Castellana, de la Universidad Distrital Francisco José de Caldas. Actual coordinadora del área de Lengua Materna del Departamento de lenguas de la Pontificia Universidad Javeriana. Intereses de investigación centrados en las Ciencias sociales y los estudios del lenguaje.

particularmente en artículos que traten o trabajen el problema del lenguaje y publicados en las últimas dos décadas.

Para abordar nuestra problemática, se elaboró un corpus de artículos de investigación en español que tienen relación con el tema del lenguaje y que se sitúan en los campos de las ciencias sociales y humanas. El proceso de elaboración del corpus constó de varias etapas. En primer lugar, se realizó una búsqueda de artículos en el índice citacional de SciELO y con los que se construyó una base de datos con la información bibliográfica de los artículos descargados. Posteriormente, se revisaron los resúmenes de todos los artículos a la luz de cinco preguntas orientadoras que definieron un porcentaje de concordancia de 73% en la clasificación realizada por cuatro evaluadores. Las preguntas fueron las siguientes: 1. ¿El artículo da cuenta de una concepción de investigación? 2. ¿El artículo aborda el lenguaje? 3. ¿Es el lenguaje un elemento central o accesorio en el ejercicio investigativo? 4. ¿Se da cuenta de una preocupación por el lenguaje? 5. ¿El problema de investigación está relacionado con el lenguaje?

Como resultado de este proceso, obtuvimos un corpus de 1 260 artículos (10.112.627 ocurrencias-tokens) publicados en español durante un periodo de 16 años (2002 a 2018) en revistas latinoamericanas. La amplitud del corpus nos exigió adoptar un enfoque que nos permitiera hacer lecturas y análisis de grandes cantidades de datos textuales, para lo cual recurrimos a la lingüística de corpus (Rastier, 2011; Mayaffre, 2005; Tognini-Bonelli, 2001) articulada al análisis del discurso y al uso de herramientas lexicométricas con programas de software libre.

En este sentido, el objetivo de la ponencia es, por uno lado, plantear la relación de la lingüística de corpus como estrategia metodológica y su vínculo con las humanidades digitales, entendidas no sólo como el uso de métodos informática para dar respuesta a preguntas de investigación en las ciencias sociales y en las humanidades (Dacos y Mounier, 2014), sino como un espacio de posibilidades heurísticas en el que convergen lo cuantitativo y lo cualitativo. Por el otro, buscamos establecer un posible vínculo entre el nivel de explicitación de las posturas asumidas por quién investiga en sus textos y sus posicionamientos frente a la actividad de investigar, es decir, una concepción de investigación. Así pues, partimos de la hipótesis de que lo que se concibe como investigación puede estar ligado a los mecanismos lingüísticos empleados por los autores para construir y construirse en el discurso de la investigación.

Para el tratamiento del corpus, empleamos como herramienta de análisis de datos textuales el programa *Iramuteq*, el cual cuenta con un conjunto de funcionalidades que permiten realizar análisis multidimensionales de corpus textuales. Así pues, a partir de un primer tratamiento lexicométrico del corpus con el programa y sus respectivos análisis cualitativos, los resultados iniciales revelan la construcción de un discurso de experticia a partir de la elaboración de una retórica de demostración y de evidencia en una enunciación dominantemente despersonalizada (uso de verbos en tercera persona, modo verbal del presente del indicativo, pasivas reflejas,

campos semánticos de evidencia y de autoridad). De esta manera, podríamos esbozar una concepción de lo que es investigar sobre el lenguaje relacionada con el acto de demostrar, de explicar y de evidenciar lo enunciado, concepción que no distaría de lo que ha institucionalizado como ciencia desde un enfoque positivista, sus formas de enunciarla y comunicarla.

Bibliografía

Adam, J.-M. (1997). Genres, textes, discours : pour une reconception linguistique du concept de genre. *Revue belge de philologie et d'histoire*, *75*(3), 665–681. https://doi.org/https://doi.org/10.3406/rbph.1997.4188

Bazerman, C., Bonini, A. & Figueredo, D. (eds). 2009: *Genre in a Changing World*. Fort Collins and West Lafayette.

Benveniste, E. (1966). *Problèmes de la linguistique générale, Tome I.* Paris : Gallimard.

Charaudeau, P. (1992). *Grammaire du sens et de l'expression.* Paris: Hachette.

Charaudeau, P. y Maingueneau, D. (2002). *Dictionnaire d'analyse du discours*. Paris: Éditions du Seuil.

Dacos M. y Mounier P. (2014). *Humanités numériques. État des lieux et positionnement de la recherche française dans le contexte international*. Paris: Institut français/ministère des Affaires étrangères pour l'action culturelle. Disponible en : http://www.institutfrancais.com/sites/default/files/if_humanites-nume- riques.pdf.

García Negroni, M.-M. (2008). Subjetividad y discurso científico-académico. Acerca de algunas manifestaciones de la subjetividad en el artículo de investigación en español. *Signos* 41(66), 9–31.

Garric, N. y Calas, N. (2007). *Introduction à la pragmatique.* Paris : Hachette.

Grande Alija, F.-J. (2002). Aproximación a las modalidades enunciativas. León: Universidad de León.

Mayaffre, D. (2005). Rôle et place du corpus en linguistique. Réflexions introductives. In P. Vergely (Éd.), Actes du colloque JETOU'2005 (p. 5-17). Disponible en: https://hal.archives-ouvertes.fr/hal-00553742

Rastier, F. (2011). *La mesure et le grain. Sémantique de corpus*. Paris: Honoré Champion Éditeur.

Sánchez-Upegui, A. A. (2018). Consideraciones sobre el artículo científico (ac) : una aproximación desde el análisis de género y el posicionamiento. *Lingüística y Literatura*, *73*, 17-36. https://doi.org/10.17533/udea.lyl.n73a01

Tognini-Bonelli, E. (2001), *Corpus linguistics at work.* Amsterdam: John Benajmins.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

'It's not supposed to be this hard' A Corpus-based Study of the Intensifying Function of *this/that* in New Zealand English

Juan Lorente Sánchez[74]
[lsjuan2711@gmail.com]
Universidad de Málaga, Spain

Intensifiers are lexical items which function as modifiers of an adjective or and adverb in order to denote the exact degree or value of the elements that they modify (Huddleston and Pullum 2002: 585). Intensifiers are subjected to a remarkably emotional function depending on the speaker's need of expressivity and, as a consequence, it is a category permanently undergoing a process of innovation and semantic change in the quest of higher expressivity (Calle-Martín 2014: 401; Méndez-Naya 2003: 372; Lorenz 2002: 143-144). In fact, they are constantly immersed in a grammaticalization process which is often the result of primary grammaticalization showing the development 'down the cline' from adjectives/adverbs with a potential degree reading to a proper degree word (Hopper and Traugott 2003: 122; Paradis 2011: 252), as in the cases of *very*, *quite*, *fairly*, and *pretty*, among others. Despite this, in other cases, the degree meaning develops as a result of secondary grammaticalization where already grammaticalized elements acquire the intensifying function in the light of their implicit connotation of quantification. This development is particularly common in many languages and the input for the generation of new degree adverbs.

In this fashion, the intensifiers *this/that* stand out as a prototypical example of this secondary grammaticalization with a development from determiners to adverbs of degree. Their adverbial use stems from the deictic, contrastive and comparative potential of demonstratives, which turned them into 'here' and 'there' adverbs, becoming manner deictics or degree deictics (Calle-Martín 2019: 154). The phenomenon disseminated in the early-nineteenth century as a standard resource of spoken English and it has found their room in the written domain ever since, being considerably favoured in the less formal types of writing, fiction in particular, followed by magazines and newspapers (Calle-Martín 2019: 169). These intensifiers have been hitherto ignored in the literature, perhaps due to a preconceived accusation of informality, and

[74] Juan Lorente Sánchez graduated in English Studies in 2016 at the University of Málaga, and later he obtained two Master's degree at the same university: the first in Teaching English as a Foreign Language (2017), and the second in English Studies, Multilingual and Intercultural Communication (2018). He is currently working on a PhD thesis dealing with the semi-diplomatic edition and philological study of Glasgow University Library, MS Ferguson 7. His research interests lie within the fields of manuscript studies, historical Linguistics, sociolinguistics and corpus linguistics.

consequently the booster *so* has been traditionally recommended in these contexts (Quirk *et al.* 1985: 1466; Fowler 1926: 772; Swan 1980: 566).

Even though the phenomenon is attested in practically all the varieties of English worldwide, it has a variable distribution. Among the inner circle varieties, on the one hand, the construction is found to be more widespread in American English, with Canadian, Australian and British English lagging well behind it. Among the outer circle varieties, on the other hand, the phenomenon is also subject to geographical preferences. In African Englishes, for instance, *this/that* become more significant in Nigerian English, whilst in Asian Englishes Singapore and Philippines English stand out. This considered, the present paper aims to analyse the development and distribution of these booster in New Zealand English, insofar as this variety may be analysed as a yardstick to assess the distribution of the phenomenon in some inner and outer circle varieties of English. The paper then pursues the following objectives: i) a quantitative examination of the use and distribution of *this/that* in the variety under scrutiny; and ii) a qualitative analysis of the phenomenon in terms of the lexico-semantic structure of the right-hand collocates. The data used as source of evidence come from the New Zealand component of the *Corpus of Global Web-based English* (*GloWbE*), containing 1.9 billion words from 340,000 websites in 20 different English-speaking countries using a random selection of web pages and blogs, thus becoming the appropriate input for the study of the phenomenon in New Zealand English.

Keywords

intensifiers; *this/that*; New Zealand English; right-hand collocates; *GloWbE*; corpus linguistics

References

Calle-Martín, Javier. 2014. "On the History of the Intensifier *wonder* in English". *Australian Journal of Linguistics* 34.3: 399-419.

Calle-Martín, Javier. 2019. 'No cat could be that hungry!' *This/that* as Intensifiers in American English". *Australian Journal of Linguistics* 39.2: 151-173.

Fowler, H.W. 1926. *Fowler's Modern English Usage*. Oxford: Oxford University Press.

Hopper, Paul J. and Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge: Cambridge University Press.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Lorenz, Gunter. 2002. "Really Worthwhile or not Really Significant? A Corpus-based Approach to the Delexicalization and Grammaticalization of Intensifiers in Modern English". In Ilse Wishcer and Gabriele Diewald (eds.), *New Reflections on Grammaticalization*. Amsterdam and Philadelphia: John Benjamins.

Méndez-Naya, Belén. 2003. "On Intensifiers and Grammaticalization: The Case of Swiþe". *English Studies* 84.4: 372-391.

Paradis, Carita. 2011. "Reinforcing Adjectives: A Cognitive Semantic Perspective on Grammaticalization". In Ricardo Bermúdez-Otero and David Denison (eds.). *Generative Theory and Corpus Studies. A Dialogue from 10 ICEHL*. Amsterdam: Mouton de Gruyter. 233-258.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Swan, Michael. 1980. *Practical English Usage*. Oxford: Oxford University Press.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

La lengua y el COVID-19 en los medios periodísticos digitales en Argentina: un estudio de corpus contrastivo

Pablo Carpintero[75] & Martín Tapia Kwiecien[76]
[pcarpintero.winona@gmail.com | makwiecien@gmail.com]
Universidad Nacional de Córdoba, Argentina

Desde principios de la propagación del COVID-19 a nivel mundial, nuevas palabras y expresiones han aparecido y hoy forman parte de los tantos glosarios relacionados con la enfermedad, glosarios que se están formando en diferentes idiomas alrededor del mundo. A partir de la confirmación del primer caso de un paciente infectado por COVID-19 en la Argentina, el 3 de marzo de 2020, los medios del país han hecho uso del "discurso del COVID-19" sin pausa. Los medios nacionales han introducido al discurso popular nuevos términos y conceptos de uso muy frecuente en la actualidad, por ejemplo *distanciamiento social*, *casos asintomáticos*, *aplanar la curva*, y *crecimiento exponencial*, entre tantos otros. Para este estudio, nos propusimos analizar artículos periodísticos de los cuatro diarios digitales con mayor llegada a los argentinos desde el día de la confirmación del primer caso en el país hasta la actualidad. Para lograr este objetivo se ha empezado a construir un corpus digital especializado con aquellos artículos sobre el COVID-19 desde el día 3 de marzo hasta el 3 de agosto. La elección del mes de marzo como punto de partida no solo coincide con la confirmación del primer caso local, sino también, con la declaración, por parte de la OMS, del COVID-19 como pandemia global. A través de la metodología de corpus, analizaremos algunos aspectos lingüísticos referidos y relacionados con aspectos lingüísticos sobre el COVID-19, tales como: frecuencia de uso, terminología, colocaciones y combinaciones frecuentes de palabras, secuencias de palabras y expresiones multilexicales recurrentes, entre otros. Se pretende realizar un estudio descriptivo, mixto y diacrónico que observe el comportamiento de los términos más frecuentes y sus contextos de uso a los largo del periodo marzo-agosto. Por otra parte, se hará una comparación contrastiva de los resultados de este estudio con algunos corpus disponibles en inglés, como por

[75] Pablo Carpintero es Profesor de inglés, docente JTP Efectivo de las cátedras de Inglés Médico, Niveles I y II en el Departamento de Idiomas con Fines Académicos de la Facultad de Lenguas, Universidad Nacional de Córdoba. Docente Titular de Práctica Docente II y IV del Profesorado de Inglés en el ISFD Juan Zorrilla de San Martín. Se encuentra en la redacción de su tesis de Maestría en Inglés, mención Lingüística Aplicada. Investiga sobre la lectura del Inglés con Fines Específicos, el Análisis del Discurso y la Lingüística de Corpus.

[76] Martín Tapia Kwiecien es Profesor y Licenciado en Lengua y Literatura Castellana (Facultad de Lenguas-UNC). Especialista en Lingüística y Magíster en Competencias Digitales en Educación, por distintas instituciones europeas. Actualmente, se desempeña como Profesor Adjunto Interino de las asignaturas "Lengua Castellana I", sección Inglés, y "Fonética, Fonología, Morfología españolas", sección Español; y como Profesor Asistente Regular de la cátedra "Lexicología, Lexicografía españolas", sección Español (todas en la FL-UNC). También, cumple funciones de codirector de la Revista Digital de Lexicología, Lexicografía y Terminología y de secretario de redacción de la Revista Culturas y Literaturas Comparadas (ambas del CIFAL- FL-UNC).

ejemplo el *The Coronavirus Corpus* (un subcorpus del *English-Corpora.org*) y el *COVID-19 Corpus* de *Sketch Engine*, como también se utilizarán glosarios en inglés especialmente diseñados sobre el COVID-19, como el *Glossary on the COVID-19 pandemic*, para un análisis terminológico más detallado. Se detallarán los procedimientos llevados a cabo para la construcción del corpus, las estrategias metodológicas de análisis y los resultados preliminares. Este estudio pretende aportar información valiosa sobre la lengua vinculada con el COVID-19 en Argentina y sobre posibles herramientas y procedimientos para la construcción y análisis de un corpus especializado.

Palabras clave: comparación contrastiva, COVID-19, metodología de corpus, vocabulario especializado.

Bibliografía

COVID-19 Corpus de Sketch Engine. Disponible en: http://ske.li/covid_19

Davies, M. (2020). The Coronavirus Corpus. Disponible en https://www.englishcorpora.org/corona/

Glossary on the COVID-19 pandemic. Disponible en: https://www.btb.termiumplus.gc.ca/publications/covid19-eng.html

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress* (Lorient, France). 105--116.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Las formas nominales del verbo en griego y latín: retos en la anotación de un corpus

Eveling Garzón Fontalvo[77]
[eveling.garzon@uam.es]
(Universidad Autónoma de Madrid)
Berta González Saavedra[78], **J. Ignacio Hidalgo González**[79], **Iván López Martín**[80]
[bertagon@ucm.es | josehida@ucm.es | ivlopez@ucm.es]
(Universidad Complutense de Madrid)
Alberto Pardal Padín[81]
[pardal@usal.es]
(Universidad de Salamanca)
Guillermo Salas Jiménez[82]

[77] Estudió Español y Filología Clásica en la Universidad Nacional de Colombia (2009) y Filología Clásica en la Universidad Autónoma de Madrid (2011), donde, además, obtuvo el título de Máster (2012). Entre 2013 y 2017 disfrutó de una beca predoctoral FPI-UAM. Su tesis, dirigida por la Dra. Esperanza Torrego, lleva por título «Nombres de Acción y otros derivados deverbativos en latín» y fue defendida en noviembre de 2018. Actualmente es personal contratado de Investigación del proyecto COMREGLA.

[78] Doctora en Filología clásica por la UCM y licenciada en Filología Italiana por la UV. Su tesis doctoral, defendida en 2015, se titula "Expresión de la Procedencia en lenguas indoeuropeas antiguas: griego, latín e hitita". Ha trabajado como contratada de investigación en la Università Cattolica de Milán, en el proyecto Index Thomisticus Treebank, con el investigador Marco C. Passarotti en la anotación semántico-pragmática de textos latinos. Actualmente es profesora ayudante doctora en la Universidad Complutense de Madrid y colaboradora del proyecto COMREGLA.

[79] Es graduado en Filología Clásica por la Universidad Complutense de Madrid, donde también cursó el máster en Filología Clásica y el máster en Formación del Profesorado. Actualmente es doctorando en Filología Clásica por la UCM, con la tesis "La expresión léxica de la diátesis pasiva en latín: colocaciones verbo-nominales" y colabora con el proyecto COMREGLA. Además, imparte clases de latín y griego en el Instituto de Educación Secundaria Marc Ferrer (Formentera, Baleares, España).

[80] Es graduado en Filología Clásica (2015) y ha realizado el Máster Interuniversitario en Filología Clásica (2016) y el Máster en Formación del Profesorado (2017), todo ello por la Universidad Complutense de Madrid. Actualmente es beneficiario de un contrato predoctoral FPU, en el marco del cual realiza su tesis doctoral en torno a las construcciones con verbo soporte en la obra clásica la Historia Augusta bajo la dirección de los dres. J. M. Baños y Mª D. Jiménez, y es colaborador en el proyecto COMREGLA.

[81] Estudió Filología Clásica en la Universidad de Salamanca, donde se licenció en 2009 y obtuvo su master en 2010. Se doctoró en 2017 con la tesis "La interacción entre fonología y sintaxis en griego antiguo" bajo la co-dirección de Julián Méndez Dosuna y Jesús de la Villa. Ha trabajado como profesor asociado en la Universidad de Valladolid y en la de Alcalá de Henares. Actualmente es profesor ayudante doctor en la Universidad de Salamanca y colaborador del proyecto COMREGLA.

[82] Estudió Filología Clásica en la Universidad de Granada, donde se graduó en 2018, y obtuvo el Máster en Filología Clásica por la Universidad Complutense de Madrid en 2019. Actualmente cuenta con una ayuda predoctoral FPI gracias a la cual está realizando su tesis, titulada "La expresión léxica del aspecto incoativo en

[guisalas@ucm.es]
(Universidad Complutense de Madrid)
Cristina Tur Altarriba [83]
[cristina.tur@uam.es]
(Universidad Autónoma de Madrid)

Resumen

Tanto el griego como el latín han sido objeto de estudio y análisis de la lingüística computacional desde los orígenes de la misma (Winter 1999) y eso hace que distintas cuestiones que atañen la creación de un corpus anotado hayan sido tratadas desde múltiples perspectivas, como ocurre con la morfología, con los modelos del Index Thomisticus Treebank y de Perseus (Passarotti 2011). Sin embargo, hay pocos corpus anotados sintácticamente; entre ellos, PROIEL (Haug & Jøhndal 2008) y Alpheios, y con análisis semántico el Index Thomisticus Treebank –que solo anota latín–; de ahí que el proyecto COMREGLA haya tenido que enfrentarse a algunas cuestiones en su conversión de base de datos a corpus anotado que apenas habían sido tenidas en cuenta para otros corpus. Una de estas ha sido el tratamiento de las formas nominales del verbo, muy frecuentes en griego y en latín. En griego contamos con los infinitivos, los participios y los adjetivos verbales en -τεος, mientras que en latín tenemos infinitivos, participios, gerundios, gerundivos y supinos. La naturaleza de estos elementos es doble, porque por una parte actúan como elementos nominales en la complementación de verbos y sustantivos, forman parte de expresiones *multiword*, etc.; pero por otra actúan como núcleos verbales que tienen sus propias estructuras predicativas.

El proyecto COMREGLA, en el que se enmarca esta comunicación, tiene como objetivo la adaptación de la base de datos REGLA para su comunicación con otros recursos digitales. REGLA surgió en un principio como una herramienta para el análisis sintáctico-semántico de los marcos predicativos de los verbos más habituales del latín y el griego desde el marco teórico de la Lingüística Funcional desarrollado por Dik (1997) y Hengeveld & Mackenzie (2008), modelo que ha sido aplicado al latín por Pinkster (1984, 2015, 2021) y a las lenguas clásicas por el grupo de investigadores que participan en este proyecto: Torrego (1998, 2007), Baños (2003, 2009) y Jiménez (2021). Asimismo, para la representación de las relaciones sintácticas hemos tenido en consideración, sobre todo, la Gramática de Dependencias, modelo particularmente útil para la representación de tales relaciones en latín y griego, al ser estas lenguas que presentan una gran flexibilidad en el orden de palabras.

latín: colocaciones verbo-nominales", bajo la dirección de los profesores José Miguel Baños (UCM) y María Dolores Jiménez (UAH) y es colaborador del proyecto COMREGLA.

[83] Es graduada en Lingüística y Lenguas Aplicadas (especialidad de Lingüística Computacional) y doctora en Filología Clásica, ambas por la Universidad Complutense de Madrid. Su tesis, titulada "Sintaxis y semántica de los nombres de sentimiento en latín: empleos adverbales y colocaciones" y codirigida por los dres. José M. Baños, Mª D. Jiménez y E. Torrego, fue defendida en noviembre de 2019. Actualmente trabaja como personal técnico contratado en el proyecto COMREGLA en la Universidad Autónoma de Madrid.

Los objetivos iniciales de REGLA hacían que la herramienta de trabajo fuera una base de datos relacional, con ítems independientes para cada instancia de cada uno de los verbos analizados en un corpus seleccionado de obras griegas y latinas. El objetivo de COMREGLA, por su parte, es adaptar esta base de datos relacional a una notación de texto en XML, en la línea de lo desarrollado por otros corpus anotados como los señalados en párrafos anteriores o el Prague Dependency Treebank (o PDT). En esta conversión, por un lado, es necesaria la anotación de elementos a los que no se atendía en REGLA; por otro lado, las formas nominales del verbo, objeto de esta presentación, requieren un tratamiento diferenciado que se aleja en lo técnico del practicado en REGLA (donde cada forma verbal tenía su ítem propio), pero que desborda el tratamiento recibido en otros corpus debido a la riqueza de datos del análisis sintáctico-semántico realizado en la base de datos de partida.

Uno de los contextos en que las formas nominales del verbo plantean dificultades es el de las construcciones perifrásticas donde un participio y un verbo auxiliar forman una estructura compleja con un significado unitario, como *interfectus est* ("fue asesinado"). Los problemas que conlleva el análisis de estas expresiones son variados. Entre ellos están la posibilidad de que los elementos que la constituyen puedan aparecer separados en la oración (y, por lo tanto, tokenizados de manera independiente: *interfectus* // *est*); los fenómenos de elipsis, que hacen que el verbo auxiliar (*est*) pueda omitirse (apareciendo únicamente el participio); o el hecho de que, si bien el auxiliar es morfológicamente un presente, la construcción perifrástica expresa tiempo de perfecto.

Otro contexto complejo es el de los participios concertados, como (1).

(1)	ταῦτ'	εἰπόντες	καὶ	ἀκούσαντες	ἀπῆμεν
	taut'	eipóntes	kaì	akúsantes	apêmen
	eso.AC.PL.NOM	decir.PTCP.AOR.ACT.NOM.PL	y	escuchar.PTCP.AOR.ACT.NOM.PL	marcharse.PRS.IND.1PL

"Tras decir y escuchar esto, nos marchamos (Pl. *Prt.* 362s)

En REGLA se ha tendido a reconstruir todos los participantes elípticos de un elemento predicativo con objeto de determinar su estructura argumental, pero en ítems separados donde cada uno se analizaba con su verbo (εἰπόντες, ἀκούσαντες). Esto trae consecuencias evidentes al pasar a un análisis de texto completo etiquetado: por ejemplo, los participios concertados nunca presentan un sujeto explícito, dado que hacen referencia a otro constituyente de la oración con el que concuerdan morfológicamente y que, además, con frecuencia está elidido.

Finalmente, también ha supuesto problemas el hecho de que los participios no solamente hacen referencia a la entidad con la que concuerdan morfológicamente, sino también al evento del que esta forma parte.

(2) *currens* *Lepta* *uenit*
correr.PTCP.PRS.ACT.NOM.SG Lepta.NOM.SG.F venir.PRF.IND.3SG
"Lepta llegó corriendo" (Cic. *fam.* 3,7,4)

Así, en la oración de (2) el participio *currens* puede analizarse en relación con las propiedades léxicas de su referente, *Lepta* (una entidad humana, animada, concreta), y con las características semánticas del evento que expresa, es decir, su estructura aspectual (un evento dinámico, atélico, durativo…) y el grado de control de sus participantes (una situación controlada por el sujeto). La principal dificultad radica en que en la base de datos de REGLA ambas perspectivas están subsumidas en una única categoría denominada "Léxico", por lo que sus análisis son mutuamente excluyentes.

En la comunicación esperamos no solamente presentar estos problemas a los que se ha enfrentado el proyecto COMREGLA en el proceso de conversión de REGLA, sino también dar a conocer, de manera detallada, las propuestas que se han adoptado y las herramientas que se han empleado para abordarlos. En este sentido, creemos que el III Congreso Internacional de Lingüística Computacional y de Corpus es un contexto excelente para exponer y discutir tales cuestiones.

Palabras clave: griego antiguo; latín; sintaxis; semántica; lingüística computacional.

Bibliografía

Baños, J. M. (2009) (coord), *Sintaxis del latín clásico*, Madrid, Liceus.

Baños, J. M. *et al.* (2003) (ed.), *Praedicativa I. Complementación en griego y latín*, Santiago de Compostela, Universidad de Santiago de Compostela.

Dik, S. C. (1997), *The Theory of Functional Grammar. Part I: The structure of the clause*, Berlin-New York, The Gruyter.

Haug, D. & Jøhndal, M. (2008), "Creating a parallel treebank of the old Indo-European Bible translations". En *Proceedings of the Language Technology for Cultural Heritage Data Workshop* (LaTeCH 2008), ELRA, 27–34. URL http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf#page=31

Hengeveld, K. & Mackenzie, J. L. (2008), *Functional Discourse Grammar: A typologically-based theory of language structure*, Oxford, Oxford University Press.

Jiménez, Mª D. (2021) (coord.), *Sintaxis del griego antiguo* (en prensa).

Passarotti, M. (2011), "Language Resources. The State of the Art of Latin and the Index Thomisticus Treebank Project". En *Corpus anciens et Bases de données*, Presses universitaires de Nancy, 301–320.

Pinkster, H. (1984), *Latijnse syntaxis en semantiek*, Amsterdam, B. R. Grüner Publishing.

Pinkster, H. (2015), *The Oxford Latin Syntax. Vol 1, The Simple Clause*, Oxford, Oxford University Press.

Pinkster, H. (2021), *The Oxford Latin Syntax. Vol 2, The Complex Sentence and Discourse*, Oxford, Oxford University Press (en prensa).

Torrego, Mª E. (1998) (ed.), *Nombres y funciones. Estudios de sintaxis griega y latina*, Madrid, Ediciones Clásicas-UAM.

Torrego, Mª E. *et al.* (2007) (ed.), *Praedicativa II. Esquemas de complementación verbal en griego antiguo y en latín*, Zaragoza, Universidad de Zaragoza.

Winter, T. N. (1999), "Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance", *Faculty Publications, Classics and Religious Studies Department*,70, 3-21

Recursos electrónicos:

Alpheios, disponible en https://texts.alpheios.net/ (última visita 10/06/2020)

Index Thomisticus Treebank, disponible en https://itreebank.marginalia.it/ (última visita 10/06/2020)

PROIEL Treebank, disponible en https://proiel.github.io/ (última visita 10/06/2020)

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Las oraciones condicionales y la factualidad en español: Un estudio basado en corpus

Leyre Barrios Vicente[84] **& Glòria Vázquez García**[85]
[leyreb594@gmail.com | gloria.vazquez@udl.cat]
Universitat de Lleida, España

La factualidad (o certeza) es la noción que se refiere al compromiso del hablante en relación con la veracidad de una situación descrita. Para ello los hablantes nos servimos de distintos marcadores lingüísticos de tipo léxico, sintáctico, semántico y discursivo. En este trabajo pretendemos analizar qué marcadores pueden usarse para determinar la factualidad de las oraciones condicionales. Este estudio se enmarca dentro de un proyecto (Autores 1) cuyo objetivo es analizar y automatizar la interpretación de la factualidad en español.

Una oración condicional es aquella formada por dos cláusulas, una principal y una subordinada, denominadas asimismo prótasis y apódosis respectivamente. Estas cláusulas están unidas por una conjunción de carácter hipotético y que aporta un carácter no asertivo a la oración (Sweetser 1990, Dancygier 1993, Schwenter 2001). En español, la conjunción condicional prototípica es *si*, pero también encontramos *por si (acaso), como si, solo si, cuando*, etc. Existen estudios en los que se analiza la interpretación de la prótasis de estas oraciones en términos de factualidad, pero no de la apódosis, cuyo valor factual no siempre es coincidente con el de la prótasis.

Nuestra hipótesis de partida es que, en general, es posible formalizar la interpretación de la factualidad en las dos cláusulas de las oraciones condicionales usando marcadores lingüísticos. En este trabajo nos hemos centrado en analizar este aspecto de las oraciones condicionales cuya prótasis incluye tiempos del subjuntivo: pretérito imperfecto o pretérito pluscuamperfecto. Nos hemos focalizado en el uso de la conjunción *si*. Para ello se han extraído 200 oraciones (100 para

[84] Estudiante de doctorado en la Universitat de Lleida, he cursado el máster en Lengua Española: Investigación y Prácticas Profesionales de la Universidad Autónoma de Madrid (España) y el grado en Lengua y Literatura Españolas de la Universidad de La Rioja (España). Entre mis trabajos destacan "La sustitución del pretérito imperfecto de subjuntivo por el condicional en Logroño", publicado en Novas perspectivas na lingüística aplicada (pp. 83-93) y "La extensión del verbo 'chupar'", presentado en el V Congreso Internacional en Lengua y Literatura, Universidad de La Rioja, Logroño).

[85] Profesora titular en el área de Lingüística General de la Universidad de Lleida, miembro del Grupo de Investigación Interuniversitario en Aplicaciones Lingüísticas (GRIAL). Ha realizado aportaciones respecto a los patrones sintácticos verbales del español y su significado construccional. Destaca también su trabajo en la descripción de las perífrasis verbales de esta lengua. Actualmente está investigando sobre el grado de factualidad que se expresa en las oraciones. En total, tiene más de cincuenta publicaciones, entre las cuales destacan libros como Clasificación verbal. Alternancias de diátesis y Las construcciones con "se" en español y capítulos en manuales de lingüística de corpus y lexicografía. En estos ámbitos ha participado también confeccionando recursos a gran escala a partir de su colaboración en proyectos financiados. Actualmente está dirigiendo el proyecto TAGFACT.

cada tiempo) del subcorpus *Now*, el cual forma parte del *Corpus del español* de Mark Davis (2018). Estas oraciones han sido extraídas de noticias sobre política o deporte que han sido publicadas en los últimos años en periódicos de ámbito nacional. Este análisis empírico nos ha permitido completar algunas combinaciones de tiempo verbales (esquemas temporales) que no están previstos en la bibliografía y aportar los valores factuales para las apódosis.

En los últimos años se han propuesto múltiples clasificaciones de este tipo de oraciones, tanto formales como semánticas. En cuanto a las clasificaciones semánticas, destacan para el inglés las llevadas a cabo por Sweetser (1990) y Dancygier (1993). Para el español son conocidas básicamente 3 propuestas: la de Veiga y Mosteiro (2006), quienes se basan en el modo de la prótasis para diferenciar entre condicionales reales e irreales; la de Rodríguez Rosique (2008), autora que diferencia entre condicionales de contenido, epistémicas e ilocutivas; y la clasificación semántico-formal de Montolío (1999), en la cual se basa este estudio debido a que es la que más ahonda en aspectos formales.

Esta autora distingue entre condicionales reales, potenciales e irreales. Las condicionales reales son aquellas que hacen referencia a hechos relacionados entre sí en el presente o en el pasado (*Si yo he aprendido es porque me han enseñado* (Montolío, 1999: 3663)) o de cumplimiento probable en el futuro (*Si saco la oposición me dedicaré al, al desempeño del cargo* (Montolío, 1999: 3665)). Las condicionales potenciales son aquellas en las que el hablante no se compromete del todo con la realización del evento y están enfocadas al futuro (*Me sorprendería mucho si resultara que pudiéramos encontrar alguna forma de ayudar a los pacientes afásicos* (Montolío, 1999: 3668)). Por último, las condicionales irreales son aquellas que hacen referencia a sucesos que no han tenido lugar en el pasado (*Si hubiese escuchado a mi hermano a estas horas estaría en Hollywood* (Montolío, 1999: 3672)).

En nuestro trabajo, se amplía la casuística y se utiliza una terminología distinta de la empleada por esta autora, ya que se ha priorizado el uso de etiquetas habituales en el campo de la anotación de corpus en este ámbito: 'factual', cuando se hace referencia a una situación que acontece o ha acontecido, 'no factual', cuando la situación descrita se presenta con cierto grado de incerteza o probabilidad; 'contrafactual', cuando hace referencia a un evento que no ocurre o que no ha ocurrido; y 'no aplica', cuando hace referencia a una situación imaginada sobre el futuro basada en deseos o conjeturas básicamente, sobre la cual no es relevante determinar la factualidad.

En nuestro análisis, en primer lugar, hemos comprobado qué combinaciones temporales verbales de las 6 recogidas en la bibliografía (2 con el imperfecto de subjuntivo en la prótasis y 4 con el pluscuamperfecto) tienen representación en el corpus analizado (ciñéndonos al tipo de condicionales objeto de nuestro estudio). Estas son: <*Si* + imperfecto de subjuntivo + condicional simple>, <*Si* + imperfecto de subjuntivo + imperfecto de indicativo>, <*Si* + pluscuamperfecto de subjuntivo + pluscuamperfecto de subjuntivo o condicional compuesto>, <*Si* + pluscuamperfecto de subjuntivo + condicional simple>, <*Si* + pluscuamperfecto de subjuntivo + imperfecto de indicativo> y <*Si* + pluscuamperfecto de subjuntivo +

pluscuamperfecto de indicativo>. Hay que advertir que esta última no ha sido documentada en el corpus manejado, seguramente debido a su carácter dialectal.

En segundo lugar, se han buscado en el corpus nuevas posibles combinaciones temporales del tipo de condicionales estudiadas y hemos obtenido 6 esquemas no citados en la bibliografía manejada, 5 con el imperfecto de subjuntivo en la prótasis y 1 con el pluscuamperfecto: <*Si* + imperfecto de subjuntivo + condicional compuesto>, <*Si* + imperfecto de subjuntivo + presente de indicativo>, <*Si* + imperfecto de subjuntivo + futuro>, <*Si* + imperfecto de subjuntivo + pretérito pluscuamperfecto de subjuntivo, <*Si* + imperfecto de subjuntivo + presente de subjuntivo> y <*Si* + pluscuamperfecto de subjuntivo + presente de indicativo>. En total, por tanto, se han analizado 11 esquemas, 4 con pluscuamperfecto de subjuntivo en la prótasis y 7 con imperfecto. Para la mayoría de estos esquemas se han manejado entre 15 y 20 ejemplos en cada caso. Para 3 de estos esquemas solo hemos podido recopilar entre 6 y 10 ejemplos por lo que interpretamos que son esquemas poco frecuentes.

A la hora de llevar a cabo el estudio de la factualidad de estas oraciones se ha analizado, por un lado, la factualidad de las prótasis, para comprobar si coinciden con los valores propuestos en la bibliografía. Según Montolío (1999), el pretérito imperfecto de subjuntivo en la prótasis condicional puede ser potencial e irreal (en nuestra terminología no factual o no aplica), mientras que el pluscuamperfecto de subjuntivo solo puede ser irreal (contrafactual). Esta propuesta se corrobora en nuestro estudio.

Por otro lado, se ha analizado la factualidad de las apódosis para estudiar cómo se expresa la factualidad en dichas cláusulas, ya que esta cuestión no ha sido analizada hasta ahora. Ello nos permitirá obtener el análisis completo de la oración condicional. Por un lado, en el caso de las apódosis con imperfecto en las prótasis, el valor factual de las primeras tiende a ser 'no aplica', cuando el tiempo es futuro, imperfecto de indicativo, presente de subjuntivo o condicional simple. En cambio, es 'contrafactual', cuando el tiempo es condicional compuesto o pluscuamperfecto de subjuntivo. Por último, cuando el tiempo es presente de indicativo suele ser factual. Por otro lado, en las oraciones con pluscuamperfecto en la prótasis, las apódosis tienden a ser contrafactuales, salvo en los casos con presente e imperfecto de indicativo, en las que también pueden ser factuales o no aplicar.

El resultado de este estudio es la propuesta de una serie de reglas lingüísticas para formalizar la interpretación de la factualidad de las oraciones condicionales en español estudiadas. Concretamente, para 8 de los 11 esquemas temporales analizados (que cubren dos cláusulas cada uno) se ha podido establecer un patrón para determinar el valor factual de la apódosis y la prótasis a partir de los tiempos verbales usados. En los otros 3 esquemas, que coinciden con presentar el imperfecto del subjuntivo en la prótasis, a través de los rasgos formales no se ha podido llegar a una única interpretación final y se juega con 2 posibles valores.

Como trabajo futuro pretendemos, aplicando la misma metodología, analizar más oraciones condicionales con otras combinaciones temporales así como con diferentes conjunciones para llegar a realizar un estudio lo más completo posible de este tipo de oraciones.

Palabras clave: condicionales, factualidad, hipoteticidad, prótasis, apódosis.

Bibliografía

Dancygier, B. (1993): "Interpreting conditionals. Time, knowledge and causation", Journal of Pragmatics, 19, pp. 403-434.

Davis, Mark (2018). *Now. Corpus del español*. Disponible en: https://www.corpusdelespanol.org/

Montolío, E. (1999) "Las construcciones condicionales", en I. Bosque & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (vol. 3, pp. 3805-3878). Madrid: Espasa-Calpe.

Rodríguez Rosique, S. (2008). Pragmática y Gramática: condicionales concesivas en español. En *Studien Zur Romanischen Sprachwissenschaft Und Interkulturellen Kommunikation* (Vol. 47). Frankfurt am Main: Peter Lang.

Schwenter, S. (2001): Expectations and (in)sufficiency: Spanish *como* conditionals. Linguistics 39-4, pp. 733-760

Sweetser, E. (1990). From etymology to pragmatics. Metaphorical and Cultural Aspects of Semantic Structure. Cambridge University Press / Peking University Press.

Veiga, A. y Mosterio, M. (2006). El modo verbal en cláusulas condicionales, causales, consecutivas, concesivas, finales y adverbiales de lugar, tiempo y modo

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Las tecnologías del lenguaje y las lenguas indígenas mexicanas: constitución de un corpus paralelo amuzgo-español

Antonio Reyes Pérez[86]
[areyes98@uabc.edu.mx]
Universidad Autónoma de Baja California, México
H. Antonio García Zúñiga[87]
[hamlet_garcia@inah.gob.mx]
Instituto Nacional de Antropología e Historia, México

En México, además del español, coexisten más de 60 lenguas indígenas que son reflejo de riqueza cultural y social, así como de cosmovisión e identidad. De acuerdo con los censos de población y vivienda, así como de lo plasmado en el Catálogo de las lenguas indígenas nacionales: variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas (Instituto Nacional de Lenguas Indígenas, 2008), se sabe que México cuenta con 11 familias lingüísticas, de las cuales se desprenden 68 agrupaciones con 364 variantes. No obstante esta riqueza lingüística, es muy evidente que desde la perspectiva de las tecnologías del lenguaje hay una carencia de recursos, herramientas e, incluso, materiales lingüísticos para la gran mayoría de estas lenguas. En este marco, este documento describe un trabajo en activo para la creación de herramientas para lenguas con escasos recursos, en particular, para la lengua amuzgo.

El amuzgo o *jnó^{n} ^{n}dà* es una lengua de la familia otomangue perteneciente a la rama amuzgo-mixtecana del tronco otomangue occidental (Campbell, 1997). Se habla en diferentes localidades del sur de México. A pesar de ser una lengua poco conocida, cuenta con alrededor de 60,000 hablantes (Instituto Nacional de Estadística y Geografía, 2015). En términos gramaticales, se distingue por un repertorio extenso de clases léxicas y complejidad verbal (Smith y Tapia, 2002; Apóstol, 2014), un conjunto amplio de pronombres personales (Buck, 2015; Palancar y Feist, 2015), así como por la presencia de tonos fonológicos para marcar

[86] Profesor titular de tiempo completo en la Facultad de Idiomas y Coordinador del Doctorado en Ciencias del Lenguaje de la Universidad Autónoma de Baja California. Es doctor en Lingüística por la Universidad Politécnica de Valencia. Es especialista en procesamiento del lenguaje natural. Entre sus publicaciones se encuentran “On the difficulty of automatically detecting irony: beyond a simple case of negation” y “Emotions and irony per gender in Facebook”, ambas de 2014

[87] Profesor investigador titular de tiempo completo en el Instituto Nacional de Antropología e Historia. Cuenta con estudios de maestría en Lingüística hispánica y metodología de la ciencia. Se especializa en lingüística histórica y lenguas amerindias mexicanas. Entre sus publicaciones más recientes están "Relaciones inanimadas y codificación de rasgos en la forma su/sus en el español de la Ciudad de México" (2020) y "Posesión y otras relaciones semánticas en amuzgo de San Pedro Amuzgos " (en prensa).

distintos significados morfológicos, como la posesión (Hernández, Mora y García, 2017; García, Hernández y Mora, en prensa).

El objetivo de este trabajo consiste en describir la construcción de un pequeño corpus paralelo amuzgo–español que sirva como herramienta inicial para el desarrollo de recursos más amplios, así como una fuente de información disponible para fines de investigación en distintas áreas.

Dado este objetivo, la organización del trabajo atiende a la descripción de las siguientes etapas y desafíos: i) obtención de datos en amuzgo mediante entrevistas grabadas, ii) transcripción de las entrevistas; iii) limpieza de datos; iv) creación de glosas al español y; v) traducción de datos al español. Asimismo, se detalla el proceso para la integración de los datos en una plataforma de acceso libre que permita la consulta del corpus.

Finalmente, es importante destacar que la creación de este recurso, además de ser un aporte para aumentar la atención a las lenguas escasamente representadas e, incluso, en peligro de extinción, permitirá el desarrollo de nuevos recursos que pueden nutrirse del conocimiento integrado en el corpus. Por ejemplo, el hecho de aportar las glosas puede generar que, en un ámbito como la traducción automática, se puedan mejorar los procesos de alineación entre segmentos del texto origen y el texto meta o, por otro lado, desarrollar sistemas de extracción de información sustentados en las características intrínsecas de la lengua.

Palabras clave

corpus paralelo, amuzgo, español, lenguas indígenas mexicanas.

Referencias

Apóstol, J. (2014). *Clases flexivas verbales en el amuzgo de Xochistlahuaca (Guerrero)*. Tesis de maestría. Ciudad de México: Centro de Investigaciones y Estudios Superiores en Antropología Social.

Buck, M. (2015). *Gramática del amuzgo de Xochistlahuaca*. Ciudad de México: Instituto Lingüístico de Verano.

Campbell, Lyle. (1997). *American Indian languages: the historical linguistics of Native America*. Oxford: Oxford University Press.

García, H., Hernández, N. y Mora, A. (en prensa). Posesión y otras relaciones semánticas en Amuzgo de San Pedro Amuzgos (otomangue). En Estrada, Z. *Dependencias simétricas y asimétricas: Dominios semánticos y motivaciones*. Hermosillo: Universidad de Sonora.

Hernández, N., Mora, A. y García, H. (2017). Estructura de la frase nominal posesiva en amuzgo (otomangue). UniverSOS. *Revista de Lenguas Indígenas y Universos Culturales*, 14: 63-82.

Instituto Nacional de Lenguas Indígenas, (2008). *Catálogo de las lenguas indígenas nacionales: variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*. En Diario Oficial de la Federación, 14 de enero de 2008.

Instituto Nacional de Estadística y Geografía, (2015). *Encuesta intercensal 2015*. En línea en: http://cuentame.inegi.org.mx/hipertexto/todas_lenguas.htm

Palancar, E. y Feist, T. (2015). Agreeing with subjects in number: The rare Split of Amuzgo verbal inflection. *Linguistic Typology*, 93(3): 337-383.

Smith, T. y Tapia, F. 2002. Amuzgo como lengua activa. En Levy, P. (Ed.) *Del cora al maya yucateco. Estudios lingüísticos sobre algunas lenguas indígenas mexicanas*, pp. 81-129. Ciudad de México: Universidad Nacional Autónoma de México.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Learning phonology through audiobooks: a parallel corpus-based approach to ESL

Michael Lang[88]
[michael.lang@rai.usc.es]
University of Santiago de Compostela, Spain
Xavier Gómez Guinovart[89]
[xgg@uvigo.gal]
University of Vigo, Spain

Our presentation aims to demonstrate the potential of the audio-textual English-Spanish parallel literary corpus LITTERA in language learning by providing practical examples of how it can be used by students to explore English phonology through Data-Driven Learning, and more generally, through the framework of corpus-based language pedagogy. We will demonstrate how an audio-textual English-Spanish parallel corpus can be exploited by means of the search tools available through CLUVI (http://sli.uvigo.gal/CLUVI) and SensoGal (http://sli.uvigo.gal/ SensoGal) corpora collections in order to examine the segmental and suprasegmental aspects of English phonology that frequently cause difficulty for Spanish-speaking students such as vowel reduction, co-articulation (elision and assimilation) and the importance of stress words in connected speech. This type of information is essential for students whose aim is natural, native-like speech and to avoid the excessively "polished" speech that doesn't reflect the reality of spoken English.

Corpora have become increasingly relevant in language pedagogy over the last few decades, not only in the creation of didactic material (e.g. dictionaries, reference books and text books), but also in its direct use by teachers and students. This direct, hands-on interaction with the corpus is known as Data-Driven Learning (DDL). The underlying idea is rather simple: students find and examine linguistic patterns in the corpus data and formulate rules based on their observations (usually under the guidance of a teacher). One of the pioneers of DDL, Tim Johns (1997), described it succinctly as “every student a Sherlock Holmes”. Since then, there has been much research coming out of this innovative approach thanks in large part to improvements in technology, not only in corpus linguistics, but computational linguistics as well. This, coupled with the fact that Internet access is becoming ever more widespread and each generation seems

[88] PhD student from the Department of English and German Philology at the University of Santiago de Compostela. His current research is centered on designing and developing data-driven pedagogical resources and methodologies for teaching ESL to Spanish university students.

[89] PhD degree in Computational Linguistics from the University of Santiago de Compostela in 1996. He is currently a professor at the University of Vigo. His research has been mainly in language technologies and resources, machine translation, corpus linguistics and computational lexicography.

more tech-savvy than the last, has showed that DDL deserves more consideration as a useful classroom resource.

The central tool in DDL is the corpus. The vast majority of corpora that have been implemented in DDL studies and activities have been almost exclusively monolingual text corpora, which is not surprising given the high costs and the amount of time needed to create a multimedia corpus, along with the limited technological capacity in the field's early days. It was not until the turn of the 21st century that the first pedagogical multimedia corpora came onto the scene, such as the ELISA corpus[90] (Braun, 2005) or the SCOTS corpus[91] (Anderson *et al.*, 2007). After making a call to the DDL community on the need for more speech corpora in DDL, Campbell el al. (2007) created the FluenCi corpus in order to study prosody and intonation in interactions between native speakers of English. FluenCi does not appear to be publicly available and, to our knowledge, no DDL studies have carried out using the corpus. Years later, Hasebe (2015) created the TED corpus[92], which Aston (2015) used in a DDL study on phraseological items in speech. The corpus contains hundreds of TED talks, mainly in English, although there are other languages available. Aston's study and the TED corpus are the most relevant to our own research objectives described below. A couple other multimedia corpora with pedagogic potential are the SACODEYL corpus[93] (Hoffstaedter & Kohn, 2009) and the VEIGA corpus[94] (Sotelo Dios & Gómez Guinovart, 2012). Nevertheless, audio-textual corpora continue to exist as a small minority compared to strictly text corpora.

Due to this lack of audio-textual corpora in DDL, we have created the LITTERA corpus[95], an audio-textual English-Spanish literary parallel corpus. It is comprised of 25 English language literary texts from various genres (novel, short story, young adult literature and children's literature). Each text has been aligned with its Spanish translation at the level of translation units (TUs), and each TU is accompanied by the corresponding audio from the audiobook. The corpus contains 1,968,609 words (982,115 in English and 986,494 in Spanish) and 63,693 translation units, which also means 63,693 corresponding audio segments. LITTERA exists as a sub-corpus within the CLUVI corpus[96]. CLUVI is a collection of parallel corpora with nearly 50 million words and includes texts and translations in Galician, Spanish, English, Catalan, Latin, Basque, Portuguese, French, Italian, German and Chinese throughout a variety of open access sub-corpora on its website.

LITTERA was created with pedagogical aims for Spanish-speaking university students of English. The corpus' search interface allows for a variety of search options. Searches may be carried out in English, Spanish or both simultaneously, and with the use of regular expressions

90 http://universal.elra.info/product_info.php?cPath=25&products_id=1835
91 https://www.scottishcorpus.ac.uk
92 http://yohasebe.com/tcse/
93 https://www.um.es/sacodeyl/
94 http://sli.uvigo.gal/CLUVI/index.php?corpus=23&lang=en
95 http://sli.uvigo.gal/CLUVI/index.php?corpus=24&lang=en
96 http://sli.uvigo.gal/CLUVI/index.php?lang=en

to make the search more flexible. It is also possible to limit the number of results that are displayed, making them more manageable and less overwhelming. With each translation unit in the results, the user can listen to the English audio with the option to rewind two seconds in order to listen to the same fragment as many times as necessary. The spoken English variety can also be specified (American or British) in the search menu.

LITTERA has been automatically processed using multiple NLP tools (FreeLing[97], IXA pipes[98] and UKB[99]) in order annotate each relevant word with its lexeme, part of speech and its *sense* according to the lexical synsets from WordNet[100]. This semantically annotated version of LITTERA can be accessed through SensoGal[101], a collection of parallel corpora totaling nearly 36 million words that have been semantically tagged at the lexical level, lemmatized and annotated with translation equivalences. The SensoGal corpus[102] was built up from a selection of sub-corpora from CLUVI with the purpose of studying the semantic processing of parallel corpora in research, pedagogy and in the development of language technologies. The SensoGal interface allows the user to search by form (specifying the word or lemma, with the additional option of filtering for part of speech) or by meaning (specifying the *sense* of the queried lexeme in WordNet).

While there are many possible applications of LITTERA in pedagogy, this presentation will focus specifically on English phonology. We will demonstrate how an audio-textual English-Spanish parallel corpus can be exploited by means of the search tools available through CLUVI and SensoGal in order to examine the segmental and suprasegmental aspects of English phonology that frequently cause difficulty for Spanish-speaking students such as vowel reduction, co-articulation (elision and assimilation) and the importance of stress words in connected speech. This type of information is essential for students whose aim is natural, native-like speech and to avoid the excessively "polished" speech that doesn't reflect the reality of spoken English.

Summing up, our presentation aims to demonstrate the potential of the audio-textual English-Spanish parallel literary corpus LITTERA in language learning by providing practical examples of how it can be used by students to explore English phonology through DDL, and more generally, through the framework of corpus-based language pedagogy.

Bibliography

Anderson, J., Beavan, D., & Kay, C. (2007). SCOTS: Scottish corpus of texts and speech. In J. Beal, K. Corrigan & H. Moisl (Eds), *Creating and digitizing language corpora* (pp. 17-34). London: Palgrave Macmillan.

97 http://nlp.lsi.upc.edu/freeling/

98 http://ixa2.si.ehu.es/ixa-pipes/

99 http://ixa2.si.ehu.es/ukb/

100 https://wordnet.princeton.edu/

101 http://sli.uvigo.gal/SensoGal/index.php?corpus=24&lang=en

102 http://sli.uvigo.gal/SensoGal/index.php?lang=en

Aston, G. (2015). Learning phraseology from speech corpora. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 65-84). Amsterdam: John Benjamins.

Braun, S (2006). ELISA–a pedagogically enriched corpus for language learning purposes. In *Corpus technology and language pedagogy: new resources, new tools, new methods* (pp. 25-47). Frankfurt: Peter Lang.

Campbell, D. F., McDonnell, C., Meinardi, M., & Richardson, B. (2007). The need for a speech corpus. *ReCALL*, 19 (1), 3-20.

Gómez Guinovart, X. (2019). Enriching parallel corpora with multimedia and lexical semantics: From the CLUVI Corpus to WordNet and SemCor. In I. Doval & M. T. Sánchez Nieto (Eds.), *Parallel corpora for contrastive and translation studies: new resources and applications* (pp. 141-158). Amsterdam: John Benjamins.

Hasebe, Y. (2015). Design and implementation of an online corpus of presentation transcripts of TED talks. *Procedia: Social and Behavioral Sciences,* 198 (24), 174-182.

Hoffstaedter, P., & Kohn, K. (2009). Real language and relevant language learning activities: insights from the SACODEYL project. In A. Kirchhofer & J. Schwarzkopf (Eds.), *The workings of the anglosphere. Contributions to the study of British and US-american cultures* (pp. 291-303). Trier: WVT.

Johns, T. (1997). Contexts: The Background, Development and Trialling of a Concordance-based CALL Program. In A. Wichmann [T][SEP]*et al.*[T][SEP] (Eds.), *Teaching and language corpora* (pp. 100-115). London: Longman.

Sotelo Dios, P., & Gómez Guinovart, X. (2012). A multimedia parallel corpus of English-Galician film subtitling. In A. Simões, R. Queirós & D. da Cruz (Eds.), *1st symposium on languages, applications and technologies*, (pp. 255-266). Saarbr

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Methodology for the treatment of multimodal corpora data: the C-ORAL-BRASIL corpora proposal

Camila Barros[103] **& Heliana Mello**[104]
[mila.ab98@gmail.com]
Federal University of Minas Gerais, Brazil

The research reported here proposes a protocol for the treatment of multimodal data through the concatenation and interplay of the sound signal and its corresponding gestuality in multimodal spontaneously produced data. The main objective was to connect information structure organization, as it is treated through the Language into Act Theory – L-AcT (Cresti 2000; Cresti & Moneglia, 2010), with the concept of "idea unit" as found in Kita and Özyürek (2003). The novelty of this methodological proposal stems from the crucial role prosody plays in the definitional categories found in L-AcT. In this presentation our focus is placed on how the parenthetical information unit (PAR), a metaelocutive unit that provides a comment about its host utterance content (Tucci, 2009), is gesturally mapped in multimodal data. The interplay between speech categories and their prosodic counterpart has not yet been sufficiently discussed and understood. Some pioneering research in the area can be found in Kendon (2004), McNeill (1992) and Loehr (2004). The gist behind the orchestrated correspondence between information units (and its corresponding prosodic units) and gestures comes from the actional nature of both activities (Kita & Özyürek, 2003). The steps followed in the reported research were organized in sequential phases. The first phase comprised the compilation of the corpus of multimodal speech monologues through the use of high-quality equipment that afford recordings to support fine grained analysis of both the speech signal and gestures. The second phase covered the transcription, segmentation into tonal units and information structure annotation following the guidelines of the C-ORAL corpora family (Cresti & Moneglia, 2005; Raso & Mello, 2012; Cantalini, 2018) and a simplified gestural annotation scheme based on Bressem, Ladewig and Müller's (2013). The software employed were Praat (Boersma & Weernick, 2020) and ELAN (2019). The third phase focused on data analysis, through the extraction of the relevant descriptive and analytical elements found in parenthetical information units and their gestural mappings. Results of four multimodal files analyzed point to a high percentage correspondence between gestural unit and tone unit boundaries and gestural evidence supporting the metaelocutive role of parenthetical units, i.e., gestures found seem to create a semantic domain that pertains to an interpretational level distinct from that found in the other textual units that make up the utterance content. The next step in the development of this research is to microanalyze each speech/gesture token found in the whole corpus in order to systematize the range of variability in the correspondence mapped between parentheticals and their

[103] Master's student at the Graduate Program in Linguistic Studies (PosLin/UFMG and Fapemig).
[104] Full Professor at the Federal University of Minas Gerais, affiliated to CNPq and Fapemig.

corresponding gestural units. The corpus comprises ten excerpts up to three minutes long, annotated taking into account the parameters of movement, position, direction and hand form along with the sound signal segmentation following the L-AcT guidelines. As an ongoing project, the minicorpus is still being compiled and, so far, it features 10 minutes of video that have gone through multimodal treatment, corresponding to roughly 200 annotated gestures and gathered in 2,000 words. This project is a pilot study within the C-ORAL-BRASIL research initiative aiming at the compilation and treatment of a larger multimodal corpus.

Keywords

Corpus Linguistics. Gestures. Speech. Parentheticals.

References

Boersma, Paul & Weenink, David (2020). *Praat:* doing phonetics by computer [Computer program]. Version 6.1.15, retrieved 20 May 2020 from http://www.praat.org/

Bressem, Jana & Ladewig, S. H. & Müller, Cornelia. (2013). Linguistic annotation system for gestures (LASG). In: Müller *et al.* (Eds.). *Body – Language – Communication (HSK 38.1)*. de Gruyter Mouton. p. 1098–1124

Cantalini, Giorgina. (2018). *La gestualità co-verbale nel parlato spontaneo e nel recitato.* Doctoral Dissertation. Università degli studi Roma Tre.

Cresti, Emanuela. (2000). *Corpus di Italiano Parlato.* Firenze: Accademia della Crusca.

Cresti, Emanuela & Moneglia, Massimo. (2005). *C-ORAL-ROM:* Integrated Reference Corpus for Spoken Romance Languages. John Benjamins Publishing Company.

Cresti, Emanuela & Moneglia, Massimo. (2010). Information Patterning Theory and the Corpus based description of Spoken language. The compositionality issue in the Topic Comment pattern. In: Moneglia, Massimo & Panunzi, A. *Bootstrapping Information from Corpora in a Cross Linguistic Perspective.* Firenze: FUP. p. 13-46.

ELAN (Version 5.8) [Computer software]. (2019). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Kendon, Adam (2004). *Gesture*: Visible Action as Utterance. Cambridge University Press.

Kita, Sotaro & Asli Özyürek. (2003). What Does Cross-Linguistic Variation in Semantic Variation. *Journal of Memory and Language,* 48, p. 16–32.

Loehr, Daniel. (2004). *Gesture and Intonation*. Doctoral Dissertation. Georgetown University.

McNeill, Daniel. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

Raso, Tommaso & Mello, Heliana. (2012). *C-ORAL-BRASIL I:* Corpus de Referência para o Português Falado Informal. Editora UFMG.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

On the Standardisation of Short Monosyllables in Early Modern English

Javier Calle Martín[105]
[jcalle@uma.es]
Universidad de Málaga, Spain

Middle English is elsewhere defined to display a high level of dialectal variation as a result of the absence of a written standard in the period according to which the same lexeme was bound to appear with a number of orthographic realisations depending on the scribes' origin. This variation is found at all language levels, not only on the grounds of spelling where the same item may appear with a plethora of variant forms (i.e. THOUGH '*thigh, theygh, youh, þoȝth, thau, þawe*, *þai, þay, yeȝe, þeght, þeȝþ, thought,* etc.), but also in morphology (i.e. the expression of the third person plural pronominals) and lexis (*child* vs *barn* 'child'), among many others. The Early Modern English period (1500-1700) brought some sort of order to the chaotic spelling system of Middle English with the progressive standardisation of the English language. Standardisation is here viewed as a process according to which users opt for particular linguistic forms in cases of linguistic competition relying on authoritative texts, which explains the outstanding role of Chancery writing in the propagation of standard forms. The selection of the standard forms was not an arbitrary process as it pursued the use of complex structures "because of their sense of the prestige and difference of formal written language" (Hope, 2000: 53; also Wright, 2000: 6). It was not until 1650 when we can properly talk about a standard system of spelling in printed documents.

The present paper assesses the spelling competition of short monosyllables in the history of English, paying particular attention to Early Modern English when their rendering was eventually standardised. The present investigation aims at accounting for the fate of the declining CVtt forms such as *lett*, *nott*, *butt* in favour of the CVt forms such as *let, not, but* in the period. Even though the former declined somewhat early (c. 1600), their constrained distribution in-between 1500-1600 stands out as a reliable clue as to their chronological stages of development, the genres that maintained these forms longest, and the speaker groups that were the last to use these forms. In the light of all this, the present paper has been conceived with the following objectives: a) to analyse the decline of these forms in the history of English to provide a convincing date for the standardisation of the CVt forms; b) to evaluate their distribution across legal, medical and speech-related texts in the assumption that this spelling competition evolved at a lower pace across the different text types; and c) to assess the sociolinguistic dimension of the phenomenon to determine whether there are social classes with

[105] El Dr. Javier Calle es Catedrático de Universidad en el Departamento de Filología Inglesa, Francesa y Alemana de la Universidad de Málaga. Su investigación se centra en la Historia de la Lengua Inglesa, el inglés contemporáneo y las distintas variedades de inglés en el mundo.

a more positive attitude to the declining forms. The proposal therefore evaluates the rendering of these words from the perspective of historical, genre, text type and sociolinguistic variation. In view of this, the historical analysis is based on the evidence provided by the corpus of *Early English Books Online*. The genre- and sociolinguistic analyses rely on *Early English Medical Texts*, *English Witness Depositions*, the *Corpus of English Dialogues* and the *Corpus of Early English Correspondence*.

References

Calle-Martín, J. and L. Esteban Segura. (forthcoming 2020). "Early Modern English Standardisation Revisited. New Insights into Early Modern English Standardisation". *International Journal of English Studies*.

Hope, J. (2000). "Rats, bats, sparrows and dogs: Biology, Linguistics and the Nature of Standard English". In L. Wright (Ed.), *The Development of Standard English, 1300–1800: Theories, Descriptions, Conflicts*. Cambridge: Cambridge University Press. 49–56.

Wright, L. (2000). "Introduction". In L. Wright (Ed.), *The Development of Standard English, 1300–1800: Theories, Descriptions, Conflicts*. Cambridge: Cambridge University Press. 1–8.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Online Corpora: an innovative approach in Colombia in EFL teaching and learning

Hernando Andrés Jiménez Rocha[106] **& Angélica María Ruiz Briceño**[107]
[hernandojimenez@usantotomas.edu.co | angelicaruiz@usantotomas.edu.co]
Universidad Santo Tomás, Colombia

EFL learners in monolingual countries unfailingly face some issues in their learning process related to a lack of contact with the language in context, especially because EFL is usually guided by prefabricated language books and samples. Nowadays, students' superficial development of abilities such as critical thinking, problem-solving, curiosity and autonomy is evident. This situation makes necessary the development and encouragement of strategies targeting contact with real-life language samples and virtual and technological approaches that are already an essential part of their lives. Although contact with foreign languages is today much more common and more necessary, in contexts such as Colombian that contact remains prefabricated, especially at an educational level. Then, it becomes necessary to ensure the direct contact of the student with the target language, so the latter can use it and reflect on it in a more friendly and practical way.

Bearing this in mind, the implementation of corpus linguistics directly faces difficulties related to grammar and vocabulary, as the corpora contain references in a real context and different language registers. This characteristic can allow students to get closer to the language they want to master more efficiently. McCarthy & Carter (2001) state that the Corpus-Based Approach to Language Teaching and Learning is very useful because it has "authenticity, sampling and representativeness" as its data demonstrates how language is really used, sampling and contrasting registers and periods. Thus, a Corpus can show a vast number of examples in real communication contexts and can be described as a collection of texts used for a specific purpose. In teaching or research, it can be applied to encourage students in their learning

[106] Teacher of languages; his research and subjects of investigation are on discourse analysis, sociolinguistics, morphosyntax, pragmatics, corpus – based approach and foreign languages teaching. He completed his Master´s degree, in the English and Linguistics faculty at Bangor University in the United Kingdom and during his MA studies in Linguistics was rewarded an awarded merit. He received his Bachelor's degree from the faculty of Science Education in the course of his BA in foreign languages, English, French and Spanish at La Salle University. He is currently a member of USTA– LEARNING research group and a full-time English teacher at Santo Tomas University.

[107] Professional with over 10 years' experience in foreign languages teaching (English -EFL, French-FLE and Spanish-ELE); her work focuses on the development of communicational skills, interculturality and research in English, French and Spanish. Graduated from the Pedagogical University of Colombia with a B.A in Spanish and Foreign Languages (English and French), she has obtained two master's degrees in France in Arts, Literature and Languages and in Sociolinguistics of Languages and Cultures from the University Jean Monnet of Saint-Étienne. She is currently a member of USTA-LEARNING Research group and a full-time languages teacher at Santo Tomas University.

processes by promoting curiosity, PBL, critical thinking and autonomy what makes the target language more reliable, natural and contextualized for them. Liu and Jiang (2009) affirm that by using corpora the instructors easily realized their students' grammar difficulties. Correspondingly, Aston (2001) argues that the Corpus-Based Approach to Language Teaching and Learning is very effective because its usage involves students to have a learning active process called "discovery by learning". In addition, Mishan (2002) suggests that it promotes cognitive functions, for instance, data-driven learning, problem-solving, analytical skills and language learning strategies. Liu (2011) suggests that the use of corpora in learning and teaching has been important in universities because of its ability to foster discovery by learning, grammar empowering and contextualization of language. Likewise, it promotes critical understanding and the analysis of patterns and rules. Also, Vyatkina (2013) states that the usage of corpora is an appropriate resource based on a complex exercise, in contrast to textbooks, which offer much less help.

This purpose constitutes the analysis of a study case, carried out at the Languages Institute Fray Bernardo de Lugo O.P, Santo Tomas University in Bogotá Colombia with a population of four groups of EFL intermediate level. The participants were 50 university students from four groups of pre-intermediate English levels (B1-B1+). These courses are offered as part of students' academic programs considering that the certification of a B1 or B2 CEFR level of English as a Foreign Language (EFL) is required to graduate, according to the Colombian Ministry of Education. It is framed in the principles of action-research and was held during an academic period corresponding to four (4) months.

The three corpora used in the development of this research were free online corpora, in order of use: Word Frequency Data, Corpus of Contemporary American English (COCA) and IWeb corpus. Moreover, students were aimed to use any dictionary of their preference, as a support to the activities proposed. The choice of COCA corpus is justified by two main reasons: it offers one of the widest amounts of data and it has the most recent updates, until 2019, compared to the British National Corpus (BNC). The use of IWeb and Word Frequency Data corresponds to the need for more specific contents that were not directly available on COCA. In fact, IWeb also offers the possibility of watching a wide variety of videos that allow users to check pronunciation and the use of the word/lemma in context; while Word Frequency Data displays a list of the top 5.000 words used in English, based on COCA and IWeb corpora. All of the corpora applied are based on the use of American English.

The main objectives rely on three main aspects: the introduction of alternative procedures to improve learners' process (use of linguistic corpora) and their ulterior analysis; fostering the use of IT tools in the classroom, and the development of 21st-century skills through Project-Based Learning, such as critical thinking and autonomy, among others cognitive functions. Based on the nature of Project-Based Learning, each activity was prepared to make students getting familiarized with one or more tools of the corpora through practice. Hence, every task was conceived following the same sequence: first, an inductive exercise developed by the instructor as an introduction to the topic, followed by a series of examples (two or three) to

reinforce the procedure and, finally, a product made by the students to consolidate its use. The tasks were also cohesively prepared, growing in difficulty and demanding the development of different abilities leading to critical thinking and argumentative analysis based on data.

This study follows the parameters of a mixed analysis method in which quantitative and qualitative data converge to allow a wider in-depth analysis of the study case. The data obtained is issued from three research tools: (1) students' project products, (2) researchers' logs and (3) a survey composed of a Likert question section supported by an open-ended question section. The main outcome of this research is to show the positive incidence of corpus linguistics in intermediate and advanced EFL teaching-learning courses. The findings made in this study can help EFL teachers not only to discover linguistic corpora but also to encourage their students to face their learning difficulties - mainly related to grammar and vocabulary - by using reliable technological and virtual tools. It can also foster the development of teaching strategies and didactics by strengthening the contact with real-context language samples, and the comprehension of the target language in context. By this means, students can develop deeper cognitive functions linked to analysis, critical thinking, PBL, discovery by learning and autonomous learning.

Keywords

English as a Foreign Language (EFL), Corpus linguistics, Corpus-Based Approach to Language Teaching and Learning, Project-Based Learning (PBL)

Bibliography

Aston, G. (Ed.). (2001). *Learning with corpora*. Athelstan.

Liu, D., & Jiang, P. (2009). Using a corpus-based lexicogrammatical approach to grammar instruction in EFL and ESL contexts. *The Modern Language Journal*, *93*(1), 61-78.

Liu, D. (2011). Making grammar instruction more empowering: An exploratory case study of corpus use in the learning/teaching of grammar. *Research in the Teaching of English*, 353-377.

McCarthy, M., & Carter, R. (2001). Size isn't everything: spoken English, corpus, and the classroom. *Tesol Quarterly*, *35*(2), 337-340.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Mishan, F. (2004). Authenticating corpora for language learning: a problem and its resolution. *ELT Journal*, *58*(3), 219-227.

Vyatkina, N. (2013). Discovery learning and teaching with electronic corpora in an advanced German grammar course. *Die Unterrichtspraxis/Teaching German*, *46*(1), 44-61.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Orthography in Online English: The *-our/-or* Alternative in English Worldwide

Marta Pacheco-Franco[108]
[mpachecof95@gmail.com]
Universidad de Málaga, Spain

Linguistic competition is the phenomenon in which at least two synonymous forms contend for a specific distributional domain. In the history of the English language, countless patterns and structures have undergone competition, often resulting either in the specialisation of one variant or in its ultimate extinction (Aronoff, 2019: pp. 41-2). The push towards standardisation in Early Modern English (henceforth EModE) made the period a crucial stage in the unfolding of this process (Nevalainen, 2006). Among many other developments, the period saw the struggle between the derivational suffixes *-our* and *-or*, which would conclude with the re-arrangement of the morphological and the orthographic paradigms (Pacheco-Franco and Calle-Martín, 2020: in press). In consequence, there exist in Present-day English two distinct spellings for words like *honour* or *labour*: the one hereby adopted is characteristic of British English (henceforth BrE) and the other, *honor* and *labor*, is representative of American English (AmE) (Peters, 2004: pp. 397-8; *OED* s.v. *-our*, *-or*, suffixes). Grammarians and lexicographers have duly noted that there exist exceptions to such a rule: the *-our* rendering is preferred in the latter variety for items like *saviour*, *glamour* and *honour* (Fowler, 1926; Greenbaum and Whitcut, 1988; Gramley and Patzold, 2004). Nonetheless, data from contemporary American sources seem to disprove such a claim, since these nouns occur more frequently regularised (that is, ending in *-or*) than not. Such a situation suggests that the Present-day English (henceforth PDE) spelling system is not rigid and that the process of competition attested in EModE might indeed continue in PDE at the level of orthography.

This paper thus aims to present a corpus-driven analysis on the distribution of the spelling variants *-our* and *-or* in PDE. The status of English as a global language entails that enquiry into linguistic change extend beyond BrE and AmE. Lying on the assumptions that most varieties of English worldwide are modelled after BrE and that AmE must be taking a toll on orthography so much as in other areas of culture around the globe, this study will analyse more than 1.5 million instances of the phenomenon in some inner and outer circle varieties as delimited by Kachru (2009). This means that the scope of the present investigation is three-

[108] Marta Pacheco-Franco is a PhD student at the Department of English, French and German of the University of Málaga (Spain). She graduated in English Studies and, after completing a Master's degree on Teaching Spanish as a Foreign Language and spending a year in the University of Guelph as an intern, she obtained her Master's degree in English Studies. Her MA dissertation was titled 'Suffixes in Competition: The -our/-or Alternative in English', which clearly shows that her interests lie within variation in the English language, a topic on which she plans to continue research.

fold. First, the data for the inner circle will be presented and analysed, including BrE, AmE and the varieties of Canada, Ireland, Australia and New Zealand. Since there already exists literature on the topic, the initial part of the analysis enables the contrasting of the information to the studies of Peters (2009), Fritz (2010), Heffernan *et al.* (2010) and Gonçalves *et al.* (2018), among others. The conclusions thereby reached will lay the foundation for the next section of the paper. Secondly, the status of both variants in the outer circle will be examined, with a special focus on the major varieties from Africa and from Asia.[109] This section will be the core of the paper as it provides insight into the issue of orthography in the outer circle, which has often been overlooked. Finally, the absolute data will be presented in order to ascertain whether the spelling of English worldwide is indeed changing and towards which standard it might be leaning.

The data for such an analysis will be drawn from the corpus of *Global Web-Based English* (henceforth *GloWbE*) (Davies, 2013). The *GloWbE* is a 1.9 billion-word corpus that provides a sizeable input for diatopic analysis since it contains texts from twenty varieties of PDE –most of which are to be analysed in the paper–. As the name of the corpus indicates, the texts featured come from up to 340,619 websites, including general webs (around 40%) and personal blogs (around 60%). The texts are characterised by their participation in the medium: they have been produced for online communication. This means that –other than blogs– texts from newspapers, magazines, governmental websites and company websites have been included. As for the number of words per variety, these differ in a significant manner: from over 300 million words in the case of AmE to 35 million words of Tanzanian English. This effectively raises the need for normalising the raw results, which were obtained in the following fashion. First, the entire list of items allowing for the aforementioned variation was retrieved. Then, the twenty-five topmost frequent words were selected as the input for the study. These amounted to 1,510,672 tokens. The number of occurrences for each item was then recorded, as well as the variety in which they were produced and the information regarding the text type (namely, whether they were produced in general websites or in personal blogs). Next, the raw data was normalised to a common base of 100,000 words, therefore enabling the researcher to compare the data from different varieties in the analysis of the results.

The standard orthographic realisations in BrE and in AmE have become, as suggested above, norm-providing, which means that hey are not as prone to change as others –yet-to-be-codified– varieties. In turn, the data obtained from BrE and AmE as regards the distribution of *-our* and *-or* will function as the models against which the remaining Englishes ought to be compared. With a historical perspective always at hand, and in consideration of Schneider's Dynamic Cycle (2007, 2010), the data will be interpreted in search for what will be considered as orthographic innovations (that is, the adoption of the *-or* spellings in typically British-influenced countries, or vice versa). Nonetheless, the very nature of the texts included in the *GloWbE* corpus requires further analysis from the perspective of English online. The features

[109] These are the Englishes from Ghana, Kenya, Nigeria, South Africa and Tanzania and from Bangladesh, Hong Kong, India, Malaysia, Pakistan, the Philippines, Singapore and Sri Lanka, respectively.

of these texts and this medium will also be considered in drawing the conclusions of the study. In this endeavour, Crystal's works (2001, 2009, 2011) will prove to be essential as they laid the foundations on this area of study.

Keywords: competition; suffixes; inner circle; outer circle; orthography

References

Aronoff, M. (2019). Competitors and Alternants in Linguistic Morphology. In F. Rainer, W. U. Dressler, F. Gardani, & H. C. Luschützky (Eds.), *Competition in Inflection and Word-Formation* (pp. 39–66). https://doi.org/10.1007/978-3-030-03550-2

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

Crystal, D. (2009). *Txtng the Gr8 Db8*. Oxford : Oxford University Press.

Crystal, D. (2011). *Internet Linguistics: A Student Guide*. New York: Routledge.

Davies, M. (2013). *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE)*. Web < https://www.english-corpora.org/glowbe/>.

Fowler, H. W. (1926). *A Dictionary of Modern English Usage*. Oxford: Clarendon.

Fritz, C. W. A. (2010). A Short History of Australian Spelling. *Australian Journal of Linguistics*, *30*(2), 227–281.

Gramley, S., and Pátzold, M. (2004). *A survey of Modern English*. London: Routledge.

Greenbaum, S., Whitcut, J., & Quirk, R. (1988). *Longman Guide to English Language*. Harlow: Longman.

Heffernan, K., Borden A. J., Erath A. C. & Yang, J. (2010). Preserving Canada's 'honour': Ideology and diachronic change in Canadian spelling variants. *Written Language and Literacy*, 13(1), 1–23.

Kachru, B. B. (2009). World Englishes in World Contexts. In H. Momma & M. Matto (Eds.), *A Companion to the History of the English Language* (pp. 567–580). Chicester: John Wiley & Sons.

Nevalainen, T. (2006). *An Introduction to Early Modern English*. Edinburgh: Edinburgh University Press.

Pacheco-Franco, M., & Calle-Martín, J. (2020). Suffixes in Competition: On the Use of –our and –or in Early Modern English. *IJES: International Journal of English Studies*, in press.

Peters, P. (2004). *The Cambridge Guide to English Usage*. Cambridge University Press.

Peters, P. (2009). Australian and New Zealand English. In H. Momma & M. Matto (Eds.), *A Companion to the History of the English Language2* (pp. 389–399). Chicester: John Wiley & Sons.

Schneider, E. W. (2007). *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press.

Schneider, E. W. (2010). Developmental patterns of English: similar or different? In A. Kirkpatrick (Ed.), *The Routledge Handbook of World Englishes*. New York: Routledge.

Simpson, John and Edmund Weiner. 1989. *Oxford English Dictionary*. Web www.oed.com

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

¿Por qué nos comemos la r?: la elisión de la consonante percusiva en posición final de verbos infinitivos en el corpus Preseea-Valledupar

Luis Terraza Turizo[110]
[luiisedua@gmail.com]
Universidad Popular del Cesar, Colombia

RESUMEN

En esta investigación se analiza la variación fonética de la consonante percusiva alveolar sonora /ɾ/ en posición final de verbos infinitivos en la comunidad de habla de Valledupar, con el fin de determinar qué factores lingüísticos y extralingüísticos condicionan la variación de la realización de /ɾ/. El estudio se hizo a partir de 36 entrevistas del corpus PRESEEA-Valledupar (2005)[111], conformado por 54 grabaciones de habla espontánea y distribuidas equitativamente en las variables sexo, nivel generacional (edad) y grado de instrucción (PRESEEA, 2003). Además, el corpus de PRESEEA-Valledupar es el primer corpus de habla de esta ciudad y uno de los primeros corpus recogidos en Colombia y en el Caribe, por lo que toda investigación del habla de Valledupar debe considerarlo como un referente. Es necesario señalar, además, que este es un corpus que ha sido infrautilizado puesto que, hasta ahora, el único antecedente investigativo en variación fonética es el estudio de Olmos y Gómez (2012). En este sentido, el presente estudio sobre el corpus PRESEEA-Valledupar permite ampliar la caracterización de esta comunidad de habla y establecer comparaciones con otros corpus del Caribe y de Colombia. En este estudio, se analizaron las variantes [ɾ] simple o plena (vibrante con oclusión), [ɹ] aproximante (vibrante sin oclusión) y [∅] elidida (cero fonético) de la consonante percusiva en posición final de verbos infinitivos, a partir de las variables lingüísticas categoría gramatical y posición del fonema /ɾ/, y contexto fónico siguiente. Las variantes analizadas en esta investigación están correlacionadas con otros estudios similares, en los cuales se han caracterizado las vibrantes del español a partir de análisis acústicos y se ha podido determinar la manifestación de cada una de la variantes de las vibrantes (Blecua, 2001; Ortiz de Pinedo, 2017; Quilis, 1993; Ugueto y González, 2013). Las variables extralingüísticas que se consideraron

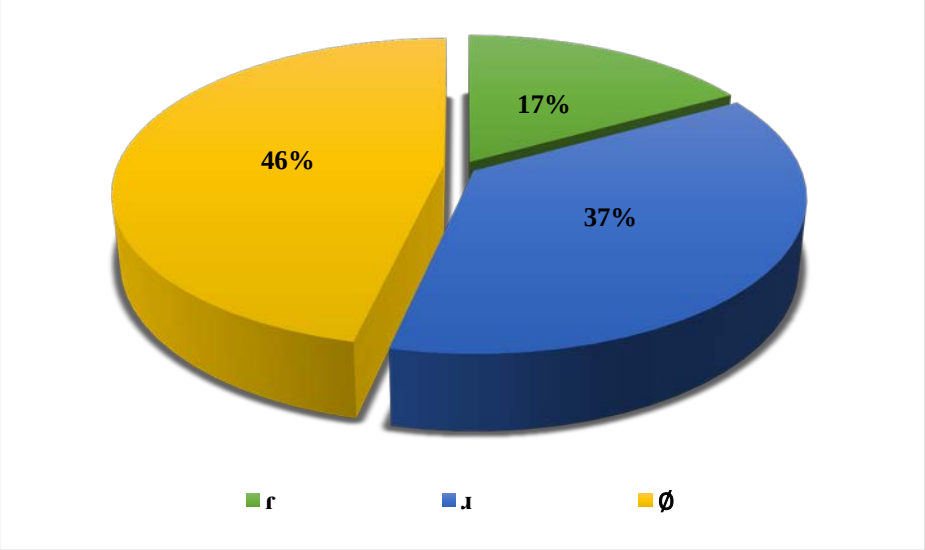
[110] Mi nombre es Luis Terraza, tengo 23 años. Hablo inglés en un nivel avanzado, y portugués en un intermedio. Amo la lingüística y la investigación. He tenido experiencia en la enseñanza de la lengua inglesa. Actualmente, yo curso el último semestre en la Licenciatura en Lengua Castellana e Inglés de la Universidad Popular del Cesar; y ejerzo mis prácticas profesionales en una institución educativa de Valledupar, Cesar. Hago corrección de estilo a tesis y doy asesorías a proyectos de investigación.

[111] El corpus fue recogido y coordinado por Donald Calderón como parte del Proyecto PRESEEA (2005).

fueron la edad, el sexo y el nivel de instrucción. El estudio y la distinción de las variantes de /ɾ/ se desarrolló a través de dos tipos de análisis: el perceptivo-acústico, que se realizó con el programa para el análisis acústico del habla PRAAT (6.1), usado en lingüística; y el segundo, el estadístico, que se hizo a través del programa Excel. Los resultados muestran que la variante de la consonante percusiva en posición final de verbos infinitivos más frecuente es la elisión [∅], seguida por la aproximante [ɹ], y la variante simple plena [ɾ] es la menos frecuente (véase el gráfico 1). En esta investigación, se establece que la variación está condicionada tanto por variables lingüísticas como extralingüísticas, sin embargo, en las realizaciones de /ɾ/, no se presentan diferencias significativas desde la variable lingüística contexto fónico siguiente y de acuerdo con la variable tipo de palabra y posición del segmento, se puede afirmar que, aunque, la posición de la consonante percusiva se encuentre interna o al final de los verbos infinitivos, la variación fónica se presenta, no por la variable distribucional, sino por la variable funcional. Es decir, la realización de /ɾ/ está más influenciada por el tipo de palabra que por la posición del segmento; hay variación independientemente de que se trate de verbos con o sin pronominalización. Las variables sociales que más inciden en la variación de la percusiva son la edad y el nivel de instrucción; la influencia de la variable nivel de instrucción es diciente y significativa, en cuanto que se ha comprobado que los hablantes de nivel bajo son los que más eliden el fonema en cuestión (ver gráfico 2); no obstante, la elisión es un fenómeno presente en todos los niveles de instrucción.

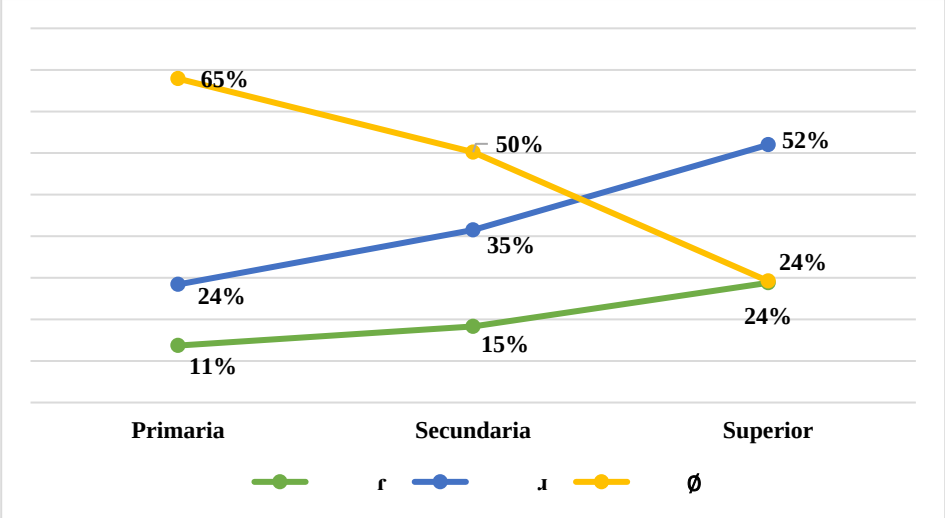
PALABRAS CLAVE: lingüística de corpus, variación lingüística, variación fonética, vibrante simple.

Gráfico 1. Porcentaje general de las realizaciones de /ɾ/



La distribución general de 3600 casos de las variantes de la consonante percusiva en posición final de verbos en infinitivo se presenta de la siguiente manera: la retención del segmento [ɾ] se manifiesta con una frecuencia de 608 casos, equivalentes al 17% de la muestra. La frecuencia de la aproximante [ɹ] es de 1319 casos, que equivalen a 37% y la vibrante elidida [Ø] se muestra con una frecuencia de 1673 casos, equivalentes a 46%.

Gráfico 2. Patrón de las variantes de /ɾ/ según el nivel de instrucción



BIBLIOGRAFÍA

Blecua, B. (2001). La vibrante del español: manifestaciones acústicas y procesos fonéticos (Tesis Doctoral). Universitat Autònoma de Barcelona, Barcelona, España.

Olmos, A. y Gómez, T. (2012). La aspiración y la elisión de /s/ implosiva en la comunidad de habla de Valledupar, Colombia. Universidad Popular del Cesar, Valledupar, Colombia.

Ortiz de Pinedo, N. (2017). Análisis acústico de las vibrantes del español en habla espontánea (Tesis Doctoral). Universidat de Barcelona, Barcelona, España.

PRESEEA. (2003). Proyecto Para el Estudio Sociolingüístico del Español de España y de América. Metodología general. Recuperado de http://preseea.linguas.net/Portals/0/Metodologia/METODOLOG%C3%8DA%20PRESEEA.pdf

Quilis, A. (1993). Tratado de fonología y fonética españolas. Madrid: Arcos/Libros, S.L.

Ugueto, M, y González, J. (2013). Análisis acústico de /r/ en posición final de palabra en el habla de Caracas. Letras, 55(89), 19-47. Recuperado de http://www.scielo.org.ve/scielo.php?script=sci_arttext&pid=S0459-12832013000200002&lng=es&tlng=es

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Pronunciation of consonant clusters in Spanish speakers based on the Czech read speech corpora

Kateryna Pugachova[112] **& Jitka Veroňková**[113]
[pugachova.katerynka@gmail.com | jitka.veronkova@ff.cuni.cz]
Charles University, Czech Republic

Czech and Spanish differ in the syllable structure, especially in the nature of consonant clusters. A coherent Czech text (838 words long) containing 75 target consonant clusters was created and sequentially recorded by 13 Spanish speakers. The goal was to carry out perceptual analysis in order to examine which types are difficult for non-native speakers and what types of sound changes are most frequent.

Introduction

Recently, the Czech Republic has been host to an increasing number of Latin American and Spanish people who usually study at universities or work. Proficiency in the local language is essential for the successful integration of a foreigner into a new community. The difficulties L2 students may encounter stem mainly from the influence of the L1 features on the target language. Sound characteristics distinguishing Czech and Spanish include syllabic structure and consonant groups. The primary difference lies in the number of consonants within a single syllable, their frequency, and combinatorics, e.g., CCVCC is the longest Spanish syllable type (occurrence of 0.01% [1] is very low); the same syllable type in Czech occurs in 0.26% [2]. The CCCV syllable type is not present in Spanish, unlike Czech (0.44% of syllables [2]). Difficulties with the pronunciation of consonant groups in Spanish speakers are generally known, especially in English as L2 [3]. In this paper, the pronunciation of consonant clusters in L2 Czech speakers with Spanish as L1 is presented. The goal is to examine which types are difficult for non-native speakers and what types of segment changes are the most frequent.

[112] Graduated in Linguistics from a Charles University in Prague. The field of her studies was Czech linguistics, Translation, and intercultural communication. She spent one year in Spain on the Erasmus program attending Hispanic studies at the University of Valencia. In Spain, she started research for a master thesis about the acquisition of Czech as a second language for speakers with Spanish as mother tongue, successfully defending a thesis later at Charles University.

[113] Institute of Phonetics, Faculty of Arts, Charles University in Prague. She focus on the sound structure of language and speech, especially the Czech language. As part of her pedagogical activities, she led courses mainly for the study programs of phonetics and Czech for foreigners. She has participated in phonetics courses and corrective pronunciation exercises for the Summer School of Slavonic Studies. Principal research interests include speech prosody, especially speech rate, and analyses of speeches elocution, and covers various aspects related to foreign accents in Czech.

Method

Material. Based on [1] and [4], such consonant clusters were selected, for which the pronunciation differs in Czech and Spanish, or such that may cause difficulties for L2 Czech speakers with Spanish as L1. Due to a large number of these groups, another selection followed. The selected clusters were then systematically supplemented, e.g., by combinations containing voiced/voiceless counterparts. Altogether, 23 clusters were included in the experiment: 2-consonant clusters – a) clusters [ps] (with the special subset of [psɪ]) and [pt]; b) bilabials [p], [b] and velars [k], [ɡ], each combined with nasals [n] and [ɲ]; c) consonant [s] combined with unvoiced obstruents, sonorants and [v] (S+cons); 3-consonant clusters – [psk] and [pst]. Note: In the following text, capital letters, i.e., [ps] PS are used, and [ɲ] is written as Ň. It was determined which clusters would be examined in the initial (I), middle (M), and final (F) positions.

A list of words containing the selected clusters in defined positions was created. In order to ensure that any errors would be a matter of personal pronunciation and not a case of ignorance of orthoepic rules, in the S+cons groups, only words, in which the graphic form and pronunciation of a target cluster did not differ due to voicing assimilation, were used. Trying to examine as many cases as possible while avoiding excessive text length, an average of 2–3 words per group and position were chosen; the number of words examined also depended on the number of suitable candidates, so, e.g., the GŇ cluster was not tested in the end. A total of 75 different words were selected, and the coherent text was created (838 words long). In order to prevent the spread of the analyzed group across the word boundary, the cluster in I was preceded by a vowel, and a vowel followed the cluster in F, or it was assumed that a pause would be realized.

Speakers. A group consisted of 13 speakers (10 Latin Americans, 3 Spanish; 10 males, 3 females) who were either from the author's circle of acquaintances or responded to requests posted on Facebook (social media). All speakers interested in the experiment were recorded. Women showed significantly less interest, which resulted in groups not being balanced by sex. The speakers' length of residence in the Czech Republic ranged from 1.5 years to 9.5 years, and the length of studying Czech ranged from 1 month of self-study to a graduate from a Czech university. A speaker, possibly considered bilingual, was also included in the research as his speech displayed similart endencies as the rest of the speakers.

Recordings were taken in a sound-treated and sound-proofed room (AKG C 4500 B-BC microphone, sample rate 32 kHz, 16-bit depth).

Perception analysis. Perception analysis was performed using Praat software [5]. Target words were transcribed, and the following aspects were examined: 1. Presence or absence of intonation juncture (pause) / liaison of the target word and adjacent words; 2. Concerning the target word itself: a) the degree of fluency (scale 0–4), b) acceptability, and intelligibility (scale 1–3); 3. Concerning the target consonant group: a) correctness (yes/no; a small group of uncertain

cases was decided upon repeated listening), b) acceptability and intelligibility (scale 1–5), c) type of error (sound change), d) affected segments.

Results

Global results. The original set of 975 realizations (target clusters) was analyzed: 66.2% of items were pronounced successfully, 26.9% contained an error, and 7.0% of items were affected by slips of tongue or dysfluency and were excluded from a detailed analysis.

Concerning the position within a word, the success in M and F was similar (M: 71.3%, F: 69.2%); it was a little bit lower (60.7%) in case of I. The I, M, F positions did not differ in the number of excluded cases, ranging from 6.4% to 7.2%.

Concerning individual speakers, the number of correct forms ranged from 45.3% to 84.0%, the number of errors ranged from 12.0% to 52.0%. Speakers also differed in the number of excluded cases ranging from 1.3% to 13.3%. Three speakers continuously studying Czech for more than 6 years and allegedly using Czech at work and in their everyday life displayed the lowest number of errors.

Sound changes. Three types of sound changes occurred with highest frequency of repetition: substitution 43.4%, elision 22.4%, and prothesis 20.6% (other types 13.5%). Fig. 1 shows their distribution in individual groups of clusters. The main findings are: a) the highest number of substitutions occurred within the BN/BŇ (86.2%), KN/KŇ (81.3%) and GN (75.0%) groups. It occurred in almost 50% of the PN group; b) the highest number of elisions was found in the PT (57.7%), PS (51.9%), PST/PSK (41.2%) and PN/PŇ (37.0%) groups; c) The prothesis occurred only in the S+cons group (unlike PS, PSY/PSI, PT, PN and GN) and it accounted for 51.8% (of instances); d) in the GN group, only substitution and elision were present.

The most frequent substitution was the change [s] → [z] (20.5% of all substitutions), mostly voicing assimilation in front of the sonorant ([slunt‿sɛ] → [zlunt‿sɛ]), and the substitution of nasal consonants [ɲ] → [n] (16.5%), ([barokɲi:] → [barokni:]). Most of the elisions affected the vowel [p] (73.0% of all elisions). It occurred frequently in PT ([pta:k] → [ta:k]), PS ([psɛudogotɪckɛ:ɦo] → [sɛudogotɪckɛ:ɦo]) and PN ([pnula] → [nula]). The insertion of a prothetic [ɛ] ([smutna:] → [ɛsmutna:]) is evenly distributed within S+cons group.

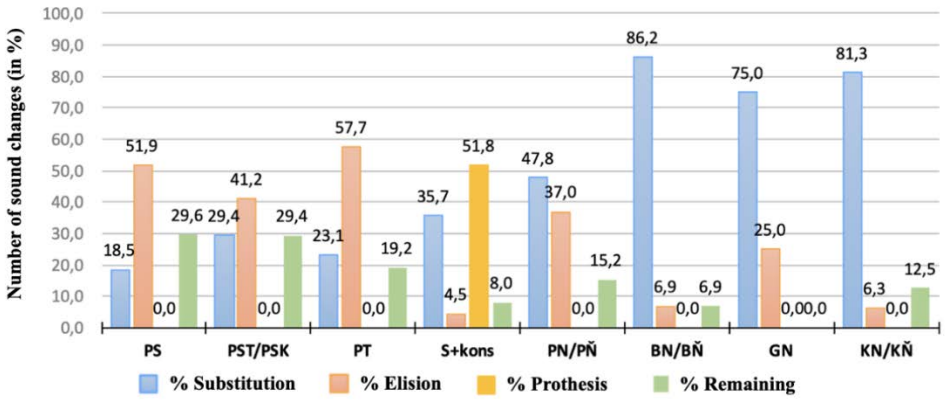


Fig. 1. Distribution of sound changes within consonant cluster groups (in %).

Conclusion

It has been proved that pronunciation of consonant clusters cause difficulties for Spanish speakers, including the advanced ones. On average, 2/3 of realizations were correct, 1/3 contained errors or slips of tongue or dysfluency. The results suggest that the initial position is more difficult, but the difference from medial and final positions is not that obvious. Substitution, elision and prothesis represent almost 90% of sound changes. Among them, the most frequent is substitution, which affected all clusters considered for the analysis. In BN/BŇ, KN/KŇ and GN, it was obviously the dominant sound change, as it occurred at least in ¾ of realizations. The occurrence of substitution is also significant for PN (almost half of the changes) and S+consonant (about 1/3 of changes). Elision appeared in both 2-syllable and 3-syllable clusters, starting with the consonant [p]; it was this consonant that was elided. In clusters PN, PS and PT, the elision is regularly found in Spanish, often even recommended [3]. Unlike most of the other sound changes, prothesis was present only in S+consonant clusters, and it accounted for more than half of all changes in this group. The reason for this may be the fact that S+consonant group is widely spread in Spanish but is standardly divided into two syllables with an additional vowel preceding the S+consonant cluster. In contrast, Czech allows this cluster as one syllable, which makes its pronunciation more complicated for a non-native speaker.

Acknowledgement. This research was supported by the Czech Science Foundation project No. 18-18300S "Phonetic properties of Czech in non-native and native speakers' communication".

Keywords

consonant, consonant clusters, Czech as L2, Spanish as L1, second language acquisition, sound change, substitution, elision, prothesis

References

[1] Baptista, B. O., & Watkins, M. A. (eds.) (2006). *English with a Latin Beat: Studies in Portuguese/Spanish-English Interphonology.* Amsterdam: John Benjamins.

[2] Těšitelová, M., & kol. (1985). *Kvantitativní charakteristiky současné češtiny*. Praha: Academia.

[3] Hidalgo Navarro, A., & Quillis, M. (2012). *La voz del lenguaje: Fonética y fonología del español*. Valencia: Tirant Humanidades.

[4] Quillis, A. (1993). *Tratado de fonología y fonética españolas*. Madrid: Gredos.

[5] Boersma, P., & Weenink, D. (2005). *Praat: Doing Phonetics by Computer* [online]. http://www.praat.org version 6.0.25 (2019).

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Recolección y análisis preliminar de un corpus de texto escrito en la producción de textos argumentativos universitarios

Karime Vargas Cáceres[114]
[karimevargascaceres@gmail.com]
Universidad Industrial de Santander, Colombia

Resumen

En la ponencia se pretende mostrar cómo se llevó a cabo el proceso de recolección y análisis preliminar de un corpus de lengua escrita que constituye el corpus de análisis central de un proyecto de tesis doctoral. El objetivo central del proyecto es identificar y analizar la influencia de las variables sociales: estrato, género y edad en los procesos de construcción de textos académico-argumentativos de estudiantes de una licenciatura de una universidad pública de Santander. Se busca identificar la incidencia de variables sociales en la selección de la clase de argumentos, la densidad argumentativa y los tipos de disputa en las prácticas de escritura argumentativa.

Para la recopilación de información, se solicitó la redacción de un texto argumentativo a partir de una pregunta genérica alrededor de la situación de profesionalización docente. La caracterización del corpus y análisis preliminar se hizo mediante los programas WordSmith Tools, versión 5 y TermoStat Web 3.0. Se recopiló un corpus con 33 muestras de texto, integrado por 13946 tokens, 2669 types, 410 oraciones y 41 párrafos. Con ayuda del Software TermoStat Web 3.0, se conocieron otros datos relevantes del corpus como la frecuencia de las palabras más recurrentes, la especificidad, las variantes ortográficas o lemas y las categorías gramaticales.

De igual modo, con WordSmith Tools, se deteminó de qué manera aparecen ciertas palabras en un texto en relación con otras. En el caso de los textos argumentativos, las clases o tipos de argumentos suelen estar introducidos por ciertas partículas o palabras que funcionan como marcadores discursivos. Por ejemplo, la presencia de un marcador como "sin embargo" marca una restricción a lo que se ha anunciado previamente. La presencia de un "según", podría indicar la presencia de un argumento de autoridad o el marcador discursivo "por ejemplo" pueden introducir argumentos por ejemplificación. Estas herramientas informáticas, entonces, nos permiten identificar con mayor precisión las partículas que introducen tipos de argumentos.

Los resultados de este ejercicio académico permitieron consolidar el corpus de análisis y hacer unas aproximaciones preliminares referidas a la selección de la clase de marcadores discursivos que introducen tipos de argumentos como ejemplificaciones o autoridades.

[114] Docente cátedra de la Escuela de Idiomas de la Universidad Industrial de Santander, Bucaramanga, Colombia. Magíster en Semiótica. Estudiante del Doctorado en Lingüística de Universidad de Antioquia.

Palabras clave: lingüística de corpus, discurso argumentativo universitario, variables sociales, argumentación.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Representar colocaciones verbales en recursos terminológicos mediante el uso de corpus

Melania Cabezas-García[115] **& Pamela Faber**[116]
[melaniacabezas@ugr.es | pfaber@ugr.es]
Universidad de Granada, España

Las colocaciones verbales (p. ej. *to absorb energy*) son unidades fraseológicas muy habituales en la lengua general y el discurso especializado. Conocerlas resulta esencial para expresarse de forma idiomática y construir y difundir el conocimiento en el discurso científico-técnico. En este sentido, los recursos lexicográficos figuran entre las herramientas más empleadas para resolver este tipo de cuestiones de expresión. La Lingüística de Corpus ha facilitado el análisis y la incorporación de datos lingüísticos en los recursos lexicográficos, siendo las colocaciones una de las combinaciones más estudiadas (Sinclair 1991; Corpas Pastor 1996; Montero Martínez 2008; Buendía Castro 2013). Sin embargo, a pesar de la utilidad de encontrar información fraseológica en los recursos terminológicos, este tipo de datos no siempre se incluye o, cuando lo hace, no se presenta de forma sistemática (L'Homme & Leroyer 2009; Buendía Castro 2013).

Nuestro estudio sigue las premisas de la Terminología Basada en Marcos (Faber *et al.* 2005, 2006; Faber 2012) y se centra en las colocaciones verbales del discurso científico mediante el análisis de un corpus comparable de artículos de investigación sobre protección medioambiental. El corpus está compuesto por un subcorpus en inglés de 1 142 609 palabras y un subcorpus en español de 992 135 palabras, que incluyen textos extraídos de revistas especializadas. Dado el abundante material sobre protección medioambiental que se produce cada día en distintas lenguas, para la comunicación internacional resulta fundamental su correcta comprensión y traducción. De este modo, nuestro objetivo consistía en la extracción

[115] Doctorado en Traducción e Interpretación por la Universidad de Granada en 2019. Es Graduada en Traducción e Interpretación por la Universidad Pablo de Olavide y Máster en Traducción Profesional por la Universidad de Granada, donde imparte docencia en el Grado en Traducción e Interpretación. Pertenece al grupo de investigación LexiCon y contó con una beca de investigación predoctoral otorgada por el Ministerio de Educación de España para realizar su tesis doctoral, en la que estudió la formación, traducción y representación de los términos compuestos en inglés y español. Sus áreas de investigación son la terminología, la lingüística de corpus y la traducción especializada. Ha publicado artículos en revistas internacionales sobre lingüística, terminología y traducción especializada, y forma parte del comité científico de congresos y revistas internacionales, como LREC y el International Journal of Lexicography.

[116] Catedrática de Traducción e Interpretación en la Universidad de Granada desde 2001, donde imparte clases sobre Terminología, Traducción Especializada, Semántica Léxica y Lingüística Cognitiva. Es la directora del grupo de investigación LexiCon, con el que ha llevado a cabo varios proyectos de investigación del plan nacional español. Uno de los resultados de estos proyectos, así como la aplicación práctica de su Terminología Basada en Marcos, es EcoLexicon (ecolexicon.ugr.es), una base de conocimiento terminológica sobre el medio ambiente. Cuenta con más de 100 publicaciones y pertenece al comité científico y editorial de distintas revistas, como Terminology o el International Journal of Lexicography.

de información fraseológica del corpus para enriquecer la representación de las colocaciones verbales en la base de conocimiento terminológica multilingüe EcoLexicon (Faber 2012; San Martín *et al.* 2017). Para ello, seleccionamos los verbos más frecuentes mediante la herramienta de extracción TermoStat (http://termostat.ling.umontreal.ca/, Drouin 2015). A continuación, estos verbos se clasificaron según su significado en los dominios léxicos propuestos por Faber & Mairal (1999). Por ultimo, exploramos su estructura argumental en ambas lenguas utilizando Sketch Engine (https://www.sketchengine.eu/, Kilgarriff *et al.* 2004).

En nuestros resultados destaca la riqueza conceptual y fraseológica que puede extraerse de los corpus. Esta información se utilizó para poblar el módulo fraseológico que se ha desarrollado recientemente en EcoLexicon (San Martín *et al.* 2017). Además, se observó que los verbos que pertenecen a un mismo dominio léxico suelen presentar estructuras argumentales similares, lo que facilita la inferencia de nuevos verbos pertenecientes al mismo dominio. Por otro lado, la comparación de la combinatoria de los verbos en inglés y español reveló un comportamiento similar de los predicados equivalentes en ambas lenguas (p. ej. *affect* y *afectar*). Sin embargo, estos a menudo muestran estructuras preferentes que difieren en la otra lengua (p. ej. la preferencia por la voz activa o pasiva, o por determinadas preposiciones). Por ello, los recursos terminológicos deben incluir información sobre este tipo de coocurrencias, mejorando así la experiencia de los usuarios. Por ultimo, observamos la influencia de la estructura argumental en la formación de términos compuestos (p. ej. *absorción parcial de energía*), que constituyen uno de los principales mecanismos de formación de términos en inglés y español, y pueden ser problemáticos en los textos especializados (Cabezas García 2019, 2020).

Palabras clave: colocación verbal; fraseología; estructura argumental; corpus; terminología

Referencias

Buendía Castro, Miriam (2013). *Phraseology in Specialized Language and its Representation in Environmental Knowledge Resources*. Tesis doctoral. Granada: Universidad de Granada.

Cabezas García, Melania (2019). *Los compuestos nominales en terminología: formación, traducción y representación*. Tesis doctoral. Granada: Universidad de Granada.

Cabezas García, Melania (2020). *Los términos compuestos desde la Terminología y la Traducción*. Berlín: Peter Lang.

Corpas Pastor, Gloria (1996). *Manual de Fraseología Española*. Madrid: Editorial Gredos.

Drouin, Patrick (2015). TermoStat. En *TermoStat*. http://termostat.ling.umontreal.ca/

Faber, Pamela & Mairal, Ricardo (1999). *Constructing a Lexicon of English Verbs*. Berlín: Mouton de Gruyter.

Faber, Pamela (2012). *A cognitive linguistics view of terminology and specialized language*. Berlín, Boston: De Gruyter Mouton.

Faber, Pamela; Márquez Linares, Carlos & Vega Expósito, Miguel (2005). Framing Terminology: A process-oriented approach. *Meta: Journal des traducteurs / Meta: Translators' Journal, 50*(4).

Faber, Pamela; Montero Martínez, Silvia; Castro Prieto, Rosa; Senso Ruiz, José; Prieto Velasco, Juan Antonio; León Araúz, Pilar; Márquez Linares, Carlos & Vega Expósito, Miguel (2006). Process-oriented terminology management in the domain of Coastal Engineering. *Terminology, 12*(2):189-213.

Kilgarriff, Adam; Rychlý, Pavel; Smrz, Pavel & Tugwell, David (2004). The Sketch Engine. En *Proceedings of the 11th EURALEX International Congress*, 105-115. Lorient: Euralex.

L'Homme, Marie-Claure & Leroyer, Patrick (2009). Combining the semantics of collocations with situation-driven search paths in specialized dictionaries. *Terminology*, *15*, 258–283.

Montero Martínez, Silvia (2008). Constructional Approach to Terminological Phrasemes. En *Proceedings of the XIII EURALEX International Congress*. Barcelona: Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra) / Documenta Universitaria.

San Martín, Antonio; Cabezas-García, Melania; Buendía, Miriam; Sánchez-Cárdenas, Beatriz; León-Araúz, Pilar & Faber, Pamela (2017). Recent Advances in EcoLexicon. *Dictionaries: Journal of the Dictionary Society of North America, 38*(1):96-115.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Subcorpus GITA_VOT. Análisis de las consonantes obstruyentes oclusivas sordas en personas con Enfermedad de Parkinson

Patricia Argüello-Vélez[117]
[patricia.arguello@udea.edu.co]
Universidad de Antioquia-Universidad Santiago de Cali, Colombia
Tomás Arias Vergara[118]
[tomas.arias@udea.edu.co]
Universidad de Antioquia, Colombia y *Friedrich-Alexander University*, Germany
Maria Claudia González-Rátiva[119]
[mclaudia.gonzalez@udea.edu.co]
Universidad de Antioquia, Colombia
Juan Rafael Orozco Arroyave[120]
[rafael.orozco@udea.edu.co]
Universidad de Antioquia, Colombia y *Friedrich-Alexander University*, Germany
Elmar Nöth[121]
[elmar.noeth@fau.de]
Pattern Recognition Lab, *Friedrich-Alexander University*, Germany
Maria Schuster[122]
Department of Otorhinolaryngology, Head and Neck Surgery, Ludwing-Maximilians University, Germany

Resumen

El corpus PC-GITA consiste en una base de datos desarrollada a partir de grabaciones de personas con Enfermedad de Parkinson (EP) y sus respectivos controles. Consta de una

[117] Fonoaudióloga- Universidad Santiago de Cali, Candidata a doctora en Lingüística Universidad de Antioquia.
[118] ngeniero Electrónico de la Universidad de Antioquia, Candidato a Doctor en Ingenieria, Investigador del grupo GITA.
[119] Doctora en lingüística de la Universidad de Antioquía, Investigadora en el Grupo de Estudio Sociolingüísticos GES.
[120] Doctor en Ciencias de la computación Friedrich-Alexander. Universität Erlangen. Nürmberg, Germany. Investigador del grupo GITA.
[121] Professor for Applied Computer Science at the Friedrich-Alexander. Universität Erlangen. Nürmberg, Germany.
[122] PhD. Médica especialista en foniatría y audiología pediátrica.

muestra de 100 personas hablantes nativos de español colombiano, distribuidas en 50 personas con EP y 50 personas sanas. En el caso de las personas con EP se tiene en cuenta la distribución de variables clínicas obtenidas a partir del diagnóstico neurológico de acuerdo al tiempo de diagnóstico de la enfermedad y las escalas MDS-UPDRS III y H&Y (Orozco-Arroyave, Arias Londoño, Vargas-Bonilla, González-Rátiva, & Nöth, 2014; Orozco-Arroyave, 2016).

El subcorpus GITA_VOT surge desde la perspectiva lingüística, especialmente fonética acústica, para la caracterización segmental de las consonantes obstruyentes oclusivas sordas, durante la producción de diadococinecias (DDK). Este subcorpus retoma la distribución general de variables clínicas, sociodemográficas y lingüísticas consideradas en el PC-GITA, además de especificar componentes acústicos propios de las consonantes oclusivas: *voice onset time* (VOT), duración de la consonante, duración de la oclusión, punto de máxima concentración de energía en la barra de explosión, y la descripción acústico-perceptual de los fenómenos glóticos y supraglóticos en la producción alternada de sílabas.

La consolidación de este subcorpus GITA_VOT se constituye por aproximadamente 3000 ocurrencias de DDK [pa ˈ taka], y cada una de ellas se complementa con cuatro especificaciones acústicas que dan cuenta de la integridad fonética de los elementos segmentales. Este proceso se realizó mediante la elaboración de un protocolo de segmentación y etiquetado, transcripción, registro computarizado de los datos, y manejo de archivos .Collection que, contienen por producción, la grabación, el etiquetado y el corte del espectro de cada una de las consonantes.

El subcorpus GITA_VOT presenta los parámetros acústicos en una distribución que facilita la adaptación a metodologías automatizadas.

Sílaba	Repetición	Inicio (VOT)	Final (VOT)	Duración de la consonante	Duración de la oclusión	Máxima energía	f. glótico	f. supraglótico

Los fenómenos glóticos y supraglóticos se abordan desde un enfoque acústico-perceptual, con el establecimiento de variables categóricas. Los fenómenos glóticos describen el comportamiento laríngeo por medio de la presencia, ausencia o dispersión de los pulsos glotales en el oscilograma; y los fenómenos supraglóticos se refieren a la descripción espectrográfica de la barra de explosión.

La metodología con la que se consolidó el subcorpus GITA_VOT actualmente está siendo utilizada en el proceso de segmentación, etiquetado y análisis manual fonético acústico de las consonantes obstruyentes oclusivas sordas en alemán y checo para su adaptación a métodos computacionales y automatizados que permitan detectar y monitorear la EP.

Palabras clave: corpus, fonética clínica, fonética acústica.

Bibliografía

Ackermann, H., Hertrich, I., & Hehr, T. (1995). Oral Diadochokinesis in Neurological Dysarthrias. *Folia Phoniatrica et Logopaedica*, *47*(1), 15–23. https://doi.org/10.1159/000266338

Baum, S. R., & Ryan, L. (1993). Rate of Speech Effects in Aphasia: Voice Onset Time. *Brain and Language*, *44*(4), 431–445. https://doi.org/10.1006/brln.1993.1026

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, *23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

Hoehn, M. M., & Yahr, M. D. (1967). Parkinsonism: onset, progression and mortality. *Neurology*, *17*(5), 427–442. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6067254

Martínez Celdrán, E. (2013). *Los sonidos obstruyentes en la cadena hablada*. M. A. Penas (Ed.), Panorama de la fonética española actual (pp.253-290). Madrid: Arco/Libros.

Martínez Celdrán, E. (2016). *En torno al concepto de aspiración o sonido aspirado*. (Elvira-García & Roseano, Eds.). Barcelona: Laboratori de Fonètica de la Universitat de Barcelona. Retrieved from http://stel.ub.edu/labfon/amper/homenaje-eugenio-martinez-celdran/homenaje.html

Montaña, D., Campos-Roca, Y., & Pérez, C. J. (2018). A Diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of Parkinson's disease. *Computer Methods and Programs in Biomedicine*, *154*, 89–97. https://doi.org/10.1016/J.CMPB.2017.11.010

Orozco-Arroyave, J., Arias, J., Vargas, J., González Rátiva, M., & Nöth, E. (2014). New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. *LREC 2014. Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 342–347. Retrieved from http://www.lrecconf.org/proceedings/lrec2014/summaries/7.html

Orozco-Arroyave, Juan Rafael. (2016). *Analysis of Speech of People with Parkinson's Disease*. Logos Verlag Berlin.

Wang, Y.-T., Kent, R. D., Duffy, J. R., & Thomas, J. E. (2009). Analysis of diadochokinesis in ataxic dysarthria using the motor speech profile program. *Folia Phoniatrica et Logopaedica : Official Organ of the International Association of Logopedics and Phoniatrics (IALP)*, *61*(1), 1–11. https://doi.org/10.1159/000184539

Wong, M. N., Murdoch, B. E., & Whelan, B.-M. (2012). Lingual kinematics during rapid syllable repetition in Parkinson's disease. *International Journal of Language &*

Communication Disorders, *47*(5), 578–588. https://doi.org/10.1111/j.1460-6984.2012.00167.x

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

The "Small World of Words" Free Association Norms for Rioplatense Spanish

Álvaro Cabana[123], **Camila Zugarramurdi**[124]
[czugarramurdi@psico.edu.uy]
Universidad de la República, Uruguay
Simon De Deyne[125]
[simon2d@gmail.com]
University of Melbourne, Australia

Word association norms have been useful in psycholinguistic research both as tools for experimental design (for measuring and controlling association strength) and as devices for creating useful semantic representations. The "Small World of Words" project (De Deyne *et al.*, 2019) is a large-scale scientific study that so far has produced sizeable free association norms in major languages such as English, Dutch, Mandarin and Spanish. Until now, Spanish norms were available for Iberian Spanish only; given the differences in vocabulary use between Iberian and Latin American varieties, norms for major Spanish dialects are needed. In this work we present Rioplatense Spanish word association norms for over 13,000 cues, collected from over 70,000 participants. Data was gathered online in Argentina and Uruguay following the "Small World of Words" methodology. A graph-theoretic analysis of the data shows the common "small world" signature characteristic of semantic graphs in other languages. We also investigated the external validity of these norms as a way of studying a variety of word processing tasks by comparing them to Iberian Spanish. The results show that the new Rioplatense norms, although largely compatible with Iberian Spanish data, provide a better account of lexical decision and semantic similarity tasks, especially when confronted with data from rioplatense native speakers.

[123] Álvaro Cabana is a cognitive neuroscientist at the Center for Basic Research in Psychology, at the Universitdad de la República in Uruguay. He is interested in the use of mathematical modeling and new data analysis techniques to neurophysiological and behavioral data, in order to understand how the mind/brain understands and produces language and culture.

[124] Camila Zugarramurdi is a cognitive neuroscientist at the Center for Basic Research in Psychology, at Universidad de la República in Uruguay. Her research interests are centered in the neural and cognitive precursors of reading acquisition in children. She uses behavioural and neurophisiological data to approach this problem.

[125] Simon de Deyne is a cognitive scientist at the Computational Cognitive Science Lab and the Human Complex Data Hub at the School of Psychological Sciences, at the University of Melbourne. He is interested in how the mind acquires and represents word meaning through our experiences with the world and the use of language. Current work addresses questions about how connotative meaning varies across different languages and cultures, and how bilinguals learn and represent word meaning.

Keywords: free association, psycholinguistics, semantic similarity

Bibliography

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior research methods*, *51*(3), 987-1006.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Translation correspondences of 'in fact', 'indeed' and 'de hecho': a corpus-based study

Mª Estefanía Avilés Mariño[126]
[meaviles@ucm.es]
Universidad Complutense de Madrid, Spain

This paper is focused on the analysis of three discourse and speech elements, two English ones 'indeed' and 'in fact', and one Spanish one 'de hecho'. What motivates this analysis is the fact that 'indeed' and 'in fact' belong to a group of markers addressed by Halliday and Matthiessen (2015) in their taxonomy as 'Verificative' (VMs) but which no other academics seem to classify in such a category. It is true, though, that these VMs seem to appear under other labels such as markers of Corroboration (Aijmer, 2013) or pragmatic Markers, whereas other academics seem to ignore them, as it is the case of Mª Ángeles del Saz and her study of markers of reformulation.

Provided that the study of Discourse markers in translation is small in comparison to other linguistic elements, specially on the area of parallel corpora (Aijmer, 2013) and the fact that parallelisms between English and Spanish are scarcer, their study seemed of offer an interest for both functional linguistics and translation.

The choice of these three markers has been motivated due to the frequency of usage of each of them in the corresponding language and the fact that both English markers tend to be mostly translated as 'de hecho' and 'de hecho' seems to be mostly translated as either 'indeed' or 'in fact'. These results have been possible due to a process of annotated translation and the usage of the online tool Sketch Engine, with different parallel corpora so as to include as many genres and styles possible, the usage of back translation was meant to confirm the bilateral correspondence between the two English Markers and the Spanish one which also demonstrates the closeness in meaning between 'indeed' and 'in fact':

- OpenSubtitles (Hummel,1995): indeed (8.41 per million), in fact (9.17 per million), de hecho (34.62 per million).
- Opus2 (2019): indeed (28.5 per million), in fact (21.54 per million), de hecho (48.11per million).

[126] PhD in English Linguistics (Universidad Complutense de Madrid). Master in english teaching for higher education (Universidad a Distancia de Madrid (UDIMA)) BA in English Studies (Universidad Complutense de Madrid). Major in English Linguistics. Major in English Literature. UCM: Associate professor. CENP: Centro español de nuevas profesiones (University). Universidad Complutense de Madrid: Lenguas Modernas y Traductores (IULMyT).Textlink2018 TOULOUSE: 27ESFLC PROCEEDINGS VOLUME. Lavid, J. and Avilés, Estefanía. "Investigating Discourse Markers in English and Spanish: The case of elaborating connectives". Member of FUNCAP research group UCM https://www.ucm.es/funcap/el-grupo.

- EMEA (Hummel,1995): indeed (1.82 per million), in fact (0.91 per million), de hecho (-).
- Europarl7 (2019): indeed (271.38 per million), in fact (217.86 per million), de hecho (115.65 per million).

The results of the parallel annotated translation have been the following:

(1) A corpus-based translation studying the frequency of the latter VMs in English and Spanish (in fact, indeed, de hecho) through their translation correspondences, on the basis of their translation equivalents in the target language.

(2) The finding of peripheral meanings for the markers as well as pragmatic implicatures which can be derived from them.

- **Indeed** , they represent a celebration of the indomitable spirit of our people→ Connector.
- I am pleased to say that we have **indeed** done much to implement → Emphasis Adverb.
- Yes, **indeed** → Reafirmation, pragmatic.
- **Indeed**! → Unique speech act.

(3) Constructing contrastive (English-Spanish) semantic fields or translation lexical domains of VMs on the basis of their translation equivalents (Hummel, 1995).

(4) Distinguishing which is the best genre and register for their translation equivalents.

It is expected that the reported work opens the way for future research on the creation of bilingual resources for applications in Translation Studies (TS), Machine Translation (MT) and Statistically- Based Machine Translation (SBMT) (Newmark, 1988). The latter is expected to be achieved by providing a list of suitable translation equivalents according to context and function (Van Dijk, 2006), avoid literal translations which do not investigate the communicative purpose of discourse and speech while working with SMT (Statistical Machine translation).

Bibliography

Aijmer, K. (2013). *Understanding Pragmatic Markers*. Edinburgh.

Europarl7 English corpus: https://www.sketchengine.eu/europarl7-english-corpus/ (2019).

Halliday, M and Matthiessen, M.I.M (2004, 2015): Introduction to Functional Grammar. Hodder Head- line Group. London, Great Britain.

Hummel, K.M. (1995). "Translation and second language learning". Canadian Modern Language Review. utpjournals.press

Newmark, P. (1988). A Texbook of Translation. NY: Prentice Hall.

Opus2 English corpus: https://www.sketchengine.eu/opus2-english-corpus/ (2019).

Van Dijk. T. A. (2006). "Discourse, context and cognition". Discourse studies.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Un estudio de la terminología utilizada para definir el concepto de "lengua artificial" a partir del análisis de textos científicos

Leticia Gándara Fernández[127]
[leticiagf@unex.es]
Universidad de Extremadura, España

La búsqueda de una lengua perfecta ha sido desde siempre una de las tareas que con encomiable esfuerzo han acometido estudiosos e investigadores, lingüistas y no lingüistas, preocupados por solventar los problemas derivados de la comunicación entre naciones. El intento por conseguir la soñada lengua utópica no solo forma parte de la cultura europea, sino que el tema de la confusión de lenguas, así como el intento de remediarla a través de la recuperación e invención de una lengua común, ha sido siempre un tema recurrente en la historia de todas las culturas (Eco, 1994: 12). Los primeros sistemas lingüísticos de elaboración artificial aparecen en el siglo XVII con el "deseo de garantizar la transparencia, racionalidad y univocidad de la comunicación científica mediante nuevos sistemas con proyección universal" (Galán, 2018: 75). Desde entonces, el movimiento de creación de lenguas artificiales ha recorrido todas las etapas de la historia, acompañado inevitablemente de una evolución interna en sus motivaciones y en sus resultados, consecuencia del marco socio-cultural de cada periodo histórico (Martínez, 2016: 78). Los propósitos que guiaron la construcción de los primeros esquemas lingüísticos poco tienen que ver con la finalidad por la que actualmente se siguen diseñando lenguas. Esto ha motivado a su vez un aumento en el número de términos utilizados para designar dichos proyectos y en que estos no siempre se empleen adecuadamente.

En el área de la Lingüística, más concretamente en el campo de la Interlingüística, existe un amplio abanico de denominaciones para nombrar los sistemas semióticos de elaboración artificial. Algunos de estos son: "lengua universal", "lengua auxiliar", "lengua artificial", "lengua internacional", "lengua del mundo", "lengua artística", "lengua construida", "ideolengua", etc.

Las expresiones mencionadas forman parte de la "terminología técnica" utilizada para los diseños lingüísticos artificiales. Dichos términos se emplean indistintamente y con fines dispares en trabajos académicos, enciclopedias, diccionarios y obras de autores, lo que conlleva ciertos errores y ambigüedades. En este sentido, conviene advertir que la Interlingüística es una rama de la Lingüística relativamente joven, por lo que no presenta un enfoque unificado, ni tampoco

[127] Doctora en Lenguas y Culturas con Mención Internacional y Premio Extraordinario por la Universidad de Extremadura y Máster en Enseñanza de Español como Lengua Extranjera por la Universidad Pablo de Olavide. Su formación se complementa con el Máster Universitario en Investigación en Artes y Humanidades y el Máster Universitario en Formación del Profesorado. Su línea de investigación principal se centra en el estudio de las lenguas artificiales.

una nomenclatura precisa. Por ello, en esta investigación proponemos un estudio sobre la terminología empleada para designar estos sistemas mediante el análisis de textos científicos en los que se han utilizado. Para ello, tomaremos como referencia la clasificación expuesta por Blanke en su trabajo "The Term "Planned Language"" (1997) para proyectos construidos hasta mediados del siglo XX. Con respecto a las lenguas diseñadas desde mediados del XX, realizaremos un estudio de los textos reales en los que se emplean estos términos. En suma, este trabajo nos permitirá ofrecer, por un lado, una muestra del uso de estos vocablos en los diferentes textos científicos analizados y, por otro, una clasificación de los mismos en función de su significado y uso.

Palabras clave

lengua artificial; problemas terminológicos; análisis de términos; textos científicos.

Bibliografía

Albani, Paolo y Berlinghiero Buonarroti (2010), *Dictionnaire des langues imaginaires*, Paris, Les Belles Lettres.

Ball, Douglas (2015), "Constructed languages", en Jones, Rodney H. (ed.), *The Routledge Handbook of Language and Creativity*, Londres, Routledge, pp. 143-157.

Blanke, Detlev (1997), "The term 'Planned Language'" en Tonkin Humphrey (ed.), *Esperanto, Interlinguistics, and Planned Language*, Oxford, University Press of America, Center for Research and Documentation on World Language Problems, pp. 2-20.

Blanke, Detlev (1997), "The term 'Planned Language'" en Tonkin Humphrey (ed.), *Esperanto, Interlinguistics, and Planned Language*, Oxford, University Press of America, Center for Research and Documentation on World Language Problems, pp. 2-20.

Eco, Umberto (1994), *La búsqueda de la lengua perfecta*, Barcelona, Crítica.

Galán Rodríguez, Carmen (2018a), «Género, sexo y lenguas artificiales», *BSEHL*, 12, pp. 75-93.

Guérard, Albert L. (1921), *A short history of the international language movement*, Nueva York, Unwin.

Martínez Gavilán, María Dolores (2016), «La contribución de Caramuel a la creación de lenguas artificiales: características universales, lenguas filosóficas y lenguas secretas», *Revista de Investigación Lingüística*, 19, pp. 77-106.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Un Gold Standard sobre factualidad para el español

Hortènsia Curell[128], **Ana Fernández Montraveta**[129]
[hortensia.curell@uab.cat | ana.fernandez@uab.cat]
Universidad Autónoma de Barcelona, España
Glória Vázquez[130], **Leyre Barrios**[131]
[gloria.vazquez@udl.cat]
Universitat de Lleida, España
Irene Castellón[132]
[icastellon@ub.edu]

[128] Profesora titular en el Departamento de Filología Inglesa en la Universidad Autónoma de Barcelona. Sus áreas de especialidad incluyen la lingüística contrastiva (especialmente inglés-catalán), pragmática intercultural y lingüística cognitiva, áreas en las que ha escrito artículos en una amplia variedad de publicaciones (además de dirigir 3 tesis doctorales). En los últimos tiempos ha empezado a trabajar en el campo de la lingüística de corpus, con un énfasis en las perífrasis verbales y la expresión de la factualidad. Es miembro del Grupo de Investigación Interuniversitario en Aplicaciones Lingüísticas (GRIAL), y actualmente participa en el proyecto TAGFACT.

[129] Doctorado en Lingüística y Comunicación y un MA en Lingüística Computacional. Es profesora titular en la Universitat Autònoma de Barcelona en el Departamento de Filología Inglesa y Germanística e investigadora principal del grupo de investigación interuniversitario en aplicaciones lingüísticas-GRIAL (http://grial.edu.es). Su investigación se centra en el análisis de las construcciones, especialmente las relacionadas con la factualidad, la modalidad y la aspectualidad. Ha participado en la creación de diversos recursos, como el WordNet 3.0 para el español y varios corpus anotados con información lingüística. Es coautora de los libros Clasificación verbal. Alternancias de diátesis y Las construcciones con 'se' en español, y cuenta con más de cincuenta publicaciones.

[130] Profesora titular en el área de Lingüística General de la Universidad de Lleida, miembro del Grupo de Investigación Interuniversitario en Aplicaciones Lingüísticas (GRIAL). Ha realizado aportaciones respecto a los patrones sintácticos verbales del español y su significado construccional. Destaca también su trabajo en la descripción de las perífrasis verbales de esta lengua. Actualmente está investigando sobre el grado de factualidad que se expresa en las oraciones. En total, tiene más de cincuenta publicaciones, entre las cuales destacan libros como Clasificación verbal. Alternancias de diátesis y Las construcciones con “se” en español y capítulos en manuales de lingüística de corpus y lexicografía. En estos ámbitos ha participado también confeccionando recursos a gran escala a partir de su colaboración en proyectos financiados. Actualmente dirige el proyecto TAGFACT.

[131] Actualmente estudiante de doctorado en la Universitat de Lleida, he cursado el máster en Lengua Española: Investigación y Prácticas Profesionales de la Universidad Autónoma de Madrid (España) y el grado en Lengua y Literatura Españolas de la Universidad de La Rioja (España). Entre sus trabajos destacan “La sustitución del pretérito imperfecto de subjuntivo por el condicional en Logroño”, publicado en Novas perspectivas na lingüística aplicada (pp. 83-93) y “La extensión del verbo ‘chupar’”, presentado en el V Congreso Internacional en Lengua y Literatura, Universidad de La Rioja, Logroño.

[132] Licenciada en Filología Románica (1986) y doctora en Filología Románica (1992). Actualmente, es profesora titular de Lingüística General en el Departamento de Filología Catalana y Lingüística General de la Universidad de Barcelona. Es especialista en Lingüística Computacional y Procesamiento del Lenguaje Natural, campo en el que ha dirigido un total de 8 tesis doctorales en esta área. Entre otras, ha publicado en revistas como DigitalScholarship In The Humanities, Glottometrics, Linguistics and Linguistic Theory, Procesamiento del Lenguaje Natural, Linguamática, Sintagma. Algunos de los proyectos desarrollados son KNOW2, Skater, EurowordNet, Multilingual Central Repository, Sensem. Actualmente està trabajando en el desarrollo del proyecto TAGFACT.

Universidad de Barcelona, España

En este trabajo presentamos los resultados de la anotación de la factualidad en un corpus periodístico en español, enmarcado en un proyecto subvencionado por el Ministerio en el que se está creando una herramienta automática de etiquetación. La anotación manual, a partir de una preetiquetación automática morfológica y sintáctica de una parte del corpus, constituye el Gold Standard (GS). Este será posteriormente utilizado para evaluar la herramienta mencionada. Para la anotación de dicho corpus se ha usado un editor creado *ad hoc* para el proyecto (Autores 1).

El corpus GS está formado por más de 10.000 palabras, extraídas de 22 noticias de 6 periódicos españoles de distinta ideología: *ABC*, *El Diario, El Periódico, La Razón, La Vanguardia* y *Público*. La media de palabras por noticia es de 553,7 (la noticia más corta tiene 181 palabras y la más larga 1.157). Aproximadamente el 70% del total de palabras anotadas pertenecen a textos sobre política y el resto a otros temas (sociedad, economía, deporte, tecnología, estilo de vida o sucesos internacionales).

El punto de partida para la anotación han sido 1.616 predicados verbales, provenientes del resultado obtenido en la etiquetación morfológica de Freeling (Padró, L. & Stanilovsky, E. 2012). Esta herramienta identifica como predicados verbales algunas palabras que no lo son, como participios con función adjetival. En nuestro corpus hemos contabilizado 124 de estos casos (7,67%) y no han sido anotados. Tampoco han sido anotados los predicados usados para describir las situaciones enfocadas al futuro presentadas dentro del eje de la irrealidad, es decir, la que incluye situaciones deseadas o hipotéticas, así como casos de prohibiciones, permisos y mandatos (Autores 2). El número de predicados identificados en este grupo asciende a 301 (18,63%). Así pues, en la práctica, de los 1.616 predicados iniciales se han anotado 1.191 (73,70%), que son a los que nos referiremos a partir de ahora.

Nuestra aproximación a la descripción de la factualidad es multidimensional: incluye el tiempo, el grado de certeza con el cual se presenta la situación, la polaridad y la categoría eventiva del predicado (Autores 3). Además, como información adicional relacionada con el grado de certeza, se identifica la fuente (denominada "voz" en nuestro proyecto), es decir el emisor de cada predicado. En textos periodísticos, la voz principal es el narrador, o sea, el periodista, cuya visión de los hechos puede variar según el medio o su posicionamiento personal. En otras ocasiones, son los protagonistas de la situación descrita los que directamente la presentan con mayor o menor certeza, ya que las noticias incluyen muchas oraciones de discurso directo e indirecto. Como en otros proyectos (FactBank (Saurí y Pustejovsky 2009), FactA (Minard *et al.*, 2016), DARPA DEFT (Prabhakaran *et al.* (2015), entre otros), partimos de la base de que no hay un solo valor de certeza posible asociado a una situación, sino que esta se puede presentar como segura por una voz y como posible por otra, e incluso como falsa, en el otro extremo.

En cuanto al eje temporal, se han diferenciado los predicados teniendo en cuenta si se refieren al presente, al pasado o al futuro. En la mayoría de los casos (67,67%) se relatan

situaciones pasadas, lógicamente, puesto que se trata de noticias. En segundo lugar, se sitúan las situaciones presentes (26,11%), que, en muchos casos, se relacionan con la expresión de estados. Finalmente, las situaciones futuras, una vez excluidos los casos de irrealidad mencionados, son minoritarias (6,21%), e incluyen predicciones, planificaciones, previsiones e intenciones.

Respecto al grado de certeza con que se presenta la situación, se ha distinguido entre certeza, que son la gran mayoría (95,97%) o posibilidad/probabilidad (4,03%). Los elementos de la frase que suelen expresar posibilidad/probabilidad son auxiliares perifrásticos, adverbios o expresiones modales.

Una tercera etiqueta utilizada es la polaridad. La negación de una situación (la mayoría de veces expresada con un elemento negativo, pero no siempre) permite describir la realidad en tanto que determina lo que no forma parte de ella. La tendencia clara que se observa, sin embargo, es narrar lo que sucede o puede suceder (el 93,03% de situaciones son afirmativas).

Claramente, la tendencia observada en los textos analizados es describir situaciones pasadas en afirmativo con certeza, ya que se corresponden con casi la mitad de los predicados anotados. Por el contrario, no hay ninguna situación futura en negativo descrita como posible/probable, solo 1 para el pasado y solo 3 para el presente.

Se ha incluido una cuarta etiqueta en la descripción de la factualidad, la categoría eventiva del predicado, con cinco valores: evento (73,81%; el más frecuente, lo cual es esperable dada la naturaleza del registro estudiado), mental (6,38%), verdades absolutas (0,17%) y estados (19,65%).

La anotación del GS se ha llevado a cabo por 6 anotadoras expertas que previamente habían participado en un experimento sobre el grado de acuerdo y cuyo resultado fue exitoso: el valor medio de la kappa para la información temporal fue de 0,64, para la polaridad 0,55 y para el tipo de evento 0,35.[133] Se partió de una etiquetación predefinida con los valores siguientes (que se observó que eran más frecuentes en el experimento): pasado, positivo, evento y certeza, lo cual facilitó mucho la tarea. La anotación manual se ha realizado de forma secuencial por las anotadoras, cada una de las cuales determinó los valores de una o dos categorías: la primera seleccionó los predicados y decidió si el predicado debía ser anotado con respecto a los valores que definen la factualidad, así como el tipo eventivo del predicado. La siguiente anotadora determinó la polaridad, la tercera, el tiempo y, por último, se anotó el grado de certeza. Además, *a posteriori*, se han realizado revisiones entre las diferentes anotadoras y se han ido comentado y resolviendo los casos problemáticos identificados durante el proceso.

En este artículo hemos presentado un corpus Gold Standard para el español anotado con valores de factualidad, que se pondrá a disposición de la comunidad científica desde la página web de nuestro grupo de investigación. De los datos extraídos y presentados en este estudio

[133] Para el valor de certeza no se puede calcular por no existir diversidad en la anotación en algunos pares.

podemos concluir que en el registro periodístico la mayoría de hechos se presentan como ciertos y se expresan mediante frases afirmativas (polaridad positiva) que narran eventos situados en el pasado. Lo más destacable de nuestro trabajo es, sin duda, la constitución de un corpus para el español anotado manualmente con información multidimensional sobre factualidad, ya que los recursos de este tipo para esta lengua son escasos. Según nuestro conocimiento, hasta el momento solo existe el de Wonsever *et al.* (2016).

Palabras clave: factualidad, certeza, irrealidad, anotación de corpus

Referencias

Minard, E., Speranza, E., & Caselli, T. (2016). The EVALITA 2016 Event Factuality Annotation Task (FactA). In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016)*

Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. En *International Conference on Language Resources and Evaluation-LREC2012* (pp. 2473-2479). Istanbul.

Prabhakaran, V., By, T., Hirschberg, J., Rambow, O., Shaikh, S., Strzalkowski, T., ... & Dalton, A. (2015): "A new dataset and evaluation for belief/factuality" A new dataset and evaluation for belief/factuality. En *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics* (pp. 82-91).

Saurí, R., & Pustejovsky, J. (2009): FactBank: a corpus annotated with event factuality. En *Language resources and evaluation*, 43(3), pp. 227-268.

Wonsever, D., Rosá, A., & Malcuori, M. (2016). Factuality Annotation and Learning in Spanish Texts. En *Proceedings of the Tenth International Conference on Language Resources and Evaluation-LREC 16* (pp. 2076-2080). Portorož, Slovenia.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Where is the subtitler's thumbprint hidden? A case study on a brazilian subtitles translator style

Janailton Silva[134]
[janailtonm@gmail.com]
Federal Institute Goiano, Brazil

Literature has shown that studies on the translator's style are still recent and have mostly sought to understand how the translator's thumbprint materializes in the translated text (Baker, 2000; Munday, 2008; Saldanha, 2011a; 2011b). Within the subtitling context, "the subtitler's style is the subtitler's way of expression in comparison to another subtitler's way of expression. This form of expression is characterized by a set of linguistic habits observed in more than one work by the same translator, who, subjected to the stylistic influences of subtitling, shows consistent patterns of stylistic choices motivated by certain purposes and capable of generating multiple effects." (Silva, 2018: 43). Following this premise, this paper aims to show how the style of a Brazilian professional subtitler is characterized in the subtitles of some episodes of the TV Series *Star Trek: Enterprise*. The study thus presents the way she employs pronoun placement and discourse markers (DMs) in these subtitles and discusses the implications of such linguistic habits in comparison to the work done by other subtitlers. By drawing on Corpus-Based Translation Studies (Baker, 1993; 1995; 1996; 2000; Berber Sardinha, 2004; McEnery; Wilson, 2001; Saldanha, 2011a), this research has made use of software *WordSmith Tools©, version 7* (Scott, 2018), to detect some pronouns and DMs, and observe their use in context. To investigate the Brazilian subtitler's style, two corpora have been compiled. The study corpus is made up of subtitles files available on Netflix, in Brazilian Portuguese, for episodes 10 and 11 (season 1) and 16 and 24 (season 2) of *Star Trek: Enterprise*, whereas the reference corpus is composed of other subtitles files of the same seasons made by various subtitlers. As far as the use of pronoun placement and discourse markers in the subtitles are concerned, results point to two directions. On the one hand, the Brazilian subtitler shows preference for enclisis, while the other subtitlers regularly opt for proclisis, including absolute proclisis. On the other hand, she employs DMs that most clearly conform to Netflix standards, such as *"não é?"* and *"certo",* instead of other DMs more commonly used by the other subtitlers, such as *"né?"* and *"tá(!)".* Considering these results, the research concludes that the subtitler's thumbprint is hidden in the subtle, almost untraceable language choices of pronouns and DMs

[134] Portuguese and an English professor at IF Goiano - Campus Campos Belos, where he teaches high school, technical, technological, and higher education courses, as well as supervises extension and research projects. He holds an MA in Translation Studies from University of Brasília (UnB) and a BA in Modern Language (English) from Federal University of Campina Grande (UFCG). He has experience in Modern Languages and Translation Studies, specifically in i) Audiovisual translation, ii) Corpus Linguistics, iii) Corpus-Based Translation Studies, iv) English as a Foreign Language (EFL).

in the subtitles. The subtitler's form of expression consists of different linguistic registers, especially the formal register of the Brazilian Portuguese language.

Keywords

Subtitler's style. Linguistic registers. Pronoun placement. Discourse markers.

Bibliography

Baker, M. (1993). Corpus linguistics and translation studies: implications and applications. In Baker, M. *et al.* (Ed.), *Text and technology: In honour of John Sinclair* (pp. 233-250). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Baker, M. (1995). Corpora in translation studies: an overview and some suggestions for future research. *Target*, 7(2), 223-243.

Baker, M. (1996). Corpus-based translation studies: the challenges that lie ahead. In Somers, H. (Ed.), *Terminology, LSP and translation: studies in language engineering in honour of Juan C. Sager* (pp. 177-186). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target*, 12(2), 241-266.

Berber Sardinha, T. (2004). *Lingüística de corpus*. [Corpus Linguistics] Barueri: Manole.

McEnery, T.; Wilson, A. (2001). *Corpus Linguistics:* an introduction. 2. ed. Edinburgh: Edinburgh University Press.

Munday, J. (2008). *Style and Ideology in Translation:* Latin American Writing in English. New York: Routledge.

Saldanha, G. (2011a). Translator style: methodological considerations. *The Translator*, 17(1), 25-50.

Saldanha, G. (2011b). Style of translation: the use of foreign words in translations by Margaret Jull Costa and Peter Bush. In Kruger, A. et. al. (Eds), *Corpus Based Translation Studies: Research and Applications* (pp. 237-258). London/New York: Continuum.

Scott, M. (2018). *WordSmith Tools manual*. Stroud: Lexical Analysis Software Ltd.

Silva, J. M. V. (2018). *Que espaço a legendista ocupa? Um estudo sobre estilo do tradutor.* [What space does the subtitler occupy? A study on translator style.] MA Thesis. Brazil: UnB.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

You: giving an identity to the audience of oral presentations

Ricardo Nausa[135]
[ra.nausa20@uniandes.edu.co]
Universidad de los Andes, Colombia
University of Birmimngham, UK

Studies on oral presentations (OPs) in academic contexts show how presenters use self-mentions (*I, we*) to engage their audience and project academic identities as reflected in their positions towards existing disciplinary knowledge (authorial stance) (e.g. Morton & Rosse, 2011; Zareva, 2013, Author, 2016). Other studies contrast presenter self-mentions (*I, we*) and audience mentions (*you*) focusing on the differences of use between native and non-native speakers of English (Rowley-Jolivet & Carter-Thomas, 2005b, 2005a; Vassileva, 2002), disciplines (Rowley-Jolivet & Carter-Thomas, 2005a), and OPs and related written texts (Webber, 2005). Until recently (Fernández-Polo, 2018), only one study has exclusively focused on *you*. These studies show how *you* is used to address the audience and perform specific discourse functions related to academic identity construction. In the comparisons, *you* is shown to be a versatile polysemic/polyfunctional pronoun but which can turn out to be problematic for novice scholars and NNSs. These studies have focused on conference presentations in contexts where both the audience and the speaker are members of the same disciplinary community (e.g. medicine, geology) and have mainly approached *you* as a corollary to *I* and *we*. There are no studies that focus on contexts where English is used as a foreign language and the audience is composed of people from other disciplines, which is the case of EAP programs whose purpose is to prepare researchers for writing for publication and speaking in international conferences.

This study analyses *you* in the oral presentations of an EAP class for PhD researchers from the different departments of a private university in Colombia. The study demonstrates how PhD researchers use *you* to assign real or imagined academic identities to their non-expert audience that are "mirror images" of the identities that they would project for themselves with self-mentions. These identity projections with *you* are demonstrated to be correlated to the PhD researchers' disciplines (hard vs soft), their levels of oral performance (high, medium, and low), and the composition of the audience (researchers from different disciplines) and not to individual differences.

Corpus techniques, significance tests (Log-likelihood) and effect size scores (BIC factor) are used in a 72811-token corpus of 88 oral presentations to identify and select frequent uses of *you* expressing different types of audience-identities (or discourse functions). These cases are analysed from the territory of information (Kamio, 2001), dynamic-inferential view of

[135] B. A. in Philology and Language from Universidad Nacional de Colombia; M. A. in Applied Linguistics to TEFL from Universidad Distrital Francisco José de Caldas. Ph. D. in Applied Linguistics and English Language from the University of Birmingham. Professor of Corpus Linguistics and EAP at Universidad de Los Andes.

communication (Gast, Deringer, Haas, & Rudolf, 2015) and stance projection (Tang & John, 1999) theoretical models to identify how audience mentions and the identities assigned to them differ depending on the presenters' disciplines (hard vs soft) and levels of oral achievement (high, medium, and low). Performances are not compared to NS standards.

Findings show that hard disciplines researchers construe their audience as research-apprentices or users of their findings or by-products while soft-discipline researchers construe them as equals in the construction of knowledge. In the comparisons by level of oral performance, high achievers more consistently use *you* with a sense of audience, which is reflected in *you*-projections to guide the audience through the OP or to construe them as users of their findings and creations. These audience identity role projections, their associated linguistic patterns, and discourse functions are discussed as well as the limitations, implications, and perspectives for future research.

References

Fernández-Polo, F. J. (2018). Functions of "you" in conference presentations. English for Specific Purposes, 49, 14–25. https://doi.org/10.1016/j.esp.2017.10.001

Gast, V., Deringer, L., Haas, F., & Rudolf, O. (2015). Impersonal uses of the second person singular: A pragmatic analysis of generalization and empathy effects. Journal of Pragmatics, 88, 148–162. https://doi.org/10.1016/j.pragma.2014.12.009

Kamio, A. (2001). English generic "we", "you", and "they": An analysis in terms of territory of information. Journal of Pragmatics, 33(7), 1111–1124. https://doi.org/10.1016/S0378-2166(00)00052-7

Morton, J., & Rosse, M. (2011). Persuasive presentations in engineering spoken discourse. Australasian Journal of Engineering Education, 17(2), 55–66.

Rowley-Jolivet, E., & Carter-Thomas, S. (2005a). Scientific conference Englishes. In G. Cortese & A. Duszak (Eds.), Identity, community, discourse: English in intercultural settings (pp. 295– 320). Bern: Peter Lang.

Rowley-Jolivet, E., & Carter-Thomas, S. (2005b). The rhetoric of conference presentation introductions: context, argument and interaction. International Journal of Applied Linguistics, 15(1), 45–70. https://doi.org/10.1111/j.1473-4192.2005.00080.x

Tang, R., & John, S. (1999). The 'I' in identity: Exploring writer identity in student academic writing through the first person pronoun. English for Specific Purposes, 18, S23–S39. https://doi.org/10.1016/s0889-4906(99)00009-5

Vassileva, I. (2002). Speaker-audience interaction: the case of Bulgarians presenting in English. In E. Ventola *et al.*, The Language of Conferencing. Frankfurt am Main: Peter Lang (pp. 255– 276).

Webber, P. (2005). Interactive features in medical conference monologue. English for Specific Purposes, 24(2), 157–181. https://doi.org/10.1016/j.esp.2004.02.003 Zareva, A. (2013). Self-mention and the projection of multiple identity roles in TESOL graduate student

presentations: The influence of the written academic genres. English for Specific Purposes, 32(2), 72–83. https://doi.org/10.1016/j.esp.2012.11.001

Humanidades Digitales

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

3DePict: memorizando vocabulario de inglés a través del uso de geoinformación y realidad aumentada

Alexia Larchen Costuchen[136]
[larchen0501@gmail.com]
Universidad Politécnica de Valencia (UPV), España

Resumen:

La propuesta 3DePict afronta el problema de la escasa eficacia de la memorización del léxico inglés en los alumnos universitarios, cuando están presionados por el tiempo y tienen unas fechas límite. El 3DePict repercute en la memoria a través de la vista, oído y las acciones-movimiento (impacto VAC: visual, audio y cinestésico). Las personas que supuestamente se beneficiarían de la propuesta son alumnos de diferentes departamentos de la Universidad Politécnica de Valencia cuyo objetivo será la memorización del vocabulario especifico en inglés a través de una ruta segura con objetos conocidos y la aplicación de nuevas tecnologías, así como métodos mnemotécnicos. Uno de los objetivos más importantes de la propuesta es dar una alternativa a los enfoques pasivos durante el proceso de estudio antes de los exámenes y fomentar, por tanto, los enfoques activos: paseos al aire libre, utilizando rutas ya establecidas, o movimientos de los alumnos por sus propias casas, por ejemplo, pasando de "estar sentado enfrente del ordenador o junto a un libro" a "estar en movimiento-ruta con el uso de dispositivos móviles." Desde el punto de vista didáctico se sustituye el aprendizaje pasivo por el aprendizaje activo, añadiéndole el aprendizaje significativo desarrollador que adquiere y constituye un nuevo aprendizaje (Vigotsky, 1984, 1995, 1996; Luria, 1983; Ausubel *et al.*, 1990; Good y Brophy, 1995, Diaz Barriga y Hernandez, 2002; Agra *et al.*, 2019).

La Aplicación 3DePict plantea ofrecer una solución lúdica y experimental al aprendizaje del vocabulario específico en inglés, según las exigencias de las programaciones didácticas en los diferentes departamentos de la Universidad Politécnica de Valencia, aplicando métodos mnemotécnicos de asociación de palabras o frases con objetos ya conocidos a través de una determinada ruta. Según varios autores (Yates, 1966, 1999; Carney y Levin, 1998; Moè y De Beni, 2005; Huttner y Robra-Bissantz, 2017), el uso del Palacio de Memoria implica seleccionar una serie de distintos objetos a lo largo de un entorno familiar, crear, posteriormente, una imagen para cada palabra clave correspondiente al concepto y, finalmente, asociar la nueva

[136] Alexia Larchen Costuchen es licenciada en filología inglesa y alemana; tiene dos másteres en educación y en estos momentos está cursando un doctorado en el departamento de Lingüística Aplicada de la Universidad Politécnica de Valencia (UPV). Su experiencia profesional incluye más de doce años de enseñanza de inglés como lengua extranjera en colegios, universidades y empresas multinacionales. Sus intereses de investigación abarcan la enseñanza de segundas lenguas, los métodos de enseñanza, la lingüística cognitiva y la tecnología educativa.

información con los objetos físicos seleccionados. Al volver mentalmente sobre la ruta, se recuperan todos los componentes de tal manera que se facilita la memorización entre ellos, así que, en un recorrido de este tipo, cualquier objeto como una silla, armario o mesa (o escultura, árbol, fuente) pueden servir como parte del Palacio de Memoria.

La App propuesta, desarrollada a través de Wikitude, ofrece la utilización de un espacio conocido y seguro (ruta de paseo por el Campus de la UPV, ruta por el antiguo cauce del rio paseando un perro, etc.) y el uso de dispositivos móviles que a través de esta App puedan transformar palabras o frases introducidas por los usuarios en palabras 3D que se activen como Realidad Aumentada al lado de objetos físicos determinados (propuesta básica con la RA). Habiéndose mirado distintos servidores en la nube, finalmente nos decantamos por el servicio de Amazon (Amazon Web Services, Inc., AWS), al ser el más versátil en nuestro caso de implementación de la Inteligencia Artificial (IA). También se prevé considerar otros servidores Cloud Computing como es el de Microsoft y el de IBM. Son las opciones que se verán para obtener datos cualitativos y cuantitativos relacionados con el funcionamiento de la demo de 3DePict, teniendo en cuenta las mejores opciones económicas.

Palabras clave: memorización de vocabulario, palacio de memoria, geoinformación, realidad aumentada, inteligencia artificial

Referencias bibliográficas:

Agra, G., Formiga, N. S., Oliveira, P. S. de, Costa, M. M. L., Fernandes, M. das G. M., & Nóbrega, M. M. L. da. (2019). Analysis of the concept of Meaningful Learning in light of the Ausubel's Theory. *Revista Brasileira de Enfermagem*. NLM (Medline).

Ausubel, D., Novack., J & Hanesian. (1978). Educational psychology: a cognitive view. New York: Holt, Rinehart and Winston.

Carney, R.N., Levin, J.R. Pictorial Illustrations Still Improve Students' Learning from Text. Educational Psychology Review 14, 5–26 (2002).

Díaz Barriga, F. y Hernández, G. (2002). Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista (2a. ed.). México: McGraw Hill.

Good., T & Brophy., J. (1990). Educational Psychology. A realistic Approach. New York: Longman.

Huttner, J.-P., & Robra-Bissantz, S. (2017). An Immersive Memory Palace: Supporting the Method of Loci with Virtual Reality. 23rd AMCIS, 1-10

Luria, A. R. (1983). La mente de un mnemónico: Un Pequeño Libro Sobre Una Gran Memoria. Madrid: Taller de Ediciones.

Moe, A., de Beni, R., 2005. Stressing the efficiency of the loci method: oral presentation and the subject-generation of the loci pathway with expository passages. Appl. Cogn. Psychol. 19, 95-106.

Vigotsky, L. (1984). *Infancia y Aprendizaje*, Akal, Madrid.

Vigotsky, L. (1995). *Pensamiento y lenguaje*, Ediciones Fausto, Buenos Aires.

Vigotsky, L. (1996). *Artículos Clásicos*, Nauka, Moscú.

Yates, F. A. (1966). *The art of memory*. London: Routledge & Kegan Paul.

Yates, F. A. (1999). *The Art of Memory*. New York: Routledge.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Acerca del proyecto DHistOntology: una muestra de la modelación del dominio de la salud y la enfermedad

Itziar Molina Sangüesa[137]

[itziarmolina@usal.es]

Universidad de Salamanca, España

Esta comunicación pretende mostrar las fases que se han llevado a cabo para la creación de una ontología aplicada a la lexicografía diacrónica, denominada *DHistOntology*. Esta herramienta se ha diseñado con el objetivo de vincular, en la web semántica, los datos consignados en dos repertorios lexicográficos —que se publican y actualizan progresivamente— en soporte digital: por un lado, el *Nuevo Diccionario Histórico del Español* (*NDHE*) de la Real Academia Española y, por otro, el *Tesoro della Lingua Italiana delle Origini* (*TLIO*), desarrollado por el instituto de investigación del CNR "Opera del Vocabolario Italiano" (OVI) - Accademia della Crusca. En particular, analizaremos y expondremos la modelación de los datos correspondientes a las primeras fases de desarrollo del proyecto *DHistOntology*, basadas en el dominio de la salud y la enfermedad.

La metodología que hemos empleado para la elaboración de esta ontología parte de la guía de Noy &McGuinness (2001). Asimismo, nos hemos servido de la última versión del editor *Protégé* para redactarla, tal y como ejemplificaremos en esta ponencia. En suma, esta iniciativa y propuesta de comunicación nace del interés por compartir y gestionar el conocimiento que ofrecen un par de diccionarios paradigmáticos —en su concepción y tipología— del español y del italiano.

Palabras clave: lexicografía histórica; diccionarios digitales; ontologías; terminología médica.

Referencias bibliográficas:

NDHE - Real Academia Española / Instituto de Investigación Rafael Lapesa (2015-): *Nuevo Diccionario Histórico del Español* [en línea]: http://ndhe.frl.es/ [consulta: 11/06/2020].

[137] Licenciada y Doctora en Filología Hispánica por la Universidad de Salamanca, ha completado su formación en el Instituto de Lexicografía "Rafael Lapesa" de la Real Academia Española, en el marco del proyecto del Nuevo Diccionario Histórico del Español (NDHE), como redactora de voces del léxico médico. Asimismo, ha participado en los sucesivos proyectos I+D+i del Diccionario de la Ciencia y de la Técnica del Renacimiento (DICTER), como coordinadora técnica y como redactora del área de matemáticas (aritmética y álgebra). Entre sus principales líneas de investigación destacan: la lexicografía y lexicología históricas, el estudio del léxico científico y la confección de diccionarios en soporte digital.

Noy, N. F. y D. L. McGuinness (2001): Ontology Development 101*: A Guide to creating your first ontology*. Stanford (CA): Stanford Knowledge System Laboratory Technical Report KSL-01-05 y Stanford Medical Informatics Technical Report SMI-2001-0880 [en línea]:<https://protege.stanford.edu/publications/ontology_development/ontology101.pdf> [consulta: 01/03/2020].

TLIO – Squillacioti, P. (dir.) (2008-): *Tesoro della Lingua Italiana delle Origini* [en línea]: http://tlio.ovi.cnr.it/TLIO/ [consulta: 11/06/2020].

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Análisis de las redes sociales de personajes en tres obras teatrales de Galdós

Teresa Santa María[138] **& María Ángel Somalo**[139]
[teresa.santamaria@unir.net | mangeles.somalo@unir.net]
Universidad Internacional de La Rioja, España

Introducción

Las redes sociales entre personajes que los grafos y otras herramientas de las Humanidades Digitales (HD) nos permiten visualizar aportan perspectivas nuevas que avalan en algunos casos las interpretaciones críticas tradicionales, pero también incluyen enfoques diferentes que enriquecen el estudio de dichos textos (Fernández *et al.*, 2015, Fisher *et al.*, 2017). Dentro de la Literatura Española ya tenemos algunas investigaciones que han recurrido al estudio de esas redes sociales de los personajes, destacando en este caso que nos compete, los que se refieren al ámbito del teatro de la llamada Edad de Plata (Martínez Carro, 2019) o a dos de los *Episodios Nacionales* (Isasi, 2017). En este año en que se cumple el centenario de la muerte del escritor español Benito Pérez Galdós resulta interesante estudiar las redes sociales que establecen los personajes de tres piezas teatrales: *Doña Perfecta* que fue estrenada en el Teatro de la Comedia el 28 de enero de 1896, *Electra,* que se estrenó en 1905 en el Teatro Español y *Casandra* que se representó por primera vez en ese mismo teatro en 1910. Las tres piezas forman parte del corpus teatral digitalizado en TEI de la *Biblioteca Electrónica Textual del Teatro en Español de 1868-1939 (BETTE)* (Jiménez *et al.* 2017 y Martínez Carro y Santa María, 2019), que se ha integrado en el *Drama Corpora Project* bajo el nombre de *Spanish Drama Corpora* (*DraCor*).

Cuestiones iniciales y metodología

Las preguntas que nos planteamos son diferentes y nos sirven para comprobar la capacidad de las HD para dar respuesta a cuestiones que la crítica literaria no ha podido abarcar, así como aportar nuevos matices y perspectivas. Nuestro objetivo, por tanto, ha sido comprobar si los

[138] Teresa Santa María es Directora académica y profesora del Máster Universitario en Didáctica de la Lengua y la Literatura en Educación Secundaria y Bachillerato y Coordinadora del área de Lengua en los Posgrados de la Facultad de Educación de la Universidad Internacional de La Rioja. Investigadora Principal del proyecto HDATEATROUNIR que está digitalizando en TEI la Biblioteca Electrónica Textual del Teatro Español, 1868-1939 (BETTE), y miembro del Grupo de Estudios del Exilio Literario (GEXEL). Otras líneas de investigación son el teatro español contemporáneo y la pervivencia de los mitos grecolatinos en la dramaturgia del exilio español de 1939.

[139] María Ángel Somalo es Doctora en Filología Hispánica por la Universidad de La Rioja. Ejerce como docente en la Universidad Internacional de La Rioja, donde también investiga dentro del grupo HDATEATROUNIR. Su especialidad en literatura, teatro y cultura la ha llevado a centrar sus intereses en las carteleras teatrales, tanto comerciales como de aficionados que constituyen la actividad cultural (teatros, cines, publicaciones, música, etc.) en la ciudad de Calahorra (La Rioja) entre 1840 y 1910, o Logroño, entre 1901 y 1950.

grafos y datos cuantitativos de las tres piezas que encontramos en *DraCor* y *Shiny Dracor* que pueden generarse mediante Gephi a partir de los mismos textos en TEI que proporciona BETTE responden a preguntas como las siguientes:

1. ¿Tienen mayor peso e importancia los personajes masculinos que los femeninos en las obras teatrales de Galdós?
2. ¿Son siempre los personajes femeninos que dan título a las obras los que más importancia tienen?
3. ¿Se diferencia la red social en *Electra* –obra pensada desde su origen para la escena – de las de las otras dos obras, que parten de una novela anterior?
4. ¿Podríamos encontrar una red social de personajes característica de Galdós y que se diferencie de la de otros dramaturgos, como Valle-Inclán o Lorca?

Avance de las respuestas

Gracias a los grafos que podemos generar con Gephi, sobre los tres textos que se encuentran en BETTE, podemos responder a la primera pregunta respecto a la preponderancia de personajes de un sexo u otro en Galdós que no hay diferencias elevadas entre ellos, como podemos apreciar en la siguiente figura, con la separación por sexos de los personajes de *Electra* (Fig. 1), pero que también recibe el mismo comportamiento en las dos obras de Galdós incluidas en dicho corpus y que se basan en novelas previas:

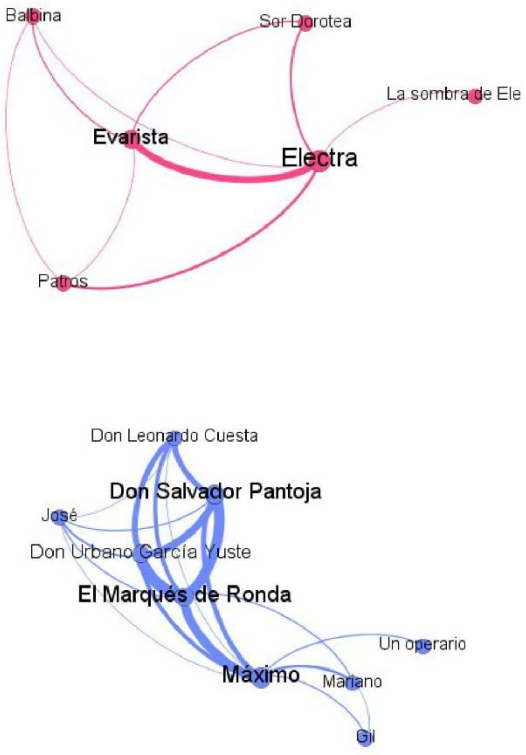
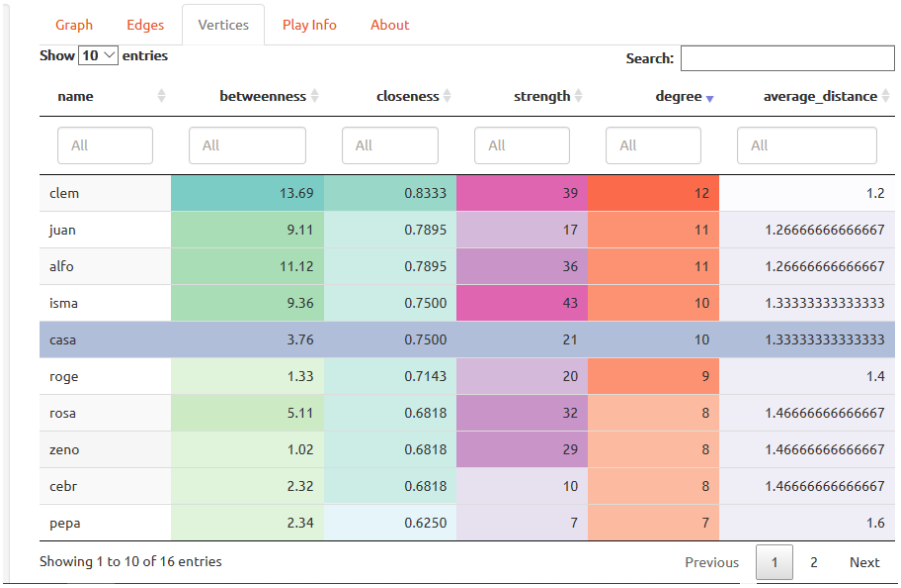


Fig. 1. Grafos creados por Gephi, según el sexo de los personajes, de *Electra* de Galdós

Podemos comprobar también a partir de la información sobre los vértices o conexiones entre los diferentes personajes (*Vertices*) que genera *Shiny DraCor* que el personaje que da título a la obra no tiene, como en el caso de *Casandra*, que ser el que más veces intervenga o tenga mayor peso dentro de las escenas de la obra. En la Figura 2 podemos ver a los personajes ordenados por grado (*Degree*) y observar cómo el personaje de Casandra (casa) ocupa el quinto lugar, no superando esa posición si analizamos los vértices en ninguna de las otras cuatro categorías (*betweeness* o centralidad de intermediación, *closeness* o centralidad de cercanía, *strength* o fuerza y *average distance* o distancia media):



name	betweenness	closeness	strength	degree	average_distance
All	All	All	All	All	All
clem	13.69	0.8333	39	12	1.2
juan	9.11	0.7895	17	11	1.26666666666667
alfo	11.12	0.7895	36	11	1.26666666666667
isma	9.36	0.7500	43	10	1.33333333333333
casa	3.76	0.7500	21	10	1.33333333333333
roge	1.33	0.7143	20	9	1.4
rosa	5.11	0.6818	32	8	1.46666666666667
zeno	1.02	0.6818	29	8	1.46666666666667
cebr	2.32	0.6818	10	8	1.46666666666667
pepa	2.34	0.6250	7	7	1.6

Fig 2. Datos de los vértices de los personajes de *Casandra*, visualizados en *Shiny DraCor*

Por último, en respuesta a las dos últimas preguntas, podemos observar cómo los grafos de las tres y la distribución de las escenas constituyen unas redes sociales que se comportan de manera semejante entre sí y que difieren de las de otros autores cuyas obras aparecen en estos corpus –como Valle-Inclán o García Lorca –; así como también por la cantidad de personajes que aparecen sobre la escena y que resulta bastante homogéneo en Galdós, ya que suele esta entre quince y dieciséis personajes, mientras en las piezas valleinclanescas oscilan entre los treinta y siete de *Divinas palabras* y los setenta y uno de *Águila de blasón*.

Inclán, Ramón María del Valle Wikidata: Q311001	Águila de blasón Comedias bárbaras - 2 Wikidata: Q6172602	71	1907 1907 1907
Inclán, Ramón María del Valle Wikidata: Q311001	Cara de Plata Comedias bárbaras, 1 Wikidata: Q5748755	45	1922 1967 1922
Inclán, Ramón María del Valle Wikidata: Q311001	Divinas Palabras Tragicomedia de aldea Wikidata: Q8561242	37	1920 1933 1920
Inclán, Ramón María del Valle Wikidata: Q311001	Luces de bohemia Esperpento Wikidata: Q3312512	55	1924 1924
Inclán, Ramón María del Valle Wikidata: Q311001	Romance de lobos Comedia bárbara Wikidata: Q6111385	47	1908 1970 1908

Fig 3. Página inicial del *Spanish Drama Corpora* con la inclusión en la tercera columna del número de personajes de cada una de las obras de Valle-Inclán

Existen relevantes estudios que ahondan sobre la importancia de los personajes en el teatro de Galdós desde una perspectiva de su relevancia en cuanto al tipo de parlamentos que pronuncian y a la gradación del dramatismo de las escenas (Menéndez Onrubia, 1983, pp. 264-271); o bien relacionan los caracteres con la importancia simbólica e histórica (Ávila, 1991) y destacan la relevancia de los personajes femeninos (Jiménez Gómez, 2018). Sin embargo, las herramientas y aplicaciones de las HD permiten valorar y reflejar de manera clara y objetiva datos cuantitativos que permiten cuestionar algunas de las afirmaciones que se han hecho tradicionalmente.

Bibliografía

Ávila, J. (1991). El personaje femenino del teatro de Galdós. (Una aproximación al simbolismo histórico del escritor). Tesis. Madrid: Universidad Complutense de Madrid.

Calvo, J., Jiménez, C y Santa María (2017). Biblioteca Electrónica Textual del Teatro en Español (BETTE). Logroño: UNIR. https://github.com/GHEDI/BETTE

Calvo, J., Jiménez, C., Martínez, E. y Santa María (2018). ¿Existe correlación entre importancia y centralidad? Evaluación de personajes con redes sociales en obras teatrales de la Edad de Plata en Girón, J. y Galina, I. (eds), Revista de Humanidadades Digitales 2018 Puentes. Libro de resúmenes, 494-498. México: Red de Humanidades Digitales. Recuperado de: https://de2018.adho.org/wp-content/uploads/2018/06/dh2018 abstracts.pdf

Fernández, M., Peterson, M. y Ulmer B. (2015). Extracting social network from literatura to predict antagonist and protagonist. Recuperado de: https://nlp.stanford.edu/courses/cs224n/2015/reports/14.pdf

Fisher, F., Göbel, M., Kamplaskar, D., Kittel, C. y Trilcke, P. (2017). Network dynamics, plot analysis: approaching the progressive structuration of literary texts. Digital Humanities 2017. (Montréal, August 8-11, 2017). Montréal: McGill University/Université de Montréal, 437-441. Recuperado de https://dh2017.adho.org/abstracts/071/071.pdf

Iglesias, J. C. (2006). Anagnórisis en la Electra de B. P. Galdós. Bulletin Hispanique, 108 (2), 459-474.

Isasi, J. (2017). Acercamiento al análisis del sistema de los personajes en la narrativa escrita en español: el caso de Zumalacárregui y Mendizábal de Pérez Galdós. Caracteres. Estudios culturales y críticos de la esfera digital, 6 (2), 107-137.

Jiménez Gómez, C. (2018). La construcción de los personajes femeninos galdosianos desde una instancia receptora de mujer (Tesis doctoral). Córdoba: Universidad de Córdoba.

Martínez, E. (2018). Una interpretación digital de dos tragedias lorquianas: Yerma y Doña Rosita la soltera. Caracteres. Estudios culturales y críticos de la esfera digital, 7 (2), 240-267.

Martínez, E. (2019). Aproximación al teatro lorquiano desde la teoría de las redes sociales: La casa de Bernarda Alba. Art Nodes. Revista de Arte, Ciencia y Tecnología, 24, 134-141.

Martínez, E. y Santa María, T. (2019). Biblioteca electrónica textual del teatro español (1868-1936) e investigación con grafos, Revista de Humanidades Digitales, 3, 23-45. Recuperado de: http://revistas.uned.es/index.php/RHD/article/view/23144

Menéndez Onrubia, C. (1983). Introducción al teatro de Benito Pérez Galdós. Madrid: CSIC.

Moretti, F. (2013). Operationalizing: or, the function of measurement in modern literary theory. The New Left Review, 84, 103-119. Recuperado de https://stanford.io/2tfU2jt

Santa María, T. (2018). Una interpretación del teatro de Galdós, a partir de obras en XML-TEI. En Álvarez, E. y Blasco, J. (eds.), Humanidades digitales. Recursos y nuevas propuestas, Valladolid: Agilice digital. Recuperado de: http://lithumadigital.blogs.uva.es/files/2018/11/Humanidades-Digitales-p%C3%A1ginas.pdf

Spanish Drama Corpora. https://dracor.org/span

Shiny Dracor. https://shiny.dracor.org/

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Análisis del epistolario del coronel Anselmo Pineda con Python: Una mirada al proyecto coleccionista y al territorio desde las redes sociales y el aprendizaje automático

Santiago Alejandro Ortiz Hernández[140]
[santfloyd@gmail.com]
Red Humanidades Digitales, Colombia

El campo de las humanidades digitales (HD) abre la valiosa posibilidad de traer metodologías novedosas para responder a las preguntas más tradicionales de las ciencias sociales y las humanidades. Esta propuesta versa sobre el uso de potentes herramientas computacionales, escritas por el ponente en el lenguaje de programación Python, para procesar los datos recabados a partir del epistolario del coronel antioqueño Anselmo Pineda conservado en la Biblioteca Nacional y en otros archivos dispersos en Colombia. El propósito principal es caracterizar el alcance social y territorial del importante proyecto coleccionista de la elite intelectual del siglo XIX neogranadino, en el que el coronel Pineda se perfiló como eje fundamental para la formación del archivo documental más robusto y profuso del siglo: el fondo Pineda de la Biblioteca Nacional de Colombia.

Antes de adentrarse en la metodología que incorpora aprendizaje automático, procesamiento de lenguaje natural, sistemas de información geográfica y ciencia de datos, es oportuno describir brevemente la figura histórica del militar, político, colonizador interno y coleccionista Anselmo Pineda, así como referirse al coleccionismo en tanto que proyecto ilustrado de la élite granadina y que fue central en la formación del Estado-nación tras el periodo de independencia. Anselmo Pineda nació en El Santuario en 1805, en una región de producción exclusivamente agrícola y aurífera. En 1822 es enviado por su padre a Bogotá a estudiar jurisprudencia en el Colegio Seminario de San Bartolomé, pero abandona en 1825, sin embargo allí se destacó en las cátedras de José Félix de Restrepo y este le ofrece el cargo de oficial archivero de la tesorería general de la provincia de Antioquía. Una vez en contacto con las reservas documentales del Estado, Pineda inicia su colección, pero se ve envuelto en la conmoción de la conspiración en contra de Simón Bolívar en septiembre de 1826 y huye junto a su amigo íntimo Mariano Ospina Rodríguez. En 1829, Pineda se une a las huestes del general José María Córdova en contra de Simón Bolívar, en cuya representación fue enviado el general Daniel O´Leary para hacer frente a Córdova, quien perece en la confrontación de El Santuario,

[140] Historiador, especialista en Ciencia de Datos y máster en Sistemas de Información Geográfica, orientado al uso de las TIC para la exploración de nuevos métodos de investigación en humanidades basados en el uso de herramientas digitales para la adquisición, procesamiento y visualización de datos históricos y espaciales.

por su parte, Pineda sobrevive con varios impactos de bala, más tarde, por orden de O´Leary, recibe el indulto.

Se enrola en el ejército en 1832 e inicia una carrera de ascenso militar culminada con su nombramiento como coronel por su desempeño en la Guerra de los Supremos en 1839-1841, conflicto en el que terminó por enfrentarse a Salvador Córdova, su antiguo aliado, amigo personal y hermano menor de J. M. Córdova. En 1843 fue nombrado gobernador de la provincia de Panamá, allí impulso las escuelas dominicales, la fabricación del sombrero panameño y negoció por primera vez la construcción del canal de Panamá sin mucho éxito. En 1845, Pineda es nombrado por Tomás Cipriano de Mosquera como prefecto del recién formado departamento de Caquetá, pero en 1846 fue nombrado como gobernador de Túquerres donde se enfocó en la *civilización de los salvajes;* construcción de caminos; impulso de la educación popular con ayuda de Simón Rodríguez, maestro de Simón Bolívar, e incursionó en el reconocimiento de la geografía inexplorada y de posibles productos de exportación.

En 1848, Pineda es elegido cono representante de Antioquia en la cámara de representantes y en ese mismo año comienza la elaboración del primer índice de su colección documental y que presenta al congreso, junto con su oferta de venta, en 1849. Finalmente, en 1851 solicita su baja militar y consigue venderle al Estado colombiano su colección, pero se ve envuelto una vez más en las complicaciones políticas típicas del siglo XIX, en esta ocasión terminó preso por su supuesta participación en la revuelta conservadora en Guasca, Cundinamarca promovida por Mariano y Pastor Ospina Rodríguez. Ya en 1852 fue liberado y colocado como custodio de las colecciones de la Biblioteca Nacional de Colombia. En 1854 lideró la guerra en contra del mandato de José María Melo, al lado de Tomás Cipriano de Mosquera y Pedro Alcántara Herrán. Los siguientes años fueron dedicados a su actividad como masón e ideólogo del partido conservador durante las décadas de 1860-1870.

Al respecto del coleccionismo del coronel Pineda, vale la pena señalara que este concibió su colección documental como un monumento que contiene la historia de la nación. Es bajo este entendido que cobran sentido los diferentes lugares de sujeto que Pineda ocupó en tanto que militar, político y coleccionista-intelectual, pues desde cada uno de esos lugares buscó oportunidades para concretar su proyecto ilustrado, que fue percibido por su red de amigos intelectuales y colaboradores, como imprescindible para la constitución del Estado-nación y la realización de los fines últimos de la comunidad de letras neogranadina que, en últimas, buscaba traer el progreso, materializar la idea ilustrada de felicidad para los ciudadanos y llevar la civilización a los espacios precariamente dominados institucionalmente, como Panamá, Túquerres o Caquetá. En estos lugares de frontera Pineda hizo presencia y se sirvió de ello para su colección o para la colección de algunos de sus conocidos. De manera que, el coleccionismo fue indudablemente un pilar en los complejos procesos políticos, sociales y culturales del siglo XIX.

Una vez acotada la figura histórica de Pineda y señalada la importancia del coleccionismo en la formación de Estado-nación, es tiempo de referirse a los datos y a la metodología de la investigación del problema histórico desde las humanidades digitales. Los datos recogidos de las

fuentes primarias son fundamentalmente las fechas de producción del documento epistolar, los corresponsales, la descripción del documento elaborada por el investigador y las transcripciones de una buena cantidad de cartas elaboradas por el ponente y por terceros durante años de investigación. Otras características son producidas in situ para enriquecer el marco de datos con empleo de técnicas de transformación propias de la ciencia de datos para hacer posible, entre otras, la *espacialización* de los mismos, generar mapas de la distribución espacial de la red epistolar del coronel y determinar entidades, así como similitudes dentro del corpus documental con procesamiento de lenguaje natural (NPL).

Adicionalmente al enriquecimiento de los datos, se realizan otros procesos necesarios para la inspección, interpretación y caracterización de dichos datos. En primer lugar, se visualizan las redes epistolares para analizar los patrones sociales y realizar la extracción de indicadores de estas estructuras sociales para determinar el grado de centralidad e intermediación de los distintos corresponsales. Ello con el fin de cuantificar la importancia de cada individuo en el flujo de información y documentos dentro de la red. En segundo lugar, se lleva a cabo un procesamiento de lenguaje natural (NLP) de las descripciones de los documentos y las transcripciones que consiste en darle la capacidad a la máquina de leer, interpretar y derivar sentido al lenguaje humano, a través de la implementación de modelos necesarios para dicho procesamiento conocidos como *Bag of Words, TF-IDF, Tokenization, Lemmatization* y limpieza de caracteres especiales y de *stop words,* esto con la finalidad de identificar, por ejemplo, temas relevantes del corpus, convergencias temáticas entre documentos o similitudes y relevancia de términos presentes en los documentos. En tercer lugar, se toman como insumos los pasos anteriores para producir un modelo de aprendizaje automático que permite identificar, por una parte, las agrupaciones espaciales de corresponsales en el territorio y también clasificar a los colaboradores de Anselmo Pineda en el proyecto coleccionista a partir de las transcripciones, descripciones y localizaciones de remitentes en el espacio.

De esta manera el enfoque HD de esta propuesta contribuye a desentrañar los patrones subyacentes a los mecanismos que hicieron posible la colección Pineda, su relación con en el proyecto colonialista del siglo XIX y a descubrir aquellos actores y factores que hicieron posible su formación.

Palabras clave

Machine Learning, Humanidades Digitales, Procesamiento Lenguaje Natural, Ciencia de Datos, Sistemas de Información Geográfica, Coleccionismo, Historia

Bibliografía

Baudrillard, Jean. *El Sistema de los Objetos*. México: Editorial Gallimard, 1969.

Benjamin, Walter."*El coleccionista*" en Libro de los Pasajes. Madrid: Editorial Akal, 2005.

Bourdieu, Pierre. "*Las formas del capital. Capital Económico, capital cultural y capital social*". En: Poder, derecho y clases sociales. Barcelona: Editorial, Desclée, 2000.

Brown, Matthew. "*The end of Bolivarian Networks*". En *The Struggle for Power in Postindependence Colombia and Venezuela*. New York: Editorial Palgrave Macmillan, 2012.

Castillo Gómez, Antonio. "*Del tratado a la práctica. La escritura epistolar en los siglos XVI y XVII*" En: La correspondencia en la historia. Modelos y prácticas de la escritura epistolar. Madrid: Editorial Calambur, 2002.

Derrida, Jaques. *Mal de Archivo*. Una impresión freudiana. Madrid: Editorial Trotta. 1996.

González Stephan, Beatriz. "*Coleccionar y exhibir: la construcción de patrimonios culturales*" en Revista de Literatura Vol. 29, No. 86 (2000).

Imízcoz Beunza, José María y Arroyo Ruíz, Lara. "*Redes Sociales y Correspondencia Epistolar. Del Análisis Cualitativo de las Relaciones Personales a la Reconstrucción de Redes Egocentradas*". Revista: Redes. Revista Hispana para el Análisis de Redes Sociales. Vol. 21. Número 4. (2011).

König, Hans-Joachim. *El Camino Hacia la Nación: nacionalismo en el proceso de formación del Estado y de la Nación de la Nueva Granada, 1750 a 1856*. Bogotá: Editorial Banco de la República, 1994.

Loaiza Cano, Gilberto. S*ociabilidad, Religión y Política en la Definición de la Nación: Colombia, 1820-1886*. Bogotá: Editorial Universidad Externado de Colombia, 2001.

Melo, Jorge Orlando. *La conformación del patrimonio bibliográfico colombiano: algunas notas sobre su pasado y propuestas para hoy.*

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Digital tools for personal knowledge building in the humanities

Jacek Tomaszczyk[141] **& Anna Matysek**[142]

[jacek.tomaszczyk@gmail.com | anna.matysek@us.edu.pl]

University of Silesia in Katowice, Poland

The work of humanity scholars relies heavily on literature reading, analyzing and synthesizing. It's very difficult to remember huge amounts of information for a long time so researchers take notes that contain important ideas, quotations and bibliographic references. However, in many cases such notes are only transient information chunks that are meant to be used just for one research paper. While they are useful because they temporarily support working memory, they don't add to permanent, network structure of knowledge. In fact, there's little knowledge in this kind of notes as they are not interlinked, they mainly relate only to one article, and they are static in the sense that once they are made, they don't get updated. With time, although they are written down, they fade in the researcher's memory and without good search facilities it's almost impossible to make use of them.

In our presentation, we would like to talk about building personal knowledge that involves organizing, interpreting and interlinking concept-based information that accumulates and evolves over time. This knowledge work requires proper tools, so we are going to demonstrate a few applications featuring the most useful mechanisms in structuring information and turning it into a knowledge system: outlining, internal links, bi-directional links (backlinks), transclusion, and spaced exposure to concepts.

[141] Dr Jacek Tomaszczyk is an Associate Professor, Deputy Head of the Institute of Culture Studies, University of Silesia, Poland. He has twenty years' experience in teaching information technology courses to undergraduate, graduate and PhD students. His main field of scientific inquiry focuses on personal information management.

[142] Dr Anna Matysek is an Assistant Professor in the Institute of Culture Studies, University of Silesia, Poland. She teaches a wide range of undergraduate and graduate courses in information behavior and digital tools for researchers, teachers and educators.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

El componente hispanoamericano de la Biblioteca Virtual de la Filología Española (BVFE)

Jaime Peña Arce[143] **& M.ª Ángeles García Aranda**[144]
[jaimepena@ucm.es | magaranda@filol.ucm.es]
Universidad Complutense de Madrid, España

La descripción de las obras lingüísticas difícilmente puede hacerse si antes no sabemos cuáles han existido y cuáles hay ahora. Por eso, una de las tareas de los investigadores es la catalogación de los repertorios lexicográficos, las gramáticas y las ortografías del pasado y del presente. En este sentido y hasta épocas muy recientes, el trabajo que se había hecho para el español era muy deficiente. A ello se habían dedicado, especialmente, algunos investigadores italianos; labor a la que debemos añadir los catálogos más amplios de textos filológicos (Niederehe-Esparza), cuyo antecedente más benemérito es el del Conde de la Viñaza.

La *Biblioteca Virtual de la Filología Española* (www.bfve.es) pone a disposición de los usuarios un portal fácil y único, que permite a cualquier usuario acceder libre y gratuitamente a nuestras obras del pasado que se encuentran digitalizadas en la red, aunque dispersas en multitud de lugares diferentes. Precisamente, con el título de *Biblioteca Virtual de la Filología Española* se quiere evocar el de la bien conocida *Biblioteca Histórica de la Filología Castellana,* del Conde de la Viñaza, que tanto ha ayudado durante decenios en la historiografía de la lingüística del español.

Las búsquedas en la *BVFE* son muy sencillas y pueden realizarse: 1) en el buscador de la página principal introduciendo el término de búsqueda, 2) en el buscador alfabético seleccionando la letra inicial de la obra o tipo de texto que se desea localizar (gramática, tratado gramatical, ortografía, prosodia, nomenclatura, diccionario), 3) en "búsqueda avanzada", en

[143] Jaime Peña Arce (Madrid, 1989) es licenciado en Filología Hispánica por la Universidad Complutense de Madrid y doctor en Lengua Española por esa misma institución. Entre sus trabajos de investigación destacan los estudios dedicados a las variedades del español, tanto de España como de América; su dedicación a las tareas lexicográficas, con especial atención al quehacer de la Real Academia Española; y sus incursiones historiográficas, evidenciadas mediante sus contribuciones a la Biblioteca Virtual de la Filología Española (BVFE). Es autor de 18 artículos publicados en revistas nacionales e internacionales, numerosas contribuciones en volúmenes colectivos, varias reseñas y un libro, El léxico de Cantabria en los diccionarios de la Academia. De Autoridades al DLE-2014. Igualmente, desde los inicios de su tarea investigadora ha participado en diversos congresos y seminarios especializados.

[144] M.ª Ángeles García Aranda es Doctora en Filología Hispánica por la Universidad Complutense de Madrid. Actualmente, es Profesora Titular del área de Lengua española en el Departamento de Lengua española y Teoría de la Literatura y Literatura comparada de la Universidad Complutense de Madrid. Ha investigado y ha impartido docencia en diferentes universidades españolas y europeas. Su ámbito de investigación lo constituyen la Lexicografía histórica, la Historiografía lingüísticamente, la Historia de la lengua española y, recientemente, la Lingüística misionera. Ha participado en varios proyectos de investigación competitivos y, en la actualidad, es la responsable de la Biblioteca Virtual de la Filología española (www.bvfe.es). Es autora de más de setenta publicaciones en diversas revistas y publicaciones nacionales e internacionales.

donde se pude filtrar por obra, fecha de publicación, impresor, lugar de impresión, lenguas de publicación, periodo cronológico, etc. Una vez finalizada la búsqueda, solo hay que pinchar en el título de la obra para acceder a los datos completos del registro y al ejemplar.

Esta comunicación tiene como objetivo presentar el origen, el funcionamiento, el estado actual y la proyección futura de la *BVFE*, una herramienta útil y recomendable para cualquiera que desee o necesite consultar las fuentes filológicas de nuestro pasado, con especial atención al componente hispanoamericano (autores, lugares de impresión de las obras indexadas y bibliotecas en las que se conservan los ejemplares físicos). Además, desde una perspectiva historiográfica más amplia y menos centrada en nuestros registros, se pretende realizar una aproximación general comparativa al número de trabajos sobre la lengua castellana compuestos en cada una de las dos orillas del Atlántico (España frente a América o viceversa), así como a los detalles sobre la tipología, la extensión o la transcendencia de los títulos compuestos en cada uno de estos dominios.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Exploring linguistic representation of ethnicity in YouTube

Mariela Andrade[145]
[mariela.andrade@gmail.com]
Universidad de Sevilla, España

Abstract:
This presentation explores the semiotic modes and their representation of ethnicity, particularly the building of a Latino[146] identity in the US using corpus linguistics and discourse analysis. My starting point is to observe the patterns of representation that YouTubers use to enact their Latino identity paying especial attention to the indexicality of their ethnicity. I focus on how users and producers of material on YouTube build an ingroup ethnic identity. Using quantitative and qualitative data, I analyze the discursive construction of the ethnic dimension of Latino identity as understood strictly within the contexts of identity politics in the United States (Blitvich, Bou-Franch, & Lorenzo-Dus, 2013). The quantitative portion looks at finding those words collocating together in which social identity is salient in a corpus of comments to videos on YouTube that treats representation issues. The study is later complemented with a qualitative analysis of the content categories found (thematic contents) in a corpus made up of 400 videos from a YouTube channel.

Keywords: digital discourse analysis, CMC, identity construction, YouTube, Latino

The United States is largely known for being a country of immigrants but at the same time having a hard time dealing with important racial rhetoric that classifies people into 3 major races: White, Black, and Asian. The situation of Latinos is therefore ambiguous because this tag is not considered a race in the U.S. census but rather an ethnic group. This dilemma is a strong component of the identity building in the community and it is one of the 8 thematic contents observed and analyzed in the video corpus. 400 videos from BuzzFeed's channel *PeroLike* were viewed and analyzed over a period of 10 months looking to find those thematic contents that were more salient in relation to the construction and enacting of a Latino identity.

[145] Mariela Andrade is a Ph.D. candidate in linguistics at Universidad de Sevilla and the Free University in Brussels, VUB. M.A in Linguistics and Foreign Languages Teaching, she is especially interested in the ways people make use of languages to enact identity in a variety of contexts. Most recently, she is invested in exploring this activity on internet, especially on social media.

[146] The term "Hispanic" was coined in the '70s through the US Census Bureau to address the evident arrival to the country of large populations of Spanish speakers. For purposes of this work, we will use the word "Latino" as it is the one preferred by the community

In the times we are living it is out of the question that social media and the internet in general is a viable source for analyzing discourse as many of the modern utterances take place there. This poses a challenge for discourse analysis as it has traditionally analyzed small corpus from printed material. As times have changed, innovations in the field have been done and today researchers can count on larger corpora that allow for analyzing digital context by means of corpus linguistics (Androutsopoulos J. a., 2015; Baker, 2013). In the case of this work in particular, we use the aforementioned video corpus for the qualitative analysis and a second subcorpora based on the comments from 5 of those videos (rougly 17.000 words). To analyze these comments we use the WordSketch tool from the software Sketch Engine, looking for the strongest collocates for the word *Latino*. The aim is to combine both corpora for revealing some of the ways in which Latinos construct their group, enact and perceive their *latinidad*.

Bibliography:

Androutsopoulos, J. (2010). The study of language and space in media discourse. In P. A. (Eds.), *Language and Space* (pp. 740-759). Berlin and New York: de Gruyter.

Androutsopoulos, J. a. (2015). Language and Discourse Practices in Participatory Culture. In A. a. Georgakopoulou, *The Routledge Handbook of Language and Digital Communication* (pp. 354-370). Routledge.

Baker, P. G. (2013). Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim'in the British press 1998–2009. . *Appplied linguistics, 34(3)*, 255-278.

Fuentes Rodríguez, C. (2013). Identidad e imagen social. In C. (. Fuentes Rodriguez, *Imagen Social y Medios de Comunicación* (pp. 13-21). Madrid: Arcos Libros.

Fuentes Rodríguez, C. (2015). *Lingüística Pragmática y Análisis del Discurso.* Madrid: Arco.

Garcés-Conejos Blitvich, P. &.-F. (2014). ¿! Hispano y Blanco?!: Racialización de la identidad latina en YouTube. *Discurso & Sociedad, 8 (3)*, 427-461.

Garcés-Conejos Blitvich, P., Bou-Franch, P., & Lorenzo-Dus, N. (2013). Despierten, Latinos ('Wake up, Latinos'). *Journal of Language and Politics 12:4*, 558 - 582.

Van Dijk, T. (1998). *Ideology: A multidisciplinary approach.* London: Sage.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Geolingüística digital: proyecto de un corpus de atlas lingüísticos

Carolina Julià Luna[147]

[cjulia@flog.uned.es]

Universidad Nacional de Educación a Distancia, España

Palabras clave: variación, atlas lingüísticos, corpus, humanidades digitales

La aplicación de las nuevas tecnologías al estudio lingüístico en el marco de las humanidades digitales ha promovido la creación de multiplicidad de herramientas que perfeccionan y mejoran los resultados de las investigaciones. Corpus, diccionarios, tesoros, archivos y bases de datos digitales permiten al investigador manejar una ingente cantidad de datos para llevar a cabo análisis contrastivos, históricos o variacionistas con mayor precisión (De Benito 2019). En el ámbito de los estudios sobre variación, el lingüista tiene a su disposición un amplio abanico de recursos para examinar el cambio en las lenguas desde distintos ejes (Torruella 2009). Desde la perspectiva diatópica, las nuevas tecnologías han supuesto una revolución aún mayor, si cabe, pues además de permitir la recopilación de un volumen considerable de datos, la tecnología permite la geolocalización de la información, lo que permite ofrecer datos lingüísticos a través de su distribución en el espacio y representados en mapas interactivos. Esta posibilidad conlleva un cambio en la tradicional geografía lingüística, esto es, en el «método dialectal y comparativo […] que presupone el registro en mapas de un número relativamente elevado de formas lingüísticas (fónicas, léxicas o gramaticales), comprobadas mediante encuesta directa y unitaria en una red de puntos de un territorio determinado, o, por lo menos, tiene en cuenta la distribución de las formas en el espacio geográfico correspondiente a la lengua, a las lenguas, o a los dialectos o hablas estudiados» (Coseriu 1977: 103).

La innovación tecnológica ha permitido a la geografía lingüística ir un paso más allá y pasar del mapa en papel al mapa digital (Sousa 2017). Esto constituye un salto considerable tanto en lo que se refiere a la recopilación, manejo y representación de los datos como en lo que compete a la consulta y explotación de los mismos. En este sentido, en la actualidad cabe distinguir dos grupos en el marco de investigaciones que utilizan la representación cartográfica para el estudio lingüístico: por un lado, los que siguen la línea de la geolingüística tradicional y

[147] Licenciada en Filología Hispánica por la Universitat Autònoma de Barcelona (2005) y Doctora en Filología Española por la misma (2010). Actualmente, es profesora ayudante doctora en el Departamento de Lengua Española y Lingüística General de la Universidad Nacional de Educación a Distancia (Madrid). Anteriormente, ha trabajado como profesora asociada en la Universitat Autònoma de Barcelona (2005-2018) como profesora de idiomas (español y catalán) en la Universitat Pompeu Fabra (2015-2018) y como profesora lectora invitada en la Universiteit Antwerpen (2010-2013) de Bélgica. Líneas de investigación principales: variación léxica, geolingüística, lingüística cognitiva, morfología léxica y lexicografía.

son deudores del método y, por otro lado, los que no se basan en el método de la geografía lingüística y simplemente usan los mapas como medio de representación de información lingüística.

Desde la perspectiva y el concepto tradicional de atlas lingüístico –es decir, el que se corresponde con el que Jules Gillierón ideó y materializó sobre Francia (*ALF*) a principios del siglo XX y que sigue la metodología de recogida y representación de datos de la geolingüística–, pueden diferenciarse tres grandes proyectos: (a) por una parte, aquellos que tienen como finalidad recopilar las imágenes de los mapas de atlas publicados en papel con el objetivo de conservarlos y ofrecerlos a los usuarios en internet. Son ejemplos de este tipo el proyecto de digitalización del atlas de Georg Wenker (el proyecto *DIWA: Digital Wenker Atlas*, cfr. Herrgen 2010), del atlas lingüístico francés (el proyecto CartoDialect sobre el *ALF*), del atlas dominio catalán (*ALDC* y *PALDC*) y del atlas lingüístico de Colombia (*ALEC Interactivo*), entre otros. (b) Por otra parte, los que siguen el método de recogida de datos tradicional y que se han publicado directamente en formato digital en internet (*ALeCMan, ADiM*). (c) Finalmente, los que, además de ofrecer en internet un mapa con los resultados de los cuestionarios lingüísticos, también permiten hacer consultas a las bases de datos que generan el cartografiado digital. Ejemplos de este tipo son el proyecto en el que se está trabajando sobre el *ALPI* o los magníficos datos que ofrecen tanto los *Índices do Atlas lingüístico galego* como la versión digital del *EHHA* (*Euskararen Herri Hizkeren Atlasa*), en cuyas bases de datos es posible consultar la información de los mapas por diferentes campos.

Desde una concepción más amplia y actual, son muy variadas las herramientas digitales de consulta en línea que se han creado para el estudio de la variación lingüística. Algunas de ellas emplean el término *atlas* en su denominación a pesar de que se alejan del concepto tradicional de este tipo de compilaciones desde una perspectiva metodológica. Los datos que se representan y recogen en este tipo de plataformas proceden de fuentes de diverso tipo: de encuestas, vaciados de obras o informaciones procedentes de las redes sociales. Son muestra de este tipo de proyectos el *ASINES* (*Atlas Sintáctico del Español*), el *ADiBa* (*Atlas Dialectal de Barcelona*), el *Atlas interactivo de la entonación del español* o FRONTESPO (*Frontera Hispano Portuguesa*). Destaca, entre este tipo de plataformas, @*Tweetolectology* (University of Cambridge), un perfil de la red social Twitter desde el que se publican mapas lingüísticos del Reino Unido. Los datos que se publican son de diversa procedencia "Most maps we post on this account are from four sources- the Survey of English Dialects (SED), a 1950s survey of rural speech in England; the English Dialects App (EDA), a 2016 smartphone-based survey of speech in the British" (Tweetolectology, 20 de julio de 2016).

Es entre todas estas herramientas entre las que se evidencia la necesidad de la dialectología española de poner a disposición de los investigadores y de la sociedad los materiales de la geolingüística regional publicada en la segunda mitad del siglo XX (p. ej. *ALEA, ALECant, ALEICan, ALEANR*, *ALCyL*), puesto que esta es la única forma de que no caigan en el olvido y se pierda en las bibliotecas el valor que atesoran los mapas, a los que podemos considerar verdaderos diamantes en bruto listos para ser interpretados, contrastados y comentados. Es con

este fin que se está trabajando en el desarrollo de un corpus digital de los materiales geolingüísticos regionales del español, el *Corpus de los atlas lingüísticos* (*CORPAT*).

El objetivo principal de esta comunicación es la presentación del proyecto de este nuevo corpus. *CORPAT* almacenará y gestionará la información lingüística de los mapas en bases de datos interrelacionadas que permitirán la búsqueda por ámbito semántico, concepto, variante designativa, lema, categoría gramatical, información morfológica, localización geográfica, fuente geolingüística e información adicional. En futuras fases se ha previsto ofrecer también búsqueda fonética. Se trata de una herramienta lingüística digital que complementará los datos que ofrecen los corpus textuales, los diccionarios y los atlas publicados en línea y permitirá obtener mucho más rendimiento de los datos geolingüísticos. El proyecto que se enmarca en el ámbito de los estudios humanístico-lingüísticos y se constituye en dos objetivos principales: por un lado, la conservación del patrimonio histórico-lingüístico y cultural de la lengua española y, por otro lado, la aproximación de la ciencia y sus avances a la sociedad. Para la consecución de estos dos grandes objetivos, es imprescindible partir de las posibilidades que ofrecen las nuevas tecnologías, por ello, este trabajo se ubica en el campo de las Humanidades Digitales (Iribarren 2017).

Bibliografía

ADiBa = Gallego Bartolomé, Ángel (dir.): *Atlas Dialectal de Barcelona*, Barcelona. <http://adiba.cat/2018/>.

ADiM = García Mouton, Pilar/Molina Martos, Isabel, *Atlas Dialectal de Madrid,* Madrid, CSIC, 2015. <www.adim.cchs.csic.es>.

ALCyL = Alvar, Manuel (1999): *Atlas lingüístico de Castilla y León*, 3 vols., Salamanca, Junta de Castilla y León/Consejería de Educación y Cultura.

ALDC = Veny Clar, Joan y Pons Griera, Lídia (2001-): *Atles lingüístic del domini català*, 7 vols., Barcelona, Institut d'Estudis Catalans. <http://aldc.espais.iec.cat/mapes/>

ALEA = Alvar, Manuel (1963-1973): *Atlas lingüístico y etnográfico de Andalucía*, 6 vols., Granada, Universidad de Granada.

ALEANR = Alvar, Manuel (1970-1983): *Atlas lingüístico y etnográfico de Aragón, Navarra y Rioja*, 12 vols., Madrid, La Muralla.

ALEC = Instituto Caro y Cuervo (2018): *Atlas Lingüístico-Etnográfico de Colombia.* Recuperado de <http://alec.caroycuervo.gov.co>.

ALECant = Alvar, Manuel (1995): *Atlas lingüístico y etnográfico de Cantabria*, 2 vols., Madrid, Arco/Libros.

ALeCMan = García Mouton, Pilar yMoreno Fernández, Francisco (1987-): *Atlas lingüístico y etnográfico de Castilla-La Mancha*. <http://www.uah.es/otrosweb/ alecman/>.

ALEICan = Alvar, Manuel (1975-1978): *Atlas lingüístico y etnográfico de las Islas Canarias*, 3 vols., Madrid, La Muralla.

ALF = Gilliéron, Jules y Edmond Edmont (1902-1910): *Atlas Linguistique de la France*. Paris: Honoré Champion, 12 vols. <http://lig-tdcge.imag.fr/cartodialect4/>

ALGa = García González, Constantino y Santamarina Delgado, Antón (1990-): *Atlas lingüístico galego*, 6 vols., Santiago de Compostela, Universidade de Santiago/Instituto da Lingua Galega.

ASinEs = *Atlas Sintáctico del Español* [en línea] < http://www.asines.org>.

Coseriu, Eugenio (1977): "La geografía lingüística", en *El hombre y su lenguaje. Estudios de teoría y metodología lingüística*, Madrid: Gredos, pp. 103-158.

De Benito, Carlota (2019): "Los corpus del español desde la perspectiva del usuario lingüista", *Scriptum digital* 8, pp. 1-21.

EHHA = *Euskararen Herri Hizkeren Atlasa,* Bilbo: Euskaltzaindia. < https://www.euskaltzaindia.eus/index.php?option=com_content&view=article&id=565&Itemid=466&lang=eu >.

FRONTESPO = *Álvarez Pérez, Xosé Afonso (2018-): Frontera España-Portugal:* documentación lingüística y bibliográfica, <http://www.frontespo.org/>.

Herrgen, Joachim (2010): "The digital wenker atlas (www.diwa.info): an online research tool for modern dialectology", *Dialectologia, Special Issue I,* pp. 89-95.

INTITUTO DA LINGUA GALEGA [en línea]: *Índices do Atlas lingüístico galego*, <http://ilg.usc.es/indices/>.

Iribarren Donadeu, Teresa (2017): "Humanidades digitales" en Raquel Gómez Díaz (coord.): *Fuentes especializadas en Ciencias Sociales y Humanidades*, pp. 441-473.

PALDC = Veny, Joan (2007-): *Petit atles lingüístic del domini català*. Barcelona: Institut d'Estudis Catalans

Sousa, Xulio (2017): "From field notebooks to automatic mapping: the *Atlas Lingüístico Galego* database", *Dialectologia et Geolingüistica*, 23/1, pp. 1-22.

Torruella, Joan (2009): «Los ejes principales en el diseño de un corpus diacrónico: el caso del CICA», Pascual Cantos y Aquilino Sánchez (eds.) *A survey of corpus-based research*, Murcia: Asociación Española de Lingüística de corpus, pp. 21-36.

Wenker, Georg (1881): *Sprachatlas von Nord-und Mitteldeutschland, Auf Grund von systematich mit Hülfe der Volksschullehrer gesammeltem Material aus circa 30.000 Orten*, Strassburg. <http://diwa.info/titel.aspx>

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Infraestructura e investigación en Humanidades Digitales: pilares en la búsqueda, recuperación y difusión documental para la construcción de grafos dinámicos sobre teatro de la Edad de Plata

Concepción María Jiménez Fernández[148]
[concepcionm.jimenez@unir.net]
Universidad Internacional de La Rioja (UNIR), España

Palabras clave: Humanidades Digitales / Corpus teatral / Codificación en TEI / Acceso a datos / Difusión de información / GitHub / DraCor / Investigación colaborativa.

Como es sabido, las Humanidades Digitales (HD) es una disciplina diferente al resto por contar con una de sus características más relevantes la de tener un carácter interdisciplinar. Es por ello que necesitan de infraestructuras de investigación que posibiliten el incremento del alcance científico así como la realización de proyectos de investigación de larga duración que requieran un esfuerzo común. Igualmente, se necesita promover tanto la colaboración como la interdisciplinariedad y la interacción.

El trabajo colaborativo y la interdisciplinaridad han resultado fundamentales para la búsqueda, recuperación y difusión de los datos referidos a representaciones teatrales de obras de un corpus concreto (teatro español desde 1868 a 1936) desde su estreno. Un corpus sometido a una revisión filológica para su digitalización en TEI, puesto a disposición de la comunidad investigadora a través del repositorio en abierto GitHub y que hoy forma parte del proyecto DraCor.

Antecedentes, justificación y objetivos

Desde 2015 a 2017, el grupo GHEDI (Grupo de Humanidades y Edición Digital) de UNIR, trabajó en el proyecto BETTE (Biblioteca Electrónica Textual del Teatro Español, 1868-1936). Esta línea de investigación centrada en las Humanidades Digitales se ve hoy continuada por

[148] Concepción Jiménez es diplomada en Biblioteconomía y Documentación por la Universidad de Granada, Licenciada en Documentación por la Universidad de Extremadura y Doctora en Documentación por la Universidad de Salamanca (Premio extraordinario de doctorado). En la actualidad es la directora de la revista Mi Biblioteca, editada por la Fundación Alonso Quijano. También es Directora Académica del Máster Universitario en Didáctica de la Lengua en Educación Infantil y Primaria y profesora adjunta en la Facultad de Educación de la Universidad Internacional de La Rioja (UNIR). Ha publicado varios libros y artículos en revistas científicas sobre didáctica de la lengua y la literatura así como sobre Documentación. Igualmente ha presentado comunicaciones en congresos internacionales sobre las áreas mencionadas.

HDATEATROUNIR cuyo proyecto "Visualización de redes de sociabilidad mediante grafos en el teatro español" tiene como objetivo principal proponer una metodología para el análisis con medios digitales de las relaciones sociales que se establecen entre distintos agentes que crean y protagonizan textos teatrales españoles.

El corpus de obras de teatro codificado en XML-TEI y que ya realizó el grupo GHEDI, incluye textos de autores como Galdós, Clarín, Dicenta, Valle-Inclán, Muñoz Seca, Echegaray, Valera, Unamuno y García Lorca. En total son 25 obras de estos nueve dramaturgos.

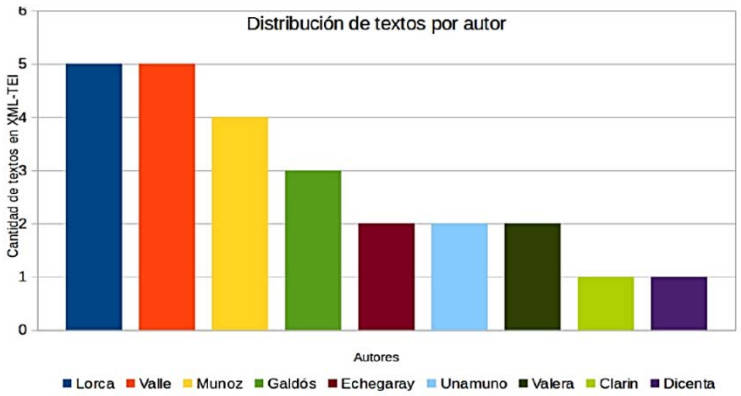


Ilustración 1: Distribución de textos teatrales según autores

La selección de autores y textos ha seguido varios criterios, entre los que se pueden destacar los siguientes:

- La existencia de alguna digitalización del texto (en formatos HTML, ePUB, PDF…)
- La canonización del autor y del texto
- Los intereses de los investigadores del grupo
- Conseguir diferentes textos del mismo autor, sin que ninguno de ellos tenga un peso desproporcionado en el corpus total
- Textos sin derechos de autor, que permitirían su publicación en abierto

A partir de esas obras digitalizadas, se hace imprescindible conocer las representaciones (vida teatral) de las mismas desde su estreno hasta diciembre de 2018. Con ello obtendríamos datos valiosos para establecer conclusiones sobre la cantidad de representaciones, su tipo de distribución, su popularidad en los escenarios hasta el día de hoy, su correlación con otros datos cuantitativos referidos al grado de canonización de esas obras (como la cantidad de reediciones). Todo ello facilita el trabajo en equipo y la investigación colaborativa así como su difusión abierta a la comunidad investigadora a través de sistemas como GitHub.

Nuestro principal objetivo es, como hemos visto, realizar una búsqueda y localización de datos sobre representaciones teatrales del corpus mencionado desde su estreno hasta 2018 (cuando se cumplen los necesarios 80 años para la pérdida de derechos de autor de ciertas obras). En este orden de cosas, varias son las pretensiones en este empeño:

a) Rastrear diferentes fuentes de información portadoras de datos sobre fechas, compañías de teatro, actores y lugares de representación de otras teatrales.

b) Crear una base de datos que sirva para clasificar y ordenar los datos obtenidos.
c) Diseñar un instrumento que facilite la organización, búsqueda y recuperación de la información de forma automatizada por parte de los miembros del grupo de investigación.

Para lograrlo comenzamos con una toma de contacto con las posibles y numerosas fuentes de información que nos facilitasen los datos que necesitábamos. Esas fuentes son muy variadas: Diarios, revistas de teatro, programas de mano, hojas de taquilla… Esta documentación se podría localizar en diferentes instituciones como Bibliotecas (Fundación Juan March, Biblioteca Digital Hispánica de la BNE), Centros de documentación (Centro de documentación teatral, centro de documentación de artes escénicas de Andalucía…), archivos, hemerotecas, manuales (como los escritos por Berta Muñoz o los de Dougherty y Vilches)… Y destacar, principalmente, el Centro de Investigación SELITEN@T (Semiótica Literaria, Teatral y Nuevas Tecnologías) de la UNED y dirigido por el Dr. Romera Castillo. En él pudimos localizar numerosos trabajos de investigación sobre la cartelera teatral en distintas ciudades españolas. Uno de los trabajos más relevantes que nos ha facilitado nuestra labor de gran manera ha sido la *Base de datos de representaciones teatrales en algunos lugares de España (1850-1900)* elaborada por la doctora Dolores Romero, de la Universidad Complutense de Madrid a través de investigaciones previas de estudiosos de diferentes lugares del país (Albacete, Ávila, Badajoz, Ferrol, León, Las Palmas de Gran Canaria, Pontevedra y Toledo).

Tras la lectura detallada de cada una de las fuentes encontradas en las distintas instituciones que se han mencionado –y que nos redirigían a otras posibles fuentes más concretas– llevamos a cabo un análisis minucioso de los datos encontrados con miras a crear un instrumento o herramienta específica como fórmula eficaz para organizar, localizar y recuperar los datos de forma automatizada por parte de todos los miembros del grupo.

No se trata, de "recopilar" y organizar datos sin más sino de contar con una base de datos que nos permita analizar información a través de correlaciones, es decir, medir la relación entre dos variables numéricas, si están asociadas o no… Por ejemplo: ¿Existe correlación entre la cantidad de veces que se representa una obra de un autor (desde 1918-1926) y la canonización de ese autor?

Otros aspectos que también se pueden valorar y que nos darían resultados muy llamativos gracias a este análisis de la vida escénica que estamos describiendo serían:

- ¿Puede ser que unas compañías (de más prestigio) representen alguna obra u obras más veces que otras?
- ¿Qué lugares son los preferidos para representar una determinada obra? ¿Y por qué será así? ¿Influye que un autor sea de una ciudad concreta para que sea más representado allí?
- ¿Qué autor ha sido más representado (o menos representada) en diferentes ciudades durante los mismos años?

- ¿Influye la posición ideológica de un autor en su éxito a la hora de ser más representado que otros?
- ¿Existe una correlación entre el número de publicaciones de investigación sobre un autor y la representación de las obras de este?
- Etc.

Somos conscientes de la dificultad que entraña este análisis y rastreo pero nos mueve el interés y la importancia de esta labor como campo de investigación que todavía está por explotar. Por ello podemos son varias las razones que nos motivan a seguir:

- Abrir una nueva línea de investigación relacionada con un proyecto anterior, BETTE
- Continuar con una línea anterior iniciada por investigadores (como el profesor Romera Castillo, la profesora Romero, el trabajo desarrollado por Muñoz Cáliz desde el CDT…) pero adaptándolo a nuestras necesidades e intereses
- Completar una información de un periodo que cuenta con mucha dispersión en sus fuentes de búsqueda
- Corregir algunos datos incorrectos creando una base de datos lo más completa y uniforme posible

Con todo lo expuesto, desde HDATEATROUNIR se continuará la labor iniciada por proyecto BETTE, con la idea de ampliar el corpus de veinticinco textos en lenguaje de marcado XML-TEI que el anterior Grupo de Humanidades y Edición Digital (GHEDI) había elaborado y puesto a disposición de toda la comunidad investigadora a través de un repositorio en abierto GitHub de dichos textos de la dramaturgia de la Edad de Plata (https://github.com/GHEDI/BETTE). Se pretende, de esta forma, ampliar el corpus que nos servirá de base para el estudio comparativo con medios digitales de las relaciones sociales de los personajes que protagonizan estas obras. Por otro lado, se propone profundizar en el estudio de las redes sociales con grafos a partir de dichos textos marcados previamente en XML-TEI, a partir de los cuales será posible extraer la información necesaria para la construcción de los grafos.

Se explicaría cómo se ha llevado a cabo la minuciosa revisión filológica de esos textos para lograr la uniformidad en el corpus; el proceso de accesibilidad a través de GitHub así como la pertenencia de BETTE al proyecto Drama Corpora (DraCor), accesible en línea (https://dracor.org). Se trata de una plataforma digital de investigación sobre teatro que incluye corpus codificados en TEI y que facilita investigaciones reproducibles. No solo se aportan grafos sino otros muchos elementos y aspectos que de manera tradicional requerirían un tiempo y un esfuerzo ilimitado. Así mismo, con la Interfaz de programación de aplicaciones (API) de DraCor se pueden implementar preguntas de investigación y se permite el acceso mucho más sencillo a los datos sin tener que lidiar con el XML subyacente en los diferentes textos. Se investigan las redes interactivamente además de ofrecer grafos dinámicos y enriquecedores para la comunidad investigadora.

Bibliografía

Calvo, J. y Santa María, T., *GHEDI/BETTE: Versión 1.0 de BETTE* (octubre 2017) (Version 1.00). Zenodo. doi:10.5281/zenodo.1010140.

Gómez, S., Calvo Tello, J., González, J. M. and Vilches, R. (2015). Hacia una biblioteca electrónica textual del teatro en español de 1868-1936 (BETTE). *Texto Digital*, 11(2), pp. 171–84.

Isasi, J., "Acercamiento al análisis del sistema de los personajes en la narrativa escrita en español: el caso de Zumalacárregui y Mendizábal de Pérez Galdós", *Caracteres*, 6 (2), (2017): 107-137, http://revistacaracteres.net/wp-content/uploads/2017/12/Caracteresvol6n2noviembre2017-galdos.pdf

Jannidis, F., Reger, I., Krug, M., Weimer, L., Macharowsky, L. and Puppe, F. (2016). Comparison of Methods for the Identification of Main Characters in German Novels. *DH2016*. Krakow: ADHO, pp. 578–82 http://webcache.googleusercontent.com/search?q=cache:LjYz88cQhboJ:dh2016.adho.org/abstracts/297+&cd=1&hl=es&ct=clnk&gl=de&client=ubuntu.

Jannidis, F., Kohle, H. and Rehbein, M. (eds). (2017*). Digital Humanities: eine Einführung*. Stuttgart: J.B. Metzler Verlag.

Jiménez, C. M., Santa María Fernández, T. y Calvo Tello, J. (2017). *Biblioteca Electrónica Textual del Teatro en Español (BETTE)*. La Rioja: UNIR. https://bit.ly/2SGtwhB el 11/01/2019.

Jiménez, C., Calvo Tello, J., Simón Parra, M., Martínez Nieto, R. B. and García Sánchez, M. (2017). BETTE: Biblioteca Electrónica Textual del Teatro en Español de la Edad de Plata. *Sociedad, Políticas, Saberes. Málaga: HDH*, pp. 88–91 http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf.

María Jiménez, C., Santa María Fernández, T. y Calvo Tello, J., Biblioteca Electrónica *Textual del Teatro en Español (BETTE),* La Rioja, UNIR, 2017. https://github.com/GHEDI/BETTE

Martínez Carro, E. y Santa María, T. (2019). Biblioteca electrónica textual del teatro español (1868-1936) e investigación con grafos. *Revista de Humanidades Digitales* 3, pp. 23-45. http://revistas.uned.es/index.php/RHD/article/view/23144

Moretti, F. (2011). Network Theory, Plot Analysis. *The New Left Review (*68), pp. 80–102.

Moretti, F. (2013). "Operationalizing": or, the function of measurement in modern literary theory. *The New Left Review* (84), pp. 103-119.

Santa María Fernández, T., Calvo Tello, J., Jiménez Fernández, C. (2020). ¿Existe correlación entre importancia y centralidad? Evaluación de personajes con redes sociales en obras teatrales de la Edad de Plata. *Digital Scholarship in the Humanities (DSH).*

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

La aplicación del lenguaje TEI al estudio de la puntuación medieval hispánica: la *General e grand estoria* de Alfonso X

Miguel Las Heras Calvo[149] **& Fernando García Andreva**[150]
[miguel.las-heras@unirioja.es | fernando.garciaan@unirioja.es]
Universidad de La Rioja, España

El conocimiento que se tiene actualmente sobre el uso de los signos de puntuación en manuscritos medievales hispánicos es todavía muy escaso. A partir del *Coloquio en París sobre frases, textos y puntuación en los manuscritos medievales* (1981), se han publicado trabajos que intentan hallar tendencias en el uso de estos signos. Una gran parte de estos estudios se adscriben al ámbito de la filología (Blecua 1984; Martín Aizpuru 2012; Fernández López 2014, 2015; o Sánchez-Prieto Borja 2017, entre otros), pero también se encuentran aportaciones desde la paleografía y la diplomática (Millares Carlo 1983: 283 o Serna Serna 2011), sin olvidar que es una aspecto que está estrechamente relacionado al ámbito de la edición (Roudil 1978 o Bédmar Sancristóbal 2006). Esta diversidad de disciplinas muestra la complejidad del estado de la cuestión y, además, la variedad de factores, que se deben tanto a condicionantes de tipo lingüístico (causas prosódicas, morfológicas o sintácticas) como extralingüístico (tipo de letra o disposición del texto en la página). Este enfoque interdisciplinario ha sido precisamente el privilegiado por los estudiosos europeos, quienes han profundizado en los valores y funciones de la puntuación medieval en sus correspondientes países.

Asimismo, a lo largo de las últimas décadas, ha habido un incremento exponencial de los estudios filológicos que se apoyan en las ventajas que proporcionan las Humanidades Digitales (Bia y Sánchez Quero 2001, Fradejas Rueda 2009, Spence 2014 o Martín Aizpuru 2016), pero que todavía no se han aplicado de manera sistemática al estudio de la puntuación medieval hispánica. Por ello, estimamos que su estudio puede verse beneficiado y enriquecido por la versatilidad, sistematicidad y precisión que proporciona el lenguaje TEI, ya que, como es bien

[149] Graduado en Lengua y Literatura Hispánica, con Premio Extraordinario, por la Universidad de La Rioja. Tras su Máster Universitario en Profesorado, continúa su formación en el Programa de Doctorado en Humanidades. Ha disfrutado de becas de colaboración de la universidad y el MECD. Líneas: E/LE, Dialectología e Historia de la Lengua. Ha intervenido en varios congresos internacionales de jóvenes investigadores y colaborado en la organización de jornadas realizadas en la misma universidad. Participa en Proyectos de Innovación Docente y en otros vinculados a las Humanidades Digitales.

[150] Profesor contratado interino de la Universidad de La Rioja. Licenciado y doctor (sobresaliente cum laude) en Filología Hispánica por la Universidad de La Rioja. Líneas de investigación: Historia de la lengua española a través de la documentación altomedieval; historia del léxico español; la oralidad en la enseñanza de español como LE/L2. Ha participado en varios proyectos de investigación nacionales e internacionales, entre los que cabe señalar The Confluence of Religious Cultures in Medieval Historiography: A Digital Humanities Project (2019-2024), sobre la General e grand estoria de Alfonso X.

sabido, permite elaborar marcas y etiquetas propias en función de las cuestiones que queramos editar, analizar o presentar.

Asimismo, para la recuperación de los datos y su posterior estudio, este lenguaje de etiquetas encuentra un complemento ideal en las hojas de estilo XSLT (*Extensible Stylesheet Language Transformations*), que, tras su correcto diseño y configuración, transforman el documento de origen en HTML, visualizando aquellos datos que nos interesen con el formato que se desea.

Por tanto, para este trabajo en concreto se pretende aplicar este lenguaje de etiquetas a los diversos factores (lingüísticos y extralingüísticos) que podrían condicionar el uso de los signos de puntuación en una obra en concreto: la *General e grand estoria* de Alfonso X el Sabio, texto fundamental dentro del *scriptorium* alfonsí y en la historiografía medieval hispánica y europea. Concretamente, se han confeccionado alrededor de 200 etiquetas diferentes, atendiendo a factores como las divisiones textuales, la disposición del texto en la página, la sintaxis, la modalidad de los enunciados e, incluso, las posibles interferencias puntuarias que podría haber con respecto a la Vulgata. Asimismo, se mostrará el diseño de las hojas de estilo que se han utilizado para la recuperación de los datos y su posterior estudio.

Palabras clave

Puntuación medieval, *TEI*, *XSLT*, *General e grand estoria*, *Alfonso X el Sabio*

Bibliografía

Bia, Alejandro y Manuel Sánchez-Quero (2001): «Diseño de un procedimiento de marcado para la automatización del procesamiento de textos digitales usando XML y TEI» en Pablo Lucio de la Fuente y Adoración Pérez (coords.), *JBIDI. Segundas jornadas de Bibliotecas Digitales,* Almagro, Ciudad Real, pp. 153-165.

Bédmar Sancristóbal, María Elena (2006): «Problemas de edición de textos manuscritos modernos: la puntuación», en Lola Pons Rodríguez (ed.), *Historia de la Lengua y Crítica Textual*, Vervuert, Lingüística Iberoamericana, pp. 127-180.

Blecua, José Manuel (1984): «Notas sobre la puntuación española hasta el Renacimiento», en *Homenaje a Julian Marías*, Madrid, Espasa Calpe, pp. 119-130.

Fernández López, M. ª del Carmen (2014): «Estudio contrastivo de hábitos de interpunción en manuscritos medievales castellanos: ¿sistematización en los usos de los copistas?», en María del Rocío Díaz y Belén Almeida (coords.), *Estudios sobre la historia de los usos gráficos en español*, Lugo, Axac, pp. 23-72.

Fernández López, M.ª del Carmen (2015): «La puntuación en los manuscritos medievales castellanos: el manuscrito evorense CXXV / 2-3 de Évora (Portugal)», *Revista de Historia de la Lengua Española*, 10, pp. 3-36.

Fradejas Rueda, José Manuel (2009): «La codificación XML/TEI de textos medievales», *Memorabilia*, 12, pp. 219-247.

Martín Aizpuru, Leyre (2012): «Cómo puntuaban los escribanos reales: el sistema de puntuación en la documentación de cancillería real del siglo XIII dirigida al Norte de Burgos», en José María García (dir.), *Actas del IX Congreso Internacional de Historia de la Lengua Española*, Madrid/Frankfurt, Iberoamericana/Vervuert, pp. 523-536.

Martín Aizpuru, Leyre (2016): «Algunos recursos informáticos al servicio de la edición de textos: la edición en XML-TEI», en Chiara Albertini y Santiago del Rey Quesada (coords.), *Hispánica Patavina. Estudios de historiografía e historia de la lengua española en homenaje a José Luis Rivarola*, Padua, CLEUP, pp. 139-154.

Millares Carlo, Agustín ([1932] 1983): *Tratado de paleografía española*, Madrid, Espasa-Calpe.

Roudil, Jean (1978): «Edition de texte, analyse textuelle et ponctuation (brèves réflexions sur les écrits en prose)», *Cahiers de linguistique hispanique médiévale*, 3, pp. 269-299.

Sánchez-Prieto Borja, Pedro (2017): «La puntuación en los códices de la *General Estoria* de Alfonso X el Sabio», *Atalaya. Revue d'études médiévales romanes*, 17.

Serna Serna, Sonia (2011): «El Becerro Gótico de Cardeña: signos de puntuación», en Elena Rodríguez y Antonio Claret (coords.), *La escritura de la memoria: los cartularios*, Huelva, Servicio de Publicaciones, Universidad de Huelva, 237-254.

Spence, Paul (2014): «Edición académica en la era digital: modelos, difusión y proceso de investigación», *Anuario Lope de Vega. Texto, literatura, cultura*, 20, pp. 48-83.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Lenguas clásicas en XML: la base de datos COMREGLA

Cristina Tur Altarriba[151]
[cristina.tur@uam.es]
Universidad Autónoma de Madrid, España
Guillermo Salas Jiménez[152]
[guisalas@ucm.es]
Universidad Complutense de Madrid, España
Alberto Pardal Padín[153]
[pardal@usal.es]
Universidad de Salamanca, España
Iván López Martín[154], **J. Ignacio Hidalgo González**[155], **Berta González Saavedra**[156]
[ivlopez@ucm.es | josehida@ucm.es | bertagon@ucm.es]

[151] Graduada en Lingüística y Lenguas Aplicadas (especialidad de Lingüística Computacional) y doctora en Filología Clásica, ambas por la Universidad Complutense de Madrid. Su tesis, titulada "Sintaxis y semántica de los nombres de sentimiento en latín: empleos adverbiales y colocaciones" y codirigida por los Dres. José M. Baños, Mª D. Jiménez y E. Torrego, fue defendida en noviembre de 2019. Actualmente trabaja como personal técnico contratado en el proyecto COMREGLA en la Universidad Autónoma de Madrid.

[152] Estudió Filología Clásica en la Universidad de Granada, donde se graduó en 2018, y obtuvo el Máster en Filología Clásica por la Universidad Complutense de Madrid en 2019. Actualmente cuenta con una ayuda predoctoral FPI gracias a la cual está realizando su tesis, titulada "La expresión léxica del aspecto incoativo en latín: colocaciones verbo-nominales", bajo la dirección de los profesores José Miguel Baños (UCM) y María Dolores Jiménez (UAH) y es colaborador del proyecto COMREGLA.

[153] Estudió Filología Clásica en la Universidad de Salamanca, donde se licenció en 2009 y obtuvo su master en 2010. Se doctoró en 2017 con la tesis "La interacción entre fonología y sintaxis en griego antiguo" bajo la co-dirección de Julián Méndez Dosuna y Jesús de la Villa. Ha trabajado como profesor asociado en la Universidad de Valladolid y en la de Alcalá de Henares. Actualmente es profesor ayudante doctor en la Universidad de Salamanca y colaborador del proyecto COMREGLA.

[154] Graduado en Filología Clásica (2015) y ha realizado el Máster Interuniversitario en Filología Clásica (2016) y el Máster en Formación del Profesorado (2017), todo ello por la Universidad Complutense de Madrid. Actualmente es beneficiario de un contrato predoctoral FPU, en el marco del cual realiza su tesis doctoral en torno a las construcciones con verbo soporte en la obra clásica la Historia Augusta bajo la dirección de los Dres. J. M. Baños y Mª D. Jiménez, y es colaborador en el proyecto COMREGLA.

[155] Graduado en Filología Clásica por la Universidad Complutense de Madrid, donde también cursó el máster en Filología Clásica y el máster en Formación del Profesorado. Actualmente es doctorando en Filología Clásica por la UCM, con la tesis "La expresión léxica de la diátesis pasiva en latín: colocaciones verbo-nominales" y colabora con el proyecto COMREGLA. Además, imparte clases de latín y griego en el Instituto de Educación Secundaria Marc Ferrer (Formentera, Baleares, España).

[156] Doctora en Filología clásica por la UCM y licenciada en Filología Italiana por la UV. Su tesis doctoral, defendida en 2015, se titula "Expresión de la Procedencia en lenguas indoeuropeas antiguas: griego, latín e hitita". Ha trabajado como contratada de investigación en la Università Cattolica de Milán, en el proyecto Index Thomisticus Treebank, con el investigador Marco C. Passarotti en la anotación semántico-pragmática de textos latinos. Actualmente es profesora ayudante doctora en la Universidad Complutense de Madrid y colaboradora del proyecto COMREGLA.

Universidad Complutense de Madrid, España
Eveling Garzón Fontalvo[157]
[eveling.garzon@uam.es]
Universidad Autónoma de Madrid, España

La presente comunicación tiene como objetivo presentar el proyecto de investigación COMREGLA, cuyo objetivo más ambicioso es la transformación de dos bases de datos relacionales para el griego antiguo y el latín en una herramienta de acceso abierto compatible con otros recursos digitales. Desde esta perspectiva, a continuación, se presentan algunos de los aspectos más relevantes que se esperan tratar en esta comunicación.

El grupo de investigación Rección y Complementación en Griego antiguo y Latín (REGLA, http://www.reglabd.org) fue creado en 1992 por un grupo de investigadores de cuatro universidades españolas (U. Autónoma de Madrid, U. Complutense de Madrid, U. de Alcalá de Henares y U. de Santiago de Compostela). A pesar de sus diversas transformaciones, el objetivo del grupo ha sido siempre la investigación en la estructura oracional del griego antiguo y el latín y, en particular, en los aspectos relacionados con la sintaxis y semántica de los constituyentes que la integran. Su enfoque teórico es la Gramática Funcional de S. Dik (1997), aplicada al latín por Pinkster (1984, 2015, 2021) y tanto al griego antiguo como al latín por el grupo de investigadores de este proyecto (Torrego 1998, 2007, Baños 2003, 2009, Jiménez 2021, entre otros). Además, esta aproximación se ha visto enriquecida con otros modelos teóricos afines, como la Gramática Cognitiva de Langacker (2008), y otros enfoques, como la tipología lingüística.

En los últimos años, el equipo ha estado trabajando en el desarrollo de dos bases de datos (BD). Las BD REGLA-Griego y BD REGLA-Latín tienen como objetivo último la obtención de un repertorio lo más completo posible de los marcos predicativos, esto es, los esquemas de complementación obligatoria, de los verbos más frecuentes en griego antiguo y latín. Están diseñadas para recoger, organizar y recuperar las apariciones de cada verbo en el corpus seleccionado con su correspondiente análisis sintáctico-semántico. No obstante, esta forma de organizar y almacenar los datos ha resultado no ser del todo efectiva, puesto que, entre otras cuestiones, no tiene en cuenta aspectos relativos al contexto en que las frases se encuentran. Así pues, se ha hecho necesario una revisión profunda del modelo de almacenamiento y gestión de la información.

Para ello, se han tomado en consideración otras herramientas y proyectos similares que se han y se siguen desarrollando para lenguas clásicas. Así pues, sin ánimo de ser exhaustivos, a

[157] Estudió Español y Filología Clásica en la Universidad Nacional de Colombia (2009) y Filología Clásica en la Universidad Autónoma de Madrid (2011), donde, además, obtuvo el título de Máster (2012). Entre 2013 y 2017 disfrutó de una beca predoctoral FPI-UAM. Su tesis, dirigida por la Dra. Esperanza Torrego, lleva por título «Nombres de Acción y otros derivados deverbativos en latín» y fue defendida en noviembre de 2018. Actualmente es personal investigador contratado del proyecto COMREGLA.

continuación se listan algunos de estos recursos: el proyecto *Perseus* (desarrollado entre Leipzig, Alemania, y Chicago, Estados Unidos), *LASLA* (Laboratoire d'Analyse Statistique des Langues Anciennes, Université de Liège, en Bélgica), *PROIEL* (Universitetet i Oslo), el *Index Thomisticus* (Università Cattolica del Sacro Cuore di Milano, Italia), *Ancient Greek and Latin Dependency Treebank* (Universität Leipzig, Alemania), *Homeric Dependency Lexicon* (Università degli Studi di Pavia, Italia) y, más recientemente, *LiLa* (Linking Latin, Università Cattolica del Sacro Cuore di Milano, Italia). Como cabe esperar, cada uno de estos proyectos tiene sus propios objetivos y métodos; con todo, parte de la información que sus herramientas analizan sobre los textos clásicos es coincidente.

Con el objetivo de convertir REGLA en un recurso compatible con algunas de las herramientas anteriormente citadas, el proyecto de investigación presentó en 2018 una solicitud para la obtención de una de las "Ayudas a Equipos de Investigación Científica" promovidas por la Fundación BBVA. Como reza la convocatoria, en el ámbito de las "Humanidades Digitales" las ayudas estaban destinadas a los proyectos de investigación orientados "al uso de las tecnologías de información y de técnicas estadísticas avanzadas para el tratamiento de objetos propios de las humanidades". El proyecto financiado por la Fundación BBVA se titula "Compatibilidad de la base de datos REGLA con otros recursos digitales (COMREGLA)". En la actualidad, cuenta con la participación de doce investigadores *iuniores* y *seniores* de diversas universidades (U. Autónoma de Madrid, U. Complutense de Madrid, U. Alcalá de Henares, U. de Salamanca, U. Cattolica del Sacro Cuore de Milán y la Pontificia Universidad Católica de Chile), además de dos personas contratadas a tiempo completo y un informático.

Este nuevo proyecto COMREGLA tiene un doble objetivo. En primer lugar, como se ha mencionado ya, pretende hacer compatible REGLA con otras herramientas informáticas existentes para el estudio del griego y del latín, lo que implica en primer lugar verter la información de las dos BD REGLA, esto es, las 45000 fichas de análisis de los verbos griegos y latinos, en la BD COMREGLA. Por otra parte, a diferencia de las anteriores BD, en esta nueva versión no se busca realizar el análisis sintáctico-semántico a partir de la ocurrencia de los verbos dentro de un corpus determinado, sino llevarlo a cabo mediante etiquetado de texto.

Para ello, a partir de los textos en griego y en latín que se obtienen del Proyecto Perseus, se ha creado una base de datos XML en la que los archivos contienen texto etiquetado con un modelo de marcado *stand off*, que permite un análisis en distintos niveles. No obstante, se han adaptado y redefinido características para ajustarnos mejor a nuestras propias necesidades.

En definitiva, con esta comunicación se pretende (i) presentar el proyecto REGLA; (ii) señalar cómo surge y a qué necesidades atiende el nuevo proyecto COMREGLA; y (iii) describir cuáles son las características (y novedades) fundamentales que presentan las nuevas BD con respecto a las BD relacionales previas.

Palabras clave: corpus lingüístico; base de datos; griego antiguo; latín

Bibliografía

Baños, J. M. (2009) (coord), *Sintaxis del latín clásico*, Madrid, Liceus.

Baños, J. M. *et al.* (2003) (ed.), *Praedicativa I. Complementación en griego y latín*, Santiago de Compostela, Universidad de Santiago de Compostela.

Dik, S. C. (1997), T*he Theory of Functional Grammar. Part I: The structure of the clause*, Berlin-New York, The Gruyter.

Jiménez, Mª D. (en prensa) (coord.), *Sintaxis del griego antiguo* (en prensa).

Langacker, R. W. (2008), *Cognitive Grammar: A Basic Introduction*, Oxford, Oxford University Press.

Pinkster, H. (1984), *Latijnse syntaxis en semantiek*, Amsterdam, B. R. Grüner Publishing.

Pinkster, H. (2015), *The Oxford Latin Syntax. Vol 1, The Simple Clause*, Oxford, Oxford University Press.

Pinkster, H. (2021), *The Oxford Latin Syntax. Vol 2, The Complex Sentence and Discourse*, Oxford, Oxford University Press (en prensa).

Torrego, Mª E. (1998) (ed.), *Nombres y funciones. Estudios de sintaxis griega y latina*, Madrid, Ediciones Clásicas-UAM.

Torrego, Mª E. *et al.* (2007) (ed.), *Praedicativa II. Esquemas de complementación verbal en griego antiguo y en latín*, Zaragoza, Universidad de Zaragoza.

Recursos electrónicos

Ancient Greek and Latin Dependency Treebank, disponible en https://perseusdl.github.io/treebank_data/ (última visita 17/06/2020).

Homeric Dependency Lexicon, disponible en https://studiumanistici.unipv.it/?pagina=news&id=12580 (última visita 17/06/2020).

Index Thomisticus Treebank, disponible en https://itreebank.marginalia.it/ (última visita 10/06/2020).

Laboratoire d'Analyse Statistique des Langues Anciennes, disponible en http://web.philo.ulg.ac.be/lasla/recherche/ (última visita 17/06/2020).

Linking Latin, disponible en https://lila-erc.eu/#page-top (última visita 17/06/2020).

Perseus under PhiloLogic, disponible en http://perseus.uchicago.edu/ (Chicago University, última visita 17/06/2020).

Praga Dependency Treebank, disponible en https://ufal.mff.cuni.cz/pdt3.0 (última visita 17/06/2020).

PROIEL Treebank, disponible en https://proiel.github.io/ (última visita 10/06/2020).

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Modelos de situación y características discursivas en la lectura de textos argumentativos

Gloria Esperanza Bernal Ramírez[158]
[gebernal@javeriana.edu.co]
Pontificia Universidad Javeriana, Colombia

Resumen

La investigación se centra en el análisis de los modelos de situación que construyen dos grupos de estudiantes universitarios a partir de la lectura de un texto científico argumentativo. Se exploran las relaciones identificadas entre los modelos de situación que emergen del análisis y las variaciones lingüísticas y discursivas introducidas en los textos. Se trata de un diseño mixto anidado, cuyo componente cuantitativo está orientado a definir los aportes a la comprensión que realizan dos estrategias de lectura diferentes: las preguntas metatextuales y la introducción de marcadores discursivos en los textos leídos. El componente cualitativo se centra en la identificac ión de los modelos de situación, en la construcción de una tipología de los mismos y en el análisis de los factores que influyen en la variabilidad de los modelos identificados. Toda la implementac ión se realiza en un ambiente computacional diseñado expresamente para esta investigación.

Palabras Clave: Modelos de situación, textos argumentativos, preguntas metatextuales, marcadores discursivos, ambiente computacional de herramientas web, estudiantes universitarios.

El problema

Una de las principales vías para proponer estrategias orientadas al mejoramiento de la comprensión de lectura corresponde a la indagación acerca de los modelos mentales o modelos de situación que construyen los lectores de un texto. Kintsch (1998) distingue entre base textual y modelo de situación, de acuerdo con el origen de las proposiciones en la representación mental del texto. El segundo corresponde a la estructura completa que está compuesta por las proposiciones derivadas del texto y las que aporta la memoria. Las variaciones y complejidades que pueda presentar este modelo surgen de la interacción entre las características del texto, los conocimientos previos del lector, la memoria, entre otros asuntos. Caracterizar los modelos de situación que construyen los estudiantes universitarios cuando leen un texto argumentativo breve permite realizar aportes al modelamiento de procesos lectores, al diseño de metodologías cuantitativas y cualitativas para realizar estos modelamientos y a intervenir adecuadamente en las pedagogías de la lectura de textos académicos. En consecuencia, la

[158] Licenciada en Filología e Idiomas (Universidad Nacional de Colombia), Magister en Tecnologías de la información aplicadas a la educación (Universidad Pedagógica Nacional) Profesora asistente Pontificia Universidad Javeriana Facultad de Educación. Coordinadora Línea Lenguajes, Trayectorias de Formación y Procesos socioeducativos Grupo In Novum Educatio.

investigación realizada se enfoca hacia la caracterización de estos modelos, surgidos en un ambiente computacional que propone dos estrategias diferentes para llegar a la construcción del modelo: la modificación del texto con marcadores argumentativos y la formulación de un conjunto de preguntas metatextuales. Desde una perspectiva cuantitat iva interesa identificar los efectos en la comprensión lectora -léase construir el modelo de situació n más cercano al que podría llamarse el ideal- de las estrategias aplicadas. Desde una perspectiva cualitativa, interesa establecer relaciones entre los modelos de situación que emergen el tipo de estrategia empleado. Todo lo anterior se apoya en una propuesta teórica acerca de los componentes que integrarían el modelo de situación de un texto argumentativo, de manera general. La ponencia que se propone se centra en la perspectiva cualitativa ya señalada, el modelo teórico propuesto y el trasfondo de aplicación, relativo al ambiente computacional diseñado para implementar las estrategias de lectura y hacer emerger los modelos de situación.

El marco teórico

Los modelos de situación son fundamentales en la explicación del procesamiento del lenguaje. Muchos autores demuestran que estos modelos tienen un papel importante en la llamada comprensión *online,* en la integración de información multimodal, en la manifestación de la experticia lectora, en la traducción y en el aprendizaje (Zwaan & Radvansky, (1998); van den Broek, Rapp, & Kendeou, (2005); van den Broek & Kendeou (2008)).

El mapeo de argumentos, por otra parte, corresponde a la línea de trabajo que permite identificar los modelos de situación que emergen de los procesos lectores. Corresponden en gran medida a los esquemas de argumentación que Walton define como patrones estereotípicos de razonamiento (Dwyer, Hogan, & Stewart (2012); van Gelder (2015); McCrudden, Gregory, Lehman, & Poliquin (2007); Walton (2016))

Metodología

Se aplica un diseño mixto anidado (Hernández Sampieri, Fernández Collado, & Baptista Lucio , 2010), conformado por un diseño cuasi experimental orientado a la medición de los niveles de lectura crítica (propia de la comprensión de textos argumentativos) antes y después de la implementación de las dos estrategias de lectura y un análisis de esquemas de argumentac ión producidos por los estudiantes. Todo se soporta en ACAELETA (Ambiente computacional para el aprendizaje de estrategias de lectura de textos argumentativos), estructurado a partir de la integración de herramientas web, que facilitan el trabajo de los estudiantes ya que vinculan Google Drive, correo electrónico y el software de modelamiento Mental Modeler, una aplicación libre, muy intuitiva y cuyo manejo lo aprenden muy rápido.

El diseño se apoya en un muestreo propositivo o no probabilístico: La muestra está conformado por dos grupos de estudiantes universitarios, que están inscritos en la asignatura Taller de Lectura, orientada por la misma maestra, para un total de 39 participantes, de diferentes programas académicos.

El uso de un diseño mixto en esta investigación se justifica en la medida en que su propósito es la caracterización y clasificación de los modelos de situación que sujetos reales construyen en una situación específica de lectura, en este caso, de un texto argumentativo muy breve. Esta característica propia de los modelos de situación, que podría prestarse para que el análisis tuviera un carácter subjetivo, se complementa con la medición de los niveles de lectura crítica antes y después de la implementación de la intervención pedagógica. De esta manera, se construyó con los estudiantes los parámetros básicos para la construcción del esquema (léase modelo de situació n para la investigación): uso de proposiciones completas, reconocimiento de hechos, antítesis, argumentos, tesis; y a la vez se aplicó estrategia diferente a cada uno de los dos grupos participantes para profundizar en su comprensión del texto y hacerla evidente en el esquema construido. El otro componente fundamental es el ambiente computacional construido para el desarrollo de la intervención: facilita el aprendizaje y posterior aplicación de la estrategia de lectura y anális is implementada con cada grupo; permite la construcción de los esquemas de argumentación bajo los parámetros planteados; es amigable y permite el registro completo de toda la información, esto es, de las actividades de análisis, representación y evaluación que desarrollan los estudiantes.

La información recolectada está compuesta por las respuestas al Cuestionario de Lectura Crítica 1 (inicial), al Cuestionario de Lectura Critica 2 (final) y a los Cuestionarios de Aplicación de Estrategia 1 (CAE1) y 2 (CAE2), y por los esquemas de Argumentacion elaborados por cada estudiante con el Mental Modeler.

Para desarrollar el análisis cualitativo se emplea el paradigma de codificación, que, dentro del enfoque de teoría fundamentada en los datos, proponen Strauss y Corbin (Soneira, 2006). Está compuesto por "distintos tipos de codificación (abierta, selectiva, axial) y por la aplicación de los métodos de comparación constante" (Soneira, 2006, pág. 160). En esta investigación, la codificación abierta corresponde al análisis dentro de los casos y la codificación axial y selectiva se emplea en el análisis entre los casos.

El análisis parte de las observaciones consignadas por la investigadora para cada uno de cuarenta esquemas elaborado por los estudiantes.

Resultados y discusión

Los resultados están conformados por el análisis estadístico, principalmente, descriptivo de la implementación de las estrategias y por el análisis descriptivo de los cuarenta esquemas de argumentación (léanse en adelante como modelos de situación) elaborados por los estudiantes. Dicho análisis concluye con la postulación de la tipología de modelos de situación. La ponencia propuesta se centra en el análisis de los modelos y la sustentación de la tipología. La discusión de resultados parte del análisis de los criterios que apoya la postulan el uso de los esquemas argumentativos como modelos de situación, las relaciones entre estos modelos y el tipo textual y el lugar de los conocimientos previos en los modelos (León, Escudero, & van den Broek (2003); Kendeou, Muis, & Fulton (2011); Diakidoy, Christodoulou, Floros, Iordanou, & Kargopoulos (2015))

Bibliografía

Diakidoy, Christodoulou, Floros, Iordanou, & Kargopoulos (2015) Forming a belief: The contribution of comprehension to the evaluation and persuasive impact of argumentative text. (T. B. Society, Ed.) *British Journal of Educational Psychology* (85), 300-315. doi:DOI:10.1111/bjep.12074

Dwyer, Hogan, & Stewart (2012) An evaluation of argument mapping as a method of enhancing critical thinking performance in e-learning environments. *Metacognition and Learning, 7*(3), 219-244. doi:10.1007/s11409-012-9092-1

Hernández Sampieri, Fernández Collado, & Baptista Lucio (2010) *Metodología de la investigación* (Quinta ed.). Mexico D. F.: McGraw-Hill/Interameriacana Editores

Kendeou, Muis, & Fulton (2011) Reader and text factors in reading comprehension processes. (U. K. Association, Ed.) *Journal of Research in Reading, 34*(4), 365-383. doi:10.1111/j.1467-9817.2010.01436.x

Kintsch (1998) *Comprehension: a paradigm for cognition.* Cambridge: Cambridge Uiversity Press.

León, Escudero, & van den Broek (2003) La influencia del género del texto en el establecimiento de inferencias elaborativas. En J. A. León, *Conocimiento y discurso: claves para inferir y comprender* (págs. 152-170). Madrid: Pirámide.

McCrudden, Gregory, Lehman, & Poliquin (2007) The effect of causal diagrams on text learning. *Contemporary Educational Psychology, 32*, 367-388. doi:10.1016/j.cedpsych.2005.11.002

van den Broek, Rapp, & Kendeou, (2005) Integrating Memory-Based and Constructionist Processes in Accounts of Reading Comprehension. *Discourse Processes, 39*(2-3), 299-316.

van den Broek & Kendeou (2008) Cognitive Processes in Comprehension of Science Texts: The Role of Co-Activation in Confronting Misconceptions. *Applied Cognitive Psychology, 22*, 335- 351. doi:10.1002/acp.1418

van Gelder (2015) Using argument mapping to improve critical thinking skills. En M. Davies, & R. Barnett, *The Palgrave Handbook Critical Thinking in Higher Education* (págs. 183-192). New York: Palgrave MacMillan US. doi:10.1057/9781137378057

Walton (2016)) A classification system for argumentation schemes. *Argument and Computation*, 1-27. doi:10.1080/19462166.2015.112377

Zwaan & Radvansky, (1998) Situation Models in Language Comprehension and Memory. *Psychological Bulletin, 123*(2), 162-185.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Multi-medial Corpora of Indigenous Languages from a Cultural Collections Perspective

Anne Ferger[159] **& Daniel Jettka**[160]
[anne.ferger@uni-hamburg.de | daniel.jettka@uni-hamburg.de]
University of Hamburg, Germany

Keywords: Information Visualization, Data Exploration, Multi-media Corpora, Metadata, Cultural Heritage, Digital Humanities

Introduction

The presented approach utilizes the visualization framework VIKUS Viewer[161] (VV, cf. Glinka *et al.*, 2017), which is originally intended for the exploration of cultural collections, for a thematic and temporal representation of the INEL Selkup Corpus 1.0[162], a linguistic corpus of an indigenous Northern-Eurasian language. Although linguistic corpora and Cultural Heritage data are created and compiled with different intentions, there can be considerable overlaps between the two, for instance in that both can have rich metadata and similar object types. A specific challenge is the integration of multiple object types into a single visualization. As the Selkup Corpus contains both manuscripts and sound objects, a supplemental media extension is implemented for VV that provides direct access to multi-page manuscripts and the auditive data. The approach aims at creating a reusable and sustainable visualization for multi-medial corpora with rich metadata. The availability of the used Selkup Corpus (CC BY-SA-NC 4.0), VV (MIT license)[163], and the integration of the necessary export facilities into the

[159] Anne Ferger is a Research Associate in the project INEL („Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages“) of the Universität Hamburg/Germany and currently mainly works on linguistic research data management. After her Bachelor and Master studies in Linguistics at Universität Göttingen/Germany and Universität Hamburg/Germany, she worked at The Hamburg Center for Language Corpora (HZSK) from 2016 to 2017.

[160] Daniel Jettka is a Research Associate in the INEL project at the Academy of Sciences in Hamburg and since 2016 mainly works on the development and optimization of methods and workflows for the creation and analysis of linguistic corpora for less-resourced languages. After finishing his Bachelor and Master studies in Linguistics, Text Technology, and Computational Linguistics at Universität Bielefeld/Germany and Trinity College Dublin/Ireland in 2012, he began to work in different research projects associated with the Hamburg Center for Language Corpora.

[161] VIKUS Viewer originates from the research project VIKUS (Visualisierung kultureller Sammlungen - Visualizing Cultural Collections) that was conducted between 2014 and 2017 at the University of Applied Sciences Potsdam whose goal was to investigate “new forms of graphical user interfaces to support the exploration of digital cultural heritage” (https://uclab.fh-potsdam.de/vikus/)

[162] http://hdl.handle.net/11022/0000-0007-E1D5-A

[163] https://github.com/cpietsch/vikus-viewer

open-source corpus processing framework Corpus Services[164] provide for a high degree of reproducibility.

Data and Software Basis

Currently, linguistic corpora for the languages Dolgan, Kamas, and Selkup have been published by the project INEL ("Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages")[165]. During its runtime (2016-2033), the project aims at publishing corpora for several more languages and varieties, which are or were spoken on the territory of the Russian Federation. The corpora follow a largely uniform design, that is based on Wagner-Nagy *et al.* (2018), a specific example of which can be seen in Arkhipov *et al.* (2019). For the prototypical visualization, the INEL Selkup Corpus 1.0 (Brykina *et al.*, 2020) is chosen, but the approach can be applied to all INEL corpora and supposedly many other linguistic corpora.

As sound objects are quite common in the INEL corpora, minor adaptations have to be made to VV, which was originally implemented with digitized physical objects in mind, e.g. paintings, drawings, and manuscripts (cf. Glinka *et al.*, 2017). For this purpose, the web-based interface was extended with a supplemental functionality which allows listening to the audio objects and viewing multi-page manuscripts directly in the visualization.[166]

The necessary transfer of data from the Selkup Corpus into VV is integrated in Corpus Services, a framework for quality control, conversion, publication, and visualization of language corpora (Hedeland/Ferger, 2020). This makes the export facility available and directly usable alongside numerous other methods for corpus creation, curation, and exploration in other contexts. It also allows the integration of the approach into existing automated corpus processing pipelines.

Automatically Visualizing a Corpus in VIKUS Viewer

The Selkup Corpus contains data of the language Selkup, an endangered Southern Samoyedic language of the Uralic family. It used to be spoken in various small settlements in Western Siberia. The data comprising the corpus are transcriptions of either (hand-written) manuscripts or recordings of speech as well as rich metadata.

The aforementioned Corpus Services software allows for a sustainable, re-usable and accessible way of generating a cultural collection visualization in Vikus Viewer. The VV software needs configuration files created for the respective resource as input. Therefore a

[164] Official releases at https://gitlab.rrz.uni-hamburg.de/hzsk-open-access/hzsk-corpus-services-release

[165] https://inel.corpora.uni-hamburg.de

[166] A link to the GitHub project will be revealed in the non-anonymous version of the contribution.

function was added to the Corpus Services to generate them automatically for INEL language corpora, which further enhances the reusability of the VV as well as of the corpus resources[167]. The exploration of manuscripts benefits from the rich metadata available in the corpus. The metadata can be mapped into the VV configuration files. For the Selkup Corpus a range of metadata information was selected to be displayed and used for filtering. Because of the temporal visualizations (see Figure 1) the time of the recording and creation of the manuscript is crucial. Keywords for filtering were generated from the time and region where the resource was created, its genre as well as the words of its title. Additionally, the dialect of the resource as well as the language code and country of creation were added for completeness. Since many of the resources are multipage manuscripts, links to access PDF files and their trans- criptions were added. For recordings, an audio player was added instead of the PDF source.

Figure 1: Global view of the INEL Selkup Corpus 1.0 in VIKUS Viewer

The created configuration files were then imported into a VV instance. Some challenges consisted in unclear time information of individual resources (encoded as e.g. “196X”), which are common in the work with archived data. Also using just the first page of a manuscript could be problematic, and there can also be multiple manuscripts for one resource, but providing a link to the complete resource should remedy that. Since recordings are displayed alongside manuscripts in VV, place-holder images for the former are inserted. VV contains a similarity function to sort manuscripts according to their visual similarity. This was also successfully applied to the manuscripts, yielding interesting results. Even though the similarity was not based on the hand-writing itself, the different textures of the paper they were written on still grouped manuscripts from the same author or time together.

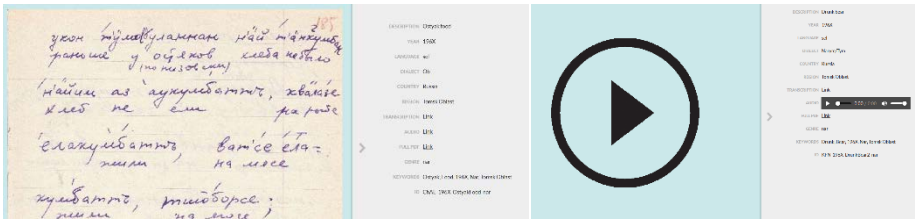


Figure 2: Detail views of a manuscript and a recording from the INEL Selkup Corpus 1.0 in VIKUS Viewer

[167] While this function was implemented for corpora in the INEL structure, a future adaption to other corpus formats is possible.

Summary

The presented approach demonstrates that the application of visualization methods from Cultural Heritage research, here in the form of the VIKUS Viewer, to linguistic corpora which have intersectional features with cultural collections is not only possible, but can provide interesting new perspectives and insights. While the extensive exploration of the resource in VIKUS Viewer is still to be discussed, it could be shown that also source types that go beyond visual artifacts (e.g. audio recordings in the case of the INEL Selkup Corpus 1.0) can be represented by adding an audio extension to the VIKUS Viewer. Thus, the incorporated thematic and temporal exploration and the visual similarity grouping functionality are complemented by direct access to multiple source types.

References

Arkhipov, Alexandre; Däbritz, Chris Lasse and Gusev, Valentin. 2019. *INEL Kamas corpus. User documentation.* Online: http://hdl.handle.net/11022/0000-0007-DE50-5 [18.06.2020]

Brykina, Maria; Orlova, Svetlana and Wagner-Nagy, Beáta. 2020. INEL Selkup Corpus. Version 1.0. Publication date 2020-06-30. Archived in Hamburger Zentrum für Sprachkorpora. http://hdl.handle.net/11022/0000-0007-E1D5-A. In: Wagner-Nagy, Beáta; Arkhipov, Alexandre; Ferger, Anne; Jettka, Daniel; Lehmberg, Timm (eds.). *The INEL corpora of indigenous Northern Eurasian languages.*

Glinka, Katrin; Pietsch, Christopher and Dörk, Marian. 2017. Von sammlungsspezifischen Visualisierungen zu nachnutzbaren Werkzeugen. In *Conference: DHd2017: Digital Humanities im deutschsprachigen Raum 2017*, Vol. 4, Bern. https://mariandoerk.de/papers/dhd2017_vikusviewer.pdf [18.06.2020]

Hedeland, Hanna and Ferger, Anne. 2020. Towards continuous quality control for spoken language corpora. In *International Journal of Digital Curation(IJDC).*

Wagner-Nagy, Beáta; Szeverényi, Sándor and Gusev, Valentin. 2018. User's Guide to Nganasan Spoken Language Corpus. In *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology*, 1.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Presencia femenina en el teatro de Lorca y de Unamuno. Una aportación desde análisis digital

Monika Dabrowska[168] **& Guadalupe Frutos Vilaplana**[169]

[monika.dabrowska@unir.net | guadalupe.frutos@unir.net]

Universidad Internacional de La Rioja, España

Resumen

El objetivo de esta propuesta es llevar a cabo un análisis comparativo de la presencia femenina y del papel de las protagonistas en el conflicto dramático en las obras escénicas de dos dramaturgos de la Edad de Plata española: Miguel de Unamuno y Federico García Lorca. En concreto, el trabajo se propone analizar la trilogía dramática de Lorca: *Yerma* (1934), *La casa de Bernarda Alba* (1936) y *Bodas de sangre* (1938) y dos piezas de Unamuno: *La esfinge* (1898) y *Fedra* (1924) con los métodos de análisis de redes y grafos, para delimitar la posición y peso de la mujer en el universo teatral de los dos dramaturgos modernistas, comprometidos culturalmente y artísticamente con la incipiente reconfiguración del rol femenino de la España de entresiglos.

Ambos dramaturgos pertenecen a la corriente renovadora de lo social y cultural, así como de la escena española de los inicios del siglo XX. Tanto al pensador bilbaíno como al poeta granadino les atañen de cerca los cambios sociales en la España del modernismo y es sus obras tejen una especie de espejo de normas culturales y morales de la época. En ambos casos puede hablarse, finalmente, de un planeamiento específico de lo femenino a través de la creación de personajes femeninos paradigmáticos.

[168] Doctora en Literatura por la Universidad de Varsovia. Profesora de la Facultad de Educación de Universidad Internacional de La Rioja. Forma parte del grupo de investigación Humanidades Digitales Aplicadas (HDAUNIR) y actualmente investiga dentro del proyecto HDATEATROUNIR. Ha sido docente en la Universidad CEU Cardenal Herrera en Valencia. Ha realizado estancia de investigación en la Universidad Veracruzana y estancia postdoctoral en la Universidad de Salamanca. Sus líneas de investigación se centran en la narrativa latinoamericana contemporánea, cultura digital, humanidades digitales, tecnologías aplicadas a la educación, estudios del cuerpo y de la mujer.

[169] Licenciada en Humanidades por la Universidad de Extremadura. Tiene formación práctica y teórica en el sector de la educación online, desempeñando diversas funciones y conocimientos: redactor, editor web (plataformas de edición), social media (conocimientos de SEO), usabilidad web, administrador de contenido web (marketing de contenidos), curación y redacción de contenidos. Actualmente trabaja como correctora-editora de contenidos académicos en el departamento de Contenidos en la Universidad Internacional de la Rioja. Personal investigador del grupo HDATEATROUNIR de la Universidad Internacional de la Rioja. Hoy en día, su investigación se centra en la aplicación de las redes sociales con grafos en el estudio del teatro español (Siglo de Oro y Edad de Plata).

La eminente presencia de personajes femeninos en el teatro lorquiano ha sido ampliamente estudiada. Sus obras ponen en escena la posición de la mujer en el contexto de la sociedad española, expresando la actitud crítica e ideales liberales del autor (Frazier 1983; Ingelmo 1994; Araújo, 2013). Varios críticos remarcan la sensibilidad de Lorca para captar las tensiones del alma femenina y la voluntad de propugnar su condición, manifestada en la amplia galería de mujeres que protagonizan sus obras escénicas, como Mariana Pineda, Bernarda Alba, Doña Rosita, Yerma, la zapatera (Degoy 1996; Nieva de la Paz 2008; Martinez Carro 2018). En cuanto a la obra unamuniana, la crítica literaria ha explorado más sus escritos ensayísticos y la narrativa que los textos dramáticos, aunque estos no quedan al margen de su visión artística (Díaz Freire 2017; Briki 2018); todo lo contrario, as figuras femeninas tanto de sus novelas: Gertrudis (*La tía Tula*), Eugenia (*Niebla*), Angela (*San Manuel Bueno, martir*), como de las obras teatrales expresan su visión de lo femenino (Sandoval Ullán, 2004)

En ambos escritores no resulta gratuita la forma dramática, por su capacidad de expresar la condición humana, sus pasiones, conflictos individuales y colectivos, la dimensión trágica de la existencia. Con el estudio contrastivo del perfil y desempeño en la escena de los personajes femeninos en sus respectivos mundos dramáticos se pretende responder las siguientes preguntas: ¿Cómo se configuran las relaciones de género en las respectivas obras dramáticas? ¿Cómo se relacionan los protagonistas de las obras con los demás *dramatis personae,* tanto masculinos como femeninos? ¿Qué perfil de personaje femenino se desprende de las analizadas? A base de los datos cuantitativos se extraen y analizan los patrones relacionados con los personajes femeninos en ambos autores.

Las obras que se examinan forman parte del corpus digital internacional de obras teatrales denominado Drama Corpora Project (DraCor), compuesto por once corpus diferentes, codificados en XML-TEI (obras de teatro clásico griego y romano, italianas, rusas, alemanas, suecas, inglesas etc). El *Spanish Drama Corpus*, al que pertenecen los textos de Unamuno y Lorca, es la colección de 25 textos teatrales de la Edad de Plata española procedentes de la Biblioteca Electrónica Textual del Teatro en Español 1868-1939 (BETTE) facilitada por el grupo GHEDI (Grupo de Humanidades y Edición Digital) de la Universidad Internacional de La Rioja (UNIR). Gracias a los metadatos como sexo (M/F), rol (protagonista, antagonista, amante y otro), importancia (principal, secundario o menor), personaje individual o intervención grupal es posible evaluar y visualizar valores cuantitativos textuales sobre cada personaje, su importancia dentro de la obra, densidad de las relaciones.

Además, Shiny DraCor posibilita el análisis de redes sociales de textos dramáticos y su visualización gráfica. Con el análisis digital mediante las redes sociales y grafos se pretende reconstruir la composición escénica y el peso de las intervenciones de las protagonistas, así como de los diálogos entre las mujeres en relación con las intervenciones masculinas, en las

piezas de ambos dramaturgos de la Edad de Plata española. Un ejemplo de visualización de estos datos muestran los siguientes gráficos:

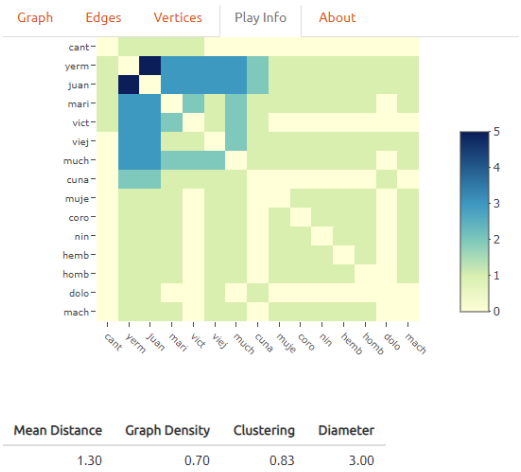


Figura 1: Relaciones entre los personajes de *Yerma* visualizados en *Shiny DraCor*

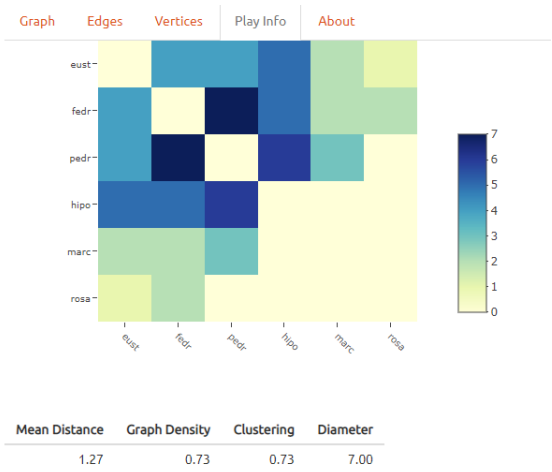


Figura 2: Relaciones entre los personajes de *La esfinge* visualizados en *Shiny DraCor*

La finalidad del análisis digital es corroborar hipótesis del protagonismo femenino y de la nueva sensibilidad de los dos dramaturgos ante la problemática femenina, en comparación con anteriores estudios textuales no computacionales. En una etapa posterior estos resultados se contrastarían con un estudio con los métodos de la lingüística de corpus, con herramienta Voyant Tools. Todo ello permite explorar las posibilidades del análisis digital (cuantitativo) contrastado con los estudios cualitativos existentes y visibilizar qué puede aportar la perspectiva digital a la filológica.

Palabras clave: Lorca, Unamuno, DraCor, Edad de Plata, teoría de las redes sociales, grafos

Bibliografía:

Briki, Maroua. (2018) La problemática de la identidad del ser en dos dramas de Unamuno: La esfinge y Sombras de sueño. Revista Científica Electrónica de Ciencias Humanas, núm 40 (año 14), 78-95.

Cerezo Galán, Pedro. (1996) Las máscaras de lo trágico. Filosofía y tragedia de Miguel de Unamuno. Trotta, Madrid.

Díaz Freire, J. J. (2017). Miguel de Unamuno: La feminización de la masculinidad moderna. Cuadernos de Historia Contemporánea, 39, 39-58.

DraCor: https://dracor.org/

Fischer F., Haaf, S., Hug, M. (2019) The Best of Three Worlds: Mutual Enhancement of Corpora of Dramatic Texts. CLARIN2019 Book of Astracts

Frazier, Brenda (1983), La mujer en el teatro de Federico García Lorca. Playor, Madrid.

GitHub Sites de Spanish Drama Corpus: https://github.com/dracor-org/spandracor

Ingelmo Fernández, Joaquín. (1994) La mujer en el teatro de Federico García Lorca. Badajoz.

Martínez Carro, Elena. (2018) Los personajes femeninos lorquianos desde una interpretación digital, en: Álvarez Ramos, Eva y Blasco Pascual, Javier (eds.). Humanidades Digitales. Retos, Recursos y Nuevas Propuestas, Agilice Digital, Valladolid

Nieva de la Paz, Pilar. (2008) Identidad femenina, maternidad y moral social: Yerma (1935), de Federico García Lorca. Anales de la Literatura Española Contemporanea (33/2), 155-176

Sandoval Ullán, Antonio. (2004) El concepto de mujer en el pensamiento de Miguel de Unamuno. Cuadernos de la Cátedra Miguel de Unamuno, (39), 27-60.

Tonra, J., Kelly, D., & Reid, L.A. (2017). Personæ: A Character-Visualisation Tool for Dramatic Texts. DH.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

The outsider art project: a transdisciplinary framework

John Roberto[170] **& Brian Davis**[171]
[john.roberto@adaptcentre.ie | brian.davis@adaptcentre.ie]
Dublin City University, Ireland

The outsider art project is a transdisciplinary research initiative that aims to deconstruct the definition of outsider art based on the computational analysis of texts and images. This is a transdisciplinary project because it fosters collaboration and knowledge exchange between artists, researchers, and citizens. The project goals involve the ontologization of the domain of Outsider Art, the computational analysis of the outsider art style of painting and the deployment of the outsider art web portal.

Capturing and codifying the knowledge of a specific domain is the first step toward understanding it. This is extremely important when dealing with socio-cultural issues, as in the case of Outsider Art, a particularly problematic and challenging domain. Thus, while traditional artistic styles (e.g. Cubism, Realism, Baroque or Abstract) are described on the basis of artistic criteria such as the use of the colour, shapes, space or techniques, non-traditional styles like Outsider Art are most frequently described on the basis of negative non-artistic criteria such as the mental condition of the artist. In other cases, where aesthetic criteria were used, these would lead to a negative assessment of the works of art. Paradoxically, in spite of this, "outsiders" are considered to be highly innovative artists and the visibility of outsider art has increased dramatically in recent years. Even more paradoxical is the fact that mainstream artists have found inspiration in the work of their marginalized peers. These conceptual inconsistencies clearly show that outsider art must be considered an extremely complex phenomenon in which different "levels of reality" are present simultaneously. Previous studies that have analysed the outsider art were not based on empirical evidence and they did not consider knowledge from each of the stakeholder groups and the complexity of the social issues involved in the domain. In many cases, researchers limited themselves to presenting their views of and concerns about outsider art. For example, David Davies (2009) proposed a theoretical characterization of the artistic status of Outsider Art on the basis of broader considerations regarding the philosophy of art. Linda Rainaldi (2015) later examined American

[170] Ph.D. in Theoretical, Computational and Applied Linguistics from University of Barcelona and Pompeu Fabra University on the topic of Polarity Analysis of Texts using discourse structure. He also completed a MA degree in Hispanic Linguistics at Instituto Caro y Cuervo. Currently, he is a post doctoral researcher at the ADAPT Centre in Dublin City University. His research focuses on Computational Linguistics for Cultural Heritage.

[171] Computer scientist and a computational linguist with core expertise in Ontology Engineering and Natural Language Processing. His interests also include Data Visualisation, Human-Computer Interaction and Natural Language Processing from social media. Dr. Davis is an Assistant Professor at the School of Computing, DCU and is affiliated with the ADAPT centre. Prior to taking up his appointment at Dublin City University, he lectured in Computer Science at Maynooth University.

and European perspectives on outsider art, focusing on biases, ideologies, and social factors, concluding that "I was no closer to articulating one comprehensive definition of outsider art." Therefore, from our point of view, in order to break through the limitations of previous studies, it is necessary to assume that outsider art is a coherent whole comprising several layers and different people involved in it have different perceptions of things. We are claiming that the outsider art cannot be properly understood in all of its complexity without using what is effectively a transdisciplinary framework that contributes to overcoming "the challenge to deconstruct the definition of Outsider Art holding just the intimate core of his meaning" (Acosta, 2015). Transdisciplinary research occurs when academics and nonacademics contribute their different expertise in order to understand a problem holistically by developing a common intellectual framework that goes beyond particular perspectives. According to Guimarães *et al.* (2019) "in a transdisciplinary research project, representatives of different disciplines, of the private and the public sectors, as well as of civil society, co- Page **2** of **2** produce knowledge on an issue". Practically speaking, transdisciplinary research involves different activities that integrate multiple scientific perspectives and the introduction of nondisciplinary knowledge from external stakeholders. In accordance with the above, the outsider art transdisciplinary project considers three well- defined activities. The **first** activity consists in developing a computational ontology of outsider art. Assuming that the existing literature on outsider art describes both the aesthetic and social issues surrounding this form of art, we will use ontology learning from texts as a methodological approach to represent part of the existing knowledge about outsider art in a machine-readable format. The **second** activity is the compilation of a big dataset of outsider art images. Currently, we are using the first version of this dataset of images to train CNN models to identify the stylistic features of outsider art paintings. The **third** activity deploys the outsider art web portal for connecting people and gathering information on outsider art. In this communication, we report on the current state of the project, which includes, among other things, its background, methodological framework, the state of the art in the semantic analysis of cultural assets and the principal results obtained to date.

Key words

Outsider Art, Ontology Development, Cultural Heritage, Machine Learning, Corpus Analysis, Computational Linguistics and Digital Humanities

Bibliography

Acosta, A. (2015). *A semantic analysis of the meaning of the word Outsider Art*. https://artslife.com/2015/07/23/a-semantic-analysis-of-the-meaning-of-the-word-outsider-art/ [Accessed 4 Jun. 2020]

Davies, D. (2009). *On the Very Idea of 'Outsider Art'*. The British Journal of Aesthetics. 49. 10.1093/aesthj/ayn056.

Guimarães, M.H., Pohl, C., Bina, O., & Varanda, M.P. (2019). *Who is doing inter- and transdisciplinary research, and why? An empirical study of motivations, attitudes, skills, and behaviours*. Futures, 112, 102441.

Harpring, P. (Ed.). (2019d). *The Getty Union List of Artist Names: Introduction and Overview*. Getty Vocabulary Program (pp. 1-143).

Le Boeuf, P., Doerr, M., Emil, Ch., and Stead, S. (Eds.). (2019). *Definition of the CIDOC Conceptual Reference Model version 6.2.7* (pp. 1–154). International Council of Museums.

Nicolescu B. (2018) *The Transdisciplinary Evolution of the University Condition for Sustainable Development*. In: Fam D., Neuhauser L., Gibbs P. (eds) Transdisciplinary Theory, Practice and Education. Springer, Cham.

Rainaldi, L. (2015). *Outsider art: forty years out (T).* University of British Columbia. Retrieved from https://open.library.ubc.ca/collections/ubctheses/24/items/1.0221495

Ramadier, T. (2004). *Transdisciplinarity and its challenges: The case of urban studies*. Futures, vol. 36, no. 4, pp. 423-439.

Schriml, L., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Félix, V., Jeng, L., Bearer, C., Lichenstein, R., Bisordi, K., Campion, N., Hyman, B., Kurland, D., Oates, C., Kibbey, S., Sreekumar, P., Le, C., Giglio, M. and Greene, C. (2019). *Human Disease Ontology 2018 update: classification, content and workflow expansion*. Nucleic Acids Research.

UNESCO. (2019). *UNESCO Thesaurus*. ISO 25964. IDENTIFIER http://vocabularies.unesco.org/thesaurus

Simposio de doctorandos

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

#BEAUTYGRAM

Rosmery-Ann Boegheholz
[rboegeholz@gmail.com]
Estudiante Doctorado en Comunicación: Universidad Austral de Chile,
Universidad de la Frontera, Chile
PhD Candidate: University of Groningen

Este proyecto de investigación busca encontrar la eventual relación entre el uso excesivo de las imágenes por parte de las *Influencers* de belleza a través de Instagram y las patologías psiquiátricas que pueden desarrollar sus seguidoras al no lograr lucir físicamente como ellas.

Para ello, es importante explorar, describir, analizar y explicar el impacto que las influenciadoras de belleza pueden generar al anunciar productos cosméticos en Instagram en la vida diaria de las jóvenes usuarias y si dicho impacto es capaz de generar algún tipo de patología psiquiátrica en las seguidoras.

Instagram se ha convertido en una parte fundamental de la vida de los jóvenes, especialmente de las jóvenes estudiantes. Esto ha abierto una nueva gama de estrategias de marketing para llegar a la juventud. Una de estas estrategias consiste en utilizar los llamados 'Influenciadores', es decir, personalidades populares online, para promover productos de belleza. Una de las consecuencias negativas del uso de influenciadores es que las jóvenes estudiantes pueden obsesionarse con parecerse a ellos. Para alcanzar este objetivo, incluso editan su apariencia física online, por ejemplo, retocando sus fotos de Instagram. Este tipo de conductas puede desencadenar diversas consecuencias psiquiátricas o emocionales, como: problemas de autoestima, autovaloración, trastornos obsesivos compulsivos, ansiedad, desórdenes alimenticios, alucinaciones, dismorfia corporal, entre otros.

Esta propuesta tiene como objetivo desarrollar una comprensión más profunda de los problemas y efectos derivados de un uso excesivo de imágenes físicas y personales en Instagram. Más específicamente examina cómo, al monitorear y utilizar esta plataforma, los *influencers* de belleza de Instagram pueden generar la aparición de las consecuencias ya mencionadas.

Marco Teórico:

Con la llegada de las nuevas tecnologías, lo que muchas personas hacen es buscar incansablemente su identidad a través de las Redes Sociales, y que este proceso es muy poderoso, tan poderoso que incluso pueden intentar, inconscientemente, atacarse a sí mismos. Al mismo tiempo, la gente busca la aprobación de otros usuarios conectados a la red. En relación con ello y, según Didi-Huberman (2009), a través de los dispositivos electrónicos, que implican el uso de aplicaciones, como las redes sociales, se ejerce un poder que sobreexpone tanto el vacío como la indiferencia, convirtiéndolo en mercancía para los espectadores.

Posiblemente, unos diez años después, la costumbre de las nuevas generaciones de exponer continuamente sus vidas y acontecimientos, les hace pensar que es normal y apropiado que las personas comuniquen incluso sus pensamientos más íntimos en la red, sin dejar espacio a las relaciones humanas de carne y hueso o sentimentales.

La red social llamada 'Instagram' es una aplicación que puede ser instalada en dispositivos móviles. Esto permite compartir y capturar fotos y videos cortos para ser publicados públicamente o sólo para los seguidores de los usuarios (Jimoh y Halima, 2017); suponiendo que cuantos más seguidores tenga una persona, más popular será. Como explicaron Jang, Han, Shih y Lee (2015), muchos adolescentes y adultos jóvenes tienden a manipular el contenido de sus fotos para recibir tantos *likes* o corazones como sea posible. Esto puede deberse a que la atención generada por el número de corazones obtenidos se ha convertido en una forma de establecer la autoestima y la autovaloración, así como de sentirse socialmente aceptados. La mayoría de las veces, los usuarios tienen un aspecto muy diferente de su apariencia real que cuando ven sus fotos publicadas.

En el contexto de esta situación, se puede decir que la belleza física está ligada a la sociedad de consumo en la que vive, "los cánones estéticos impuestos por la cultura occidental de la imagen, son para las personas una batalla contra el tiempo para ser socialmente aceptadas, produciendo problemas de inseguridad y de no aceptación del propio cuerpo" (Sossa, 2011, p. 9). Por eso hay usuarios que dedican una parte importante de su tiempo a editar las fotografías que se suben a la red. Esto parece significar que cumplen con los estándares de belleza establecidos, lo que potencialmente puede aumentar su número de seguidores. Junto con esto, sienten que necesitan aumentar su número de *likes* y popularidad, de lo contrario se sienten como si fueran castigados o rechazados por la sociedad.

Metodología:

Este proyecto utilizará un enfoque de método mixto. El principal beneficio de utilizar una metodología mixta es que ambos métodos, cualitativo y cuantitativo, pueden ser un complemento, ya que en caso de que uno de ellos deje algunas lagunas, éstas pueden ser llenadas por el otro método (Bryman, 2012).

La población que se estudiará en este proyecto son mujeres jóvenes adultas estudiantes de instituciones de enseñanza superior que son usuarias habituales de los sitios de redes sociales, más concretamente, usuarias de Instagram.

Referencias:

Bryman, A. (2012). *Social Research Methods*. Oxford University Press.

Didi-Huberman, G. (2009). *La Supervivencia de las Luciérnagas*. Abada Editores S.L.

Jang, J. Y., Han, K., Shih, P. C., & Lee, D. (2015). Generation like: Comparative characteristics in instagram. In *CHI 2015 - Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems: Crossings* (pp. 4039-4042).

Jimoh, I. y Halima, S. (2017)., Gender Analysis of Instagram Uses in Nigeria. *Plasu Journal of Communication Studies, 1*(1), 55-75.

Sossa, A., (2011). Análisis desde Michael Foucault referentes al cuerpo, la belleza física y el consumo. *POLIS, Revista Latinoamericana, 10*(28), 1-16.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

CG-Art. Una discusión estética sobre la relación entre creatividad artística y computación[172]

Leonardo Salvador Arriagada Beltrán

[l.s.arriagada.beltran@rug.nl]

PhD candidate in Humanities – University of Groningen

Dr. (c) en Filosofía con mención en Estética y Teoría del Arte – Universidad de Chile

Resumen

1. Introducción

En era de la inteligencia artificial (IA) no han sido pocos los que se han preguntado si una máquina puede crear arte. En este sentido, la investigadora cognitiva Margaret Boden (2011) ha definido un tipo especial de arte al relacionar los conceptos "creatividad" y "computación". Así, el arte generado por computador[173] (computer-generated art) es "the artwork results from some computer program being left to run by itself, with minimal or zero interference from a human being" (p. 141). Uno de los casos más populares de este tipo de arte es el cuadro Portrait of Edmond de Belamy (2018), ampliamente publicitado como la primera obra de arte de una IA en ser subastada en Christie's (Still & d´Inverno, 2019).

Que las máquinas creen arte no ha dejado indiferente al mundo artístico. En efecto, se ha tendido a criticar al CG-art porque carecería de la inspiración mística que tiene el arte humano. Como menciona Aaron Hertzmann (2018), científico principal de Adobe Research San Francisco, "it is as if only humans create art because only humans have 'souls.' Surely, there should be a more scientific explanation" (p. 1). Atendiendo a lo anterior, sostengo que el CG-art tiene un valor estético descifrable a través de dos aproximaciones, a saber, evaluación humana y maquinal (Arriagada, 2020).

2. Valor estético del CG-art

La evaluación humana del valor estético del CG-art dice relación con la recepción que el público humano tiene de las obras de arte creadas por máquinas. Investigaciones recientes han mostrado que existe un sesgo negativo hacia el CG-art (Chamberlain, Mullin, Scheerlinck, & Wagemas, 2018). Sin embargo, también se han efectuado estudios ciegos, es decir, se ha evaluado la apreciación que el público humano tiene de obras de arte creadas por humanos y por computadores, sin indicar a los observadores quién es el autor de dichas obras (Elgammal,

[172] Este resumen forma parte de una investigación doctoral del mismo título. Una versión extendida de estos tópicos ha sido publicada por la editorial Taylor & Francis en su revista *Connection Science* en el artículo "CG-Art: Demystifying the Anthropocentric Bias of Artistic Creativity" (Arriagada, 2020), cuyo copyright es dominio de Informa UK Limited.

[173] De ahora en adelante CG-art por sus siglas en inglés.

Liu, Elhoseiny, & Mazzone, 2017). En este último caso, el público humano ha asignado, sin saberlo, un puntaje superior al CG-art y menor al arte humano. En particular, el CG-art obtiene un mejor desempeño en todos los indicadores cualitativos de intencionalidad, estructura visual, comunicación e inspiración.

Por otro lado, la evaluación maquinal del valor estético del CG-art dice relación con la ponderación que una máquina hace de su propio trabajo. Este tipo de evaluaciones se encuentran íntimamente relacionadas con las redes neuronales antagónicas[174] (Generative Adversarial Networks). Estas son los algoritmos más vanguardistas a la hora de utilizar IA en todo ámbito, y el arte no es la excepción. Por ello es útil entender su funcionamiento. En particular, si queremos que una GAN cree imágenes, el proceso es el siguiente:

> Generative Adversarial Network (GAN) has two sub networks, a generator and a discriminator. The discriminator has access to a set of images (training images). The discriminator tries to discriminate between "real" images (from the training set) and "fake" images generated by the generator. The generator tries to generate images similar to the training set without seeing these images. The generator starts by generating random images and receives a signal from the discriminator whether the discriminator finds them real or fake. (Elgammal, Liu, Elhoseiny, & Mazzone, 2017, pág. 5).

Notamos entonces que una GAN incorpora una evaluación estética de su propio trabajo. De hecho, solo considera que sus creaciones están terminadas cuando su generador logra engañar a su discriminador. Entre los ejemplos populares de CG-art producido por GANs encontramos el ya referido cuadro *Portrait of Edmond de Belamy* (2018), la instalación de video *Mosaic Virus* (2019), o los álbumes musicales *PROTO* (2019) y *Hello World* (2019).

En definitiva, sostengo que los humanos sí apreciamos estéticamente el CG-art, de hecho, ante pruebas ciegas nos parece más valioso que el mismo arte humano. Por su parte, el mismo funcionamiento de las GANs hace que sus trabajos artísticos sean sometidos a constantes evaluaciones estéticas.

3. Otras críticas al CG-art

Hertzmann (2018) ha señalado que las máquinas no son "agentes sociales" incapaces de crear arte, pues el arte es social. Considero esto está errado, pues que el arte humano sea social no implica que el CG-art también deba serlo (Arriagada, 2020). Estamos hablando de creaciones de máquinas y algoritmos que tienen formas de conocer y experienciar distintas a las nuestras. Por lo demás, el CG-art se basa en *datasets* de entrenamiento, y se enriquece de la información provista por el *Big Data*. Bien se podría argumentar que el CG-art es capaz de procesar patrones sociales que para nosotros son imperceptibles. Además, la materialización del CG-art en forma de cuadros o álbumes musicales ha generado en el público humano un debate sobre su naturaleza. Lo anterior hace difícil sostener que los algoritmos no son agentes sociales.

[174] De ahora en adelante GAN por sus siglas en inglés.

Finalmente, muchos postulan que las máquinas no crean arte, sino que más bien son herramientas al servicio de programadores y artistas humanos (Hertzmann, 2018). En mi opinión esto es incorrecto por dos razones. En primer lugar, el trabajo artístico humano es social e involucra muchos agentes (Arriagada, 2020). Por ejemplo, al filmar una película existe el trabajo de dirección, de actuación, de filmación, etc. Cada uno de los agentes que participa de dicho trabajo, está contribuyendo artísticamente en su respectivo agenciamiento. La película no es patrimonio exclusivo del director, de los actores, o de los camarógrafos. El CG-art encaja perfectamente dentro de las creaciones artísticas colectivas. En los ya mencionados álbumes musicales *Hello World* y *PROTO*, los artistas humanos declaran haber trabajado en colaboración con la IA, lo que dista de la visión de los algoritmos como herramientas al servicio humano.

En segundo lugar, propongo que los algoritmos creativos son a sus programadores, como un aprendiz humano es a su maestro humano (Arriagada, 2020). Todo artista humano tiene un maestro que le provee de un necesario aprendizaje previo a su etapa creativa. Este rol de maestro puede ser desempeñado por otro humano, por sus experiencias de vida, etc. Por su parte, como vimos al revisar el funcionamiento de las GANs, el CG-art también se nutre de un aprendizaje previo, ya sea este atribuible a sus programadores o a su propia experiencia de iterados ensayos y errores. Lo fundamental acá es que los artistas humanos no ceden la autoría de sus creaciones a sus maestros y, por lo tanto, las máquinas tampoco deberían ceder la autoría del CG-art a sus programadores.

4. Conclusiones

El examen del valor estético del CG-art permite comprobar que la mayoría de sus críticas provienen de un sesgo negativo que desaparece en pruebas ciegas. Por lo demás, se constató que la incorporación de las GANs en el arte asegura una evaluación estética maquinal a la que los algoritmos someten sus creaciones. Finalmente, se enfatizó el rol colaborativo que tiene el CG-art, y se ofrecieron interpretaciones que permiten considerar a las máquinas autores de sus obras, y no meras herramientas al servicio humano.

Referencias

Arriagada, L. (2020). CG-Art: Demystifying the Anthropocentric Bias of Artistic Creativity. *Connection Science*. doi:10.1080/09540091.2020.1741514

Boden, M. (2011). *Creativity and Art: Three Roads to Surprise.* Nueva York: Oxford University Press.

Chamberlain, R., Mullin, C., Scheerlinck, B., & Wagemas, J. (2018). Putting the Art in Artificial: Aesthetic Responses to Computer-Generated Art. *Psychology of Aesthetics Creativity and the Arts, 12*(2), 177-192. doi:10.1037/aca0000136

Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (June de 2017). CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from

Style Norms. *Eighth International Conference on Computational Creativity.* Atlanta. Recuperado el 23 de Junio de 2017, de https://arxiv.org/abs/1706.07068v1

Hertzmann, A. (2018, May 10). Can Computers Create Art? *Arts, 7*(2), 18. doi:10.3390/arts7020018

Still, A., & d´Inverno, M. (2019, March). Can Machines Be Artists? A Deweyan Response in Theory and Practice. *Arts, 8*(1), 36. doi:10.3390/arts8010036

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Develando retóricas culturales represivas en el teatro de José Ricardo Morales: Una conexión con el análisis de los periódicos digitales en Chile

Juan Alfredo del Valle Rojas
[j.a.del.valle.rojas@rug.nl]
Candidato a PhD, Graduate School for the Humanities
Universidad de Groningen, Países Bajos

Resumen

Durante las últimas décadas, Chile ha sido afectado por diversos conflictos socio-políticos y económicos, tales como dictaduras, rebeliones, y también la irrupción del neoliberalismo. En este contexto, se hace necesario analizar la contingencia de Chile, especialmente ver cómo se ha ido gestando una democracia autoritaria que ha llevado a generar una gran ola de protestas y movilizaciones sociales que tiene su punto más álgido en el estallido social de octubre del 2019. Para reflexionar sobre las protestas sociales, la literatura en Chile juega un rol primordial, ya que en muchas de ellas se problematizan los conflictos sociales desde una perspectiva artística, cuyo fin es lograr que el espectador/lector reflexione sobre eventos y fenómenos socio-culturales puntuales. En especial, mi interés en este estudio se centra en el teatro en Chile, el cual ha sido utilizado como una herramienta para la transformación social (gobiernos de la Unidad Popular: Pedro Aguirre Cerda, 1938-1941 / Salvador Allende, 1970-1973) que busca crear conciencia en sus espectadores. En este contexto del teatro, emerge la figura del dramaturgo español-chileno José Ricardo Morales y su teatro vanguardista que contiene discurso predictivo que contribuye a reflexionar sobre diversos conflictos sociales en Chile, tales como el régimen militar, la manipulación mediática y las protestas sociales.

En su vasta obra, Morales posee dos obras de teatro que problematizan el poder, el autoritarismo y la manipulación de los medios de comunicación de forma muy interesante, la cual cobra vigencia si lo conectamos con lo que se demuestra actualmente en los medios de comunicación en Chile, en especial en los medios digitales. Por tanto, las obra *Cómo el poder de las noticias nos da noticias del poder* (1971) nos permite comprender el contexto chileno en cuanto a la manipulación mediática de los años '60 y '70 y contexto actual. Por otra parte, su obra *Los culpables* (1964) nos permite comprender el contexto chileno en cuanto a la emergencia de regímenes autoritarios de los años '60 y el contexto actual. Mi objetivo es analizar las retóricas culturales represivas que emergen en estas obras señalas anteriormente. Estas narrativas culturales se refieren al uso de argumento y narrativas literarias del lenguaje para enfatizar demostrar, y/o denunciar conflictos sociopolíticos y económicos en una obra literaria (Albaladejo, 2009, 2013; Mailloux, 2006). Por tanto, en este estudio me enfocaré en analizar las retóricas de la manipulación mediática en regímenes autoritarios representado en

la obra *Cómo el poder de las noticias nos da noticias del poder* (1971) y el autoritarismo político-militar representado en la obra *Los culpables* (1964) de José Ricardo Morales. Para ello, mi estudio busca responder a la siguiente pregunta: ¿Cómo las retóricas culturales represivas develadas en las obras de Morales *Cómo el poder de las noticias nos da noticias del poder* (1971) y *Los culpables* (1964) cobran vigencia al vincularse con el reciente retrato noticioso de éstas en los medios digitales de comunicación en Chile? El estudio de caso de las obras de Morales me permitirá explorar en profundidad desde múltiples perspectivas dichas obras de teatro en un contexto sociopolítico específico. A su vez, el análisis crítico discursivo de las retóricas culturales que manipulan estructuras políticas que producen conocimiento acerca de la manipulación mediática y el autoritarismo político-militar permiten involucrar a lectores con la experiencia de los contextos de las obras y retratar una mirada profunda de ellas, desde un marco histórico e interptretativo.

La obra de Morales *Cómo el poder de las noticias nos da noticias del poder* (1971) crítica a la industria de los medios de Comunicación en los años '60 y '70, en específico, en cómo la entrevista es utilizada como dispositivo de distracción en la televisión y el show business (información irrelevante y falsa a la audiencia). En esta obra, Morales se enfoca en el rol del personaje del Periodista, quien manipula las respuestas de la modelo, la cual fue invitada a su programa de televisión. Este periodista utiliza así dispositivos retóricos específicos para interpretar y conducir la entrevista, utilizando sus respuestas para reducir la presencia femenina de la modelo a la posición de un animal. En general, Morales problematiza en esta obra sobre el poder de corporaciones detrás de medios de comunicación, cuyos fines políticos y/o ideológicos buscan manipular a la audiencia. Este análisis de la obra de Morales se puede conectar con la actual cosificación de la mujer en televisión, en donde los medios de comunicación digitales han difundido mensajes en apoyo a las mujeres, como lo es el hashtag "no soy un objeto" que busca erradicar la violencia contra la mujer, abordando la cosificación del género, las etiquetas y los estereotipos que se utilizan de ellas en la sociedad. También, en la actualidad, los medios de comunicación digitales han criticado a los programas de entrevista por solamente referirse a la situación sentimental de las modelos invitadas y no de su carrera o profesión.

La obra de Morales *Los culpables* (1964) nos crítica el poder y el autoritarismo, los que problematiza en el contexto de un régimen militar. En la primera parte de la obra el personaje del Coronel en Jefe disemina información falsa para convencer a sus ciudadanos que un grupo de insurgentes liderado por el personaje de María Garcés planea un ataque terrorista. Morales se enfoca en los discursos políticos del miedo utilizado por regímenes/democracias autoritarias que dividen (bien y mal) y legitiman negación y exclusión: grupos peligrosos amenazan la población (Wodak, 2015). Además, Morales se enfoca en la diseminación de propaganda política que consiste en uso de sistema comunicativo para manipular opiniones de ciudadanos y grupos insurgentes (Bussemer, 2008), la cual utiliza para inventar unos supuestos terroristas (enemigos internos) que buscan dañar el país, con el propósito de controlar a la población. Este análisis de la obra de Morales se puede conectar con el actual montaje político por parte

del Estado chileno en contra de grupos anarquistas en Chile, el cual ha resultado ser un caso emblemático en donde el estado ha tenido que reconocer que no eran culpables. Por tanto, todos los anarquistas que supuestamente pusieron artefactos explosivos en lugares públicos de Santiago fueron absueltos.

Para finalizar, puedo señalar que este texto buscaba demostrar tres aspectos relevantes de mi investigación doctoral que se pueden relacionar con la actualidad en Chile. La vinculación de retóricas culturales represivas develadas en las obras de Morales (manipulación mediática y autoritarismo) con lo representado en los medios digitales de comunicación. La obra de teatro *Cómo el poder de las noticias nos da noticias del poder* (1971) devela los mecanismo de la entrevista como un dispositivo periodístico con estructura de preguntas y objetivos determinados que buscan distraer a la audiencia (Nash, 2016). Por su parte, la obra de teatro *Los culpables* (1964) devela los mecanismos de la propaganda como un instrumento efectivo para crear el enemigo interno, con el fin de justificar represión de movimientos sociales y controlar a la ciudadanía. Finalmente, se puede señalar que, de acuerdo a lo preentado, se presentan conexiones que se pueden entender como analogías históricas entre las obras de Morales y los eventos actuales en Chile, tales actual las movilizaciones sociales (18-Octubre) y el control de los medios digitales de comunicación por dos conglomerados: *COPESA* y *El Mercurio*.

Referencias

Albaladejo, T. (2009). La poliacroasis en la representación literaria: un componente de la Retórica cultural. *Castilla. Estudios de Literatura*, 1–26. https://doi.org/10.24197/cel.0.2009.1-26

Mailloux, S. (2006). *Disciplinary Identities: Rhetorical Paths of English, Speech, and Composition*. Modern Language Association.

Morales, J. R. (1964). Los culpables. *Anales de la Universidad de Chile*, 131(4), 65–92. https://doi.org/10.5354/0717-8883.2012.22772

Morales, J. R. (1971). *Un marciano sin objeto / Cómo el poder de las noticias nos da noticias del poder*. Editorial Universitaria. http://www.memoriachilena.gob.cl/602/w3-article-86859.html

Wodak, R. (2015). *The Politics of Fear: What Right-Wing Populist Discourses Mean*. Sage.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

El papel de la minería de datos en el rastreo de la noción de otredad en Twitter a propósito del post-acuerdo colombiano. Avance de investigación

Elizabeth Pinilla Duarte
[e.pinilla.duarte@rug.nl]
PhD Candidate: University of Groningen

El contexto nacional colombiano, después de cuatro años de la firma del acuerdo de paz, no pasa por su mejor momento. Por una parte, se ha denunciado que el gobierno de Iván Duque ha dejado de cumplir los puntos acordados en el documento. De hecho, el partido político al cual pertenece (Centro Democrático) presentó el 30 de octubre de 2020 el borrador de un decreto para para derogar la JEP (Justicia Especial para la Paz[175]) (El Tiempo, nov 1 de 2020). Por otra parte, en el interior del país ha aumentado la lucha entre grupos armados ilegales que se disputan el control del territorio para poder ejercer actividades como narcotráfico, minería ilegal, deforestación y despojo de tierras (Forero, 2020). Aparte del continuo desplazamiento forzado de poblaciones, otra de las consecuencias de esta lucha, entre muchas otras, es la persecución y asesinato de líderes sociales. Entre ellos, los más afectados son líderes comunales que, después de la firma del acuerdo, decidieron unirse a juntas de acción comunal desde las que se realizan actividades para erradicar las actividades ilegales ya mencionadas. De acuerdo con un informe especial de la revista Semana realizado en 2019, más de siete millones de personas están afiliadas a estas juntas, por lo que el escenario sobre su seguridad "resulta preocupante."

Mientras las cifras sobre los asesinatos de líderes no cesan, el gobierno se muestra reacio a admitir estos hechos y, por el contrario, afirma que estos crímenes disminuyen (Forero, 2020). En este contexto, la opinión pública en redes sociales se debate entre la denuncia de los asesinatos y la importancia de cumplir el acuerdo de paz, o entre la disconformidad frente a un acuerdo que muchos consideran ilegal y la justificación de los crímenes porque se cree que los líderes sociales no son tal, sino guerrilleros. En Colombia, el guerrillero ha sido declarado como el principal enemigo del país, tanto por actores del gobierno como por los medios de comunicación (Mesa, 2018). Y la protesta social se ha vinculado, precisamente, a la imagen que del guerrillero se tiene en el país. Ante este panorama, se teje en el imaginario social una noción sobre la diferencia y la otredad desde la perspectiva del juicio o de la indiferencia. Por ello, mi pregunta de investigación es ¿En qué medida las redes sociales como Twitter participan en la construcción de la noción de otredad en el marco del post-acuerdo colombiano? Por un

[175] La JEP es el organismo encargado de llevar cabo el proceso de justicia transicional con los exguerrilleros de las FARC, así como de juzgar a militares y otros actores que participaron en el conflicto interno del país.

lado, parto de la premisa según la cual para construir un estado de paz es vital observar los mecanismos culturales-narrativos que producen la *Otredad*. Como Quintero y Marín refieren en su investigación (2018), la noción e incorpación de la noción de otredad y cómo se ajusta esta al imaginario colectivo sobre el futuro es vital para la reconciliación en un país que atraviesa por un post-acuerdo de paz. Por otro, las redes sociales y, en particular, Twitter han sido considerados como lugares idóneos para analizar la opinión pública (Quintero y Marín, 2018), por ser posibles facilitadores para la democracia (Toepfl y Piwoni, 2015), por ser espacios que podrían contribuir a la construccion del sentido del mundo (Lee y Yang, 2010), por la manera en que influencian la opinion política (Cacciatore *et al.*, 2018) y, en el caso específico de Colombia, por cómo generan polarización (Barrios, Vega y Gil, 2019).

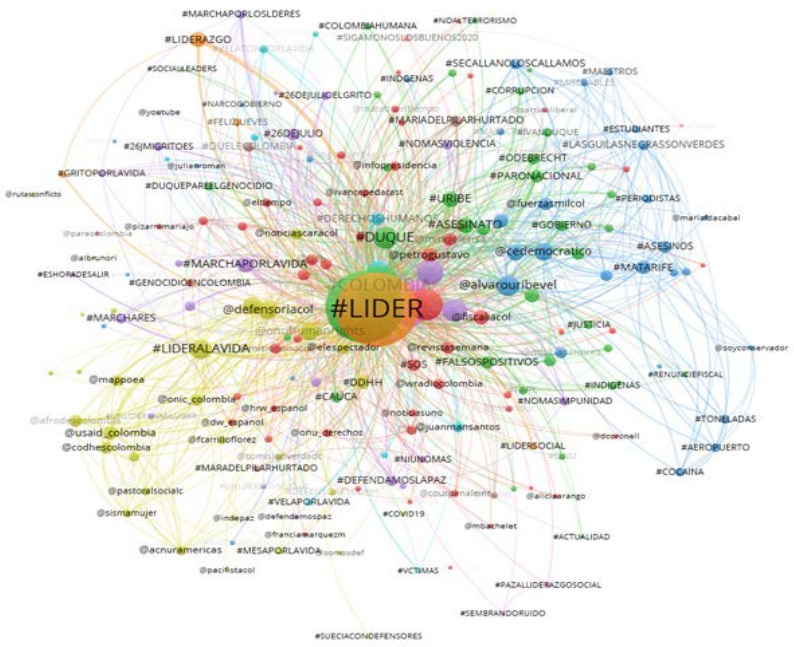
La pregunta de investigación que propongo aquí nació de una lectura previa que realicé sobre una conversación en Facebook acerca de una noticia publicada por El Espectador el 5 de julio de 2018: ¿Qué hay detrás del asesinato de líderes sociales en el país? (Osorio, 2018). La conversación, como muchas otras que ya había visto en diferentes redes sociales, era la representación de un país polarizado, en la que se acusaban los unos a los otros de guerrilleros, paramilitares e ignorantes. El uso de lo que se ha denominado lenguaje del odio resaltaba a lo largo de la conversación, así como la directa acusación hacia los líderes como guerrilleros, es decir, como asesinos, según el imaginario público. Sin embargo, mas allá de eso, pude notar el uso frecuente de metáforas como recursos discursivos para conceptualizar al interlocutor, pero también el tiempo del conflicto y el futuro. Específicamente, se usaban continuamente metáforas primarias de movimiento. Un ejemplo de diferentes extractos a continuación:

> Yo **no me voy a matar la cabeza** discutiendo con usted pierdo mi tiempo.
>
> **aquí no vengas** a defender a nadie que le ha echo daño a todo el país por todos los flancos
>
> **mi país jamás va a tener** un desarrollo el troll sos voz esa es mi opinión amo a mi país.
>
> **sigale haciendo caso al troll mentiroso.**
>
> **vaya a averiguar** la vida de estos líderes.
>
> **vaya a cocinar** no pasa nada.
>
> **Me quedo con los paracos.**

Las metáforas de movimiento fueron propuestas por Lakoff y Johnson (1980) como metáforas conceptuales primarias que indican cambio de estado. Debido a su incorporación en el lenguaje cotidiano, también han sido vistas como muertas por no proveer más un traslado conceptual importante, novedad o extensión semántica. Sin embargo, como Eva Kittay apunta, "[…] With age and use a metaphor may indeed die 20 or become standard. However, given the proper context, the metaphor may be revived" (1987, p. 299). Debido a que las redes sociales son lugares virtuales de encuentro que implican una descorporalización, mi hipótesis es que estas metáforas son mecanismos que contribuyen en la creación del sentido de movimiento y de tiempo de una manera renovada en estos contextos. En los ejemplos mostrados, los

participantes usan metáforas de movimiento, pero de manera negativa, por lo que no son indicadoras de cambio, sino de, contrariamente, no-cambio. Esta es una conceptualización que colinda con lo que María Teresa Uribe (2004) denomina como proyecto retórico nacional de un estado perpetuo de guerra. De no cambio. Y en ese contexto es que los participantes de la conversación parecían estar construyendo la noción de líder social como un *otro* que es héroe, víctima o villano de un pasado, un presente y un futuro violento.

Después de un primer análisis de la mencionada conversación en Facebook, mi pregunta es si dicha narrativización del otro y del futuro podría ser representativa de un contexto mayor o no. ¿Cómo hacer un rastreo más transparente sobre ello? Como lo dije anteriormente, Twitter es un lugar clave para el análisis de la opinión pública (De Stefano y Santelli, 2019). Por ello, el campo investigativo de las humanidades digitales se presenta como lugar de partida para el diseño metodológico de mi investigación, ya que desde allí se ha trabajado con minería de datos o análisis textual asistido con máquinas desde hace algunos años. De esta manera puedo diseñar un estudio mixto en el que el analisis cuantitativo de datos me puede llevar a realizar una lectura hermenéutica de las conexiones conceptuales del imaginario público sobre la otredad (en este caso, representada por la noción de líderes sociales). En esta vía, e inspirada por Hellsten & Laydeesdoff (2020), combino el análisis de redes sociales y el análisis de redes semánticas de dos hashtags: #LideresSociales (con un corpus de 10.207 tweets desde diciembre de 2016 hasta septiembre de 2020) y #NosEstanMatando (con un corpus de 85.000 tweets desde julio de 2018 hasta septiembre de 2020). En un primer momento, proyecto el rastreo automático de cuentas y hastaghs etiquetados en los comentarios (destinatarios locales). No me centro en autores, sino en destinatarios (que son tomados por Hellsten & Laydeesdoff también como temas). Dos razones para realizar este primer paso son: 1) hashtags y cuentas mencionadas en los mensajes podrían indicar patrones semánticos en la construcción de la cognición pública sobre los temas relacionados (líderes sociales, acuerdo de paz, denuncia, etc.). 2) teniendo en cuenta el lugar propio que la virtualidad en general y las redes sociales en particular encarnan, es posible que las relaciones que allí se generan puedan estar proponiendo una noción de la *Otredad* (en primera instancia, del interlocutor y del destinatario local) propia del medio pero no por ello menos importante para la cognición pública. Haciendo uso de dos software Pajek y VOSviewer, en la gráfica 1 presento un mapa semántico como ejemplo del análisis previo sobre del hashtag #LideresSociales. Los mapas semánticos han sido considerados como visualizaciones de una realidad cognitiva, de una organización conceptual del conocimiento (Georgakopoulos y Polis, 2018). Pajek es usado para el análisis y visualización de grandes bases de datos. Y VOSviewer es para una visualización más óptima de las gráficas.



Gráfica 1. Visualización de la co-occurencia de ítems (hashtags y cuentas) de #LideresSociales (dic 2016-sep 2020) con VOSviwer (realización propia).

La gráfica 1 muestra un análisis sincrónico del hashtag, pero el objetivo es poder realizar un análisis diacrónico en un segundo acercamiento. Aparte de la co-ocurrencia de ítems, también realizo un cosine-analysis para poder observar cómo los ítems se relacionan de manera singular y repetida en diferentes contextos oracionales. Por razones de espacio no muestro la gráfica del cosine, pero con la guía de este pretendo en un segundo momento de la aplicación metodológica realizar el análisis de contenido con NVivo 12. Con este análisis de contenido busco detectar el uso metafórico a gran escala para referirse al tiempo del conflicto y al líder social como tal. En general, este análisis que propongo aén mi investigación puede ser un punto de partida para llamar la atención sobre la cognición pública sobre líderes sociales y otredad. Servirá como caso de estudio que contribuya al entendimiento de cómo la opinión pública virtual participa en la construcción del sentido del acuerdo de paz y del futuro del país.

Referencias

Líderes sociales en Colombia: ¿quiénes los están asesinando? (enero 13 de 2019). *Especiales Semana*. Recuperado de https://especiales.semana.com/lideres-sociales-asesinados/index.html

Barrios, M.M., Vega, L.M. & Gil, L.M. (2019). When online commentary turns into violence: The role of Twitter in slander against journalists in Colombia. *Conflict & communication online, vol 18(1)*.

Cacciatore *et al*. (2018). Is Facebook making us dumber? Exploring social media use as a predictor of political knowledge. Journalism & Mass Communication Quaterly, vol. 95(2), 404-424.

Forero, J. (29 de septiembre de 2020). UNP solo admitió 16 % de las solicitudes de protección de líderes sociales. *El Tiempo*. Recuperado de https://www.eltiempo.com/politica/gobierno/cifras-de-lideres-sociales-asesinados-en-colombia-en-2020-540503

Georgakopoulos y Polis (2018). The semantic model map: State of the Art and future venues for linguistics research. *Lang Linguist Compass 12*.

Hellsten & Laydeesdoff (2020). Automated analysys of actor-topic networks on Twitter: new approaches to the analysis of socio-semantic networks. *Journal of association for information science and technology, 71(1)*, 3-15.

Lakoff, G. & Johnson, M. (1990). *Metáforas de la vida cotidiana*. España: Teorema.

Uribe, M.T. (2004, July-December). Las palabras de la guerra. Revista de estudios políticos, 25, pp. 11-34.

Lee, E. & Jang, Y. (2010). What do Others´Reactions to News on Internet Portal Sites Tell us? Effects of Presentation Format and Readers´Need for Cognition on Reality Perception. *Communication Research, 37(6)*, pp. 825-846.

Mesa, J. (2018). *Imágenes del enemigo. La construcción discursiva del enemigo en la prensa nacional colombiana 1993-2012.* Medellín: Universidad de Antioquia.

Osorio, M. (2018, 5 de julio). ¿Qué hay detrás del asesinato de líderes sociales en el país? *El Espectador*. Retrieved from: https://www.elespectador.com/noticias/paz/que-hay-detras-del-asesinato-de-lideres-sociales-en-el-pais-articulo-798403

Quintero, J.M & Marín, A.F. (2018). Proceso de paz y post-acuerdo en Colombia: expresiones de confianza en Twitter. *El Ágora USB, vol 18 (2)*, pp. 348-361.

Toepfl, F. & Piwoni, E. (2015). Public Spheres in Interaction: Comment Sections of New Websites as Counterpublic Spaces. *Journal of Communication,* 65, 465-488. Doi: 10.1111/jcom.12156.

Vea el borrador de referendo de Centro Democrático para derogar la JEP. (noviembre 1 de 2020). *El Tiempo*. Recuperado de https://www.eltiempo.com/politica/partidos-politicos/borrador-del-proyecto-del-centro-democratico-para-derogar-la-jep-546589

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Entre la ficción y la socialización jurídica: Los mass media como corpus de investigación en el campo legal, posibilidades y experiencias

Gonzalo Albornoz Barra
[g.a.albornoz.barra@rug.nl]
University of Groningen, Países Bajos

El presente resumen pretende evidenciar posibles usos de los mass media como corpus de trabajo en el campo de las humanidades digitales, a partir de la experiencia preliminar de la investigación doctoral titulada "Fictional Representations of Law, Legality and Justice, and the Construction of Legal Knowledge by Key Communities in a Chilean Prison Environment[176]".

La investigación partió de la hipótesis de que los contenidos mediáticos ficticios y no ficticios en películas, series, novelas o noticiarios (entre otros), suelen recrear situaciones que representan aspectos atribuibles a diversas dimensiones de la ley, la legalidad y la justicia (dimensiones sociales, históricas o culturales), lo que es posteriormente asimilado por los sujetos en una dinámica de *producción de los consumidores*[177] (De Certeau 1986).

En términos metodológicos, se consideraron tres *comunidades interpretativas*[178] al interior del Centro de Cumplimiento Penitenciario de Temuco (CCP Temuco[179]): reos, Gendarmes y profesionales de la prisión, a los cuales se aplicaron encuestas (N= 69) y entrevistas semi-estructuradas (N=19) con el fin de identificar: *horas de consumo mass media, prácticas de consumo compartido, tipos de géneros de ficción y no ficción, autopercepciones de conocimiento legal, identificación con personajes (a través de una pequeña historia ficticia) y preguntas de caracterización individual como género, edad y nivel educacional, entre otras (en el caso de la encuesta); y concepciones propias de ley, legalidad, justicia, identificación con personajes y aspectos experenciales y biográficos, entre otros (para el caso de las entrevistas).*

[176] Representaciones ficticias de la ley, la legalidad y la justicia, y la construcción de conocimiento legal en entornos carcelarios. (Traducción propia).

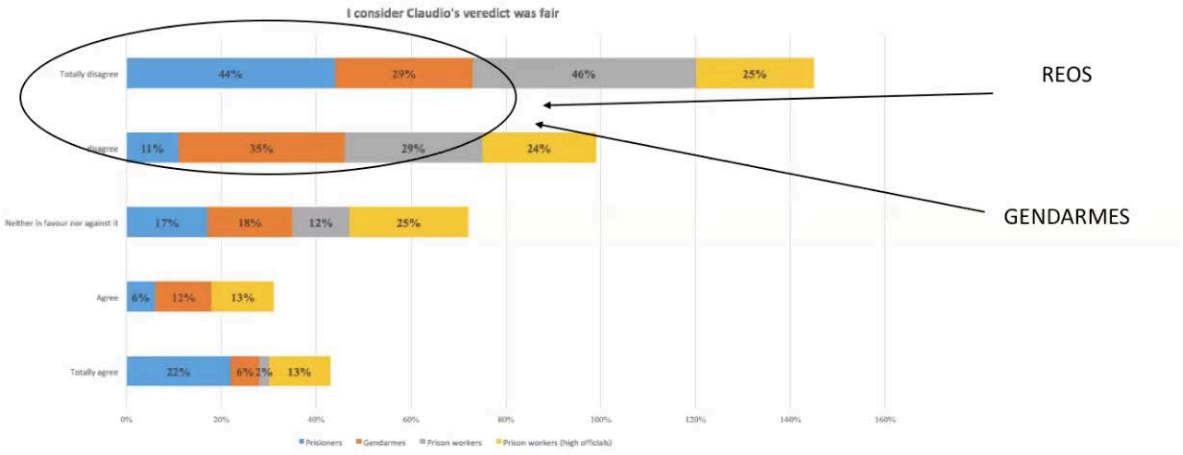
[177] El concepto de producción de los consumidores describe una dinámica en la cual los sujetos producen y reproducen información que es a su vez recibida por otros, los cuales continúan diseminando la información por diferentes medios, y agregando cada vez diferentes elementos como parte de su propia producción narrativa en la cotidianidad.

[178] El concepto de comunidades interpretativas propone que es posible identificar códigos y propiedades de interpretación en sujetos que comparten entornos socio-culturales. El concepto se enmarca dentro del campo de la interpretación de textos, pero es vinculable al campo de las narrativas culturales.

[179] El CCP Temuco es la cárcel principal de la ciudad. Es una cárcel pública, y alberga sólo a población penitenciaria masculina. En ella se encuentran alrededor de 500 internos, 100 Gendarmes y 30 profesionales externos de diferentes disciplinas.

La razón por la cual se enfocó el estudio en contextos carcelarios es que allí conviven tres formas distintas de consumo mediático, ya que por un lado, los reos tienen un consumo controlado en cuanto a contenidos y acceso a dispositivos, material y tiempos de consumo; mientras que los Gendarmes, dada su situación laboral, tendrán un consumo similar al de los reos la mitad de su jornada (mientras estén en ejercicio, viviendo en una sección del penal), y un consumo libre el resto del tiempo (mientras están de franco o días libres); lo que los diferencia a su vez de los profesionales de la prisión, quienes tienen un consumo libre todo el tiempo, tanto en términos de contenidos, como en el acceso a dispositivos, material y horarios[180].

A la luz de los resultados, se observó que los sujetos producen y reproducen diferentes valoraciones y representaciones de la ley, la legalidad y la justicia según los códigos sociales, históricos, culturales, contextuales que comparten con una *comunidad interpretativa* (Fish 1989, González 2009), en donde también influyen, por supuesto, las experiencias y biografías particulares de cada actor. Esto queda en evidencia en datos como:



El gráfico anterior identifica diferencias en las representaciones de justicia que los sujetos tuvieron en atención a un breve cuento, en el cuál, sin mayores antecedentes del caso, se les pedía que consideraran si la sentencia de uno de los personajes era "justa". Se observó que, a mayor identificación con el personaje (Claudio[181]), los sujetos consideraban de forma tajante

[180] Los diferentes contextos de consumo mediático entre estos grupos permitieron identificar algunas diferencias y similitudes en cuanto a la importancia de los contenidos mediáticos en el día a día de los sujetos.

[181] El cuento se desarrolla en un contexto post batalla. A oídos del rey, llega el rumor de que Itamar, Joel y Claudio habían dado información importante sobre la ubicación de las tropas del rey a sus enemigos, razón por la cual el reino perdió la batalla. La única información que se entrega al lector respecto de los imputados, es que Itamar y Joel son hijos de un importante general, mientras que Claudio es un trabajador común en el pueblo. Itamar y Joel

una posible injusticia en la sentencia, mientras que a menor identificación con él, las consideraciones al respecto eran más arbitrarias. Esto refuerza la idea de que la justicia:

> Es un sistema de representación: Representación de la verdad por las partes litigantes en sus argumentos, envueltos por un sistema legal basado en sus propios poderes. Sin embargo, la imagen dada no es siempre la percibida. La comprensión pública de la ley es recogida desde las representaciones culturales de justicia en las que se refleja. De hecho, las recreaciones ficticias de juicios o crímenes en contenidos como películas o caricaturas comparten una imagen de la justicia[182]. (Masson y O'connor 2007, 15).

Con todo, la idea fuerza de este apartado es que, a través de un consumo compartido de historias, es posible observar diferentes valoraciones sociales, históricas, culturales y contextuales de la ley, la legalidad y a justicia que influyen en el proceso de socialización legal. Considerando que los sujetos no comparten los mismos códigos de interpretación, se abre la posibilidades de que este tipo de corpus digitales, a través de su uso en contextos de investigación y uso compartido, emerjan una multiplicidad de formas de entender lo jurídico, que ayuden a generar programas de alfabetización legal enfocados no sólo en la búsqueda de un aprendizaje memorizado de las normas, sino que además se adapten a las dimensiones sociales y culturales de comprensión de lo jurídico, estableciendo un modelo de enseñanza-aprendizaje comunitario, horizontal y democrático, y así también, más significativo.

son declarados inocentes por el consejo de sabios, mientras que Claudio es sentenciado por el mismo tribunal. La historia no da cuenta de más antecedentes sobre el caso o los personajes.

[182] Traducción propia. Cita original: (…) is a system of representation: Representation of truth by litigating parties in their arguments and embodiment by legal system themselves of their own powers. However the image given is not always the one perceived. The public understanding of law is gleaned from cultural representations of justice which reflect, In fact, popular culture movies, caricatures, portrayal of trials by media or crime fiction shape the image of justice.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

La Retórica cultural de la Reforma y de la Contrarreforma: transferibilidad interdiscursiva y hegemonía literaria

Juan Gallego Benot
[juan.gallego@uam.es]
Rijksuniversiteit Groningen

El proyecto de investigación se encuentra en una fase inicial de desarrollo, en la que se ha llevado a cabo una revisión de antecedentes específicos, así como una primera descripción de herramientas metodológicas. El objetivo general de esta tesis doctoral es un estudio comparado de diversas fuentes, en su mayor parte ligada a las doctrinas de la Reforma protestante y de la Contrarreforma católica, con el fin de generar un análisis de las contradicciones que existieron en el conflicto religioso, social y político. La clase de análisis que se lleva a cabo se produce en los niveles de la intesionalización y extensionalización discursiva, según los términos de Janos Petöfi y se constituye en la categorización de estructuras retóricas en relación con los sucesos culturales que constituyen el objeto de estudio.

Esta tesis doctoral parte de diversas fuentes teóricas que dirigen y concretan el caso de estudio. La primera de ellas está constituida por el marco de la Retórica cultural. La Retórica cultural, enunciada por Tomás Albaladejo (1989, 2005, 2016) y enmarcada por Francisco Chico Rico dentro del contexto de la Neorretórica (2015), siendo de gran utilidad para analizar textos literarios por su capacidad reactiva en el texto, al atender al funcionamiento y construcción de los textos y discursos en relación con su intención y efecto en los oyentes y sus contextos. La base de esta aproximación teórica y metodológica tiene su origen en la afirmación de que la Poética clásica se impregnó de ciertas partes del ars rhetorica en su desarrollo, que luego fueron olvidadas en el proceso de análisis textual y limitadas a cuestiones enormemente limitadas, centradas en una enumeración monumentalista de figuras y tropos (García Berrio, 1984; Gómez Cabia, 1988). La propuesta de la Retórica cultural permite una comprensión de gran calado de los procesos semántico-extensionales de los discursos, atendiendo a la perlocución (influencia en los receptores) y considerando los textos como parte de un entramado discursivo interrelacionado, que lo conforma como sistema cultural. El análisis interdiscursivo, que parte a su vez de teorías de intertextualidad como la de Kristeva e incluso de teorías del hipertexto como la de Landlow, tiene en cuenta las dinámicas de cruzamiento que se desarrollan entre los discursos y entre los tipos de discursos.

Además, esta perspectiva se completa con las investigaciones que buscan explicar las literaturas como fenómenos que superan las perspectivas nacionales, con la consecuente ampliación de sus posibilidades analíticas y la introducción de terminología y aparataje que permiten un estudio concreto de estas características "transnacionales". Son aquí de gran importancia las aportaciones de Pablo Valdivia Martín, también director de esta tesis, y de

otros autores como Jahan Ramazani o Bill Ashcroft. Así, es posible entender estas literaturas de forma abierta, híbrida en diversos sentidos —no es posible hablar de una identidad europea en el objeto de estudio de esta tesis, pero sí acotar sus posibilidades por encima de nociones limitadas y en cierto sentido anacrónicas al ahondar en los siglos XVI y XVII, que clasifican las literaturas en patrias e lenguajes de forma exclusivista— que buscan acomodarse o enfrentarse a los discursos hegemónicos de sus lugares de origen (entendidos aquí también como "religiones" o "discursos políticos específicos" o incluso "situaciones sociales"). El aparataje metodológico que proporcionan las Humanidades digitales es de gran importancia en este proceso, con el fin de plantear un análisis de redes en el que se pretende inducir un método de comparación de metáforas estructurales (siguiendo la distinción de Lakoff y Johnson), en el que las estructuras puedan ser comparadas y clasificadas en términos relacionales a través de software de análisis de redes. Un antecedente claro de este modelo que combina lo cualitativo y lo cuantitativo es el proyecto CONNEC, de la Universidad Royal Holloway de Londres, con un proceso combinatorio de Word Embeddings, "New Institutional Theory" y análisis dinámico de redes de información.

Todo este recorrido se completa con el estudio historiográfico de base discursiva de la Reforma y la Contrarreforma y las diferentes perspectivas históricas sobre la construcción política de la división de la Iglesia a lo largo del siglo XVI y XVII. La importancia de textos canónicos de investigadores contemporáneos y los discursos teológico-políticos de la Reforma y la Contarreforma para asentar una serie de relatos sirven aquí de eje sobre el que discurrir e investigar desde la interesante posición de los textos literarios y no literarios en su convivencia e interrelación. Esto define la tesis como un estudio de las relaciones y disensiones de las formas del discurso reformista y contrarreformista; estudio en el que la relación entre los tipos de discurso y los textos literarios y no literarios cobra una importancia cardinal. La hipótesis con la que se trabaja en un principio defiende que los conflictos políticos de la Reforma y la Contrarreforma hallan su contrapunto, explicándose con mayor claridad, en unas formas específicas de persuasión y expresión desde una posición político-religiosa concreta. De forma amplia, este proyecto pretende observar esas interrelaciones, describirlas, resumir sus tipologías y analizarlas a partir del desarrollo teórico-práctico de la Retórica cultural como teoría de análisis discursivo en correspondencia con la cultura. De forma más reducida y a corto plazo, esta tesis plantea un análisis de los textos teológicos que tengan como objetivo evidenciar las diferencias para luego investigar cómo se reflejan dichas diferencias en los métodos persuasivos de textos literarios específicos.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Memorias Visuales de la plaza Aníbal Pinto y la estación de ferrocarriles de Temuco entre 1930-1950: Una reconstrucción digital basada en la Microhistoria

Paola A. Gonzalez Salas
[paola.gonzalez@ufrontera.cl]
University of Groningen / Universidad de la Frontera

Una mirada al Temuco antiguo

Temuco es la capital de la región de la Araucanía. Esta fue la última zona en anexarse al estado chileno luego de un largo proceso conocido como la "Pacificación de la Araucanía".

Dada la importancia y riqueza natural de la zona, el gobierno promovió, a finales del siglo XIX y principios del siglo XX, la llegada de inmigrantes a la región con el fin de poblar y trabajar, generando una mezcla cultural entre chilenos, mestizos, mapuches y europeos.

La plaza Aníbal Pinto y la Estación de Ferrocarriles de la ciudad, son dos emblemáticos lugares que contienen estructuras, historias y formas contrastantes entre sí. Ambas, a principios del siglo XX, eran importantes centros de reunión social, comercial y cultural. Por un lado, la plaza se convirtió en el centro económico y social de la ciudad ya que aquí existían los principales bancos (Banco Alemán, Banco Chileno-Español), la intendencia, la municipalidad, el Teatro Paramount, el Diario Austral y la Catedral. Era lugar de encuentro para familias quienes participaban de actividades como desfiles, ceremonias y bailes. Por otro lado, la estación de Ferrocarriles reunía a los comerciantes, hortaliceras (mapuche), hotelería, cantinas y hasta prostíbulos por las calles aledañas (Bello, 1992; Jorge Pinto Rodriguez, 2002). Sus edificios y fachadas se dibujaban con el estilo neoclásico que traían los europeos migrantes y que les dieron formas que caracterizaron la época entre 1930 a 1950 (Cerda Brintrup Gonzalo, 2009). La *modernidad*[183] llegaba a la ciudad a través del ferrocarril, del teatro sonorizado y la llegada de nuevos medios de comunicación como el periódico (Diario Austral), la radio y la telefonía.

En redes sociales [184] y sitios web [185] encontramos viejas fotografías en las que podemos visualizar como eran estos lugares, como vestían las personas y algunas de sus actividades. Pero el paso del tiempo, la sucesión de terremotos y la poca o nula preocupación por el cuidado o preservación patrimonial por parte de las autoridades de la ciudad de Temuco (Enríquez Eduardo, 2007), han hecho que la imagen de la plaza Aníbal Pinto y la Estación de Ferrocarriles haya sido totalmente modificada hasta el día de hoy.

[183] Refiere al proceso de crecimiento económico, concentración urbana e innovación tecnológica

[184] Facebook: Temuco antiguo en Imágenes, Chile Antiguo.

[185] www.memoriachilena.cl

Aquellas fotografías antiguas de la ciudad de Temuco, como memorias visuales[186], podríamos definirlas como un recorte de la realidad de ese tiempo y pueden verse sujetas a variadas interpretaciones siendo sumamente importantes como fuente histórica (Blanco, 2011) base para la investigación ya que pueden desencadenar una serie de memorias, asociadas a nuestro pasado y al de nuestros ancestros (Ricoeur, 2004).¿Cuáles son las narrativas detrás de las imágenes de la plaza Anibal Pinto y la Estación de Ferrocarriles de Temuco entre los años 1930 a 1950?

Muchas de estas imágenes provienen de registros familiares y que quedan sólo para la apreciación estética en sitios web y redes sociales, generando en la población, diversos comentarios y opiniones, las sensaciones y emociones que les evoca ver a Temuco de forma tan diferente a la actual. ¿Cuáles son las apreciaciones de la población de la ciudad respecto a la imagen pasada de la plaza Aníbal Pinto y la Estación de Ferrocarriles de Temuco entre los años 1930-1950? ¿Tienen recuerdos propios, de sus padres o familiares, en estos lugares y cuáles son?

Podemos profundizar el estudio hacia el análisis de archivo mediante el rescate de elementos de la microhistoria (Man, 2013) que ayuden a dar forma a los discursos cotidianos entorno a estos dos lugares de la ciudad. Los tipos de archivos pueden ser variados, desde una boleta, un registro de bautizo o matrimonio, hasta algún anuncio publicitario.

Para hablar de reconstrucción digital, es importante referirse también al concepto de museística y museística digital, en donde el primero apunta al sitio físico que reúne material histórico, arqueológico, artístico, etc. de una sociedad o cultura específica y que esta abierto a la comunidad para fomentar el estudio y el conocimiento de otras épocas o lugares (Linarez Pérez, 2008). La museística digital aborda además, el uso de herramientas tecnológicas para expandir el estudio histórico y cultural, llevándolo a nuevas formas de representaciones didácticas y lúdicas facilitando el estudio y aprendizaje (Bellido Gant *et al.*, 2014). El uso de herramientas tecnológicas, nos permite poder reconstruir espacios, elaborar narrativas entorno a objetos y elementos o prácticas culturales que pueden haber desaparecido con el paso de los años. Podríamos decir que a través de un desarrollo digital se puede cubrir el vacío que genera el devenir del tiempo y la falta de protección del patrimonio cultural, restaurando y reconectando las memorias locales a través de una simulación gráfica (Garcia Cuetos, 2004).

La realidad virtual ofrece la sensación de "presencia o estar en un lugar" al poder observar en 360º alrededor de uno mismo. Esto no es posible con otro sistema digital por lo que su riqueza consiste en la experiencia de inmersión dentro de un mundo virtual (Kim *et al.*, 2013)

[186] La memoria visual es aquella que permite percibir, codificar, almacenar y recuperar o evocar imágenes que hemos captado a través del sentido de la vista. Se trata de una memoria muy poderosa, que permite fijar recuerdos con facilidad en el cerebro. Gracias a ella somos capaces, por ejemplo, de recordar el rostro de una persona, un objeto, una situación, un entorno.

en el que se pueda interactuar con elementos hipermediales[187]. Los objetos interactivos dentro del espacio virtual, serán claves en la entrega de información para el usuario, ya que estos serán obtenidos y construidos desde el análisis microhistórico realizado previamente y podrán contener imágenes, audio o video, enriqueciendo aún más la inmersión. La reconstrucción de estos espacios, no basta con el diseño y visualización en 3d sino cuenta con un genoma digital [188]que sea capaz de sostener su estructura sobre una base de estudios previos, por lo que la construcción teórica es fundamental para el exitoso desarrollo de la estructura digital.

Esta investigación y proyecto traerá consigo una reconstrucción histórica y cultural pero también la revalorización de la ciudad, del espacio social y sus historias propias, de la mano de tecnologías que nos ayudarán a experimentar y visualizar una forma nueva de ver la cultura y el patrimonio en el sur de Chile.

[187] **Hipermedia:** Convergencia interactiva de medios y sus sustancias expresivas (imagen fija, imagen en movimiento, sonido, tipografías) con las que el receptor-lector se convierte en lectoautor

[188] **Genoma Digital:** Digitalización de alta calidad, programa hipermedia explicativo y contextualizado basado en estudios previos.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Variación terminológica horizontal en español: hacia una descripción lingüística de las unidades terminológicas desde la lingüística de corpus

*Carlos Mario Pérez Pérez
[carlosm.perez@udea.edu.co]
*Estudiante doctorado en Lingüística
Director **Gabriel Quiroz Herrera
** Doctor en Lingüística Aplicada
Universidad de Antioquia

El estudio de la terminología especializada en español se ha centrado, principalmente, en análisis de unidades nominales, relegando a un segundo plano otras unidades terminológicas como las simples (verbos, adjetivos, adverbios) y complejas (compuestos, sintagmas, frasemas, etc.). Además, las unidades metafóricas con carácter terminológico, como las de la computación ("mouse" o "ratón") u otras áreas, no han sido incluidas en ninguno de los manuales disponibles en el área, a saber: Felber (1984), Arntz & Picht (1995), Fedor de Diego (1995), Cabré (1999), Pavel & Nolet (2001), Cardero (2003). Además, en la actualidad, existe una gran cantidad de herramientas computacionales para el procesamiento automático de información lo que permitiría procesar grandes cantidades de datos en menor tiempo del empleado en los análisis manuales tradicionales. La implementación de estas herramientas, tales como Termostat, BC_Term, Multiterm, UCREL, Wordnet, no solo permitiría un estudio sistemático de distintas unidades terminológicas especializadas sino también una optimización de los recursos disponibles en los distintos proyectos de investigación.

Son habituales los estudios representados, principalmente, en tesis sobre el análisis de una sola unidad terminológica o parte de ella dentro de un área de conocimiento (preferentemente la medicina y la economía), de un determinado tipo de texto o género (generalmente textos del tipo IMMRAD), de sintagmas terminológicos (Cortés, 2004; Quiroz, 2005, 2008; Lara, 2011, 2012; Lopera, 2016; Vásquez, 2016), de colocaciones/frasemas (Rojas, 2013; Patiño, 2017), de adjetivos y participios (Folguerá, 2004; Salazar, 2011); de verbos (Lorente, 2002; Pérez Pérez, 2018), de siglas (Giraldo, 2008; Rivas, 2017); por no mencionar otros estudios cuya unidad de análisis no es clara o es difusa (casos de unidades poliléxicas). Podríamos definir estos estudios normalmente como inductivos, pero no los podríamos encuadrar como estudios de variación horizontal ya que ninguno analiza unidades terminológicas en varias áreas del conocimiento. Asimismo, algunos de estos estudios usan como corpus de contraste varios diccionarios especializados de diferentes áreas para validar las unidades encontradas, pero no en corpus textuales diversos.

La presente investigación se apoya en la lingüística computacional y de corpus para la descripción lingüística y estadística de unidades terminológicas en distintas áreas del conocimiento. En este sentido, nuestro trabajo se enmarca en un tipo de investigación básica, descriptiva y empírica. En nuestro contexto colombiano, los trabajos de este tipo son poco frecuentes, a pesar de ser bastante productivos en otros contextos de investigación a nivel mundial.

Investigación básica en cuanto se pretende partir de la adquisición de conocimientos elementales de las distintas técnicas de Lingüística de Corpus que permitan la construcción de los corpus de trabajo al igual que la destreza en el uso de las diferentes herramientas necesarias para los análisis lingüísticos a implementar. Investigación descriptiva pues a partir de los datos compilados y procesados se podrá analizar, describir y explicar el fenómeno de la variación horizontal, sin incurrir en aspectos prescriptivistas y apartándonos de las tendencias casuísticas que existen en varios estudios, principalmente en Latinoamérica. Investigación empírica, al emplear métodos computacionales, estadísticos y lingüísticos en la implementación de la metodología de corpus en el análisis del fenómeno estudiado.

Se ha compilado un corpus textual constituido por artículos de divulgación científica en 5 áreas del conocimiento y recopilados de revistas académicas en los países cuyo idioma oficial es el español (exceptuando el caso de Puerto Rico). Inicialmente, se determinaron los campos de estudio de los cuales se compilaría el corpus teniendo como criterios básicos: (1) contar con áreas de gran impacto y, en consecuencia, áreas muy analizadas; (2) áreas de menor impacto y, por ende, de menor análisis; y, por último, (3) áreas de importancia social. Para esto, se retomó la clasificación tradicional del conocimiento entre Ciencias Sociales y Humanas, Ciencias Básicas, Filosofía, Tecnología y Arte. Así, se estableció medicina, lingüística, veterinaria y zootecnia, computación y filosofía como áreas de trabajo. El corpus final quedó constituido por un total de 538 artículos (3.057.973 tokens) con el cual se llevarán a cabo los análisis a nivel sintáctico y semántico teniendo como base estructuras tanto nominales como otras estudiadas previamente en investigaciones como las de Quiroz (2008), Giraldo (2008), Lopera (2016), Patiño (2017), Galeano (2017), Vásquez (2018) y Pérez Pérez (2018).

Referencias

Las referencias aquí presentadas están incluidas dentro del texto referencia citada o son parte de la construcción conceptual del tema, aunque no estén citadas.

Arntz, Reiner; Picth, Heribert. (1989). Einführung in die Terminologiearbeit. Traducción del alemán: De Irazazábal, Amelia *et al.*, *Introducción a la terminología*. Madrid: Fundación Germán Sánchez Ruipérez, Pirámide.

Biber, Douglas. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4), 243-257.

Burgos, Diego (2014). *Towards an Image-Term Co-occurrence Model for Multilingual Terminology Alignment and Cross-Language Image Indexing*. Tesis de doctorado Dir. Leo Wanner. Barcelona: IULA, UPF.

Cabré, María Teresa. (1999). *La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Cabré, María Teresa. (1999e). "Hacia una teoría comunicativa de la terminología: aspectos metodológicos". En: *La terminología. Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, (Sèrie Monografies 3). 129-150.

Fedor de Diego, Alicia. (1995). *Terminología. Teoría y práctica*. Caracas, Equinoccio-Unión Latina. 159 pp.

Felber, Helmut y Picht, Heribert. (1984). *Métodos de terminografía y principios de investigación terminológica*. Madrid: Instituto Miguel de Cervantes.

Galeano, Marta. (2017). *La terminología de la ganadería*. Tesis de maestría, Dir. Gabriel Quiroz. Medellín: Universidad de Antioquia.

Giraldo, John. (2008). *Análisis y descripción de las siglas en el discurso especializado de genoma humano y medio ambiente*. Tesis de doctorado. Dir. Teresa Cabré. Barcelona: Universitat Pompeu Fabra.

Giraldo, John. (2014). Some Criteria for a Spanish Initialisms Recognition System. En G. Quiroz & P. Patiño (Eds.), *LSP in Colombia: advances and challenges* (pp. 19-32) Peter Lang Verlag.

Lopera, Gustavo. (2016). *Sintagmas nominales con premodificación compleja con guiones en un corpus lexicográfico especializado bilingüe*. Tesis de maestría, Dir. Gabriel Quiroz. Medellín: Universidad de Antioquia.

Patiño, Pedro. (2014). Towards a Definition of Specialized Collocations. En G. Quiroz & P. Patiño (Eds.), *LSP in Colombia: advances and challenges* (pp. 119-133) Peter Lang Verlag.

Patiño, Pedro. (2017). *Description and Representation in Language Resources of Spanish and English Specialized Collocations from Free Trade Agreements*. Ph.D. Thesis, Dir. Gisle Andersen. Bergen: NHH.

Pérez Pérez, Carlos. (2018). *Metodología para la caracterización de patrones [verbo + objeto directo] sobre pobreza en la prensa colombiana para propósitos de Análisis de Sentimientos*. Dir. Gabriel Quiroz. Medellín: Universidad de Antioquia.

Pérez Pérez, Carlos, Quiroz Herrera, Gabriel, & Tamayo Herrera, Antonio (2020). ¿Tiene carácter predictivo la estructura predicativa [verbo + objeto directo]? hacia una caracterización sintáctico-semántica para propósitos de análisis de sentimientos. *Lingüística Y Literatura*, 41(78), 11-35.

Quiroz, Gabriel. (2008). *Los sintagmas nominales extensos especializados en inglés y en español: descripción y clasificación en un corpus de genoma*. Tesis de doctorado. Dir. Mercè Lorente.

Quiroz, Gabriel & Arroyave, Alejandro. (2014). On Premodified Terms in Five Specialized Dictionaries. En G. Quiroz & P. Patiño (Eds.), *LSP in Colombia: advances and challenges* (pp. 119-133) Peter Lang Verlag.

Quiroz, Gabriel & Arroyave, Alejandro. (2018). Sintagmas nominales con premodificación compleja: algunos aspectos de su traducción del inglés al español. En Colín, Marisela (Coord.). *Traducción de textos especializados. Nuevos enfoques, nuevas metodologías*. 196-255. México DF: Universidad Nacional Autónoma de México. ISBN 978-607-30-0706-1.

Rivas, Natalia. (2017). *Siglas dentro de siglas: el fenómeno de las siglas anidadas en el discurso científico y su comportamiento en las traducciones*, Tesis de maestría. Medellín: Universidad de Antioquia.

Rojas, José Luis. (2014). *Las unidades fraseológicas en textos especializados*. Tesis de maestría. Dir. Gabriel Quiroz. Medellín: Universidad de Antioquia.

Vásquez, David. (2018). *Sintagmas nominales con premodificación compleja con guiones en un corpus de medicina*. Tesis de maestría. Dir. Gabriel Quiroz. Medellín: Universidad de Antioquia.

Zuluaga, Felipe; Tamayo, Antonio; Quiroz, Gabriel. (2017). La especialización del discurso en los medios en Colombia: el caso de los fenómenos de reducción léxica en la prensa de la pobreza. In *Poverty, Language and Media*. (Ana B. Chiquito and G. Quiroz, Eds) Linguistic Insights series. Peter Lang Verlag.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

WoPATeC

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Análisis y resumen automático de políticas de privacidad

Miguel Valenzuela, Stephanie Riff, Rodrigo Alfaro[189], Alan Bronfman & René Venegas

[miguel.valenzuela.p@mail.pucv.cl | stephanie.riff.c@mail.pucv.cl | rodrigo.alfaro@pucv.cl | rene.venegas@pucv.cl]

Pontificia Universidad Católica de Valparaíso, Chile

Cada vez que descargamos una aplicación, creamos un usuario en una red social o utilizamos alguna plataforma informática, debemos aceptar las políticas de privacidad de las mismas. Muchas veces son extensas, confusas y repiten términos, lo que lleva a los usuarios a aceptarlas sin leer. ¿Cómo podemos generar información resumida acerca de las políticas de privacidad de diferentes aplicaciones, que permita a los usuarios tener conocimiento acerca del tratamiento que se hará de sus datos? Para responder esta pregunta se utilizarán técnicas de Procesamiento de Lenguaje Natural para el resumen automático de textos: Extractivo, abstractivo, aprendizaje sin supervisión, documento único, documento múltiple, optimización, dominio y tiempo real. A partir de lo anterior, será posible obtener un resumen de las políticas de privacidad de diferentes aplicaciones y, por lo mismo, un costo reducido de lectura humana y mayor comprensión de las mismas. Además, se identifican los modelos más adecuados utilizando evaluaciones de métricas ROUGE y de profesionales del área de jurídica e informática, para medir la precisión, recuperación, información redundante y organización de la información generada. Para el análisis se consideran los marcos legales de Europa, Estados Unidos y Chile, así como el posible impacto en la sociedad, países, organizaciones, empresas y comunidades a las que esta situación afecta.

[189] Nací en Valparaíso, Chile en 1973, soy Ingeniero Civil Industrial y Magister en Ingeniería Industrial por la PUCV, Chile, y Ph.D.(c) en Ingeniería Informática de la UTFSM-Chile. Soy académico de la PUCV-Chile, donde enseño e investigo. Más información en www.rodrigoalfaro.cl

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Asignación de hiperónimos para sustantivos polisémicos en una taxonomía de inducción automática: propuesta metodológica a partir de las similitudes de coocurrencia verbal en hipónimos de segundo grado

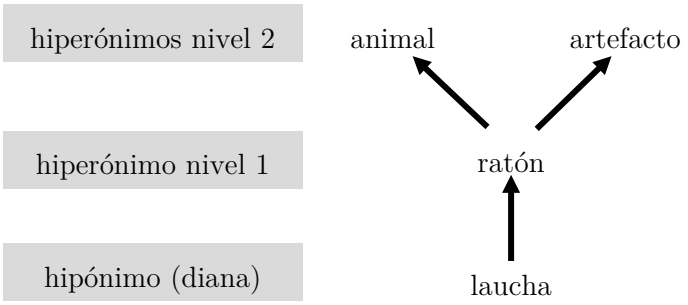
Javier Obreque[190] **& Rogelio Nazar**[191]
[j.obrequezamora@gmail.com | rogelio.nazar@pucv.cl]
Pontificia Universidad Católica de Valparaíso, Chile

La polisemia en los sustantivos podría definirse como la situación en la que un mismo sustantivo tiene, potencialmente, más de un hiperónimo o sustantivo definidor. La asignación de un hiperónimo para un sustantivo es una operación común y natural para el ser humano. En la mayoría de las ocasiones es el contexto sintagmático o pragmático-discursivo el que permite desambiguar entre todos los candidatos posibles a hiperónimo. Sin embargo, esta tarea se vuelve problemática en el marco de la conformación de una taxonomía semasiológica generada por un sistema automático (Klapaftis y Manandhar, 2010). Uno de estos casos dificultosos bien podría ser el manifestado en una tríada de relación hiperonímica como la siguiente: *laucha* ⇒ *ratón* ⇒ *animal/artefacto* (Figura 1). En un caso como este, el hiperónimo de *laucha* es *ratón* sin que exista posibilidad de ambigüedad; ahora bien, el problema se presenta luego, porque *ratón* es un término polisémico, y bien podría entenderse a *ratón*, y, por consiguiente, su hipónimo y nuestra palabra diana *laucha,* como un tipo de *animal* o como un *artefacto*, en tanto que también se conoce así al hardware de una computadora. En estos casos, el error de la automatización de una taxonomía léxica se puede producir en la serie de los nodos de hiperonimia, porque si el sustantivo diana es *laucha,* entonces el hiperónimo de segundo orden o nivel no puede ser *artefacto*, sino *animal*. Un sistema automático podría, pues, realizar erróneamente el vínculo *laucha* ⇒ *ratón* ⇒ *artefacto.*

Figura 1: Ejemplo de tríada hiperonímica.

[190] Docente de Castellano y Comunicación. Licenciado en Lengua y Literatura Hispánica. Ambos grados obtenidos en la Pontificia Universidad Católica de Valparaíso (Chile). Actualmente, ejerce docencia en un establecimiento de educación superior y forma parte del equipo técnico en anotación lexicográfica del proyecto Fondecyt 1191204 (Comisión Nacional de Investigación Científica y Tecnológica, Chile) dirigido al análisis semiautomático de corp

[191] Profesor del Instituto de Literatura y Ciencias del Lenguaje de la Pontificia Universidad Católica de Valparaíso. Se especaliza en Procesamiento del Lenguaje Natural y Lingüística Computacional. Su interés principal es el léxico, la terminología y el discurso. Actualmente dirige el grupo tecling.com, dedicado al desarrollo de tecnologías lingüísticas.



En estos términos, la dificultad explicitada es un problema de clasificación, uno de los tipos de dificultades más comunes manifestados en la construcción de taxonomías automáticas de sustantivos (Bordea *et al.*, 2015). Según Nazar y Renau (2016), actualmente las estructuras taxonómicas de poblamiento automático lidian con problemas de estructura y polisemia que las hacen alcanzar un rango de precisión que fluctúa entre el 60% y el 80%. Esta medida está por debajo de lo esperado y la conformación de esta propuesta metodológica, que considera los principios de la semántica léxica y de la lingüística distribucional aplicados en la estadística de corpus, pretende transformarse en una posibilidad para solucionarla.

En función del fenómeno planteado, y gracias al financiamiento del proyecto Fondecyt n° 1191204, el objetivo de esta investigación ha sido construir un método que, utilizando la información semántica proporcionada por un hipónimo de un sustantivo polisémico, logre seleccionar automáticamente para él un sustantivo hiperónimo entre dos o más posibles sentidos. Esto, a partir de una comparación de similitud estadística extraída de una amplia muestra de coocurrencias de verbos asociados cotextualmente al hipónimo de segundo grado (ej. *laucha*) y a los posibles hiperónimos de segundo nivel (ej. *animal* y *artefacto*). En concreto, estas muestras experimentales de tríadas polisémicas fueron extraídas de la base de datos de WordNet en castellano (Vossen, 2004) y los contextos de aparición de coocurrencias verbales desde el corpus esTenTen (Kilgarriff y Renau, 2013), versión 2011 (de 9.500 millones de palabras).

Los resultados del estudio fueron significativos en cuanto al tratamiento de la polisemia procesada con herramientas de la lingüística computacional, considerando que se trata de un método relativamente simple. Se construyó un código fuente implementado en lenguaje Perl (disponible en www.tecling.com/hat) con los algoritmos que fueron capaces de sistematizar la tarea de selección y asignación de hiperónimos para el 86% de los 26 casos del estudio. Con ello, la conclusión más significativa del estudio fue que la coocurrencia verbal entre sustantivos abre un camino funcional como elemento clasificador y predictor del significado en una relación taxonómica problemática, donde el sustantivo intermedio de una tríada taxonómica (hiperónimo de primer nivel o grado) es polisémico. No obstante lo anterior, y como trabajo futuro, es necesario reproducir este experimento con muestras de datos más amplias para analizar, entre otros fenómenos, cómo afectarían las diferencias de frecuencia a los resultados. Para ello es fundamental lograr previamente producir un método que consiga extraer una mayor cantidad de muestras experimentales.

Referencias

Bordea, G., P. Buitelaar, S. Faralli, y R. Navigli. 2015. SemEval-2015 Task 17: Taxonomy extraction evaluation (texeval). *En Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 902-910. ACL.

Kilgarriff, A. y I. Renau. 2013. esTenTen, a vast web corpus of peninsular and american spanish. *Procedia - Social and Behavioral Sciences*, 95, pp. 12 - 19.

Klapaftis, I. P. y S. Manandhar. 2010. Taxonomy learning using word sense induction. En *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the* ACL, pp. 82-90, Los Angeles, California, Junio. ACL.

Nazar, R. y Renau, I. (mayo, 2016). A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), pp. 1485-1492.

Vossen, P. 2004. Eurowordnet: A multilingual database of autonomous and language-specic wordnets connected via an inter-lingual-index. *Special Issue on Multilingual Databases, International Journal of Linguistics*, 17(2), pp. 161-173, 06.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Clasificación de movidas discursivas de tesis: representación y aprendizaje automático profundo

Humberto González[192], Héctor Allende[193] & René Venegas[194]
[humberto.gonzalez.a@mail.pucv.cl | rene.venegas@pucv.cl]
Pontificia Universidad Católica de Valparaíso, Chile

La tarea de organizar y estructurar las ideas de manera que se presenten de forma ordenada y coherente es una dificultad latente al momento de desarrollar el trabajo final de grado o tesis, esto es un problema que aqueja normalmente a los estudiantes. Para poder afrontar esta situación se debe acudir a ayuda personas externas, aunque es poco probable que encuentren todos los errores en el documento. Para ello se ha planteado un sistema automático que pueda comprobar si la tesis presenta una estructuración coherente y adecuada en base a un historial de tesis de su mismo género.

El sistema hace uso de la clasificación automática de las movidas discursivas, estas son unidades de procesamiento que cumplen una función adicional a nivel semántico-pragmático, controlando la situación de enunciación, organizando la información discursiva y guiando la interpretación del lector. De esta manera se podrá subdividir el texto de forma que se pueda encontrar patrones característicos que corresponden a una tesis.

Con el fin de desarrollar un algoritmo que utilice técnicas de Deep Learning para la clasificación de funciones discursivas, se definen la siguiente metodología en busca de alcanzar una configuración para un sistema semiautomático de apoyo a la escritura en trabajos finales de grado en carrera de ingeniería civil informática.

---

[192] Magister en Ingeniería Informática e Ingeniero Civil Informático titulado de la Pontificia Universidad Católica de Valparaíso, Chile. Se ha desempeñado como Ingeniero Analista y Full Stack Web Developer para proyectos Fondecyt, Corfo y desarrollo de proyectos multidisciplinarios enfocados principalmente al análisis de corpus y NPL. Entre otros proyectos que ha destacado es en programas de robótica que se han enfocado en acercar a la comunidad universitaria y secundaria a la robótica.

[193] Ingeniero Civil Informático, Magister en Ciencias de la Ingeniería Informática y Doctor en Ing. Informática de la Universidad Técnica Federico Santa Maria. Actualmente es profesor de la Escuela de Ingeniería Informática de la Pontificia Universidad Católica de Valparaíso. Es presidente de la Asociación Chilena de Reconocimiento de Patrones desde el año 2017. En sus intereses de investigación están Machine Learning, Reconocimiento de Patrones, Análisis de Series de Tiempo y Procesamiento de Lenguaje Natural.

[194] Doctor en lingüística y profesor titular del Instituto de Literatura y Ciencias del Lenguaje de la Pontificia Universidad Católica de Valparaíso, Chile. Realiza docencia de pregrado y posgrado en lingüística. Su línea de investigación es el discurso académico y profesional, utilizando lingüística de corpus, lingüística computacional y análisis de géneros discursivos. científicos. Es investigador responsable de múltiples proyectos de investigación y de innovación tecnológica. Actualemente dirige el Rhetorical-Discoursive and Lexical-grammar Analysis Group (REDILEGRA) y el Núcleo de Investigación en Procesamiento del Lenguaje Natural Aplicado (#NiPLNA). Para más información visite: www.renevenegas.cl; www.redilegra.com

Se utilizaron datos correspondientes a 60 tesis de pregrado de la carrera de Ingeniería Civil Informática de la universidad Pontificia Universidad Católica de Valparaíso. Cada tesis fue analizada retórico-discursivamente por 3 etiquetadores humanos, entrenados en la identificación de los propósitos discursivos del modelo retórico-discursivo (Venegas, Zamora y Galdames, 2016). Se consideró un 80% de acuerdo entre los etiquetadores en la identificaron los fragmentos de texto que pertenecían a cada paso.

Cada tesis fue dividida en su macromovida correspondiente. Está fue etiquetada, manualmente, de acuerdo a los fragmentos textuales de los pasos retóricos, que permiten instanciar el propósito comunicativo.

La representación de los textos considera como atributos tres unidades lingüísticas: a) lexemas, b) lemas y c) categorías morfológicas. Estas representaciones se han considerado relevantes, pues corresponde a información lingüística de fácil extracción con un analizador morfosintáctico a partir de los textos.

La red neuronal profunda se entrena mediante el uso de conjuntos de datos etiquetados que se mencionaron anteriormente. La red neuronal aprende directamente a partir de los datos, sin necesidad de una extracción manual de características.

Los resultados obtenidos superan los valores obtenidos en una representación clásica a través de categorías morfológicas detalladas, por ejemplo, N_com_sing para sustantivo_común_singular, presentan un valor promedio para los atributos F-POS (del inglés Fine Grained POS tag) de F= 0,794. En tanto, para I-POS (Impoverish POS Tag), i.e. N para sustantivo, el valor promedio fue de F=0,717. Por otra parte, la representación denominada "bolsa de palabras" (del inglés Bag of Words), muy común en este tipo de tareas, tuvo un valor promedio de F=0,738, menor que F-POS, aunque con mejor rendimiento para la clasificación de supergénero y macrogéneros. Los resultados demuestran una mejora en cuanto a los resultados con métodos tradicionales, según los autores, que a menor cantidad de clases de género, la representación del texto puede ser más concreta (solo palabras), en tanto que a mayor cantidad, se logra mayor abstracción lingüística y especificidad.

Aunque Deep Learning es una forma especializada de aprendizaje automático. El flujo de trabajo de Machine Learning empieza con la extracción manual de las características relevantes de las tesis etiquetadas. Estas características se utilizan entonces para crear un modelo que categoriza las funciones semánticas. Con un flujo de trabajo de Deep Learning, las características relevantes se extraen directamente de las etiquetas.

Además, el aprendizaje profundo realiza un "aprendizaje completo", es decir, se proporcionan datos sin procesar y una tarea que realizar, como es en este caso la clasificación, a una red, la cual aprende cómo hacerlo automáticamente. De esta manera se podría brindar un apoyo mucho más cercano a la retroalimentación proporcionada por un experto, lo que significaría un aumento en la calidad de escritura de los estudiantes al momento de presentar su tesis.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Clasificación de sustantivos abstractos y concretos utilizando el adjetivo como variable predictiva

Valentina Ravest Córdova[195]
[vale.ravest.c@gmail.com]
Pontificia Universidad Católica de Valparaíso, Chile

Distintas gramáticas del español, tales como la *Nueva gramática de la lengua española* (Real Academia Española, 2009) y la *Gramática descriptiva de la lengua española* (Bosque y Demonte, 1999), han mencionado las problemáticas que existen para clasificar sustantivos en las categorías de abstractos y concretos. Entre estas se encuentra la dificultad para definir las nociones mismas de concreto y abstracto, la falta de criterios morfológicos y la falta de acuerdo entre los distintos gramáticos sobre sus características y sus definiciones. Estas obras utilizan sus propios criterios para definir lo que consideran un sustantivo abstracto y uno concreto, aunque reconocen que existen obstáculos para clasificarlos correctamente debido a que las definiciones suelen encontrarse con casos de sustantivos que las desafían. Esto ha generado dificultades dentro del campo de la lexicología, ya que aún no existen criterios suficientes para clasificar sustantivos en estas categorías. Hasta hoy, solo se clasificaron con base en criterios semánticos, pero estos no se pueden considerar suficientes. Por ejemplo, la idea de que los sustantivos abstractos designan entidades separadas de las cosas mismas, tales como características o propiedades referidas a la forma (Bello, 1847), lo que no logra abarcar las singularidades de los distintos sustantivos, dificultando la clasificación.

A partir de esta falta de criterios objetivos para la clasificación, este trabajo pretende establecer un método de clasificación automática, empírica y objetiva de los sustantivos en las categorías de abstracto y concreto utilizando como variable predictiva los adjetivos con los que coocurren en su contexto de aparición. Para la consecución de este objetivo se formula la hipótesis de que el adjetivo es el elemento predictivo que permite esta clasificación. El método de clasificación propuesto se dividió en cinco partes. La primera consistió en la constitución de un conjunto de datos experimentales. Por un lado, la selección de una muestra aleatoria de 262 sustantivos en inglés y 248 en español, divididos en abstractos y concretos, subdivididos en las subcategorías de actividades, doctrinas y sentimientos, dentro de los abstractos, y de armas, frutos y minerales dentro de los concretos, a partir de los diccionarios *Diccionario de uso del español de América y España, DUEAE,* en su versión en CD-ROM (Battaner, 2003),

[195] Licenciada en Lengua y Literatura, con mención en Lingüística Aplicada, de la Pontificia Universidad Católica de Valparaíso. Desempeño como Asistente técnica en proyecto FONDECYT "lnducción automática de taxonomías de marcadores discursivos a partir de corpus multilingües", en las siguientes áreas: Investigación en el campo del procesamiento de lenguaje natural. Análisis gramatical y lingüístico. Desarrollo de sistemas de clasificación léxica. Como continuación de tesis de licenciatura, desarrollo de sistema de clasificación automático para sustantivos abstractos y concretos.

y el *Diccionario de la lengua española* (RAE, 2014). Por otro lado, se constituyó un corpus textual de trabajo, creado a partir de una muestra también aleatoria de 5000 contextos de aparición de cada uno de esos sustantivos en el corpus EsTenTen (Renau y Kilgarriff, 2013). En la segunda parte del proceso, se procedió a la extracción de los adjetivos que coocurren con los sustantivos dentro del corpus, utilizando para ello un script Perl que aprovecha el etiquetado morfosintáctico que ya trae el corpus, con el objeto de identificar los adjetivos. En la tercera parte, se procedió a ordenar los datos de sustantivos y adjetivos dentro de una matriz, en la que se ubicaron los sustantivos en las filas y los adjetivos en las columnas, asignando valores binarios a cada celda para señalar la coocurrencia de adjetivos y sustantivos. En cuarto lugar, se aplicó la técnica estadística de clustering aglomerativo usando el programa R, en su versión 3.5.1 (Ihaka y Gentleman, 2018), que grafica los resultados en forma de un dendrograma donde se refleja la similitud de los sustantivos en función de los adjetivos compartidos. Finalmente, utilizando los 50 adjetivos con mayor frecuencia de aparición para sustantivos concretos y para abstractos, se construyó el clasificador, en el cual es posible ingresar cualquier sustantivo sin categorización y clasificarlo inmediatamente.

El clasificador fue capaz de cumplir su objetivo correctamente tanto con sustantivos abstractos como concretos, en inglés y en español. Los sustantivos abstractos, en ambos idiomas, se clasificaron con un 0.81 de precisión, mientras que los concretos con un 0.89 de precisión.

Estos resultados sugieren la posibilidad de clasificar los sustantivos abstractos y concretos a través de procedimientos mecánicos y criterios objetivos, lo que plantea nuevas posibilidades para la lexicografía y promueven nuevas vías de investigación en lexicología.

Una página web ofrece una demo del clasificador y otros datos del proyecto. La demo acepta un sustantivo cualquiera como entrada y como salida ofrece una clasificación como concreto o abstracto. La dirección es la siguiente: http://www.tecling.com/chungungo

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Detección automática de verbos del español en corpus y su aplicación para la detección de neología léxica

Ana Castro Páez[196]
[ana.castro@pucv.cl]
Pontificia Universidad Católica de Valparaíso, Chile

En esta propuesta, se presentan los resultados de una investigación que consiste en una propuesta metodológica para la detección de verbos del español que han quedado registrados en un corpus general de lengua castellana, con especial atención en verbos neológicos. Este estudio se propone ofrecer un método que, a través de una serie de algoritmos, permite detectar automáticamente verbos que tienen estabilidad en el uso y que están registrados en el diccionario, verbos que están registrados en el diccionario, pero cuya baja o nula frecuencia indica que podrían estar en desuso (Algeo, 1993) y verbos que no están registrados en el diccionario y que podrían considerarse candidatos a neologismo para la actualización de este tipo de fuentes lexicográficas.

Para tales fines, el método propuesto contempla dos etapas: en la primera etapa, el algoritmo busca extraer los verbos registrados en el corpus a partir de su estructura morfológica y contrastarlos con una fuente de referencia, que en este caso fue el *Diccionario de la lengua española (DLE)* (RAE, 2014); y en la segunda, el algoritmo busca eliminar el ruido que pudiese haber en la detección de verbos neológicos y, para ello, somete los verbos detectados en la etapa anterior que no aparecen registrados en el diccionario a una serie de pruebas para determinar si se trata efectivamente de verbos neológicos o si son errores ortográficos.

Para la primera etapa, la detección se hace a partir de la distinción entre la raíz del verbo y su flexión (RAE y ASALE, 2009). Así, el algoritmo busca en un listado de formas extraídas del corpus todas aquellas cuya terminación coincide con una desinencia de un verbo regular. Cuando reconoce una raíz asociada a múltiples desinencias diferentes, reconstruye la forma de infinitivo y la busca en el corpus. Si la encuentra, asume la existencia de un verbo. Por ejemplo, al detectar la raíz *suprim-*, asociada a desinencias como *imos*, *iendo*, *iera*, entre otras, supone la existencia de un verbo y lo comprueba reconstruyendo el infinitivo que se corresponde con la raíz. En este caso, el algoritmo prueba la combinación **suprimar*, **suprimer* y *suprimir.* Al hallar la existencia de la forma *suprimir* en el corpus, comprueba su hipótesis inicial y clasifica esta como un verbo. Finalmente, el listado de verbos detectados en el corpus mediante el

[196] Profesora de Castellano y Comunicación y licenciada en Lengua y Literatura hispánica, titulada de la Pontificia Universidad Católica de Valparaíso. Desde el año 2018 ejerce como docente universitaria en Universidad Viña del Mar y recientemente se incorporó como profesora agregada en la Pontificia Universidad Católica de Valparaíso. Además, actualmente se desempeña como personal técnico en proyecto Fondecyt sobre polisemia regular de sustantivos del español dirigido por la Dra. Irene Renau Araque; como diseñadora instruccional en consultora Mastercap Spa y como miembro del grupo de investigación Tecling.

método descrito se compara con una fuente lexicográfica, lo que constituye el criterio para discriminar verbos neológicos y no neológicos en esta etapa inicial, de acuerdo a los criterios propuestos por Cabré (1999), quien propone que una palabra puede considerarse un neologismo si su uso es reciente, si no figura en los diccionarios, si presenta inestabilidad formal y si es percibida como nueva por los hablantes.

Para la segunda etapa, todos aquellos supuestos verbos detectados en la primera etapa que no se encontraron registrados en la fuente de referencia fueron sometidos a una serie de pruebas para detectar errores ortográficos y reunir únicamente candidatos a verbos neológicos. Estas pruebas, estratégicamente organizadas de acuerdo al coste computacional de cada una, atienden a diferentes aspectos de la forma verbal candidata a neologismo, tales como extensión, prefijación, omisión y sustitución de letras y similitud distribucional.

En este estudio se procesaron 2.954.973 formas extraídas del corpus EsTenTen (Kilgarriff y Renau 2013). Tras la primera etapa del método, se obtuvieron 10.034 candidatos a verbo, de los cuales 5.685, es decir, el 56.65%, se encontraba registrado en el DLE y 4.349, es decir, el 43.34% no lo estaba, por lo que fue sometido a la segunda etapa del método. Una vez aplicadas las pruebas para la eliminación de errores ortográficos en la segunda etapa, quedaron 774 verbos (el 17.79% de la lista) como candidatos a neologismo. Estos son producto de diferentes mecanismos de formación de palabras (Cabré, 2015), tales como prefijación (*desactualizar*), sufijación (*chocolatear*), parasíntesis (*desvirtualizar*) y préstamo (*loguear*). En suma, la precisión en la detección de verbos neológicos fue de un 88% y la cobertura de un 76%.

La aparición de unidades neológicas y su respectivo análisis morfológico suele representar una dificultad en tareas de procesamiento automático del lenguaje como, por ejemplo, el etiquetado morfosintáctico (Cook, 2010; Janssen, 2009; Renouf, 2016). El método propuesto tiene la particularidad de utilizar la estructura morfológica de los verbos y ofrecer múltiples posibilidades de reducir el ruido que suele haber en los análisis morfológicos que se proponen detectar verbos neológicos, particularmente en la aplicación del criterio lexicográfico, debido a la gran cantidad de errores de digitación u ortográficos que se encuentran en un corpus. Por ejemplo, el método propuesto nos permite determinar que unidades como **habrir* o **contatar*, que evidentemente no están registradas en el diccionario, no son verbos neológicos, sino verbos escritos con errores ortográficos. En este sentido, este método puede servir, por ejemplo, para la recopilación sistemática de nuevos verbos que podrían, eventualmente, ser incorporados en algunas fuentes lexicográficas. En definitiva, se considera un aporte en el campo del procesamiento del lenguaje natural, la lexicografía computacional y la detección automática de neologismos.

Bibliografía

Algeo, J. (1993). Desuetude among New English Words. *International Journal of Lexicography*, *6*(4), 281-293.

Cabré, M. T. (1999). *Terminology. Theory, Methods and Applications.* John Benjamins.

Cabré, M. T. (2015). Bases para una teoría de los neologismos léxicos: primeras reflexiones En Alves, I. y E. Simões (eds), *Neologia das Línguas Românicas* (pp. 79-107). CAPES, Humanitas.

Cook, C. (2010). *Exploiting Linguistic Knowledge to Infer Properties of Neologisms* [tesis de doctorado, Universidad de Toronto].

Janssen, M. (2009). Detección de neologismos: una perspectiva computacional. *Debate Terminológico*, 5, 68-75.

Kilgarriff, A., y Renau, I. (2013). esTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences*, 95, 12-19.

Real Academia Española (RAE) y Asociación de Academias de la Lengua Española (ASALE). (2009). *Nueva gramática de la lengua española*. Madrid: Espasa Calpe.

Real Academia Española. (2014). Diccionario de la lengua española (23.ª ed). Espasa.

Renouf, A. (2016). Big Data and its Consequences for Neology. *Neologica*, 10, 15-37.

Rey, A. (1976). Néologisme, un pseudo-concept? *Cahiers de Lexicologie*, *28*(1), 3-17.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Enriquecimiento semántico para la comparación de textos cortos durante la evaluación de competencias laborales

Laynier Antonio Piedra Diéguez[197], Surayne Torres López & José Eladio Medina Pagola
[laynier@uci.cu]
Universidad de las Ciencias Informáticas, Cuba

Las organizaciones que gestionan proyectos según crecen y se complejizan, necesitan agilizar y/o automatizar la mayor cantidad posible de procesos. Un proceso clave a automatizar es la evaluación de competencias laborales dentro del área de la Gestión de Recursos Humanos. Esta es una de las áreas para las que se han desarrollado modelos computacionales con mayor o menor aplicabilidad según el entorno, aprovechando estos grandes cúmulos de datos organizados.

Un ejemplo de automatización de este proceso se ha utilizado sobre la herramienta de gestión de proyectos XEDRO-GESPRO, donde a partir del análisis de las evidencias, se genera un modelo evaluación de competencias laborales en proyectos, aprovechando técnicas de similitud textual (similitud coseno) propias del enfoque vectorial de representación de textos. Esto es posible ya que tanto las evidencias del desempeño (tareas realizadas), como las competencias esperadas para el rol que el trabajador desempeña (competencias y sus dimensiones) se encuentran registradas en forma de texto dentro de la herramienta.

Este modelo resultó ser superior a los métodos tradicionales en cuanto a eficiencia, sin embargo, el enfoque vectorial presenta algunos problemas inherentes al procesamiento del lenguaje, sobre todo para textos cortos como el caso mencionado, el cual relaciona encabezados de tareas con alrededor de una decena de palabras o menos, con dimensiones de competencias laborales para un rol en el proyecto (una o dos decenas de palabras). Las limitaciones para manejar textos de esta longitud afectan negativamente la eficacia del modelo para apoyar la evaluación de competencias laborales, haciendo que:

- Quedan tareas sin evaluar de acuerdo a la competencia adecuada o quedan fuera del análisis tareas que sí demuestran determinada competencia [un grupo de tareas no alcanza el valor de similitud suficiente].

[197] Graduado en 2013 como Ingeniero en Ciencias informáticas. Máster en Gestión de Proyectos Informáticos desde 2019. Ha trabajado en el desarrollo de objetos de aprendizaje para la enseñanza virtual y la Internet de las Cosas (IoT). Su actividad fundamental es el desarrollo de software para la Gestión de Proyectos y las Arquitecturas Empresariales. Sus ponencias se han presentado en eventos como el 1er Congreso Internacional Cibersociedad 2017, la Conferencia Científica Internacional UCIENCIA de la Universidad de las Ciencias Informáticas en 2018, y la Conferencia Iberoamericana de Ingeniería de Software(CIbSE 2019). Se desempeña fundamentalmente en los campos relacionados con la Gestión de Proyectos, la Gestión de Recursos Humanos y el Procesamiento de Lenguaje Natural(PLN).

- Se realiza el cálculo de indicadores de rendimiento, calidad, eficiencia, etc., de acuerdo con una correspondencia incorrecta entre tarea y competencias [la RNA calcula los indicadores en el sistema en base a pares tarea-competencia incorrectos].
- Quedan competencias demostradas por el trabajador sin considerar por el sistema debido a que no detecta correspondencia con tarea alguna [efecto secundario del primer punto].

La introducción de estas imprecisiones termina en que el sistema proponga una evaluación de desempeño incorrecta para el trabajador. Se ha planteado por tanto como problema científico para esta investigación: ¿cómo mejorar la eficacia en la comparación de textos cortos en el proceso de evaluación de competencias laborales dentro de la gestión de proyectos?

Siendo el objeto de estudio, los textos cortos generados por las herramientas de gestión de proyectos, se ha planteado como objetivo general: desarrollar un método de enriquecimiento semántico para aumentar la eficacia de la comparación de textos cortos durante la evaluación de competencias laborales en herramientas de gestión de proyectos.

Algunos modelos basados en redes neuronales artificiales utilizados convenientemente son capaces de detectar relaciones semánticas entre las palabras, analizándolas a través de sus contextos y de esta forma, de lidiar de forma efectiva con estos problemas de similitud de textos cortos. Se planteó, por tanto, como hipótesis para esta investigación que: la aplicación de un método de enriquecimiento semántico de textos cortos aumentará la eficacia de su comparación durante la evaluación de competencias laborales en herramientas de gestión de proyectos.

El método de enriquecimiento semántico se aplicó como se muestra en la siguiente figura:

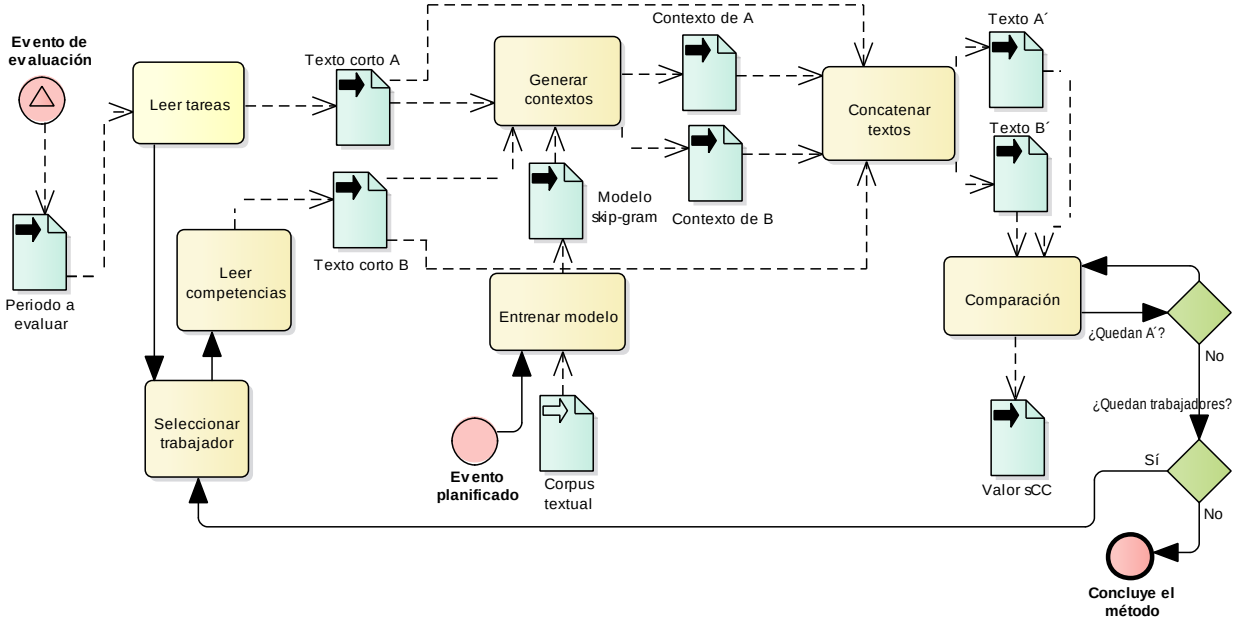


Figura 1. Enriquecimiento semántico aplicado a la evaluación de competencias laborales.

Los resultados experimentales validaron la hipótesis planteada demostrando que la aplicación del método de enriquecimiento semántico consigue elevar sustancialmente la eficacia de la comparación de textos cortos durante la evaluación. Las pruebas aplicadas demostraron un rendimiento muy superior al método actualmente utilizado en la herramienta XEDRO-GESPRO, con una exactitud, precisión y exhaustividad que en algunos casos duplica o triplica la medida anterior, como puede apreciarse en la Figura 2.

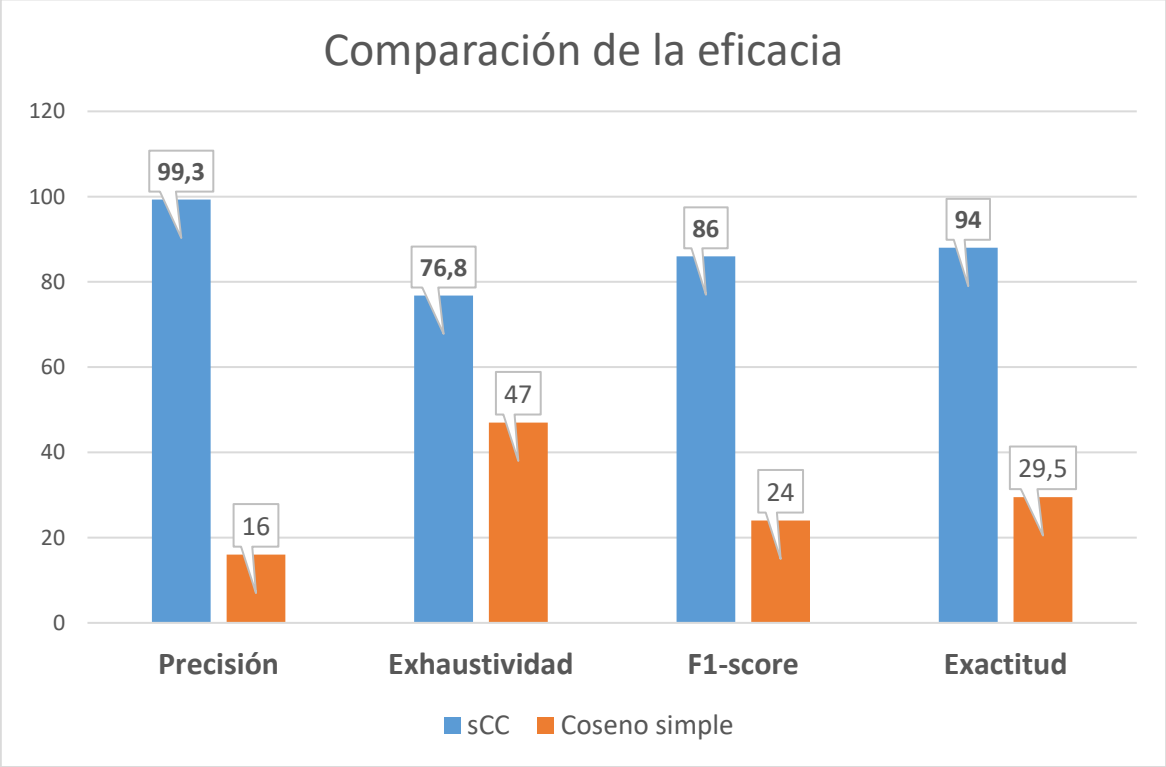


Figura 2. Comparación del método propuesto(sCC) con el método anterior (Coseno simple).

A los valores de similitud obtenidos por ambos métodos se les aplicaron pruebas estadísticas resultando que las diferencias son significativas estadísticamente. Debido a lo anterior se puede concluir que ll método presentado puede lidiar efectivamente con la similitud de textos cortos, mejorando su longitud y su carga semántica, así detecta efectivamente la relación entre los textos que compara, cubriendo ambas posibilidades de similitud: estructura y significado.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Estilector.com: herramienta de ayuda a la redacción en castellano

Nicolás Acosta[198]
[nmac.1996@gmail.com]
Universidad Nacional de Cuyo, Argentina
Rogelio Nazar[199]
[rogelio.nazar@pucv.cl]
Pontificia Universidad Católica de Valparaíso, Chile

En la presentación expondremos una descripción de la última versión de Estilector.com, un *software* de asistencia a la redacción de textos académicos en castellano. Este programa se orientó, en sus inicios, como una herramienta de ayuda a estudiantes universitarios que recién comienzan sus estudios. La idea era proporcionarles una forma de guía para la redacción y corrección de sus primeros trabajos académicos, en relación con aspectos de estilo y, en menor medida, ortografía y gramática. Sin embargo, y de manera inesperada, el resultado fue que otros usuarios se añadieron a estos, y con perfiles muy variados. La Figura 1 muestra la cantidad de textos procesados por Estilector en cada año desde su puesta en marcha. En 2020, el programa está procesando cerca de 250.000 trabajos al mes, que son enviados por usuarios provenientes de decenas de países distintos. La Tabla 1 exhibe el listado de los 15 países que más lo utilizan y el número total de textos enviados hasta ahora desde cada uno. Esta nueva situación de uso de la herramienta obliga a un replanteo general del tipo de necesidades que debe atender, aunque siempre con el mismo objetivo inicial de asistir al aprendizaje y mejoramiento de las habilidades de escritura.

[198] Estudiante de Licenciatura en Letras, orientación Lingüística y Profesorado de Grado Universitario en Letras en la Universidad Nacional de Cuyo, Mendoza, Argentina. Es miembro del grupo de investigación Tecling, perteneciente al Instituto de Literatura y Ciencias del Lenguaje (PUCV, Chile). Su trabajo está enfocado en el desarrollo de soluciones informáticas para procesamiento de lenguaje natural.

[199] Profesor del Instituto de Literatura y Ciencias del Lenguaje de la Pontificia Universidad Católica de Valparaíso. Se especaliza en Procesamiento del Lenguaje Natural y Lingüística Computacional. Su interés principal es el léxico, la terminología y el discurso. Actualmente dirige el grupo tecling.com, dedicado al desarrollo de tecnologías lingüísticas.

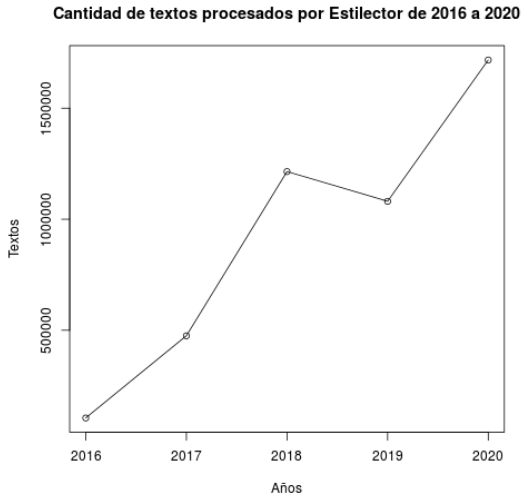


Figura 1. Gráfica de la cantidad de textos procesados por año (enero 2016-septiembre 2020)

Tabla 1. Lista de países que más usan Estilector.com (periodo enero 2016-septiembre 2020)

País	Cantidad de textos procesados hasta el año 2020
Colombia	1.208.134
México	862.759
Ecuador	488.560
Chile	459.822
España	345.427
Perú	304.795
Argentina	197.637
Venezuela	165.953
El Salvador	122.113
Guatemala	93.882

Costa Rica	62.600
Estados Unidos	48.252
República Dominicana	45.701
Panamá	43.539
Bolivia	40.137
...	...

Estilector.com es una página web que acepta un texto como entrada y lo procesa en distintos niveles. En un primer nivel, desglosa el texto para extraer los nombres propios y contar la cantidad de palabras, caracteres y párrafos. A partir de allí, el *input* es filtrado por módulos que corresponden a las reglas de estilo, ortografía y gramática. El resultado del procesamiento del texto se exhibe en pantalla con un encabezado que exhibe los resultados del análisis cuantitativo y una anotación al margen en el que se presentan, una tras otra, las marcas de los errores detectados, los consejos para mejorar cada aspecto y la correspondiente fundamentación bibliográfica. La ventaja principal del *software* es que los problemas señalados por estos módulos no son considerados por los principales procesadores de texto, como el doble espaciado, la repetición de vocabulario, el uso incorrecto de puntuación, la deficiente organización de párrafos, el dequeísmo, el mal uso de mayúsculas, el uso excesivo de la primera persona, entre varios otros aspectos.

Estilector cuenta con una arquitectura modular, es decir, está compuesto por un *script* central que ejecuta distintos módulos, para así optimizar desarrollo y ejecución. En esta estructura, primero se presenta la configuración y las variables globales, que son transversales a todos los niveles de funcionamiento. Luego se presentan, organizados en categorías, los distintos conjuntos de reglas. Esto representa una ventaja, ya que cada módulo del programa ofrece un nivel de complejidad distinto en cuanto a programación. Gracias a un sistema de plantillas, personas con conocimientos menos avanzados de informática pueden igualmente colaborar con la implementación de reglas, por lo que se facilita de esta manera el trabajo en equipo. Por último, existe el módulo para la web, que traduce los resultados del programa en HTML para su exhibición en un navegador. Esta última separación modular también simplifica la tarea porque así un diseñador gráfico puede crear una nueva interfaz sin necesidad de preocuparse por el resto del código.

En este momento del desarrollo de la herramienta existen líneas de trabajo presente y futuro. Un nuevo sistema de recepción de *feedback* del usuario nos ha mostrado un panorama

completo de las tareas urgentes. Por un lado, la situación actual exige la implementación de una clasificación del escrito por género, ya que no es posible analizar todos de la misma forma. Además de esto, estamos trabajando en un mejor marcado de los títulos, tarea que consideramos un desafío en lingüística computacional. Más adelante abordaremos otros aspectos como, por ejemplo, índices de riqueza léxica, extracción de información, detección de similitudes entre las distintas partes del texto, análisis de sentimiento para advertir si se presenta un tono muy negativo e identificación de la bibliografía, con el objeto de detectar posibles inconsistencias en las citas.

Estilector.com es una herramienta que sigue en elaboración. En los últimos cuatro años se ha logrado proporcionar un *software* útil para detectar errores comunes en la redacción de textos. Sin embargo, como se ha enumerado en el párrafo anterior, se presentan dificultades que aún hay que resolver, como los falsos positivos del sistema y los aspectos que todavía no son detectados por el programa. Su crecimiento constante y sostenido, en desarrollo y en uso, permitirá que estos problemas puedan tener solución en el mediano plazo.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Formación de neologismos jurídicos en maya yucateco: una aproximación

César Antonio Aguilar[200]
[cesar.aguilar72@gmail.com]
Pontificia Universidad Católica de Chile, Chile
H. Antonio García Zúñiga[201]
[agartzea@gmail.com]
Instituto Nacional de Antropología e Historia, México

En el presente trabajo se describen y analizan los procedimientos lingüísticos mediante los cuales los hablantes del maya yucateco (México y Belice) crean neologismos (Ahmad, 2000; Sablayrolles, 2002; Mejri y Sablayrolles, 2011; Kerremans, 2015; Matiello, 2017), con el propósito de referir de una manera más completa la realidad en la que se encuentran insertos en la actualidad.

El estudio se hace en uno de los ámbitos con mayor interés y mayor producción académica de la última década: el jurídico. En efecto, dada la situación carcelaria que se vive en México, en la que varios presos no cuentan con un juicio porque no hablan español –entre los numeroso vacíos legales que se pueden mencionar–, desde el 2003 se ha establecido la *Ley General de Derechos Lingüísticos de los Pueblos Indígenas*, junto con la creación del *Instituto Nacional de Lenguas Indígenas* (INALI: www.inali.gob.mx), se está impulsando la capacitación, certificación y acreditación de intérpretes, así como el ejercicio de la traducción de normas, leyes, acuerdos y reglamentos, tanto nacionales como internacionales. Cabe destacar que, como consecuencia de su poca diversidad interna, el maya es la lengua indígena en México con el mayor número de hablantes (859,000). Además, goza de cierto prestigio social, el cual le ha permitido que consolide un movimiento literario y encabece la profesionalización de intérpretes y traductores. Por esto último, existe una gran cantidad de material que se dispone para revisar las innovaciones léxicas.

La investigación toma en cuenta exclusivamente los procedimientos morfológicos y semánticos involucrados en los neologismos, siguiendo en esto propuestas como las de Bastuji

[200] Doctor en Lingüística por la Universidad Nacional Autónoma de México. Es profesor asistente de la Pontificia Universidad Católica de Chile. Su interés está en la procesamiento de lenguaje natural, concretamente en el desarrollo de ontologías y taxonomías basadas en información textual, la lexicografía computacional y la semántica formal.

[201] Lingüista por la Escuela Nacional de Antropología e Historia con estudios de maestría en Lingüística Hispánica y Metodología de la Ciencia. Miembro honorario de la Academia de la Lengua y Cultura Mayas de Quintana Roo. Tiene interés en el cambio histórico, la educación y las ciencias cognitivas. Trabaja lenguas mayas, español de comunidades afromexicanas y romances monorizados.

(1974), Mejri (2011) y Guerrero (2017), por mencionar algunos. Por ejemplo, para el término *derecho* se han desarrollado tres términos:

a. A'almajt'aan

b.Tojbe'enil

c. U páajtalil

En principio pareciera que estas alternativas tienen un contexto bien definido. Sin embargo, su formación destaca varias particularidades. En (1a) está compuesto por los siguientes elementos: decir-RESULTATIVO-lengua (*el idioma que resulta de decir cosas*); por su parte, (1b) se comprender como: derecho-hacer-RELACIONAL (*lo que hace lo recto*); finalmente, (1c) se entiende como una secuencia del tipo: POSESIVO poderRELACIONAL (*lo que se puede*). Más que las diferencias, lo que aquí se propone es una clasificación de estructuras con base en su integración, nivel de composición y relieve semántico (perfil/base).

Los resultados se justifican con base en una revisión de los parámetros mediante los cuales en esta lengua amerindia se desarrollan neologismos frente a la posibilidad de incorporar préstamos del español, o bien, recuperar léxico nativo registrado durante el periodo colonial.

Para registrar estos datos en un formato electrónico, con miras a configurar un corpus de prueba que sea procesable a través de recursos computacionales, se empleará la *suite* de herramientas **Toolbox** (https://software.sil.org/toolbox/), desarrollada por el *Instituto Lingüístico de Verano* (SIL: www.sil.org). Dicha *suite* ha demostrado ser útil para el análisis computacional de lenguas minoritarias (Robinson, Aumann y Bird, 2007). Asimismo, se contrastará tanto la transcripción e interpretación de los datos registrados con los que han compilado otros investigadores de maya yucateco, haciendo uso de los **Open Language Archives** (http://olac.ldc.upenn.edu/), uno de los mayores repositorios de acceso libre con información detallada de lenguas indígenas de distintas partes del mundo.

Con este estudio se espera poder contribuir al uso de las tecnologías computacionales para el estudio y la conservación de las lenguas amerindias.

Referencias

Ahmad, K. (2000): "Neologisms, Nonces and Word Formation", en Heid, U.; Evert, S.; Lehmann, E.; y Rohrer, C. (eds.), *Proceedings of the 9th EURALEX International Congress*, Stuttgart, Universitat Stuttgart. Vol II: 711-730.

Bastuji, J. (1974): "Aspects de la néologie sémantique", *Langages*, No. 36: 6-19. Guerrero, G. (2017): "Nuevas orientaciones de la terminología y de la neología en el ámbito de la semántica léxica", *RILCE*, 33(3): 1385-1415.

Kerremans, D. (2012): *A Web of New Words: A Corpus-Based Study of the Conventionalization Process of English Neologisms*, Frankfurt am Main, Peter Lang.

Mattiello, E. (2017): *Analogy in Word-formation*, Berlin, Mouton de Gruyter.

Mejri, S.; y Sablayrolles, J.-F. (2011): "Néologie, nouveaux modèles théoriques et NTIC", *Langages*, No. 183: 3-9.

Mejri, S. (2011): "Néologie et unité lexicale: renouvellement théorique, polylexicalité et emploi", *Langages*, No. 183: 25-37.

Robinson, S.; Aumann, G.; y Bird, S. (2007): "Managing Fieldwork Data with Toolbox and the Natural Language Toolkit", *Language Documentation & Conservation*, 1(1): 44-57.

Sablayrolles, J.-F. (2002): "Fondements théoriques des difficultés pratiques du traitement des neologisms", *Revue française de linguistique appliquée*, 7(1): 97-111.

Recursos electrónicos

Field Linguist's Toolbox: https://software.sil.org/toolbox. Herramienta para el anotado y análisis de datos lingüísticos en varios niveles (fonético, morfológico, sintáctico y semántico).

Open Language Archives: http://olac.ldc.upenn.edu. Repositorio de datos lingüísticos pertenecientes a distintas lenguas del mundo. Administrado por el **Linguistic Data Consortium**, de la **University of Pennsylvania**.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

GARCÍA SILICIO: aplicación computacional de análisis métrico sílabas, versos y poemas

Ricardo Martínez-Gamboa[202]
[ricardomartinezg@gmail.com]
Universidad Diego Portales, Chile

Introducción

El análisis computacional de la literatura permite el desarrollo de métodos para el procesamiento del lenguaje natural en contextos y géneros estéticos, lo que facilita posteriormente la interpretación de los mismos y puede dar cuenta de las propiedades estilométricas de la narrativa, el drama o la lírica. Estos procedimientos abordan los textos de intención poética de manera que sería inabordable por el trabajo manual con los mismos y, en consecuencia, dan paso al establecimiento de patrones, tendencias y características robustamente establecidas.

Método

Se presenta García Silicio, una aplicación computacional implementada en Perl que tiene por objetivo el análisis descriptivo de la métrica de sílabas, versos y poemas, en especial del castellano de Chile y de la poesía chilena. García Silicio es la continuación del proyecto SICAM ("Silabizador del Castellano Moderno"), desarrollado por Sadowsky y Martínez (2003- 2006), que, mediante reglas generativas propias, realiza la separación en sílabas de palabras castellanas. García Silicio se inspira en un personaje de la novela "Los Detectives Salvajes", escrita en 1997 por Roberto Bolaño, donde García Madero conocía todas y cada una de las formas de versificación de la tradición hispana. García Silicio, opera siguiendo rigurosos sistemas reglares y, en el estado actual, es capaz de: 1) separar en sílabas palabras individuales, 2) contar las sílabas gramaticales de un verso individual y grupos de versos que forman estrofas y poemas más extensos, así como el realizar el conteo que añade una sílaba virtual a los versos que finalizan con una palabra aguda, o restar una sílaba a los versos que terminan con una palabra esdrújula, 3) realizar la separación poética de sílabas aplicando las licencias métricas como sinalefas, dialefas, hiatos, sinéresis, diéresis, 4) determinar el axis rítmico de los versos, siguiendo el modelo de Ibarra que calcula ritmos troqueos y yambos, y, asimismo, determinar si las sílabas que componen un verso son rítmicas, extrarrítmicas o antirrítmicas, 5) determinar el tipo de verso de acuerdo con las clasificaciones tradicionales (por ejemplo: pentasílabo adónico o endecasílabo heroico), 6) determinar la rima de cada verso individual, 7) determinar el tipo de rima que se usa en una estrofa o poema completo (por ejemplo: rima

[202] Profesor Universidad Diego Portales, PhD en Lingüística (PUCV), MSc en Estudios Cognitivos (U Chile), Licenciado en Lingüística (U Chile)

consonante, rima semiconsonante, rima asonante, verso libre), y 8) determinar el tipo de estrofa (por ejemplo: redondilla, serventesio) y el tipo de poema (por ejemplo: soneto, décima).

Resultados

La aplicación García Silicio permite el análisis pormenorizado y descriptivo de los poemas de manera masiva, por lo que será utilizado para el establecimiento de patrones cuantitativos poéticos métricos de un corpus masivo de poemas chilenos de los siglos XX y XXI, respondiendo a interrogantes como: ¿cuáles son los ritmos preferidos de la poesía de Pablo Neruda? o ¿cuáles son los tipos de rima más utilizados por la Generación del 50? o ¿cuáles son las palabras más típicamente antirrítmicas en los poemas de Violeta o Nicanor Parra?

Conclusiones

La herramienta García Silicio se pondrá a disposición de las y los investigadores de la literatura y en especial de la lírica y la métrica, como un programa de fuente abierta, tanto para el análisis puntual de sílabas, versos y poemas, como para el trabajo con corpora de poetas específicos, periodos literarios y escuelas líricas.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Identificación automática de la elipsis nominal en español

Hazel Barahona[203] **& Walter Koza**[204]
[hhgg09@yahoo.es | walter.koza@pucv.cl]
Pontificia Universidad Católica de Valparaíso, Chile

Introducción

El presente trabajo describe la modelización computacional de la elipsis nominal del español. Para ello, se parte de los supuestos teóricos de la gramática generativa como los estudios de Merchant (1999); Sag y Hankamer (1984); Culicover y Jackendoff (2019); pero, en especial, la propuesta de Saab (2008, 2019). Concretamente, se describe un algoritmo desarrollado para el reconocimiento automático de la elipsis y la reposición del elemento elidido en textos de lenguaje natural. La elipsis consiste en un mecanismo gramatical evitativo de la redundancia léxica (Brucart, 1999), en el que un elemento de la oración no se pronuncia, es decir, se elide, bajo ciertas restricciones sintácticas. Al respecto Saab (2008) y Merchant (2001) señalan que, en este fenómeno, el elemento silenciado se reconoce por medio de una relación de identidad estructural con un elemento antecedente que permite identificar dicho elemento y su posición dentro de la oración.

Este fenómeno ha sido abordado desde diferentes perspectivas, las cuales comprenden posturas sintácticas (Chomsky 1965, 1981, 1995; Lobeck, 1995; Lasnik, 1999; Merchant, 1999) y semánticas (Sag y Hankamer, 1984; Merchant, 2001). También se encuentran trabajos dentro del marco de la Head-Driven Phrase Structure Grammar (Ginzburg y Miller, 2019) y Simpler Syntax (Culicover y Jackendoff, 2019). En relación con las investigaciones ligadas a la lingüística computacional, los trabajos de Hardt (1993), Mitkov (2002), Nielsen (2004), Rello, Baeza-Yates y McShane y Babkin (2015), se han encargado de analizar la elipsis a través del etiquetado de corpus desarrollando metodologías para la identificación de la elipsis y la recuperación del elemento elidido. Sin embargo, estas propuestas no suelen tener en cuenta los

[203] Estudiante del Doctorado en Lingüística en la Pontificia Universidad Católica de Valparaíso. Magíster en Lingüística en la Universidad de Costa Rica y Licenciada en Literatura y Lingüística con énfasis en español en la Universidad Nacional de Costa Rica. Personal técnico en el Proyecto FONDECyT 1171033 (Comisión Nacional de Investigación Científica y Tecnológica, Chile) dirigido a la traducción automática del conocimiento médico. Áreas de interés en lingüística computacional y gramática del español.

[204] Doctor en Humanidades y Artes con mención en Lingüística, egresado de la Universidad Nacional de Rosario (Argentina). Profesor Auxiliar del Instituto de Literatura y Ciencias del Lenguaje, de la Pontificia Universidad Católica de Valparaíso, Chile. Es Investigador Responsable del Proyecto FONDECyT 1171033 (Comisión Nacional de Investigación Científica y Tecnológica, Chile), orientado al desarrollo de técnicas para la traducción automática del conocimiento médico. Sus intereses se centran en la lingüística computacional, la gramática del español y las alteraciones del lenguaje, áreas en las que ha publicado diversos artículos.

mecanismos sintácticos implícitos. A tales efectos, el presente trabajo se propone formalizar la elipsis nominal, a partir del procesamiento de información morfosintáctica.

La metodología fue probada en textos especializados del dominio médico extraídos de revistas especializadas. Los resultados obtenidos (93% de precisión; 83% de cobertura; 88 % de medida f) demuestran que el algoritmo desarrollado es útil para el reconocimiento de elipsis en textos de lenguaje natural, a la vez que demuestra que las propuestas teóricas de corte generativista son adecuadas para el análisis de este fenómeno.

Metodología

Descripción formal de la elipsis

Se analizaron los contextos de aparición de los elementos elididos y se establecieron las siguientes posibilidades para la EN:

Elisión del núcleo de un sintagma nominal (SN) complemento de un sintagma determinante (SD) coordinado: *el hijo de Juan y el hijo de Pedro*

Elisión de un núcleo de SN complemento de SD argumento de un predicado: *El presidente de Perú homenajeó al presidente de Francia*.

Elisión de un núcleo de SN complemento de SD argumento de un predicado ubicado en una oración subordinada: *Los jugadores de Boca corrieron a protestarle al árbitro, mientras los jugadores de Ríver festejaban*.

Elisión de un núcleo de SN complemento de SD argumento de un predicado preposicional periférico a la oración: *Según los jugadores de Boca, los jugadores de Ríver festejaron desmesuradamente.*

Una vez analizados los tipos de EN, se propone la siguiente descripción formal que ejemplifica la representación de la gramática en NooJ. Para los fines particulares de esta presentación, se ejemplifica la formalización de los casos de A.

A: $[[\text{PRENS}]_1$ NS $[\text{POSNS}]_1]$SDI Coord $[[\text{DETELIP}]_2$ N $[\text{POSELIP}]_2]$SDII

Elaboración de reglas de recuperación

Sobre la base de la descripción formal, se proponen las siguientes reglas de recuperación.

Dada una coordinación A^B, donde B es un SD de estructura equivalente a A, y presenta elisión del núcleo de su SN:

Cópiese raíz del núcleo y género del SN complemento de A

Cópiese número del Det de B

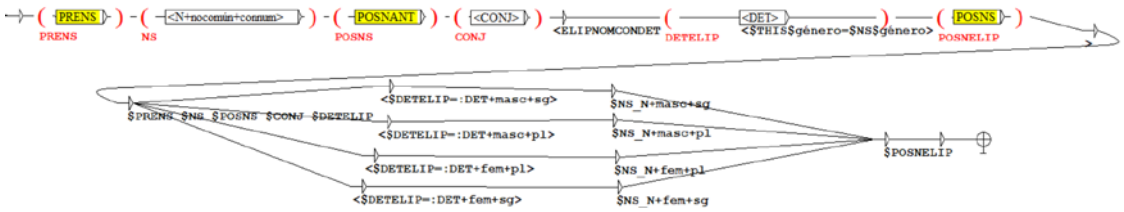
Dada una EAP A: P (α, β, γ, etcétera), con un argumento ubicado a la derecha de α que presenta una elisión del núcleo de su SN:

Cópiese raíz del SN complemento de α

Cópiese rasgos de número del determinante de β o γ, etcétera (según sea el argumento elidido)

Modelización computacional

Para el trabajo computacional, se recurrió a NooJ (Silberztein, 2016), que posibilita la elaboración de diversas herramientas de análisis lingüístico (diccionarios electrónicos y gramáticas morfológicas y sintácticas). En dicho programa, se creó un diccionario electrónico, el cual contiene los lemas del *DRAE*, a los que se les asignó información morfosintáctica. Posteriormente, se desarrollaron gramáticas informáticas para el reconocimiento y la generación de EN, como la que se muestra a continuación:



Las gramáticas estructuran un conjunto de variables (indicadas entre paréntesis rojos), que pueden ser categorías gramaticales (N, de nombre), o bien una gramática incrustada (indicada en color amarillo). Posteriormente se incluyen reglas de reescritura y una etiqueta de agrupamiento (ELIPNOMCONDET), encargada de etiquetar las expresiones que incluyen una EN.

Resultados

La metodología descrita se probó en un corpus compuesto por textos del dominio médico extraídos de revistas especializadas, en los cuales había un total de 145 elipsis nominales correspondientes a la descripción formal de A. Por lo que, los siguientes resultados se ajustan a ese reconocimiento automático.

Cobertura: 83%

Precisión: 93% de precisión

Medida F: 88%

Los errores y las omisiones identificados, en los cuales se trabaja, se deben principalmente a unidades no listadas en el diccionario electrónico, restricciones gramaticales relativas a la concordancia del género y el número.

Referencias

Brucart, J. (1999). La elipsis. En: Ignacio Bosque & Violeta Demonte (eds.) *Gramática Descriptiva de la lengua español* (vol. 1, cap. 43, pp. 2787-2863). Espasa-calpe.

Chomsky, N. (1695). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N. (1981). *Lectures on Government and Binding*. Foris.

Chomsky, N. (1995). *The Minimalist Program*. MIT Press.

Culicover, P. y Jackendoff, R. (2019). Ellipsis in simpler syntax. En Jeroen van Craenenbroeck and Tanja Temmerman (eds.) *The Oxford handbook of ellipsis* (pp. 172-187) Oxford University Press.

Ginzburg, J., Miller, P. (2019). Ellipsis in Head-Driven Phrase Structure Grammar. En Jeroen van Craenenbroeck and Tanja Temmerman (eds.) *The Oxford handbook of ellipsis* (pp. 75-121) Oxford University Press.

Hardt, D. (1993). *Verb Phrase Ellipsis: Form, Meaning, and Processing*. [Tesis de Doctorado, Universidad de Pennsylvania].

Lobeck, A. (1995). *Ellipsis: Functional Heads, Licensing and Identification*. Oxford University Press

McShane, M. y Babkin, P. (2016). Detection and Resolution of Verb Phrase Ellipsis. *LiLT*, (13), 1-36. CSLI Publications.

Merchant, J. (1999). *The Syntax of Silence: Sluicing, Islands and Identity in Ellipsis*. [Tesis de Doctorado, Universidad de Santa Cruz].

Merchant, J. (2001). *The syntax of silence: Sluicing, islands, and the theory of ellipsis.* Oxford University Press: Oxford

Mitkov, R. (2002). *The Oxford handbook of computational linguistics*. (Ed.) Oxford: Oxford University Press

Nielsen, L. (2004). Verb phrase ellipsis detection using automatically parsed text. En *COLING'04: Proceedings of the 20th international conference on Computational Linguistics*, Geneva.

Rello, L., Baeza-Yates, R. y Mitkov, R. (2012). *Elliphant: Improved Automatic Detection of Zero Subjects and Impersonal Constructions in Spanish*. Conferencia Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.

Saab, A. (2008). *Hacía una teoría de la identidad en la elipsis.* [Tesis de Doctorado. Universidad de Buenos Aires].

Saab, A. (2019). Nominal Ellipsis. En Jeroen van Craenenbroeck and Tanja Temmerman (eds.) *The Oxford handbook of ellipsis*, (pp. 526-561). Oxford University Press.

Sag, I., Hankamer, J. (1984). Toward a theory of anaphoric processing. *Linguistics and Philosophy*, (7), 325–345.

Silberztein, M. (2016). *Formalizing Natural Languages: The Nooj Approach*. Wiley-Iste

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Interfaz gráfica para la carga de eventos comunicativos en el corpus EspaDA-UNCuyo

Luis Aguirre[205] **& Nicolás Acosta**[206]
[nmac.1996@gmail.com]
Universidad Nacional de Cuyo, Argentina

El corpus EspaDA-UNCuyo (Corpus de Español de Discurso Académico de la Universidad Nacional de Cuyo) reúne textos de circulación efectiva en la universidad y busca ser una muestra representativa de los géneros que se producen en ella. Los textos del corpus son etiquetados con metadatos correspondientes a criterios externos que sirven para clasificarlos e identificarlos con la aplicación de consulta. Esta se encuentra disponible por medio de una interfaz web: <http://espada.uncuyo.edu.ar>. Por un lado, en la plataforma web se encuentra disponible para el acceso de cualquier usuario una interfaz de consulta al corpus. Por otro lado, se cuenta con un acceso restringido a una interfaz gráfica de usuario (GUI) para uso interno del grupo de investigadores. Esta posee las siguientes secciones: 1) Cargar un nuevo evento, 2) Eventos cargados, 3) Generar documento para subida al corpus y 4) Estadísticas. En esta comunicación presentaremos el desarrollo y funcionamiento de la interfaz de acceso restringido, con el objeto de analizar sus potencialidades y limitaciones para la labor de carga y registro asociadas con el corpus EspaDA-UNCuyo.

La sección "Cargar nuevo evento" constituye un formulario HTML dinámico mediante funciones JavaScript según los datos que va incorporando el usuario. Este posee una serie de campos relativos con la caracterización según criterios externos de un evento comunicativo (fecha del evento, modo, grado de interacción, ámbito, evento sometido a evaluación, palabra clave, etc.). Entre ellos, hay algunos campos obligatorios que tienen relación de dependencia, que se visualizan y se pueden completar solamente en función de la opción seleccionada en los siguientes criterios superordinados: medio (opciones: oral, escrito), ámbito (opciones: docencia, investigación, administrativo, institucional) y evento sometido a evaluación (opciones: sí, no). Una vez completado, verificado y enviado el formulario, se genera de manera automática un

[205] Doctor en Letras, Licenciado en Letras y Profesor en Lengua y Literatura por la Facultad de Filosofía y Letras de la UNCuyo (Mendoza, Argentina), donde se desempeña como profesor e investigador. Es subdirector del Instituto de Lingüística (FFyL, UNCuyo) y miembro representante del Instituto de Lectura y Escritura (FED, UNCuyo). Sus investigaciones están ligadas al estudio de la producción de textos académicos escritos, con asistencia de programas informáticos.

[206] Estudiante de Licenciatura en Letras, orientación Lingüística y Profesorado de Grado Universitario en Letras en la Universidad Nacional de Cuyo, Mendoza, Argentina. Es miembro del grupo de investigación Tecling, perteneciente al Instituto de Literatura y Ciencias del Lenguaje (PUCV, Chile). Su trabajo está enfocado en el desarrollo de soluciones informáticas para procesamiento de lenguaje natural.

identificador alfanumérico (ID) del evento. Posteriormente, el identificador y el evento cargado se registran en una base de datos SQL alojada en el servidor.

La sección "Eventos cargados", por su parte, recupera los registros de la base de datos y los exhibe en pantalla organizados en una tabla. Esta tabla puede descargarse en formato CSV para su procesamiento en aplicaciones externas (hojas de cálculo y paquetes de análisis estadístico).

La sección "Generar documento para subida" permite crear el archivo que se subirá al corpus. En esta sección el usuario dispone de una lista con el ID y la palabra clave de aquellos eventos que han sido registrados en la base de datos, pero que aún no han sido cargados para su consulta en el corpus almacenado en el servidor. De esa lista, se puede seleccionar el evento que se adaptará al formato de texto plano (UNIX con codificación UTF-8) con el que trabaja la aplicación web de consulta. La adaptación de formato requiere que el usuario pegue el texto del evento (transcripción de evento oral o texto digital escrito) en un campo de la interfaz destinado a tal efecto. Tras pegar el texto, automáticamente se generarán los metadatos a partir de los registros de la base de datos, se colocarán como encabezado del documento y se habilitará la descarga del documento para su subida al corpus.

Por último, la sección denominada "Estadísticas" permite visualizar la frecuencia y el porcentaje de eventos registrados en el corpus y de sus características (definidas según criterios externos). Las estadísticas se presentan bajo la forma de tablas y gráficos que pueden descargarse en formato CSV (tablas) y PNG (gráficos). Estas estadísticas se actualizan automáticamente con cada nueva carga de eventos realizada en el formulario.

Con el desarrollo e implementación de la interfaz gráfica se ha avanzado en la solución de la unificación de los datos, a partir de un proceso con restricciones que asegura que la carga de un documento se realice solo si se han generado los metadatos correspondientes. Esta plataforma web está programada para la identificación y clasificación de los textos del corpus. En ella, los investigadores deben completar manualmente un primer formulario de carga del evento. Luego, en un segundo formulario deben subir el texto correspondiente para que este quede incluido en la interfaz de consulta al corpus. En este último se incluirán los metadatos como encabezado en el archivo de texto plano.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

La metáfora orientacional BUENO/FELIZ ES ARRIBA – MALO/TRISTE ES ABAJO: análisis de sentimiento en 10 verbos del español con orientación vertical

Benjamín López Hidalgo[207]
[benjamin.lopez.hidalgo1@gmail.com]
Pontificia Universidad Católica de Valparaíso, Chile

La presente investigación propone un método cuantitativo para comprobar si las metáforas orientacionales BUENO/FELIZ ES ARRIBA y MALO/TRISTE ES ABAJO (Lakoff y Johnson, 1980, 1999; Lakoff, 1993; Kövecses, 2000; entre otros), en su dimensión lingüística, pueden ser observadas empíricamente en usos reales de un corpus de habla hispana.

Para llevar a cabo lo anterior, se estudió la orientación espacial ARRIBA/ABAJO a través de 10 verbos del español que, en su acepción prototípica, son definidos por los verbos *subir* (*ascender, elevar, escalar, levantar* y *trepar*) o *bajar* (*agachar, caer, derribar, descender* y *tumbar*) (Battaner, 2013). Por otra parte, los conceptos BUENO, MALO, FELIZ y TRISTE se midieron a través del análisis de la polaridad del léxico que coocurre con los verbos en una muestra de concordancias aleatorias, empleando un listado de unidades léxicas con polaridad (Martínez, 2018).

La hipótesis de investigación fue la siguiente: existe una correlación, coherente con los postulados provenientes de la lingüística cognitiva, entre orientación ARRIBA/ABAJO y los dominios BUENO, FELIZ/ MALO, TRISTE respectivamente. En otras palabras, se planteó que si la unidad verbal tiene una orientación espacial hacia ARRIBA va a presentar una mayor polaridad positiva en las unidades léxicas que cohabitan con esta y viceversa.

El corpus utilizado en esta investigación fue el corpus EsTenTen (Kilgarriff y Renau 2013), en concreto la versión Spanish Web 2011 (esTenTen11, Eu + Am) que consta de, aproximadamente, 10.000 millones de palabras, divididas entre el español peninsular y el español de Latinoamérica. De este corpus se extrajo una muestra de 5.000 concordancias aleatorias por cada uno de los 10 verbos (*ascender, elevar, escalar, levantar, trepar, agachar,*

[207] Licenciado en Lengua y literatura hispánica por la PUCV. Actualmente, es miembro del grupo de investigación Tecling y participa como personal técnico en proyecto FONDECYT 1191204 sobre polisemia regular. Anteriormente, tuvo experiencia como redactor lexicográfico en el proyecto FONDECYT 11140704 y también como co-investigador en el proyecto D.I. de Pregrado (2017) de la PUCV sobre metáfora y explotación. El ponente tiene la siguiente publicación: Renau, I.; Nazar, R.; Castro, A.; López, B.; Obreque, J. (2019). "Verbo y contexto de uso: un análisis basado en corpus con métodos cualitativos y cuantitativos". Revista Signos, vol. 52, núm. 101, pp. 878-901.

caer, derribar, descender y *tumbar*), lo que implica un total de 50.000 concordancias. El tamaño de la ventana de contexto fue de 10 palabras a la izquierda y 10 a la derecha.

La metodología de trabajo tuvo como primer paso el desarrollo de un algoritmo para detectar las unidades del lexicón de polaridad (Martínez, 2018) presentes en las concordancias de cada verbo, para luego evaluarlas junto con la polaridad final (positiva o negativa) de cada uno de los 10 verbos. El proceso se divide en tres momentos que son la lectura e instrumentalización del lexicón, un análisis a nivel local (para identificar la polaridad a nivel de concordancia) y un análisis a nivel total (para identificar la polaridad final de cada verbo). Para determinar la polaridad total del verbo se sumaron todos los resultados locales de este, es decir, el total de concordancias positivas y negativas. Si el verbo presentó más resultados locales negativos que positivos, se clasificó como negativo, y viceversa.

Los resultados fueron los siguientes: 10 de 10 casos tuvieron el resultado esperado en lo que se refiere a lo postulado en nuestra hipótesis de investigación. Es decir, los verbos de orientación ARRIBA fueron vinculados a una polaridad positiva (+) mientras que los verbos de orientación ABAJO fueron vinculados a una polaridad negativa (–) por el instrumento de medición. La probabilidad de que este resultado fuera por azar es de 0.001 de acuerdo al test binomial del software R (R Core Team, 2013).

A partir de los resultados, se puede confirmar la relación entre verbo con orientación espacial, ya sea ARRIBA o ABAJO, y la polaridad positiva/negativa respectivamente, coherente con los postulados de la metáfora orientacional (Lakoff & Johnson, 1986; Lakoff, 1993). Es decir, la orientación hacia arriba de una unidad léxica, generalmente, va a asociarse con una mayor polaridad positiva en las unidades léxicas acompañantes en contextos reales de uso, mientras que la orientación hacia abajo de una unidad léxica, generalmente, va a estar asociada con una mayor polaridad negativa en las unidades léxicas acompañantes en contextos reales de uso.

Bibliografía

Battaner, P. (2013). Diccionario de uso del español de América y España. Versión CD-ROM

Kilgarriff, A., & Renau, I. (2013). esTenTen, a vast web corpus of Peninsular and American Spanish. Procedia-Social and Behavioral Sciences, 95, 12-19.

Kövecses, Z. (2000). Metaphor and emotion. New York: Cambridge University press.

Lakoff, G. & Johnson, M. (1986). Metáforas de la vida cotidiana. Madrid: Cátedra.

Lakoff, G. (1993). "The contemporary theory of metaphor", en A. Ortony (ed.), Metaphor and thought (2a edición). Cambridge: Cambridge University Press, 202-251.

Lakoff, G. y Johnson, M. (1999). Philosophy in the flesh. The embodied mind and its challenge to western thought. Nueva York: Basic Books.

Martínez, R. (2018). La incidencia de las interacciones verbales en la configuración de la red social twitter: un análisis desde la polaridad, la novedad y el prestigio. Tesina para optar

al grado de Doctor en Lingüística. Pontificia Universidad Católica de Valparaíso, Viña del Mar.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Lenguaje ofensivo en redes sociales: definición de criterios lingüísticos para facilitar su detección automática

María José Díaz Torres[208], **Paulina A. Morán Méndez**[209]
[paulina.moranmz@udlap.mx]
Universidad de las Américas Puebla, México
Luis Villaseñor Pineda[210]
[luis.villasenor.pineda@gmail.com]
Instituto Nacional de Astrofísica Óptica y Electrónica, México

Nowadays, many online exchanges in social networking platforms are aggressive or offensive in nature, potentially leading to security problems, putting users at risk. To tackle this problem, the "Authorship and Aggressiveness Analysis in Twitter: case study in Mexican Spanish" (MEX-A3T) presented a task on aggressiveness detection in Mexican Spanish tweets in the IberEval 2018 workshop. The low performance reported by most participants confirmed the complexity of this task, thus the same task was proposed for the IberLEF 2019 and 2020 forums. After careful review of the datasets provided by the MEX-A3T, it was concluded that the corpus annotation played a very important role in the previous outcome. Therefore, the goal of the present research was to establish criteria to facilitate the identification of offensive language and allow non-experts in the subject to systematically label a corpus for a task of this nature. To this end, we defined the main linguistic features characteristic to offensive language manifested through social networks. As a result, we designed an easy-to-follow

[208] Licenciada en Idiomas Magna Cum Laude a través de la Beca de Excelencia Académica de la Universidad de las Américas Puebla. Participó en el Programa de Honores de la institución, culminando su contribución en una investigación conjunta con la Universidad de la República en Uruguay. Actualmente se desempeña como Técnico en Innovación de Contenido Digital y es fundadora del departamento de Innovación de Contenido de la empresa Faithlife de México. Asimismo, busca incursar en el campo de la lingüística computacional, particularmente en las disciplinas de aprendizaje automático y procesamiento de lenguaje natural.

[209] Licenciada en Idiomas de la Universidad de las Américas Puebla, México donde participó en el Programa de Honores con el proyecto: "Elaboración de un Diccionario del Subestándar Documentado por Corpus Electrónicos e Implementación de Herramientas Conexas de la Lingüística Computacional" creando un sistema de clasificación automática de tweets. Su área de interés es la lingüística computacional. Actualmente, se desempeña como Especialista en Procesamiento de Contenidos en BrandBastion, clasificando de manera conjunta con un sistema basado en aprendizaje automático, comentarios de redes sociales de distintas marcas con el fin de clasificarlas e incrementar la interacción con los clientes.

[210] Doctorado en Ciencias Computacionales en la Universidad Joseph Fourier de Grenoble, Francia. Su interés de investigación se centra en el tratamiento automático del lenguaje humano. Es autor de más de ciento cincuenta artículos internacionales en este campo, y es miembro del SNI (nivel 2), de la Academia Mexicana de Ciencias, de la Sociedad Mexicana de Inteligencia Artificial y de la Asociación Mexicana para el Procesamiento del Lenguaje Natural. Actualmente, es profesor en el Instituto Nacional de Astrofísica, Óptica y Electrónica.

diagram that was used to review and validate the MEX-A3T corpora. The use of this resource increased the kappa coefficient of inter-rater reliability from .58 to .92 in the training dataset, and the revised corpora contributed to a better performance of the automatic detection methods.

Keywords: aggressiveness detection, corpus, annotation, natural language processing, Spanish.

En la actualidad, las plataformas de redes sociales se han convertido en el medio de comunicación y expresión por excelencia. Lamentablemente, gran parte de estos intercambios en línea son de naturaleza agresiva u ofensiva, lo que puede implicar problemas de seguridad y poner en riesgo a los usuarios. A pesar de la relevancia y gravedad de este fenómeno, la detección automática de textos agresivos ha sido un tema escasamente estudiado en la comunidad científica. Es con esta motivación que surge el MEX-A3T "Authorship and aggressiveness analysis in Twitter: case study in Mexican Spanish", presentado originalmente en el workshop IberEval 2018. El MEX-A3T presenta una tarea de detección automática de tuits agresivos u ofensivos escritos en español de México, lo cual supone un reto extra a la clasificación al involucrar variación lingüística, pero que, al mismo tiempo, la acerca a la realidad. La dificultad de esta tarea se vio demostrada por el bajo rendimiento reportado por la mayoría de los participantes en la primera edición del MEX-A3T, y por ello se propuso exactamente la misma tarea para los foros de IberLEF 2019 y 2020.

Con esta motivación, a través de la presente investigación se buscó definir los rasgos lingüísticos principales que caracterizan al lenguaje ofensivo manifestado a través de las redes sociales, con el objetivo de establecer criterios que faciliten su identificación y así permitan a personas no expertas en el tema etiquetar sistemáticamente un corpus para esta tarea. Para poder lograr lo anterior, fue necesario primero definir claramente los conceptos de lenguaje ofensivo, agresivo y vulgar. Al tener claros los marcos de referencia, entonces se revisaron y analizaron los corpus de tuits ofensivos con los cuales cuenta el Laboratorio de Tecnologías del Lenguaje (LabTL) del INAOE. En función de las observaciones realizadas, se adecuaron los marcos actuales, con el fin de llegar a una caracterización. Durante este proceso se identificaron los elementos léxicos y semánticos más representativos de los mensajes agresivos, ofensivos o vulgares y así se diseñó una tipología que facilitó la categorización de manera clara y visual al usar un diagrama de flujo.

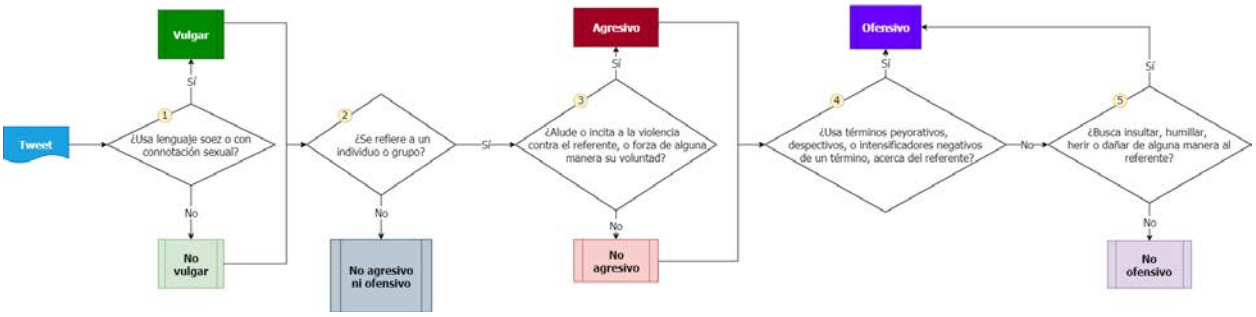


Figura 1. Diagrama de anotación para la categorización de lenguaje abusivo.

En el diagrama propuesto (Fig. 1) se incluyeron los conceptos de vulgaridad, agresividad y ofensividad como cualidades del mensaje. El diagrama comienza con la selección de un tweet a analizar, que a manera de ilustrar mejor el proceso se tomará un ejemplo del corpus: "No es que estés gorda, lo gordo se quita. Es tu cara de caballo". Inmediatamente se presenta la primera pregunta, que cuestiona si el tweet "usa lenguaje soez o con connotación sexual". Si la respuesta es positiva, entonces indica que es vulgar; de otra manera, no lo es, como es el caso del ejemplo. La siguiente pregunta lee: "¿se refiere a un individuo o grupo?". Esta pregunta busca hacer una temprana identificación de la agresividad, ya que, si el mensaje no tiene un blanco específico, carece de una característica esencial del mensaje agresivo y ofensivo. Por consiguiente, si la respuesta es negativa, ya no se seguirá con las siguientes preguntas. En el tweet ejemplo, el uso de la segunda persona da ya indicios claros de la existencia de un referente; además, las expresiones utilizadas, aluden a las características de un rostro que se entiende que es humano. Por el contrario, si la respuesta es positiva, el anotador continúa con la siguiente pregunta. Ésta, tiene la intención de confirmar si el tweet es agresivo al resaltar puntualmente características como el incitar a la violencia o forzar la conducta del referente. En nuestro ejemplo, no se encuentran estos elementos de agresividad. Posteriormente, para determinar si es ofensivo, se guía a observar si el tweet tiene la intención de humillar o insultar. Finalmente, se deberán tomar en cuenta la presencia de términos peyorativos, despectivos o intensificadores negativos y que busquen faltar al respeto usándolos. Con este último cuestionamiento termina el diagrama. Con estas últimas preguntas, el ejemplo al observar las expresiones "cara de caballo" y al resaltar la complexión de la persona es evidente que el objetivo sí es humillar al referente y que en combinación faltan al respeto. De esta manera, el anotador puede concluir que este mensaje no es vulgar o agresivo pero sí ofensivo.

En conclusión, el recurso generado sirvió para hacer una revisión y validación del corpus del MEX-A3T. Como resultado, el uso de este recurso aumentó el coeficiente kappa de confiabilidad inter-evaluador de .58 a .92 en el conjunto de datos de entrenamiento. Así mismo, para observar el efecto del nuevo etiquetado, y demostrar la consistencia del este, se pueden comparar los resultados reportados en el año 2018 y 2020 del foro de evaluación MEX-A3T. En comparación, existe una considerable mejora reportada en todos los sistemas participantes. No obstante, el efecto es evidente con los resultados reportados por el sistema desarrollado por INGEOTEC. En 2018, la primera edición del foro MEX-A3T, se utilizó el corpus etiquetado

con el conjunto de criterios original. Ese año el sistema del INGEOTEC obtuvo los mejores resultados con una medida F1 sobre la clase ofensiva de: 0.488. Los años subsecuentes, para comprobar el avance de los nuevos sistemas propuestos, se solicitó al equipo INGEOTEC que aplicará su sistema tal cual fue propuesto en el 2018. Para el año 2020, la primera vez que se usó el corpus con el nuevo etiquetado propuesto, el sistema de INGEOTEC alcanzó una medida F1 sobre la clase ofensiva de: 0.7468. Este resultado, claramente demuestra la pertinencia del nuevo esquema de etiquetado propuesto.

Referencias

Austin, J. (1962). *How to do things with words (2nd ed)*. J. O. Urmson & M. Sbisá (Eds.). Cambridge, MA: Harvard University Press.

Fromkin, V., Rodman, R., and Hyams, V. (2011). *An introduction to language, 9e*. Boston, MA: Wadsworth, Cengage Learning.

Graff, M., Miranda-Jiménez, S., Tellez, E.S., Moctezuma, D., Salgado, V., Ortiz-Bejar, J., Sánchez, C.N.: Ingeotec at mex-a3t: Author profiling and aggressiveness analysis in twitter using tc and evomsa. In: *In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval2018)*, CEUR WS Proceedings (2018).

Holmes, J. and Wilson, N. (2017). *An introduction to sociolinguistics*. Routledge.

Kumar, R., Ojha, A. K., Zampieri, M., and Malmasi, S. (2018). Proceedings of the first workshop on trolling,aggression and cyberbullying (TRAC-2018). In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

Wardhaugh, R. (2011). An introduction to sociolinguistics, volume 28. John Wiley & Sons.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017b). *Understanding abuse: A typology of abusive language detection subtasks*. arXiv preprint arXiv:1705.09899.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Reconocimiento automático de estructuras argumentales del dominio médico. Propuesta basada en el modelo de la Léxico-Gramática

Walter Koza[211] & Constanza Suy[212]
[walter.koza@pucv.cl]
Pontificia Universidad Católica de Valparaíso, Chile

Introducción

La representación de la estructura argumental de los predicados (EAP) del área médica es una de las tareas que mayor interés ha despertado en el área del análisis automático de textos (Cohen & Hunter, 2006). No obstante, la mayoría de los trabajos se focalizan en la lengua inglesa, a la vez que no existe un acuerdo respecto de cómo deberían representarse. El presente trabajo se propone describir la EAP proyectada por los verbos del área médica, junto con sus posibilidades transformacionales a fin de crear recursos para su tratamiento informático.

Marco teórico

Se recurre al modelo de la Léxico-Gramática (Gross, 1975; 1984), que propone un método para la descripción de las lenguas naturales a partir de tres condiciones: (i) la sintaxis no está separada del lexicón; (ii) la consideración de la oración simple como la unidad mínima de análisis, y (iii) una formalización a partir de un análisis distribucional y transformacional (Elia, Monteleone & Marano, 2011).

La LG propone la elaboración de tablas, que se presentan sobre la forma de una matriz que contiene: (i) en línea, los elementos de la clase correspondiente; (ii) en columna, las propiedades sintáctico-semánticas, y (iii) al cruce de una línea y de una columna el signo + o − en función de si la entrada léxica acepta o no la propiedad.

---

[211] Doctor en Humanidades y Artes con mención en Lingüística, egresado de la Universidad Nacional de Rosario (Argentina). Profesor Auxiliar del Instituto de Literatura y Ciencias del Lenguaje, de la Pontificia Universidad Católica de Valparaíso, Chile. Es Investigador Responsable del Proyecto FONDECyT 1171033 (Comisión Nacional de Investigación Científica y Tecnológica, Chile), orientado al desarrollo de técnicas para la traducción automática del conocimiento médico. Sus intereses se centran en la lingüística computacional, la gramática del español y las alteraciones del lenguaje, áreas en las que ha publicado diversos artículos.

[212] Docente de Castellano y Comunicación. Licenciada en Letras y Literatura Hispánica. Ambos títulos obtenidos en la Pontificia Universidad Católica de Chile. Actualmente, ejerce la docencia en establecimientos de educación secundaria y superior. Forma parte del personal técnico en el Proyecto FONDECyT 1171033 (Comisión Nacional de Investigación Científica y Tecnológica, Chile) dirigido a la traducción automática del conocimiento médico. Sus áreas de investigación son la sintaxis del español y la lingüística computacional.

Tabla 1. Ejemplo de Tabla LG: Análisis distribucional

Argumento 0		Lema	Argumento 1		Ejemplo
+ANIM	-ANIM		+ALIMENTO	-ALIMENTO	
+	-	comer	+	-	El perro comió un filete.
-	+	comer	-	+	El óxido comió el metal.

De este modo, es posible determinar clases de objeto (Gross, 2012) ('alimentos', 'metales') a partir del significado resultante de la estructura argumental.

Posteriormente, cada oración simple se analiza en función de sus propiedades transformacionales (voz pasiva, pronominalización, nominalización, etcétera).

Tabla 2. Ejemplo de tabla LG: Análisis transformacional

Oración	Voz Pasiva	Nominalización
El perro comió un filete.	+	+
El óxido comió el metal.	+	-

Metodología

A partir de un criterio lexicográfico y de frecuencia, se seleccionaron 100 verbos del área de la medicina (ginecología y obstetricia) del corpus CCM2009 (Burdiles, 2012). Posteriormente, en primer lugar, se llevó a cabo el análisis que consistió en determinar los argumentos seleccionados por cada uno de ellos; los cuales fueron clasificados como clases de objeto (Gross, 2012). Las clases de objeto fueron elaboradas tomando en consideración la ontología SNOMED CT (Snomed, 2016), pero adaptándolo según criterios semánticos y sintácticos. En segundo lugar, dicha información fue volcada en tablas Léxico-Gramática de análisis distribucional y transformacional.

Finalmente, se llevó a cabo una implantación computacional de las tablas LG para la detección de EAP en corpus y su generación automática. Para ello, se recurrió a NooJ (Silberztein, 2005; 2016), herramienta de análisis lingüístico para el tratamiento de textos de lenguaje natural, que posibilita elaborar diversas herramientas como diccionarios electrónicos y gramáticas morfológicas (flexivas, derivativas y productivas) y sintácticas.

Mediante la información otorgada por las clases de objeto, se enriqueció el diccionario electrónico desarrollado en el proyecto FONDECyT 1171033, incluyendo etiquetas semánticas para: (i) la clase de verbo ('diagnosticar') y (ii) la clase de objeto del sustantivo ('tumor', 'médico'):

diagnosticar,V+FLX=VERIFICAR+C1

tumor,N+FLX=ÁRBOL+conc+med+hall
médico,N+FLX=NIÑO+conc+anim+hum+med+pmed

Esto permite crear gramáticas de sintagmas nominales (SN) de la especialidad médica.
Con la información del diccionario, se elaboraron gramáticas para la generación y reconocimiento automático, que contemple tanto la oración simple 'canónica' (SVOD) como las transformaciones (se ejemplifica con la negación y la voz pasiva):

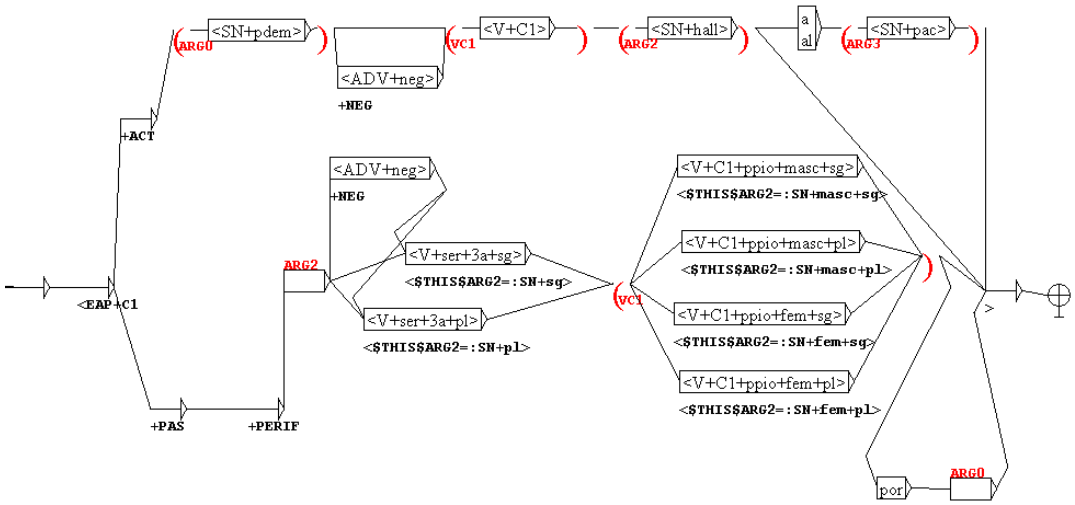


Figura 1. Ejemplo de gramática de reconocimiento y generación de EAP

Resultados

Se establecieron 41 clases de objeto, que enriquecieron el diccionario de NooJ elaborado en el Proyecto FONDECyT 1171033. Esto permitió la posterior elaboración de gramáticas que resultaron efectivas para el reconocimiento automático; los recursos elaborados se probaron en un corpus de artículos científicos del área de la ginecología y obstetricia indizados en la plataforma Scielo (5.000.000 de palabras, aprox.).

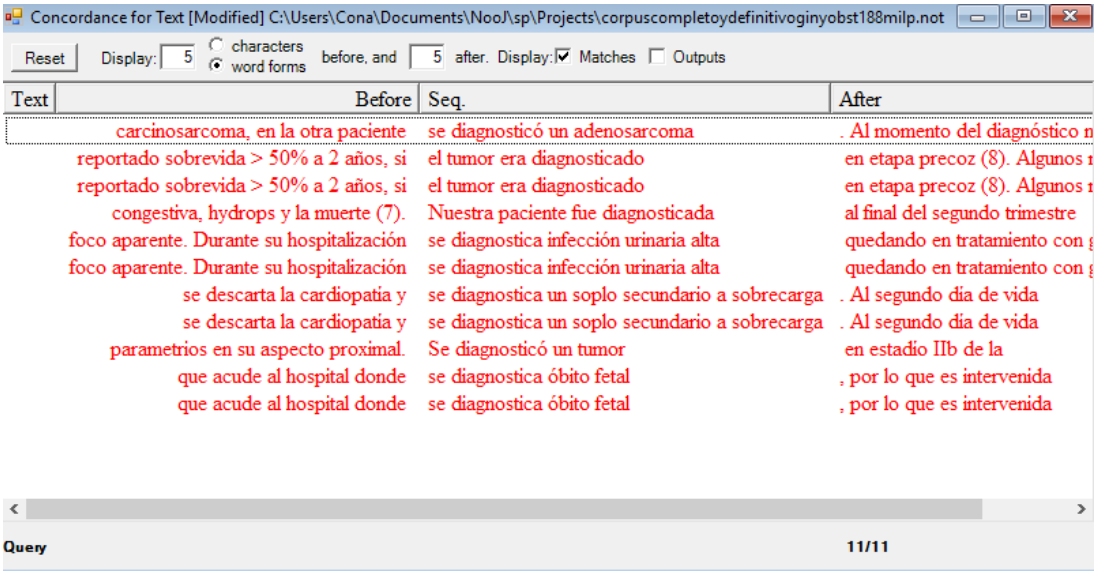


Figura 2. Reconocimiento automático de EAP

Asimismo, se generaron automáticamente posibles EAP:

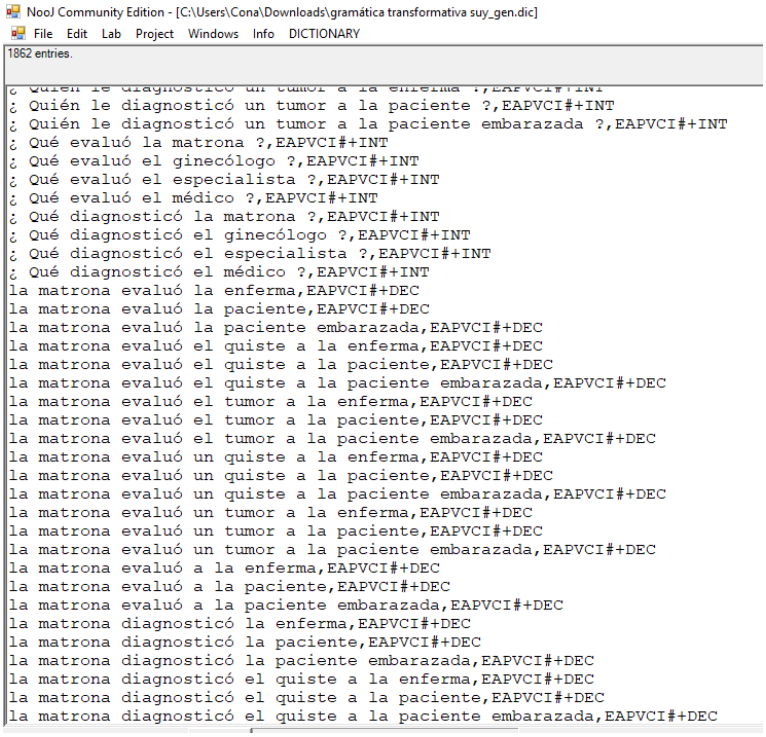


Figura 3. Ejemplo de generación automática

Los resultados obtenidos demuestran que el método es útil para el reconocimiento automático de EAP (tomando una muestra aleatoria de 2.000 oraciones, se logró 72,5% de cobertura y

100% de precisión), a la vez que las oraciones generadas cumplen las condiciones de gramaticalidad.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Sentence encoders as a method for helping users identify and improve semantic similarity in bio-medical text

Brayn Díaz, Juan Pávez, Sebastián Rodríguez, Wenceslao Palma, Héctor Allende-Cid, René Venegas & Eduardo N. Fuentes

[ef@writewise.io]

WriteWise, WriteWise Research Group Artificial Intelligence Unit,[213] Chile

Pontificia Universidad Católica de Valparaíso, Chile

BioPub, Scientific Writing Unit, Chile

Introduction

Academic writing is one of the most valuable skills a scientist can develop. A primary challenge for scientists is to coherently and concisely organize and present ideas within a manuscript. Unfortunatelly, novel computational approaches that help scientists write coherent scientific texts have been scarce and not very effective. Traditional methods that rely on bag-of-words, as well as word embeddings such as those produced by word2vec or GloVe may not be optimal due to language complexity.

Recently, positive results have been reported for sentence embeddings, which provide semantic relatedness between sentences without relying on word overlapping metrics. State of the art models of sentence embeddings are Universal Sentence Encoder and sent2vec, and the latter has an implementation trained on bio-medical data as presented in BioSentVec. Despite the rapid advancement of sentence embeddings and their applications on bio-medical text analysis, the performance of these models for helping users identify and improve semantic similarity between sentences has not been shown.

The aim of this work was to prove the effectiveness of different sentence encoders in bio-medical text to detect semantic similarity between sentences in different sections of academic texts.

Methods

a. Corpus

We used three corpora: 1) A "Gold Standard" which consists of 2116 papers; 2) a first quartile (Q1) bio-medical journal comprised of 2027 papers from the Molecular Cell journal; 3) and a fourth quartile (Q4) bio-medical journal comprised of 1230 papers. The "Gold Standard"

[213] Our group is focus on research in deep learning applied to natural language processing for scientific papers. We applied this research to develop a unique software that guides scientists on how to write scientific articles. Specifically, we are training deep neural networks that "are learning" the structure and content of well-written papers. This thanks to a unique state-of-the-art models including different type of Transformers.

consists of papers in bio-medicine that contain top scientometrics, have in-housed copy editors with rigorous editorial processes, and that present the discursive structure prototypical of well-written papers.

b. Model

We use two models: the transformer-based Univer- sal Sentence Encoder available on TensorFlow Hub, and BioSentVec, a sent2vec model trained on bio-medical text which is available on Github. Due to the lack of labeled data, no additional training is done. A section is modeled as a sequence of sentences, and the similarity between sentences is defined as 1 minus the arc cosine of the cosine similarity of their embeddings divided by π . We also define the *mean sequential similarity* (MSS) in a section as the mean of the similarities between successive sentences. Then, MSS is is used as a measure of semantic similarity in a text.

Results

First, we begin calculating the average MSS in different sections (Introduction, Results, and Discussion) in academic manuscripts in the Gold Standard. The Results section (0.72 0.02 with the USE model and 0.65 0.01 with BioSentVec) showed important differences in comparison with Introduction (0.76 0.02 with USE, 0.68 0.01 with BioSentVec) and Discussion (0.78 0.02 with USE, 0.67 0.02 with BioSentVec), which were very similar. Next, and to compare the MSS in the Gold Standard and other journals, we selected a top Q1 journal and several Q4 journals belonging in the Molecular and Cell Biology Discipline. As a control we used a shuffled version of sentences for the Gold Standard, Q1, and Q4 journals. We found that in both models, the average MSS decreased after shuffling the sentences randomly. The standard deviation increased, though the increase is more significant in the USE model. This confirms that shuffled texts are less cohesive than the original versions.

Finally, we test the effectiveness of the MSS to indicate semantic similarity between sentences. For this we used examples of different sections from new Q1 and Q4 scientific articles and plotted their sequential similarity, comparing it with the Gold Standard MSS values as a benchmark to indi-cate which sentences are outliers (i.e. present higher or lower semantic similarity in comparison with the Gold Standard). We found that this qualitative approach is successful with both USE and BioSentVec models, allowing identification of adjacent sentences that are outliers in both Q1 and Q4 papers. Finally, linguists assessed the performance of both models to test which one provides the best feedback to improve cohesion.

Conclusion

We demonstrated the effectiveness of both the USE and BioSentVec as methods for helping users identify and im- prove semantic similarity between sentences in bio-medical texts. The shared tendencies between the models support sequential similarity as a metric to evaluate a text's cohesion. With both methods outliers can be easily spotted, and then specific modifications in the sentences can be carried out depending on the type of outlier.

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

Una metodología no supervisada para la identificación de hiperonimia: experimentos en español, inglés y francés

Rogelio Nazar[214]
[rogelio.nazar@pucv.cl]
Pontificia Universidad Católica de Valparaíso, Chile
Antonio Balvet[215]
Université de Lille, France
Gabriela Ferraro[216]
DATA61 / Australian National University, Australia
Rafael Marín[217]
Université de Lille, France
Irene Renau[218]
Pontificia Universidad Católica de Valparaíso, Chile

En esta presentación describimos una combinación de algoritmos no supervisados para la identificación de hiperónimos, y que son aplicados a la tarea doble de depurar una taxonomía léxica ya existente y también repoblarla automáticamente con nuevos sustantivos mono o poliléxicos. Algunos de estos algoritmos explotan propiedades distribucionales de las unidades léxicas en grandes corpus; otros utilizan cálculo de la similitud morfológica entre cohipónimos y otros están basado en una implementación cuantitativa de patrones léxico-sintácticos. El objetivo de este sistema es clasificar semánticamente un sustantivo que se presenta como

[214] Profesor del Instituto de Literatura y Ciencias del Lenguaje de la Pontificia Universidad Católica de Valparaíso. Se especaliza en Procesamiento del Lenguaje Natural y Lingüística Computacional. Su interés principal es el léxico, la terminología y el discurso. Actualmente dirige el grupo tecling.com, dedicado al desarrollo de tecnologías lingüísticas.

[215] Profesor en la Université Charles de Gaulle Lille 3, Villeneuve d'Ascq y miembro del grupo Savoirs, Textes, Langage (STL). Se especializa en lingüística computacional y lingüística de corpus. Investiga en sintaxis, semántica, variación y lingüística cuantitativa. Ha desarrollado diversas aplicaciones de tecnologías lingüísticas.

[216] Profesora del College of Engineering and Computer Science de la Australian National University e investigadora en el Analytics Team DATA61, en Australia. Se especializa en Procesamiento del Lenguaje Natural y Lingüística Computacional e investiga en las áreas de extracción de información y aprendizaje automático para el desarrollo de aplicaciones de tecnología lingüística.

[217] Profesor en la Université Charles de Gaulle Lille 3, Villeneuve d'Ascq, miembro del grupo Savoirs, Textes, Langage (STL) e investigador del CNRS, sección 34. Se especializa en semántica e investiga en el tema del componente aspectual de la predicación, tales como incoatividad, cambios de estado y telicidad.

[218] Profesora del Instituto de Literatura y Ciencias del Lenguaje de la Pontificia Universidad Católica de Valparaíso. Su principal línea de investigación es el estudio del significado léxico y sus conexiones con la sintaxis y el contexto de uso. Dirige y ha dirigido diversos proyectos de investigación y desde 2016 es miembro correspondiente de la Academia Chilena de la Lengua.

entrada (por ejemplo, *brie*) asignando un hiperónimo directo (*queso*) pero también, de manera recursiva, otros más generales hasta llegar al nodo más general de la taxonomía (*alimento, objeto físico, entidad*, etc.).

La novedad principal de nuestra propuesta reside en su combinación de enfoques ascendentes y descendentes (bottom-up y top-down). Esto es porque, dado un sustantivo que se toma como entrada, el algoritmo comienza con un enfoque descendente haciendo las preguntas más generales: ¿se trata de una entidad, de una propiedad o de un evento? Entonces, por ejemplo, si determina que es una entidad, pasa entonces a preguntarse qué tipo de entidad es, si es concreta o abstracta. Y si determina que es concreta, se pregunta si es animada o inanimada, y así continúa hasta llegar a un punto medio de abstracción, como por ejemplo un tipo de arma o un tipo de pez. A partir de este punto, comienza con el procedimiento inverso, es decir el ascendente, tratando de encontrar los hiperónimos más específicos o inmediatos (un *fusil de asalto*, una *reineta*, etc.) que luego vuelve a buscar de manera recursiva hasta conectar la cadena hiperonímica ascendente con la descendente.

Con el objeto de evaluar la efectividad del sistema, realizamos una serie de experimentos en español, inglés y francés. De media, determinamos que en una muestra de 1000 sustantivos que clasificamos manualmente, el 82% es clasificado correctamente. También hemos comparado el desempeño del algoritmo con los de otros investigadores utilizando para ello los gold-standards de las competencias Semeval-2016 y SemEval-2018, que proporcionan material en las tres lenguas trabajadas. En la mayoría de los casos, el desempeño de nuestro sistema es moderadamente superior al informado en estas investigaciones, pero la ventaja principal es que en nuestro caso no utilizamos aprendizaje automático. Esto quiere decir que nuestro diseño metodológico es mucho más sencillo y no requiere la costosa fase de entrenamiento que caracteriza a la mayor parte de las soluciones que son populares actualmente en este campo.

El hecho de tratarse de un algoritmo no supervisado también es importante porque es una característica que lo hace inmune a los errores en los datos de entrenamiento. Y, efectivamente, hemos encontrado una cantidad no menor de errores en nuestra revisión manual de los datos de entrenamiento utilizados de manera habitual en estas competencias, lo que creemos puede explicar parte del bajo desempeño informado por algunos de los investigadores. Es por este motivo que, además del algoritmo, proponemos un nuevo set de evaluación para estas tres lenguas, el cual hemos revisado exhaustivamente para garantizar que no existan errores de ningún tipo. Se trata de cuatro categorías semánticas –peces, quesos, máquinas y armas– en un número equilibrado por clase y por lengua. Se presenta en tablas a tres columnas en las que se dispone un sustantivo, un hiperónimo y luego un valor binario que especifica si la relación es falsa o verdadera. Esto es porque, para hacer más exigente la evaluación, hemos introducido deliberadamente afirmaciones falsas con elementos que están semánticamente relacionados, pero en una relación que no es de hiperonimia, como por ejemplo *#la pizza es un tipo de queso* o *#la pesca es un tipo de pez*.

Los datos del proyecto se encuentran disponibles en línea, incluyendo una demo en la que se puede procesar listas de sustantivos en las tres lenguas, ya sean monoléxicos o poliléxicos,

generales especializados. También se incluyen los datos de evaluación, documentación y el código fuente de todos los algoritmos empleados. La dirección del sitio web es http://www.tecling.com/kind

Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquia, Medellín, 21-23 octubre de 2020.

WriteWise: software that guides scientific writing

Eduardo N. Fuentes, Hector Allende-Cid, Sebastián Rodríguez
René Venegas, Juan Pavez, Wenceslao Palma, Ismael Figueroa, Sofia Zamora,
Brayn Diaz & Ashley VanCott
[ef@writewise.io | sebastian.rodriguez.p@pucv.cl | rene.venegas@pucv.cl |
jp@writewise.cl]
WriteWise, WriteWise Research Group, Artificial Intelligence Unit[219] Chile
Pontificia Universidad Católica de Valparaíso, Chile
BioPub, Scientific Writing Unit, Chile

Introduction

Writing scientific articles is traditionally learnt through an inefficient, slow, and arduous process of trial and error. This despite the fact that writing a scientific article is not trivial and in general there is a lack of formal training for learning this task. Scientific articles are inherently complex to write, mainly due to the fact that they are highly technical and present a linguistic/stylistic format specific to the academic genre.

Currently, a scarce number of technological approaches have tried to solve this problem. Most recognizable softwares used by scientists to write papers are Word, GoogleDocs, Grammarly, Authorea, Overleaf, among others. However, none of these solutions satisfactorily provide specific feedback on academic genre, teaches about it, and aids in writing high linguistic levels (academic discourse).

To address this fundamental issue in the academic world we have developed a new software based on artificial intelligence that tackles all the aforementioned problematics; therefore, helping and teaching researchers how to write high-quality English manuscripts.

Methods

a. Corpus

Several corpora were used, but the most relevant for the development of the software were: 1) An unlabeled "Gold Standard" which consists of 2116 papers; 2) A labeled "Gold Standard" which consists of approximately 13,000 sentences. The "Gold Standard" consists of papers in

[219] Our group is focus on research in deep learning applied to natural language processing for scientific papers. We applied this research to develop a unique software that guides scientists on how to write scientific articles. Specifically, we are training deep neural networks that "are learning" the structure and content of well-written papers. This thanks to a unique state-of-the-art models including different type of Transformers.

life-sciences that contain top scientometrics, have in-housed copy editors with rigorous editorial processes, and that present the discursive structure prototypical of well-written papers.

System for tagging texts at the discursive level

A linguistic model for writing scientific texts in the life-sciences was generated. In parallel a computational system was used to subsequently tag sentences at rhetorical-discursive, structural, lexical-grammatical, stylistic, and functional elements prototypical to this genre.

Machine learning models

Several shallow and deep learning models were used for text representation (e.g. word and sentence embeddings), visualization (graphs), and extraction of the most relevant information needed for discourse segmentation, and to assess text coherence and cohesion.

Back and Front end development

Write Wise is presented as a modern web application implemented in proven technologies: Vue.js and Django for frontend and backend development, respectively. The system has a scalable architecture where the machine learning models are consumed as microservices inside the web application.

Results

The Write Wise platform can be subdivided into three modules that tackle the main problems while scientist write papers: 1) sentence length; 2) text order and organization; 3) discourse.

The first module (sentence length) provide information on the microstructure of a text, identifying the number of word per each sentence. The second module (text order and organization) automatically detects the similarity between consecutive sentences in the text. We identified an optimal N° words/sentences and semantic similarity threshold based on analysis of the unlabeled Gold Standard. With these modules the user can compare each sentence with the unlabeled Gold Standard. If the user exceeds the threshold, is notified; thus, this person can make changes in the text. Thanks to these two modules the system helps the user to transmit one idea or message per sentence using clear, precise, and concise sentences, which are consecutively connected. The third and most sophisticated module (discourse) provides the user with examples of different "discursive steps" (i.e. functional linguistic unit that fulfills a communicative purpose in a sentence). Thanks to this the user can write a logical and well-structured text at a high linguistic level.

Conclusions

WriteWise represents the first commercially available advanced platform that provides user´s help and feedback to improve scientific papers writing. This is thanks to the development of and advance textual data representation at different linguistic levels (e.g. words, sentences) through using cutting-edge machine-learning models and applied linguistics research.

Indice de autores

00 03 CONGRESO INTERNACIONAL DE LINGÜÍSTICA COMPUTACIONAL Y DE CORPUS

UNIVERSIDAD DE ANTIOQUIA

university of groningen

Fundación Universidad de Antioquia 25 Años

Tecnológico de Antioquia Institución Universitaria

Association for Computational Linguistics

UNIVERSIDAD DE ANTIOQUIA

Vicerrectoría de Docencia
Vicerrectoría de Investigación

UNIVERSIDAD DE ANTIOQUIA
Facultad de Comunicaciones

CONVERGING HORIZONS
CONICYT-Chile PIA/ANILLOS SOC180045

Universidad de Medellín
Ciencia y Libertad

CÁTEDRA Pablo Valdivia de Comunicación, Humanidades y Tecnología

DOCTORADO EN COMUNICACIÓN
Universidad de La Frontera - Universidad Austral de Chile

Udea Educación Virtual

UNIVERSIDAD NACIONAL DE COLOMBIA

Corpus Ex Machina

Universidad del Valle

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

CILAB | Laspau Affiliated with Harvard University

NETHERLANDS RESEARCH SCHOOL FOR LITERARY STUDIES

WOPATEC 2020